Where the Score Lives: A Wavelet View of Diffusion

Emma Finn **Binxu Wang** **T. Anderson Keller** **Demba E. Ba**
The Kempner Instutite for the Study of Natural and Artificial Intelligence
Harvard University, Cambridge, MA 15213

Abstract

Score-based generative models have had remarkable success over the last decade in generating a diverse set of visually plausible images. A variety of architectures including CNNs, U-Nets, and Transformers have been used as the score-approximation network in such diffusion modeling; however, to date, relatively little is known about how these architectural choices impact generative behavior. In this work, to provide insight into this area, we propose an analytically solvable parameterization of the score function using an expansion in a 2D orthogonal wavelet basis. In particular, we derive interpretable optimal score functions in terms of the moments of the data distribution. We use this parametrization to provide an architecture-agnostic, moment-based analysis that reveals which attributes of the data distribution tend to matter most for denoising. Our score machine is flexible enough to partially mimic the relevant inductive biases of multiple architectures, including U-Nets, and CNNs, taking a step towards understanding why different score architectures can exhibit distinct generative behavior. Since our score is solvable in terms of the moments of the data, we can begin to understand how the data distribution interacts with the score network to produce the behavior we observe in diffusion models.

1 INTRODUCTION

Diffusion models have rapidly advanced image generation and many other generative tasks in recent years (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Karras et al., 2022; Lou et al., 2024). However, it is still not clear what causes their remarkable ability to construct visually coherent samples which often generalize beyond their training distribution. This generalization property clearly depends both on the score network used and on the underlying properties of the data distribution; since it is clear that a diffusion model trained on a single image would not generalize well.

Recent work has explored both architectural contributions to the kind of creativity displayed in diffusion (Kamb and Ganguli, 2024) and data-determined contributions (Wang and Vastola, 2024). On the architectural side Kamb and Ganguli (2024) demonstrate that a CNN encodes certain inductive biases (namely translation equivariance) into a denoising network when used as a diffusion backbone, encouraging a certain 'patch mosaic' form of creativity; one step further, Wang and Pehlevan (2025) shows that when the backbone is linear and translation-equivariant, the diffusion model learns a stationary Gaussian process version of the data; building on prior work Li et al. (2024) which demonstrates that under certain conditions, the ideal score simply memorizes the training data.

On the data-distribution side, spatial locality has been shown to dominate denoising across time scales (Niedoba et al., 2025), and a first-order Gaussian score has additionally been shown to explain a significant amount of observed behavior (Wang and Vastola, 2024), including the spatial local dependency of the denoiser (Lukoianov et al., 2025). Generalization grows with dataset size (Bonnaire et al., 2025), and learned scores can be interpreted as shrinkage in a geometry-adapted harmonic basis (Kadkhodaie et al., 2024). Yet, it still remains unclear which distributional statistics, and which score network architectural choices, are most critical to score learning – questions central to understanding and improving diffusion models.

In this work, we propose an interpretable parameterization of the score function using an expansion in a wavelet basis, which is remarkably flexible. We then derive an ideal score function particular to our functional form, and implement a wavelet-based score machine. Empirically, we run experiments across different families of denoisers to understand which components of the data distribution are most relevant in which settings.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1.1 Background on Diffusion Models

We first provide a short overview of score-based generative modeling. Diffusion models operate by corrupting all of the input data over time according to a noising process, given by an Ornstein-Uhlenbeck stochastic differential equation (Song et al., 2021):

$$d\mathbf{X}_t = \underbrace{f(\mathbf{X}_t, t)}_{\text{drift}}\, dt + \underbrace{g(t)}_{\text{diffusion}}\, dW. \tag{1}$$

The reverse process is given by:

$$d\mathbf{X}_t = \left[f(\mathbf{X}_t, t) - \frac{1}{2} g(t)^2 \underbrace{\nabla_{X_t} \log(p_t(X_t))}_{\text{score function}} \right] dt + g(t) d\bar{W}. \tag{2}$$

We parametrize a neural network s_θ to approximate the score function, and in our setup, we use the score matching objective (Hyvärinen, 2005):

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{p_t(\mathbf{X})}[\|s(\mathbf{X}, t) - s_\theta(\mathbf{X}, t)\|_2^2] \tag{3}$$

Where λ is a time dependent weight. For simplicity, in our case, we set $\lambda(t) \equiv 1$. To further simplify, we'll divide our time interval $[0, 1]$ into N discrete steps, and treat each step independently, so the loss decouples across time. Our loss at any particular timestep is

$$\mathcal{L}^{(t)}(\theta) = \mathbb{E}_{p_t(\mathbf{X})}[\|s(\mathbf{X}, t) - s_\theta(\mathbf{X}, t)\|_2^2] \tag{4}$$

and our overall loss is simply the sum of all such $\mathcal{L}^{(t)}$. In the following, we will demonstrate how to parametrize the score function instead in a wavelet basis, and thereby find the optimal parameters exactly in closed form.

1.2 Wavelets and Denoising

Wavelets have long been used for image denoising and representation. Mallat (1989) introduced multiresolution analysis (MRA) and a fast algorithm for the discrete wavelet transform. Classical wavelet denoising by soft thresholding was developed by Donoho and Johnstone (1994), and further in (Donoho, 1995; Daubechies, 1992; Chang et al., 2000; Portilla et al., 2003), with related analytic multiscale constructions such as wavelet scattering networks (Bruna and Mallat, 2013) and the steerable pyramid (Simoncelli and Freeman, 1995). Like Fourier methods, wavelet decompositions represent square-integrable functions in a basis indexed by "frequencies," but unlike Fourier bases, wavelets are localized in both space and scale, providing joint spatial–frequency resolution.

More recently, Phung et al. (2023) have attempted to use wavelets to speed up diffusion sampling and other score-based generative frameworks, as in (Guth et al., 2022). In (Phung et al., 2023), the authors implement a modified diffusion model which leverages wavelets to increase sampling and training efficiency. Their approach applies wavelet decomposition to the input image and performs denoising in wavelet space, reducing computational cost by separating low- and high-frequency components into a more structured representation than raw pixels. This wavelet transform is used not only at the image level, but also throughout the generator's feature hierarchy.

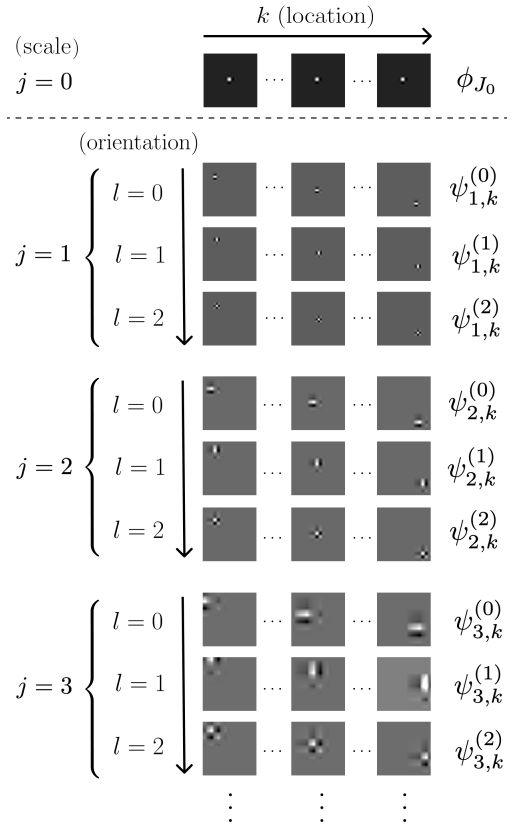


Figure 1: **Daubechies Wavelets.** Visualized across scales j, locations k, and orientations l. We see that at the coarsest scale J_0, only the scale atom is kept, while orientations are included for finer detail wavelets.

Moreover, there is an increasing awareness that wavelet approximations can be used to improve the quality of samples derived from modern image generation models, like diffusion models. For example in (Sigillo et al., 2025), the authors demonstrate that wavelets can be used to significantly increase the quality of texture and detail in very high resolution images. In particular, they propose a "wavelet-derived spatial saliency map" which uses a discrete wavelet transformation to identify where the image stores most of its structural detail. This wavelet-based approach to improve detail quality for high-resolution images has also been proposed in (Zhang et al., 2025). Finally, outside the context of image generation, papers such as (Hu et al., 2025) have proposed using wavelet transforms to help diffusion-based methods for PDE simulation better capture sharp changes and "long-term dependencies." As in (Phung

et al., 2023), the key insight is that diffusion is better behaved in the wavelet domain than in the original state space of the PDE.

This performance improvement across several different problem settings suggests that wavelets can be a useful representation for classes of data on which one might want to train diffusion models, especially when multi-resolution structure is important. Moreover, (Falck et al., 2023) show, in a multi-resolution framework, that U-Nets with average pooling implicitly learn a Haar wavelet basis representation of the data, which further motivates our use of wavelets to approximate the score functions that are typically learned by U-Nets.

We adopt compactly supported Daubechies wavelets (Fig. 1). Let ϕ denote the scaling ("father") function and ψ the wavelet ("mother") function in one dimension. In two dimensions we form tensor products to obtain one scaling atom and three detail atoms at each location/scale. The translates and dilates of these functions form an orthogonal basis of the entirety of $L^2(\mathbb{R}^2)$.

For scale $j \in \mathbb{Z}$ and translation $k = (k_1, k_2) \in \mathbb{Z}^2$:

$$\phi_{j,k}(\mathbf{u}) = 2^j\, \phi(2^j u_1 - k_1)\, \phi(2^j u_2 - k_2),$$
$$\psi^{(\ell)}_{j,k}(\mathbf{u}) = 2^j\, g^{(\ell)}(2^j u_1 - k_1,\, 2^j u_2 - k_2), \qquad \ell \in \{0, 1, 2\},$$

where $g^{(0)}(a, b) = \psi(a)\phi(b)$, $g^{(1)}(a, b) = \phi(a)\psi(b)$, and $g^{(2)}(a, b) = \psi(a)\psi(b)$. While the functions ϕ and ψ are not available in closed form for the Daubechies wavelets, for integer $N \geq 1$, the Daubechies–N wavelet $\psi(x)$ and scaling function $\phi(x)$ are defined by the two-scale relations:

$$\phi(x) = \sqrt{2} \sum_{n=0}^{2N-1} h_n\, \phi(2x - n),$$
$$\psi(x) = \sqrt{2} \sum_{n=0}^{2N-1} g_n\, \phi(2x - n).$$

With periodic boundary handling on $[0, 1]^2$ (our default), the set

$$\mathcal{B}_{J_0} \;=\; \left\{\phi_{J_0,k}\right\}_k \;\cup\; \left\{\psi^{(\ell)}_{j,k}\right\}_{j \geq J_0,\, k,\, \ell}$$

forms an orthonormal basis of $L^2([0, 1]^2)$. (Here J_0 is the coarsest scale; only the scaling atoms at J_0 are kept, while all finer detail wavelets $\psi^{(\ell)}_{j,k}$ are included for $j \geq J_0$.) At each fixed (j, k), the triple $\left(\psi^{(1)}_{j,k}, \psi^{(2)}_{j,k}, \psi^{(3)}_{j,k}\right)$ are the three orientation components often called the "horizontal", "vertical", and "diagonal" details. We refer to this per-location orientation triplet as the detail band at (j, k).

1.3 Our Contributions

Concretely, this work offers four core contributions:

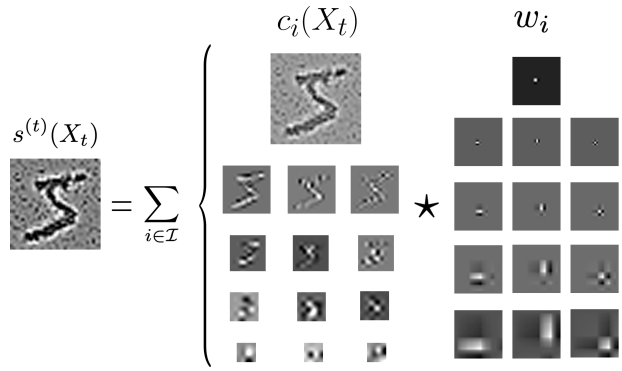


Figure 2: **Wavelet Expansion of the Score**. The score of a given image X_t can be expanded in an orthonormal wavelet basis by representing it as a sum of coefficients $c_i(X_t)$ multiplied by corresponding wavelets w_i (Equation 5). Since the wavelets come in many translated copies, we only display wavelets at a single spatial location k, and display coefficients in spatially ordered 'feature maps'. The sum over spatial locations is then implied by the convolution operator ($\star$).

- **Analytic wavelet score.** We parameterize the score in an orthonormal wavelet basis and, via Stein's identity, reduce each coefficient to a closed-form ridge least-squares estimate with moment-based right-hand sides.
- **Structured dependencies.** We introduce three interpretable classes which serve to isolate the statistics that matter across noise scales.
- **Empirical findings.** On MNIST (32×32 & 64×64) we find: higher polynomial degree lowers MSE; *local* coupling is most reliable across σ; *band-tying* sharpens edges but can raise MSE at low signal-to-noise ratio; benefits of wavelets increase with resolution.
- **Against trained baselines.** Compared to CNN/U-Net denoisers, our models narrow the gap at low–moderate noise by explicitly encoding short-range, multiscale locality—without gradient-based training and with interpretable coefficients.

2 A WAVELET EXPANSION OF THE SCORE

2.1 Expanding the Score

As motivated in Section 1.2, wavelets provide a natural multiscale basis for representing diffusion scores. We now show that, after expanding the score in this basis and modeling each coefficient as a linear function of wavelet-derived features, score learning reduces to a collection of closed-form ridge regression problems.

Let $x \in \mathbb{R}^d$ denote a discretized version of an image, obtained by sampling the underlying grayscale images on $[0,1]^2$ over the pixel grid. For each noise level t, let X_t denote the corresponding noisy image. Let $\tilde{w}_i(u)$ be the ith continuous wavelet atom and let w_i denote its sampled, vectorized version on the image grid. In the rest of the paper, unless otherwise noted, we choose to work in the discrete setting. Thus the score $s_{\text{true}}(x,t) = \nabla_x \log p_t(x) \in \mathbb{R}^d$. Recall that as a function of x, $s_{\text{true}} : \mathbb{R}^d \to \mathbb{R}^d$, but at a particular noisy image X_t, $s_{\text{true}}(X_t,t) \in \mathbb{R}^d$. For a grayscale image X_t at noise level t and score $s_{\text{true}}(\cdot,t)$, since wavelets form an orthonormal basis of L^2, the score at time t in a grayscale image can be expanded as

$$s^{(t)}(X_t) \;=\; \sum_{i\in\mathcal{I}} c_i(X_t)\, w_i, \quad c_i(X_t) = \big\langle s_{\text{true}}(X_t,t),\, w_i \big\rangle, \tag{5}$$

where $\{w_i\}_{i\in\mathcal{I}} = \mathcal{B}_{J_0}$ indexes (J_0,k) for scaling atoms and (j,k,ℓ) for detail atoms (visualized in Figure 2). The way in which we choose to model $\langle s_{\text{true}}(X_t), w_i\rangle$ determines the properties of the score function estimator. We model each coefficient by features $\varphi_i(X_t) \in \mathbb{R}^{d_i}$ and parameters $\alpha_i^{(t)} \in \mathbb{R}^{d_i}$ (visualized in Fig. 3):

$$\widehat{c}_i(X_t) \;=\; \alpha_i^{(t)\top}\varphi_i(X_t), \quad s_\theta^{(t)}(X_t) \;=\; \sum_{i\in\mathcal{I}} \widehat{c}_i(X_t)\, w_i. \tag{6}$$

Because $\{w_i\}$ is orthonormal, the population squared loss decouples across i and so taking the gradient w.r.t. $\alpha_i^{(t)}$ yields the simultaneous equations

$$\mathcal{L}^{(t)}(\theta) = \mathbb{E}_{X_t\sim p_t}\Big\| s(X_t) - \sum_i \alpha_i^{(t)\top}\varphi_i(X_t) w_i \Big\|_{L^2}^2 \tag{7}$$

$$\Rightarrow \frac{\partial \mathcal{L}^{(t)}}{\partial \alpha_i^{(t)}} = -2\mathbb{E}\Big[\varphi_i(X_t)\big(\langle s(X_t), w_i\rangle - \alpha_i^{(t)\top}\varphi_i(X_t)\big)\Big] = 0 \tag{8}$$

Solving for the optimal coefficients $\alpha_i^{(t)*}$, we find, if $\Sigma_i = \mathbb{E}[\varphi_i\varphi_i^\top]$ is invertible:

$$\underbrace{\mathbb{E}[\varphi_i\varphi_i^\top]}_{\Sigma_i}\, \alpha_i^{(t)} = \mathbb{E}\big[\varphi_i(X_t)\,\langle s(X_t), w_i\rangle\big] \tag{9}$$

$$\Rightarrow \quad \alpha_i^{(t)\star} = \Sigma_i^{-1}\,\mathbb{E}\big[\varphi_i(X_t)\,\langle s(X_t), w_i\rangle\big] \tag{10}$$

We can further simplify, by applying a more general form of Stein's Identity as in (Hyvärinen, 2005) to the expectation of the inner product. In particular, for some function f, Stein's Score Identity says

$$\mathbb{E}_{X\sim p_t}\big[\, s_t(X)\cdot f(X)\,\big] \;=\; -\,\mathbb{E}_{X\sim p_t}\big[\,\nabla\cdot f(X)\,\big] \tag{11}$$

for p_t smooth and under vanishing boundary flux (e.g., periodic boundaries). In our case, we write $\varphi_i(X_t) = [\varphi_{i,1}(X_t),\ldots,\varphi_{i,d_i}(X_t)]^\top$. For each component j, apply Stein's identity with the vector field

$$f_j(x) \;=\; \varphi_{i,j}(x)\, w_i, \qquad \|w_i\|_2 = 1.$$

Then

$$\mathbb{E}\big[\varphi_{i,j}(X_t)\,\langle s_t(X_t), w_i\rangle\big] = -\,\mathbb{E}\big[\nabla_{X_t}\big(\varphi_{i,j}(X_t)\, w_i\big)\big] \tag{12}$$

$$= -\,\mathbb{E}\big[\langle \nabla_{X_t}\varphi_{i,j}(X_t),\, w_i\rangle\big] \tag{13}$$

Stacking over $j = 1,\ldots,d_i$ gives the vector form

$$\mathbb{E}\big[\varphi_i(X_t)\,\langle s_t(X_t), w_i\rangle\big] \;=\; -\,\mathbb{E}\big[(\nabla\varphi_i(X_t))^\top w_i\big] \tag{14}$$

Thus, we find that our solution is

$$\alpha_i^{(t)\star} \;=\; -\mathbb{E}[\varphi_i\varphi_i^\top]^{-1}\;\mathbb{E}\big[(\nabla\varphi_i(X_t))^\top w_i\big] \tag{15}$$

In practice, expectations are replaced by sample averages over n training images at time t, $\mathbb{E}[f(X_t)] \approx \frac{1}{n}\sum_{r=1}^n f\big(X_t^{(r)}\big)$. We also add a ridge regularization term, because $\hat{\Sigma}_i$ can be singular or ill-conditioned with high-degree features or correlated wavelet coefficients, we stabilize and control variance by adding a ridge regularization, as follows

$$\hat{\alpha}_i^{(t)} \;=\; \big(\hat{\Sigma}_i + \gamma I\big)^{-1}\hat{b}_i, \quad \gamma > 0,$$

which guarantees an invertible system and mitigates overfitting.

Thus, for samples $\{X_t^{(n)}\}_{n=1}^N$ and ridge $\gamma \geq 0$,

$$\hat{\alpha}_i^{(t)}(\gamma) = \left(\frac{1}{N}\sum_{n=1}^N \varphi_i^{(n)}\varphi_i^{(n)\top} + \gamma I\right)^{-1} \cdot \left(-\,\frac{1}{N}\sum_{n=1}^N (\nabla\varphi_i(X_t^{(n)}))^\top w_i\right). \tag{16}$$

We estimate $\nabla\varphi_i(X_t^{(n)})$ analytically from the chosen features using the method of moments.

Diagonalizing $\Sigma_i = U\Lambda U^\top$ shows that $(\Sigma_i + \gamma I)^{-1}$ weights eigen-directions by $1/(\lambda+\gamma)$. Thus, given that the features are well suited to the data such that Σ_i in (9) is well conditioned and its columns are not highly correlated, the estimator weights eigen directions by $\frac{1}{\lambda+\gamma}$. Small-λ directions receive higher coefficients but are more noise sensitive, which the ridge regression helps to limit. Intuitively, what this says is that this score approximator emphasizes lower-variance feature directions (small λ) and down weights higher-variance directions (large λ). This aligns with the observation that natural images exhibit approximately power-law spectral decay and sparse wavelet coefficients, whereas white noise spreads energy more uniformly; consequently, informative structure often concentrates in a subset of scales and orientations. The model learns to correct more strongly along low-variance modes of the data distribution and ignore the high-variance ones.

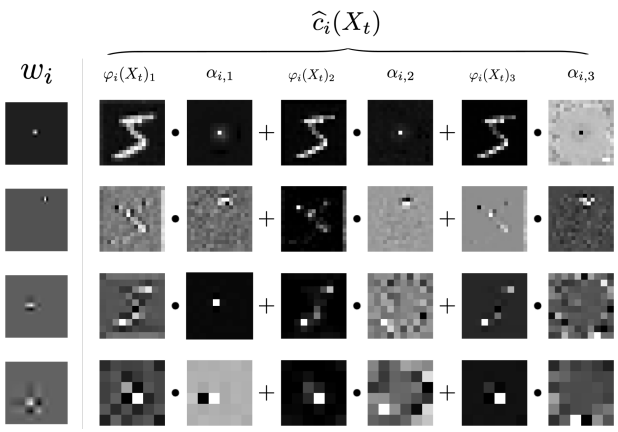


Figure 3: **Wavelet Coefficient Approximation**. We approximate the unknown true coefficients $c_i(X_t)$ of the true score function as an inner product of features $\varphi_i(X_t)$ and parameters $\alpha_i^{(t)}$. The features depicted here are *independent* degree 3 polynomial features.

2.2 Correlation Structures in the Data

Natural images exhibit structured dependencies in the wavelet domain: heavy–tailed marginals per coefficient, co-activation across orientations at a fixed location, and spatial persistence along edges and textures (Wainwright and Simoncelli, 1999). As the noise level decreases, these dependencies become more pronounced. We therefore study three families that isolate, then incrementally reintroduce, these correlations.

(i) *Independent (diagonal).* We model each coefficient $y_t^{(i)} = \langle X_t, w_i \rangle$ in isolation using degree–D monomials (or probabilists' Hermite polynomials for numerical stability). This "mean-field" score approximator is the right baseline when p_t is close to a product distribution (early time steps, near-Gaussian), and it makes failures highly interpretable: any gap to stronger models directly measures the predictive value of cross-coefficient structure that a diagonal model cannot use.

(ii) *Band-tied.* At a fixed scale and location (j, k), we couple the three detail orientations $\ell \in \{0, 1, 2\}$ with degree–D interactions (each monomial includes the target coordinate). This targets cross-orientation co-activation caused by edges and corners and mirrors the channel-mixing inductive bias of CNN/U-Net score networks at a single spatial site. Improvements here isolate the contribution of within-pixel, within-scale structure—testing whether local orientation interactions alone explain denoising gains as t decreases.

(iii) *Local-coupled.* At fixed scale and orientation, we allow the coefficient at location k to depend on neighbors $k+\delta$ within a Chebyshev ball $\|\delta\|_\infty \le r$. This realizes a small neighborhood in wavelet space and emulates the increasing receptive field of convolutional (or locally attentive) score networks. Gains that grow then saturate with r quantify how much spatial context the score actually needs for accurate denoising and globally consistent structure as noise recedes.

We use (14) to build families of models that probe how scale/orientation structure in the data influences the score. We propose three families of models which can provide insight into what properties of the data distribution are relevant for score-based diffusion. In the Daubechies wavelet setting, we have

$$s_\theta(X_t)^{(t)}(\mathbf{u}) = \sum_{k=(0,0)}^{(2^{J_0}-1, 2^{J_0}-1)} \theta_k(X_t, t)\, \phi_{J_0,k}(\mathbf{u}) + \sum_{j=J_0}^{J_{\max}} \sum_{k=(0,0)}^{(2^j-1, 2^j-1)} \sum_{l=1}^{3} \zeta_{j,k,l}(X_t, t)\, \psi_{j,k}^{(l)}(\mathbf{u}) \quad (17)$$

where J_0 is the coarsest scale of the approximation, analogous to the spatial resolution at the network's bottleneck and for an image of dimension $H \times W$ $J_{\max} = \lfloor \log_2(\min(H, W)) \rfloor$, which corresponds to the highest level of wavelet detail encoded.

2.2.1 Independent Baseline

Assume (unrealistically) that coefficients decouple across scale j, location k, and detail band l. Let $y_t^{(i)} = \langle X_t, w_i \rangle$ with moments $\mu_r^{(i)}(t) = \mathbb{E}[(y_t^{(i)})^r]$ and model our coefficients as

$$\theta_k(X_t, t) = \sum_{m=0}^{D} b_m^{(k)}(t)\, \langle X_t, \phi_{J_0,k} \rangle^m, \quad (18)$$

$$\zeta_{j,k,\ell}(X_t, t) = \sum_{m=0}^{D} d_m^{(j,k,\ell)}(t)\, \langle X_t, \psi_{j,k}^{(\ell)} \rangle^m. \quad (19)$$

Using (14) with monomial features gives the normal equations analogous to (7)for each index i and $r = 0, \ldots, D$. In particular we substitute $\phi(y) = y^m$ into Stein's identity and apply it in each coefficients to give

$$\sum_{m=0}^{D} a_m^{(i)}(t) \mu_{m+r}^{(i)}(t) = -r\mu_{r-1}^{(i)}(t), \;\; a_\bullet^{(i)} \in \{b_\bullet^{(k)}, d_\bullet^{(j,k,\ell)}\},$$

This is a Hankel system $H^{(i)}(t)\, a^{(i)}(t) = -\, h^{(i)}(t)$, where

$$H_{r,m}^{(i)}(t) = \mu_{r+m}^{(i)}(t),$$
$$h^{(i)}(t) = \left[\, 0,\; \mu_0^{(i)}(t),\; 2\mu_1^{(i)}(t), \ldots, D\mu_{D-1}^{(i)}(t) \,\right]^\top.$$

Observe that all entries of $H^{(j,k,l)}(t)$ and $h^{(j,k,l)}(t)$ are computable from clean-data raw moments of $Y_0 =$

$\langle X_0, \psi_{j,k}^{(l)} \rangle$.

$$\mu_r^{(i)}(t) = \sum_{m=0}^{r} \binom{r}{m} \bar{\alpha}_t^{\frac{m}{2}} (1-\bar{\alpha}_t)^{\frac{r-m}{2}} \mathbb{E}[\langle X_0, w_i \rangle^m] \mathbb{E}[Z^{r-m}]. \tag{20}$$

where $\bar{\alpha}$ is the same $\bar{\alpha_t}$ derived from the noise schedule $\{\alpha_t\}_t$ found in the standard DDPM variance-preserving forward process $X_t = \sqrt{\bar{\alpha_t}} X_0 + \sqrt{1-\bar{\alpha_t}} Z$ for Z a standard multivariate normal random variable. We For small D, we can easily compute $a^{(i)}(t) = -(H^{(i)}(t))^{-1} h^{(i)}(t)$ (or $(H^{(i)} + \gamma I)^{-1}$ with ridge $\gamma \geq 0$). The co-factor formula for the coefficients reads

$$\hat{\alpha}_i(t) \;=\; -\frac{1}{\det H^{(i)}(t)} \sum_{r=1}^{D} r\, \mu_{r-1}^{(i)}(t)\, C_{r,m}^{(i)}(t) \tag{21}$$

with $C_{r,m}^{(j,k,\ell)}(t)$ the (r,m)-cofactor of $H^{(i)}(t)$. This closed form can be used to investigate how higher order moments of the data distribution impact the score. Though clearly independence is an unrealistic assumption, this model is fast and interpretable, and serves as (i) an initial baseline, and (ii) a diagnostic lower bound. We can investigate the value of different kinds of correlation by measuring the difference in performance between models with different kinds of limited dependence. Allowing each coefficient to depend arbitrarily on *all* wavelet coordinates $y_t = (\langle X_t, w_1 \rangle, \ldots, \langle X_t, w_n \rangle)$—i.e., learning n functions $f_i : \mathbb{R}^n \to \mathbb{R}$ or even degree–D multivariate polynomials—leads to combinatorial parameter growth and brittle estimation. We therefore restrict to *structured*, computationally tractable, and interpretable dependencies (diagonal/independent, band-tied, and local-coupled), which retain closed-form or efficient normal-equation solvers while capturing the dominant correlations.

2.2.2 Wavelet Band Coupling

One very natural form of correlation is to allow wavelet coefficients at the same scale j and location k to depend on one another. Let $y_0 = \langle X_t, \psi_{j,k}^{(0)} \rangle$, $y_1 = \langle X_t, \psi_{j,k}^{(1)} \rangle$, $y_2 = \langle X_t, \psi_{j,k}^{(2)} \rangle$, At each scale/location in the detail bands (j,k,l), we assume a polynomial coefficient form of degree D:

$$\begin{aligned}
\zeta_{j,k,0}^{(t)}(X_t) &= C_0^{(t)} + \sum_{a=1}^{D} \sum_{b=0}^{D-a} \sum_{c=0}^{D-a-b} \beta_{a,b,c}^{(0)}\, y_0^a y_1^b y_2^c, \\
\zeta_{j,k,1}^{(t)}(X_t) &= C_1^{(t)} + \sum_{b=1}^{D} \sum_{a=0}^{D-b} \sum_{c=0}^{D-b-a} \beta_{a,b,c}^{(1)}\, y_0^a y_1^b y_2^c, \\
\zeta_{j,k,2}^{(t)}(X_t) &= C_2^{(t)} + \sum_{c=1}^{D} \sum_{a=0}^{D-c} \sum_{b=0}^{D-c-a} \beta_{a,b,c}^{(2)}\, y_0^a y_1^b y_2^c.
\end{aligned} \tag{22}$$

The constraint ensures each monomial contains the target coordinate y_ℓ at least once; cross terms are allowed but pure "other-orientation" terms are excluded.) We define our θ coefficients as polynomials with the same optimal values as in the independent case. Estimation proceeds via the same normal-equation machinery as in (14), but with mixed moments at (j,k):

$$\mu_{pqr}^{(j,k)}(t) := \mathbb{E}\big[\, y_0^p y_1^q y_2^r \,\big], \qquad p+q+r \leq D+1,$$

which, under the forward process with orthonormal $\{\psi_{j,k}^{(\ell)}\}$, expand into clean-data mixed moments and factorized Gaussian noise moments.

2.2.3 Local Coupling

Another natural choice of coupling is to allow wavelets in the same local neighborhood. For a radius $r \in \mathbb{N}$, define the neighborhood $\Delta_r = \{\delta \in \mathbb{Z}^2 : \|\delta\|_\infty \leq r,\ \delta \neq (0,0)\}$ We allow wavelets in the same neighborhood to interact. Fixing a scale j and orientation $\ell \in \{0,1,2\}$, let the oriented wavelet/detail coefficient at spatial location $k \in \mathcal{K}$ be $y_k = \langle X_t, \psi_{j,k}^{(\ell)} \rangle$. Let $D \geq 1$ be the total degree. Define $S_D = \{(d,e) \in \mathbb{N}^2 : d, e \geq 1,\ d+e \leq D\}$. We define the functional forms of θ as follows:

$$\begin{aligned}
\theta_{J_0,k}(X_t,t) = &\sum_{i=0}^{D} \alpha_i \left\langle X_t,\ \phi_{J_0,k} \right\rangle^i \\
&+ \sum_{\delta \in \Delta_r} \sum_{(d,e) \in S_D} \beta_{\delta,d,e} \left\langle X_t,\ \phi_{J_0,k} \right\rangle^d \left\langle X_t,\ \phi_{J_0,k+\delta} \right\rangle^e
\end{aligned}$$

where α_i and $\beta_{\delta,d,e}$ are the parameters over which we optimize. Similarly for ζ we define

$$\begin{aligned}
\zeta_{j,k,l}(X_t,t) = &\sum_{i=0}^{D} \xi_i \left\langle X_t, \psi_{j,k,l} \right\rangle^i \\
&+ \sum_{\delta \in \Delta_r} \sum_{(d,e) \in S_D} \omega_{\delta,d,e} \left\langle X_t, \psi_{j,k,l} \right\rangle^d \left\langle X_t, \psi_{j,k+\delta,l} \right\rangle^e
\end{aligned}$$

where ξ_i and $\omega_{\delta,d,e}$ are the parameters over which we optimize. As in the wavelet band coupling, estimation proceeds via the same machinery as in (14), but with mixed moments.

3 METHODS

In order to implement this we return to the most general framing of the problem: ridge-regression in a non-linear feature space defined by wavelet coefficients.

Preprocessing We begin by preprocessing the images by normalizing them to $[0,1]$ and resizing them. We use compactly supported Daubechies (db2) wavelets with periodized boundaries. The 2-D tensor-product basis $\mathcal{B}_{J_0}$ includes scaling atoms at the coarsest kept scale J_0 and detail atoms for $j \geq J_0$ up to $J_{\max} = \lfloor \log_2 \min(H,W) \rfloor$ (here $J_{\max} = 3$ at 32×32 in order to avoid boundary effects) with three orientations $\ell \in \{0,1,2\}$. The resulting basis forms an approxi-

Figure 4: **Visualization of noise regimes.** Clean image $\sigma = 0$ (left) to highly noised $\sigma = 4$ (right)

mately orthonormal linear operator, $B \in \mathbb{R}^{p\times d}$ where p is the number of features and $d = H \times W$ is the flattened dimension of the image. B maps vectorized images $X_t \in \mathbb{R}^d$ to wavelet coefficients.

Noise Model We add Gaussian noise according to the variance exploding regime $X_t = X_0 + \sigma Z$ for $Z \sim \mathcal{N}(0, 1)$. We consider four noise regimes in the below experiments, $\sigma \in \{1, 2, 3, 4\}$ (see Fig. 4). We note that in our theoretical derivation 20, we use the variance preserving formulation of the noising schedule. An equivalent form in terms of σ_t is also possible and just requires altering the coefficients to be in terms of the function of σ_t corresponding to $\bar{\alpha}_t$ used there. This yields noise projections $c = X_t B^T$ that serve as inputs to our non-linear feature constructions. In practice, because B is over-complete and therefore not exactly orthonormal, we use a generalized inverse of B instead.

Feature Expansion and Regression We generate the non-linear features according to the model of correlation we're interested in. In practice we solve the ridge regularized optimization

$$\widehat{W}_\gamma \;=\; \arg\min_{W} \frac{1}{N}\|AW - Y\|_F^2 + \gamma\|W\|_F^2 \qquad (23)$$

where A is a matrix containing the stacked features $\varphi(X_t)$ computed on a finite number of noised samples X_t with shape $N \times p$ and Y its the matrix of corresponding clean images whose n^{th} row is $X_0^{(n)}$. **Lemma** (Informal) The optimal W from this equation yields the same denoised predictions as our wavelet by wavelet approach in Equation 15 up to some small approximation error.

Lemma (Formal) If there exists a left-inverse $R \in \mathbb{R}^{d\times M}$ such that $R^\top B = I_d$, then letting $F : \mathbb{R}^M \to \mathbb{R}^P$ be any fixed feature map and writing $\varphi(x) := F(xB^\top)$, we have that for X_t and X_0 the two ridge problems yield the same optima.

$$\textbf{(Pixel)}\;\; W_\gamma^\star \in \arg\min_{W\in\mathbb{R}^{P\times d}} \mathbb{E}\big[\|\varphi(X_t)W - X_0\|^2\big] + \gamma\|W\|_F^2,$$

$$\textbf{(Coeff)}\;\; U_\gamma^\star \in \arg\min_{U\in\mathbb{R}^{P\times M}} \mathbb{E}\big[\|F(X_tB^\top)U - X_0B^\top\|^2\big] + \gamma\|U\|_F^2 \qquad (24)$$

Then the minimizers satisfy

$$W_\gamma^\star \;=\; U_\gamma^\star R^\top, \text{and hence}\; \varphi(x)W_\gamma^\star \;=\; F(xB^\top)U_\gamma^\star R^\top \forall x.$$

In particular, the denoised predictions produced by the pixel-space ridge and by the wavelet-by-wavelet ridge followed by synthesis agree exactly. We defer the proof of this equivalence to Appendix C.1.

We also add a ridge term to improve numerical stability (Hoerl and Kennard, 1970). We additionally experiment with the probabilists' Hermite polynomials to expand our θ and ζ with increased stability. The probabilists' Hermite polynomials $\{\mathrm{He}_n(x)\}_{n\geq 0}$ are defined by $\mathrm{He}_n(x) = (-1)^n e^{x^2/2}\frac{d^n}{dx^n}e^{-x^2/2}$ and are orthogonal for $Z \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[\mathrm{He}_m(Z)\mathrm{He}_n(Z)] = n!\,\delta_{mn}$. Using He_n in place of raw monomials makes feature vectors orthogonal in expectation under near-Gaussian coordinates $y_t^{(i)}$, reducing covariance and bringing $\Sigma_i = \mathbb{E}[\varphi_i\varphi_i^\top]$ closer to diagonal (hence a smaller condition number). In practice, this improves the conditioning of the matrix equations (9) and yields coefficients less sensitive to scaling and polynomial degree, especially when combined with a small ridge penalty. The results for this model are deferred to Appendix A.

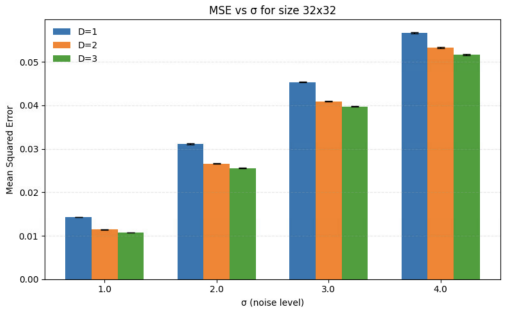
Denoising and Comparison with Models We denoise our images according to $\widehat{X_0} = A\widehat{W}$ and clamp the images back to $[0, 1]$ before computing MSE over the entire dataset. We also compare our reconstruction MSE to that of trained diffusion models with U-Net and CNN backbones. We defer the details of training those models to Appendix B.

4 EXPERIMENTS AND DISCUSSION

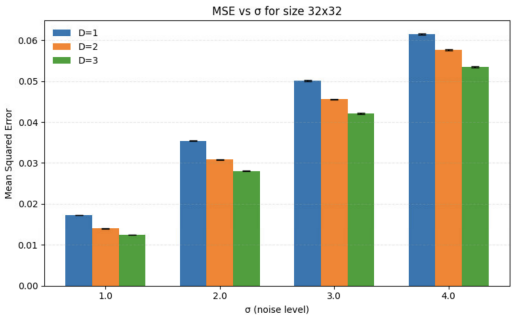
With these closed-form equations and model families in hand, we now test how much band and local coupling improve over the independent baseline across different noise levels. These models are just a few examples of the correlation structures in the data that can be probed with this approach.

Setup We evaluate all models on two copies of the full MNIST training set of 60,000 images, one set resized to 32×32 and the other resized to 64×64, denoted MNIST-32 and MNIST-64 respectively (LeCun et al., 1998).

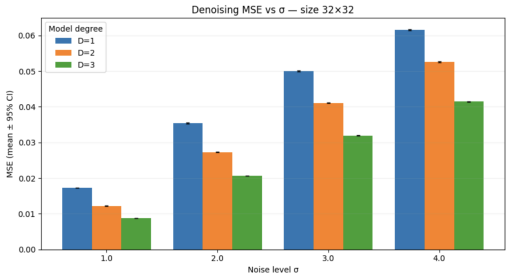
Changing the Degree of the Approximation As shown in Figure 5 and Figure 6, increasing the polynomial degree consistently improves reconstruction error by enhancing the expressivity of the feature space. The gains are especially pronounced for correlated models—most notably in the local coupling case (Figure 5c)—indicating that higher-order moments of the wavelet-projected data are particularly valuable when modeling correlation structure. Strikingly, this effect is amplified at larger image resolutions: comparing Figure 5 with Figure 6 suggests that our wavelet-based reconstruction becomes more robust as image size increases.



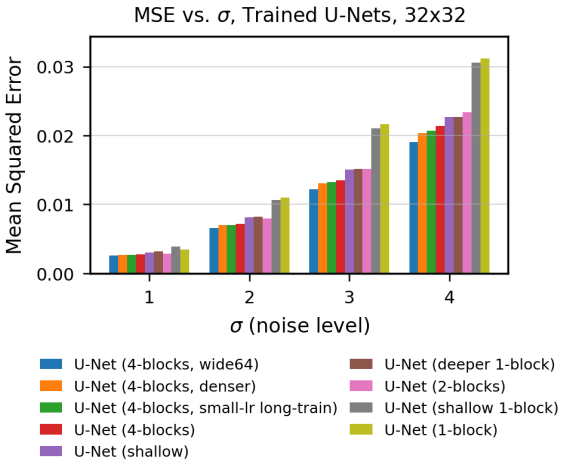
(a) Independent Features



(b) Band-Tied Coupling

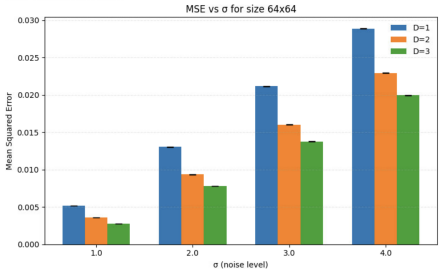


(c) Local Coupling

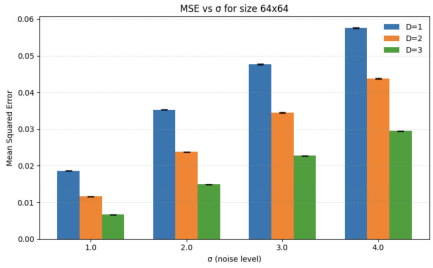


(d) Trained Diffusion Models

Figure 5: **MNIST-32.** **(a)** Denoising MSE with independent monomial features across three sets of features: degrees 1, 2, & 3. **(b)** Same features with band-tied coupling, and **(c)** local coupling. **(d)** Comparable performance of a variety of trained U-Net score function approximators.



(a) Independent Features



(b) Band-Tied Coupling

Figure 6: **MNIST-64.** Denoising MSE with **(a)** independent features, and **(b)** band-tied coupling on larger images. We see higher degree polynomials are more performant with larger images, compared with Fig. 5.

Detail Band Correlation As shown in Figure 5 and Figure 6, coupling the three orientations at a location improves visual quality (see Appendix A). Surprisingly, for both 32×32 and 64×64 images, independent features often achieve lower MSE than the band-tied counterpart. Qualitatively, however, edges and corners are sharper under band-tying, consistent with the hypothesis that cross-orientation co-activation carries most of the local predictive signal and is especially sensitive to sharp changes. This detail-band correlation can be viewed as a non-linear cross-channel interaction (e.g., quadratic products within a scale). At high noise levels ($\sigma \geq 3$), the signal to noise ratio of detail coefficients drops, so cross-orientation terms act on products of noisy coefficients, inflating variance and inducing localized overshoot/undershoot that pixel-space MSE penalizes strongly. In short, channel mixing helps perceptual sharpness but can hurt MSE when the inputs are noise-dominated. This might suggest using noise-aware mixing (e.g., down-weighting cross-orientation features or increasing ridge at large σ), or complementing MSE with structure-sensitive metrics.

Local Correlation Across image sizes (see Figure 5 and Appendix A), we observe that introducing local coupling between neighboring coefficients yields a consistent and significant improvement in denoising MSE. This provides empirical support for prior mechanistic explanations of diffusion models (Niedoba et al., 2025; Kamb and Ganguli, 2024), which emphasize that local interactions are among the most predictive features available to a denoiser. Intuitively, wavelet coefficients corresponding to nearby locations tend to co-

vary strongly, especially along edges, textures, and digit strokes (Wainwright and Simoncelli, 1999; Simoncelli and Olshausen, 2001). By capturing this short-range dependency, the model can recover local structure more effectively than when treating coefficients as independent. Notably, unlike detail-band correlation (which risks performance degradation at high noise due to channel mixing), local correlation maintains benefits across all σ regimes, suggesting that spatial neighborhoods retain informative signal even when individual coefficients are noisy.

Changing Image Size A key advantage of our approach is robustness to increasing image size and resolution. Unlike linear score approximations, the wavelet parameterization preserves locality and multiresolution structure: feature interactions remain confined to small neighborhoods and scales, so the effective design matrices retain near-block-diagonal structure and well-behaved conditioning as $H \times W$ grows. Empirically, the relative performance of our models improves at higher resolutions. Moving from 32×32 to 64×64 yields larger MSE reductions for the same polynomial degree, suggesting that additional scales provide genuinely useful predictive signal rather than simply inflating dimensionality (see Appendix A).

Comparison to Trained Models As introduced by Falck et al. (2023), U-Nets can be viewed as implicitly learning a multiresolution (wavelet-like) representation. In our comparison to trained CNN/U-Net denoisers (Fig. 5d), the learned models outperform our analytic wavelet denoisers across noise levels, as expected. However, two patterns are noteworthy. First, with local coupling and higher polynomial degree, the gap at low to moderate noise ($\sigma \in \{1, 2\}$) narrows noticeably, indicating that much of the benefit of learned models arises from exploiting short-range, multiscale locality, which is the inductive bias encoded in our features. Second, as noise increases ($\sigma \geq 3$), the trained models retain a larger advantage, consistent with their capacity to leverage deeper nonlinearity, cross-scale mixing, and data-dependent priors learned during training. Crucially, our approach achieves these results with no gradient-based training and a single closed-form ridge solve per setting, yielding interpretable coefficients and stable behavior under changes in image size.

5 FUTURE WORK AND CONCLUSION

5.1 Future Work

While this paper is only a first step toward interpreting diffusion models through a wavelet decomposition of the score, it suggests a flexible framework for doing so. Our analytic construction makes it possible to isolate how specific moments, local dependencies, and architectural biases contribute to denoising behavior, and in turn points to several natural directions for future work. On the theoretical side, an important next step is to develop approximation guarantees and finite-sample error bounds for the proposed score estimators. On the modeling side, our experiments focus on a limited family of dependencies, leaving richer forms of multi-scale structure, cross-scale interactions, and other architectural biases to be explored. Experimentally, there are also many settings left to investigate. In particular, although the framework extends naturally to RGB and other multi-channel images, we do not include color-image experiments here. Moreover, we only test at modest image resolutions, whereas prior work such as (Zhang et al., 2025) suggests that the advantages of wavelet representations may become more pronounced at very high resolutions. Finally, existing work such as (Kamb and Ganguli, 2024) derives closed-form ideal score machines that allow direct, seed-by-seed comparison between a theoretical denoiser and the output of a learned diffusion model. Developing an analogous ideal score machine in the wavelet basis could provide further insight into the behavior of diffusion models.Finally, we suggest that expanding to video data would be a natural and interesting approach, especially given the natural role of wavelets in temporal modeling.

5.2 Conclusion

We introduced an analytically tractable, wavelet–based parameterization of diffusion scores with closed-form normal equations and three structured dependency families (independent, band-tied, local). This framework isolates which attributes of the data distribution contribute to denoising across noise scales and resolutions. Empirically, higher polynomial degree improves accuracy, band-tying sharpens edges but can raise MSE at low signal to noise ratio, and local coupling yields the most reliable MSE gains across σ and image sizes. The gap to trained CNN/U-Net denoisers narrows at low–moderate noise when locality is encoded explicitly, suggesting that much of their advantage stems from short-range, multiscale interactions that our analytic model captures without gradient-based training. The method is scalable (closed-form per setting), interpretable (moment-level diagnostics), and robust to increasing resolution. Because the score coefficients are solved from moments, the framework serves as a diagnostic tool: it quantifies the value of specific correlations (orientation co-activation vs. spatial neighborhoods), motivates noise-aware channel mixing and ridge schedules, and can inform data-efficient architectural choices and initialization schemes for learned models.

6 ACKNOWLEDGMENTS

This work was supported by the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. We thank the Kempner for access to compute resources. We are also grateful to the CRISP group at Harvard SEAS for many thoughtful

conversations. EF thanks the Calvin Coolidge Presidential Foundation for support during her undergraduate studies.

References

Bonnaire, T., Urfin, R., Biroli, G., and Mézard, M. (2025). Why diffusion models don't memorize: The role of implicit dynamical regularization in training.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886.

Chang, S., Yu, B., and Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing*, 9(9):1532–1546.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Donoho, D. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455. _eprint: https://academic.oup.com/biomet/article-pdf/81/3/425/26079146/81.3.425.pdf.

Falck, F., Williams, C., Danks, D., Deligiannidis, G., Yau, C., Holmes, C., Doucet, A., and Willetts, M. (2023). A multi-resolution framework for u-nets with applications to hierarchical vaes.

Guth, F., Coste, S., Bortoli, V. D., and Mallat, S. (2022). Wavelet score-based generative modeling.

Ho, J., Jain, A. N., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoogeboom, E., Heek, J., and Salimans, T. (2023). Simple diffusion: End-to-end diffusion for high resolution images.

Hu, P., Wang, R., Zheng, X., Zhang, T., Feng, H., Feng, R., Wei, L., Wang, Y., Ma, Z.-M., and Wu, T. (2025). Wavelet diffusion neural operator.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (2024). Generalization in diffusion models arises from geometry-adaptive harmonic representations.

Kamb, M. and Ganguli, S. (2024). An analytic theory of creativity in convolutional diffusion models.

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models. In *NeurIPS*.

LeCun, Y., Cortes, C., and Burges, C. J. C. (1998). The mnist database of handwritten digits. `https://yann.lecun.com/exdb/mnist/`. Accessed: 2026-04-01.

Li, S., Chen, S., and Li, Q. (2024). A good score does not lead to a good generative model.

Lou, A., Meng, C., and Ermon, S. (2024). Discrete diffusion modeling by estimating the ratios of the data distribution.

Lukoianov, A., Yuan, C., Solomon, J., and Sitzmann, V. (2025). Locality in image diffusion models emerges from data statistics. *arXiv preprint arXiv:2509.09672*.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.

Niedoba, M., Zwartsenberg, B., Murphy, K., and Wood, F. (2025). Towards a mechanistic explanation of diffusion model generalization.

Phung, H., Dao, Q., and Tran, A. (2023). Wavelet diffusion models are fast and scalable image generators.

Portilla, J., Strela, V., Wainwright, M. S., and Simoncelli, E. P. (2003). Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351.

Sigillo, L., He, S., and Comminiello, D. (2025). Latent wavelet diffusion for ultra-high-resolution image synthesis.

Simoncelli, E. P. and Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, volume 3, pages 444–447.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, 24:1193–216.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations.

Wainwright, M. J. and Simoncelli, E. (1999). Scale mixtures of gaussians and the statistics of natural images. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Wang, B. and Pehlevan, C. (2025). An analytical theory of spectral bias in the learning dynamics of diffusion models. In *Advances in Neural Information Processing Systems*. Spotlight.

Wang, B. and Vastola, J. J. (2024). The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications.

Zhang, J., Huang, Q., Liu, J., Guo, X., and Huang, D. (2025). Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23464–23473.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, see Section 2.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, though since the point of the algorithim is interpretability note efficinecy, we only briefly discuss this in Appendix E.

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. Yes, full statements and proofs are provided.

   (b) Complete proofs of all theoretical results. Yes, some proofs are delayed until Appendix C.

   (c) Clear explanations of any assumptions. Yes, I hope.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Mostly yes, see Appendix B and the discussion of the implementation.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, see Appendix B.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, see Section 4.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, see Appendix B.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. Yes, but we only use MNIST, a very standard dataset.

   (b) The license information of the assets, if applicable. Not applicable.

   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.

   (d) Information about consent from data providers/curators. Not Applicable

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. Not Applicable

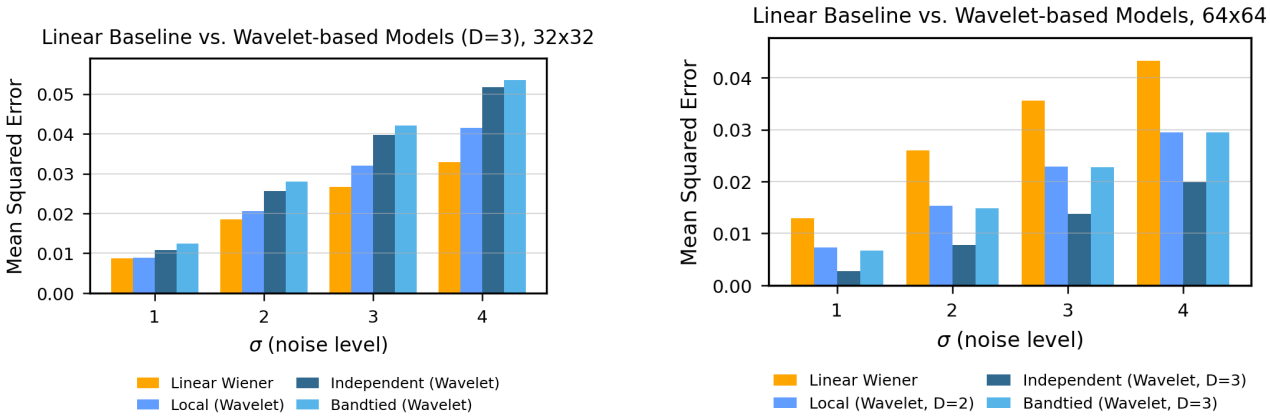
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A Additional Results

A.1 Comparison with Linear Baseline

In Figure 7 below, we compare our proposed framework with the simple linear baseline of Wang and Vastola (2024) (detailed in Section B.2). From this comparison, we see very clearly that the wavelet decomposition of the score approximator is most beneficial for higher resolution images (64x64, right) where our model significantly outperforms the linear counterpart. Conversely, on smaller images (32x32, left), the frequency-focused wavelet decomposition struggles to deal with the high frequency components induced by low spatial resolution. We note that, somewhat analogously, in prior work, it has been found that wavelet decompositions are helpful for diffusion models operating on high-resolution images, since wavelet coefficients compactly represent images as multi-channel, low-resolution feature maps (Hoogeboom et al., 2023). In future work, we intend to explore the applicability of our model to higher resolution images more thoroughly, as we think this may be a key area where this method has significant advantages over existing solutions.



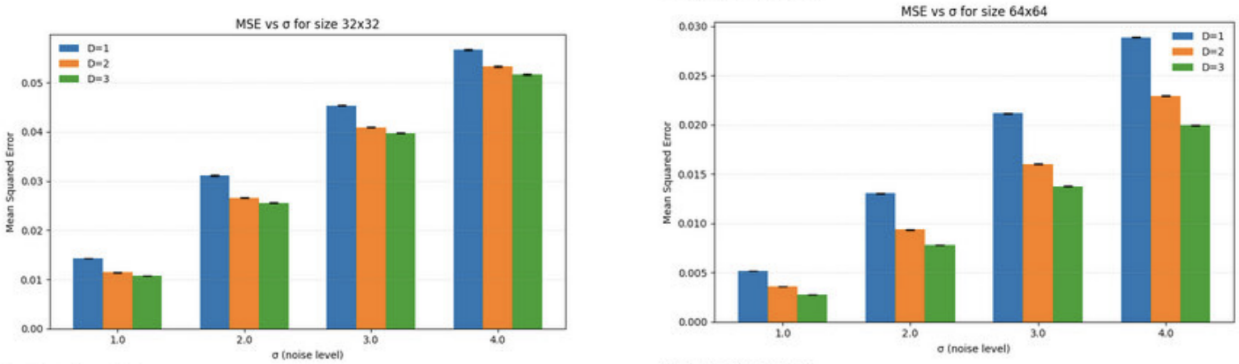
(a) MNIST-32: Linear baseline vs. wavelet models (D=3).

(b) MNIST-64: Linear baseline vs. wavelet models (D=3).

Figure 7: Comparison of linear baseline vs. wavelet models (D=3) on MNIST-32 and MNIST-64. We see the wavelet decomposition of the score significantly outperforms the linear baseline on larger images, while performing slightly worse on the smaller images.

A.2 Hermite Polynomial Results

In Figure 8 below we also include results from an independent model expanded in terms of the probabilist's Hermite polynomials, as described in Section 3. The Hermite expansion decouples features in expectation and delivers clear improvements as D grows, with the notable exception of $D = 3$. Results are plotted for both 32 by 32 images and 64 by 64 images. The setup is identical to that described in Section 3 and 4.



(a) MNIST-32: Hermite Polynomials of degrees 1,2, and 3.

(b) MNIST-64: Hermite Polynomials of degrees 1, 2, and 3.

Figure 8: Comparison of independent Hermite polynomials across different images sizes and degrees. Increasing the degree of the approximation is most helpful for larger images.

A.3 Qualitative Denoising Results

We also display qualitative denoising results below in Figure 9 for the independent model, and Figure10 for the band-tied model. We see that our model is able to denoise qualitatively well across a range of settings, and matches the quantitative results implying better performance for higher resolution images.

A.3.1 Diffusion Model Generations

While the overall quality of the generated images is low, the broad shape of features is promising, especially in the band tied high degree cases. However, the wavelet-based diffusion process appears to capture some coarse structural patterns. In particular, the band-tied, higher-degree models show a more coherent emergence of organized intensity regions, suggesting that the model begins to internalize meaningful global dependencies. These results indicate that, despite the pixel-level noise, the generative process is learning the broad spatial organization of features, which could serve as a foundation for higher-fidelity synthesis with improved architectures or longer training.

B Training Details

B.1 UNet denoiser baseline details

Preconditioning and objective We followed the same preconditioning and loss weighting scheme from Karras et al. (2022), with the code below.

```
class EDMPrecondWrapper(nn.Module):
    def __init__(self, model, sigma_data=0.5, sigma_min=0.002, sigma_max=80, rho=7.0):
        super().__init__()
        self.model = model
        self.sigma_data = sigma_data
        self.sigma_min = sigma_min
        self.sigma_max = sigma_max
        self.rho = rho

    def forward(self, X, sigma, cond=None, ):
        sigma[sigma == 0] = self.sigma_min
        ## edm preconditioning for input and output
        ## https://github.com/NVlabs/edm/blob/main/training/networks.py#L632
        # unsqueze sigma to have same dimension as X (which may have 2-4 dim)
        sigma_vec = sigma.view([-1, ] + [1, ] * (X.ndim - 1))
        c_skip = self.sigma_data ** 2 / (sigma_vec ** 2 + self.sigma_data ** 2)
        c_out = sigma_vec * self.sigma_data / (sigma_vec ** 2 + self.sigma_data ** 2).sqrt()
        c_in = 1 / (self.sigma_data ** 2 + sigma_vec ** 2).sqrt()
        c_noise = sigma.log() / 4
        model_out = self.model(c_in * X, c_noise, cond=cond)
        return c_skip * X + c_out * model_out
```

```
class EDMLoss:
    def __init__(self, P_mean=-1.2, P_std=1.2, sigma_data=0.5):
        self.P_mean = P_mean
        self.P_std = P_std
        self.sigma_data = sigma_data

    def __call__(self, net, X, labels=None, ):
        rnd_normal = torch.randn([X.shape[0],] + [1, ] * (X.ndim - 1), device=X.device)
        # unsqueeze to match the ndim of X
        sigma = (rnd_normal * self.P_std + self.P_mean).exp()
        weight = (sigma ** 2 + self.sigma_data ** 2) / (sigma * self.sigma_data) ** 2
        # maybe augment
        n = torch.randn_like(X) * sigma
        D_yn = net(X + n, sigma, cond=labels, )
        loss = weight * ((D_yn - X) ** 2)
        return loss
```

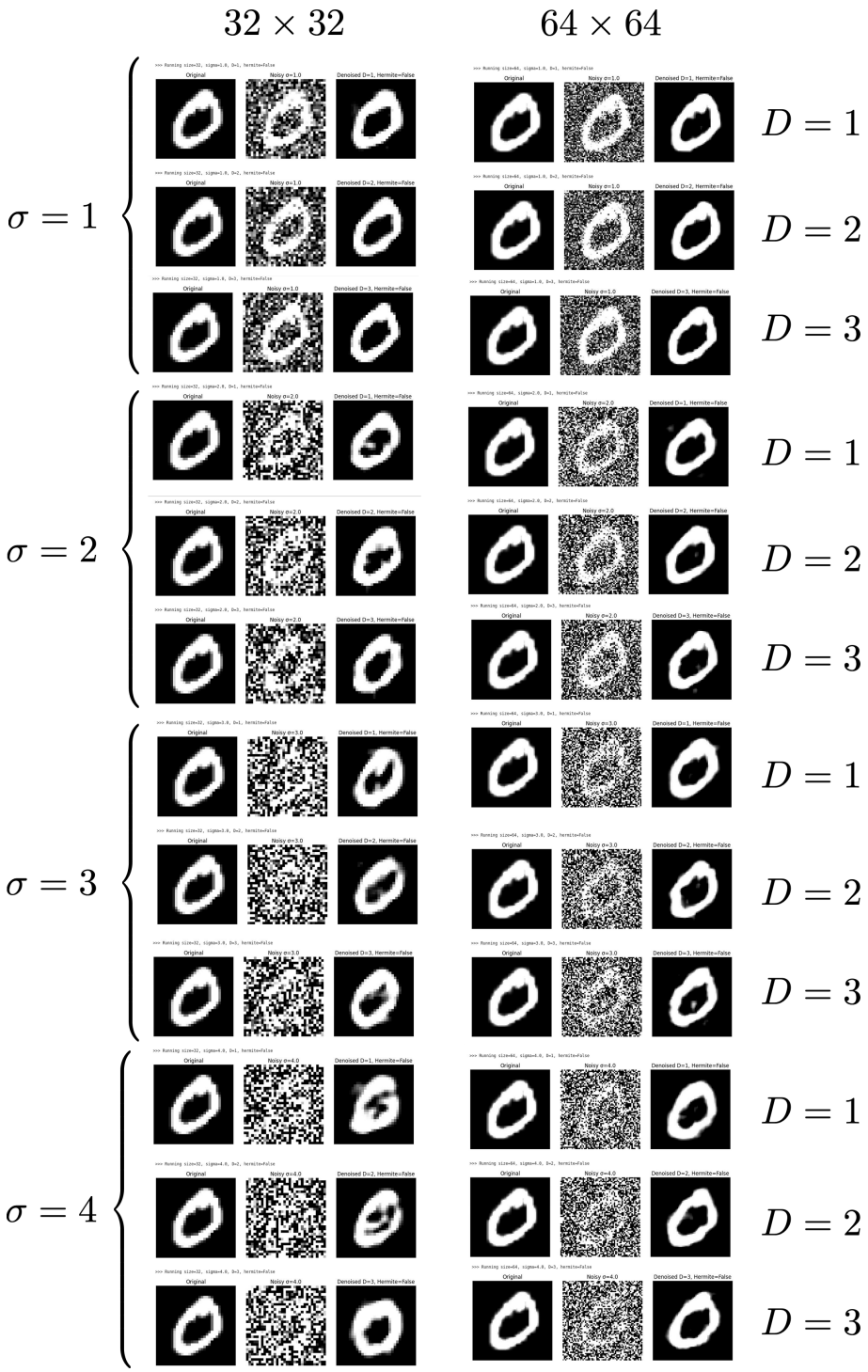


Figure 9: Denoising results for the independent monomial model on two image sizes (columns) and across ranges of noise levels (groups of 3 rows), for different degrees D.

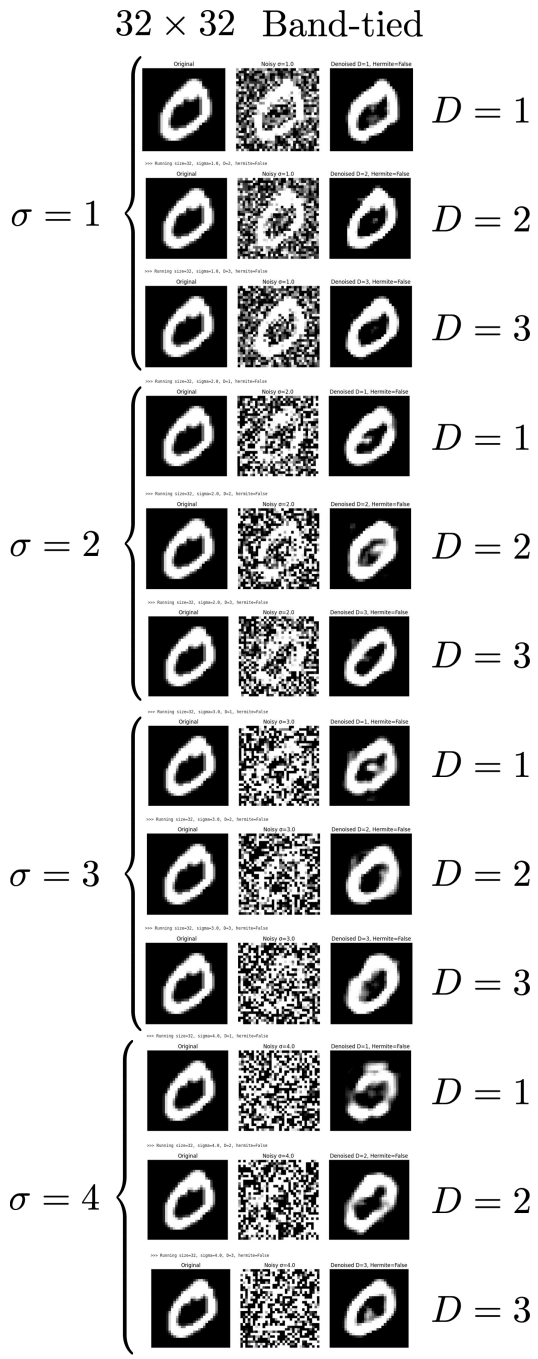
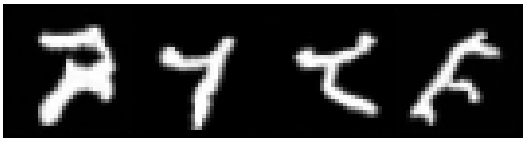


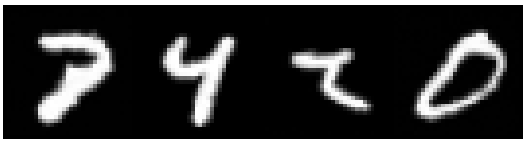
Figure 10: Denoising results for the band-tied model on 32 by 32 images, across ranges of noise levels (groups of 3 rows), for different degrees D.



(a) EDM, 1 block, no attention

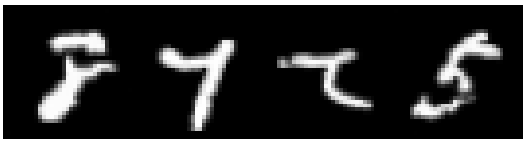
(b) EDM, 2 blocks, no attention

(c) EDM, 4 blocks, no attention

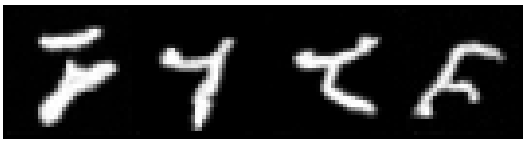


(d) EDM, 4 blocks, denser

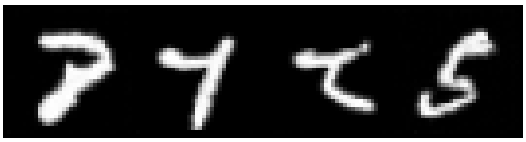
(e) EDM, 4 blocks, small LR, long training



(f) EDM, 4 blocks, width 64



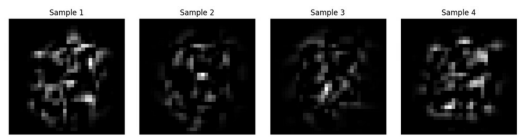
(g) EDM, shallow 1 block



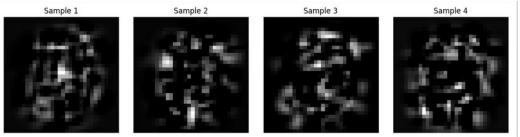
(h) EDM, shallow

(i) EDM, deeper 1 block

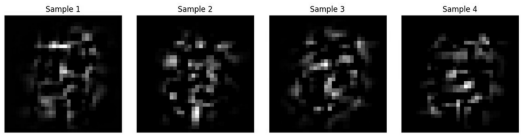
Figure 11: MNIST samples generated by different UNet-CNN EDM diffusion model variants. Each row shows images from one model configuration.



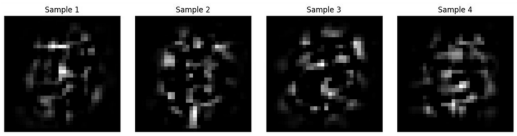
(a) Independent, degree 1



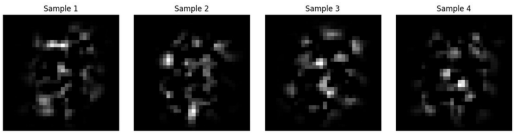
(b) Band, degree 1



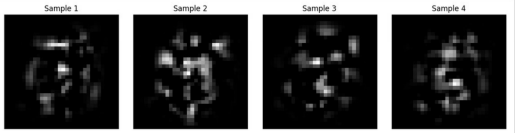
(c) Independent, degree 2



(d) Band, degree 2



(e) Independent, degree 3



(f) Band, degree 3

Figure 12: Generated samples for Independent vs. Band models at varying polynomial degrees. Each column shows one dataset; each row corresponds to a degree

Model variants and training hyperparameters We adopted the U-Net architecture from Song and Ermon (2019) as the backbone and systematically varied its architectural hyperparameters. Specifically, we trained U-Nets using the EDM objective on MNIST, varying the number of `blocks` and the base width (`model_channels`). Each encoder block halves the spatial resolution and multiplies the number of channels by `channel_mult`, while each decoder block performs the inverse upscaling by $\times 2$. We fixed `layers_per_block=1`, so each block contains one residual layer. The `channel_mult` tuple defines the width scaling across resolution levels—for example, with `base_channels=16` and `channel_mult=(1, 2, 3, 4)`, the block widths are 16, 32, 48, and 64. Unless otherwise noted, all models use `batch_size=2048`, `attn_resolutions=[]`, and `decoder_init_attn=False`, i.e., no attention is applied in the U-Net.

For training, we used Adam optimizer and constant learning rate without scheduling.

Table 1: **UNet-EDM architectures and training schedules.**
Shared hyperparameters: `layers_per_block=1`, `decoder_init_attn=False`, `attn_resolutions=[]`

Experiment	Blocks	Base_ch	ch_mult	Steps	Batch	LR	Train time
`4blocks_noattn`	4	16	1,2,3,4	10k	2048	1e-3	∼25 min
`4blocks_noattn_smalllr_longtrain`	4	16	1,2,3,4	50k	2048	1e-4	∼2h 4min
`4blocks_noattn_denser`	4	16	1,2,3,4	10k	2048	1e-3	∼37 min
`4blocks_wide64_noattn`	4	64	1,2,3,4	10k	2048	1e-3	∼77 min
`1block_noattn`	1	16	1	10k	2048	1e-3	∼28 min
`1block_wide128_noattn`*	1	32	1	10k	2048	1e-3	∼17 min
`2blocks_noattn`	2	16	1,2	10k	2048	1e-3	∼22 min

B.2 Linear denoiser baseline details

As another baseline, we computed the optimal linear denoiser, which captures the Gaussian statistics of the dataset. We computed the empirical mean and covariance of the dataset μ, Σ. Then we used the following formula for linear denoiser (Wang and Vastola, 2024),

$$\mathbf{D}_{lin}(\mathbf{x}, \sigma) = \mu + \Sigma(\Sigma + \sigma^2 I)^{-1}(\mathbf{x} - \mu) \tag{25}$$

$$= \mu + U \mathrm{diag}\left(\frac{\lambda_k}{\lambda_k + \sigma^2}\right) U^\top (\mathbf{x} - \mu) \tag{26}$$

where $\Sigma = U\Lambda U^\top$ and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ...)$ is the eigen decomposition of the data covariance. In practice, we precomputed the eigendecomposition to avoid explicit matrix inverse.

B.3 Computing infrastructure

For all DNN training runs and numerical simulation of the theory, we used an academic cluster with A100, H100 GPUs. Generally the UNet training runs on MNIST finish from 25mins to 2hrs on one A100 GPU.

C Additional Proofs

C.1 Proof of Equivalence of Code and Math

Notation Let $x \in \mathbb{R}^d$ be an image, with $d = H \times W$. Let $\{\omega_j\}_{j\in\mathcal{I}} \subset \mathbb{R}^d$ be an orthonormal wavelet basis with matrix $B \in \mathbb{R}^{d\times d}$, where each column is a wavelet ω_j. Then $B^\top B = I$. We note that in practice this wavelet basis $B \in \mathbb{R}^{D\times d}$ is over-complete and therefore not exactly orthogonal, but only approximately so. In our particular case, we compute $\|BB^\top - I\|_2 = 0.01201$ and $\|B^\top B - I\|_2 = 0.01346$, which confirms in our case that B is approximately, though not exactly, orthonormal.

We corrupt images as

$$X_t = X_0 + \sigma Z, \quad Z \sim \mathcal{N}(0, I_d).$$

In the code, we use a shared feature map

$$\Phi(x) \in \mathbb{R}^p$$

across all output coordinates. The denoiser network output is a linear readout over the features

$$\widehat{X}_0(x) = \Phi(x)W, \quad W \in \mathbb{R}^{p\times d},$$

where p is the number of features.

Lemma 1

Lemma C.1 (The pixel-level denoiser is the same as the wavelet denoiser)**.** *Let $W' = WB^\top$. Then for any $W \in \mathbb{R}^{p\times d}$, we have*

$$\mathcal{L}_{denoiser}(W) = \mathcal{L}_{wavelet}(W'),$$

i.e. $W^\star$ minimizes the denoiser loss if and only if $W'^\star = W^\star B^\top$ minimizes the wavelet loss.

Proof. The denoiser loss is

$$\mathcal{L}_{\text{denoiser}}(W) = \mathbb{E}\big[\|\Phi(X_t)W - X_0\|_2^2\big] + \gamma\|W\|_F^2.$$

The wavelet loss is

$$\mathcal{L}_{\text{wavelet}}(W') = \mathbb{E}\big[\|\Phi(X_t)W' - BX_0\|_2^2\big] + \gamma\|W'\|_F^2.$$

We say that B is approximately orthonormal if $\|B^TB - I\|_2 \le \epsilon$ and $\|BB^T - I\|_2 \le \epsilon$ for some $\epsilon > 0$. Since B is (in practice, only approximately) orthonormal,

$$\begin{aligned}\|\Phi(X_t)W - X_0\|_2^2 &\approx \|(\Phi(X_t)W - X_0)B^\top\|_2^2 \\ &\approx \|\Phi(X_t)WB^\top - X_0B^\top\|_2^2 \\ &\approx \|\Phi(X_t)W' - BX_0\|_2^2.\end{aligned}$$

The ridge term is also invariant: $\|W\|_F^2 = \|W'\|_F^2$. Taking expectations completes the proof. □

Remark C.2. By Tweedie's formula, the learned $\widehat{S}_W$ is the best (in L^2 sense) projection of the true score onto the following function space,

$$\mathcal{F} = \left\{ x \mapsto \frac{x}{\sigma^2} - \frac{\Phi(x)W}{\sigma^2} \right\}.$$

Lemma 2

Lemma C.3 (Score coefficients from the code coincide with coefficient-wise regressions.)**.** *Define the coefficient* $c_i^{(t)}(x)$ *by*

$$c_i^{(t)}(x) = \langle \widehat{S}_W(x), \omega_i \rangle = \frac{\langle x, \omega_i \rangle}{\sigma^2} - \frac{\Phi(x) W_i'}{\sigma^2}.$$

The coefficient-wise loss is

$$\mathcal{L}_{score}(\alpha_i) = \mathbb{E}\Big[\big\|\alpha_i^\top \Phi(X_t) - \langle S_t, \omega_i \rangle\big\|_2^2\Big] + \gamma \|\alpha_i\|_2^2.$$

Hence the optimal α_i *satisfies*

$$\alpha_i = -\frac{1}{\sigma^2} W_i'.$$

Proof. Because the $\{\omega_i\}$ form an orthonormal basis, the population loss decomposes as

$$\mathbb{E}\left[\big\|S_t(X_t) - \sum_i \alpha_i^\top \Phi(X_t)\omega_i\big\|_2^2\right] + \gamma \sum_i \|\alpha_i\|_2^2,$$

which decouples exactly as the sum of the losese$\sum_i \mathcal{L}_i(\alpha_i)$.

Thus, optimization decouples across coefficients, proving the claim. □

Takeaway The code is exactly equivalent to our theoretical formulation where

$$\psi_i = \Phi(x)$$

represents the complete set of features. Even if ψ_i are incomplete, the equivalence holds up to neglecting irrelevant components of the covariance.

D Extension to Color Images

In the main text, we write the theory for single channel (gray scale) images. For multi-channel (i.e. color) images, the theory apply analogously. Take RGB images as example, the wavelet filters are applied to each channel separately, resulting in three sets of wavelet coefficients (R, G, B). Then the nonlinear features can be built on top of the concatenated wavelet coefficients.

E Computational Considerations for Larger Images

In the main text, we run experiments on images of size 32×32 and 64×64, while preliminary experiments at larger resolutions (128×128 and 256×256) led to memory requirements that exceeded what our experimental setup could handle. Concretely, the wavelet representation includes detail atoms up to scale $J_{\max} = \lfloor \log_2(\min(H, W)) \rfloor$, so increasing image resolution increases the number of wavelet coefficients and the size of the associated ridge systems. Importantly, in our structured models the feature count per coefficient remains bounded: it is $D + 1$ in the independent model, $1 + \binom{D+2}{3}$ in the band-tied model, and $D + 1 + \big((2r+1)^2 - 1\big)\frac{D(D-1)}{2}$ in the local-coupled model. For $D = 3$, these are 4, 11, and 28 features per coefficient when $r = 1$. Thus, for fixed D and r, the main cost comes from having more coefficients, not from making each coefficient depend globally on all the others. By contrast, allowing each coefficient to depend arbitrarily on all n wavelet coordinates via a degree-D polynomial would require $\binom{n+D}{D}$ monomials per coefficient, leading to combinatorial parameter growth and increasingly brittle regression systems. This motivates the use of more restricted and structured coupling schemes, such as extensions of our band-tied and local-coupled models, in order to control both computational and statistical complexity while preserving locality and multiscale structure. At the same time, we believe the wavelet parameterization is more amenable to higher-resolution images than naive linear estimators, since it preserves locality and multiscale structure. Empirically, Figures 5, 6, and 7 suggest that the relative advantage of the wavelet-based model becomes stronger as image size increases.