POLYNOMIAL DENOISERS ON THE IMAGE MANIFOLD: VOLTERRA NETWORKS FOR SELF-CONTAINED PIXEL DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural images are often conjectured to lie on a low-dimensional manifold within the high-dimensional pixel space. Recent work leverages this to build self-contained pixel-space diffusion models without specialized machinery or pretrained components. A plain Transformer on raw image patches suffices, provided it predicts the clean image rather than the noise and compresses each patch through an aggressive bottleneck. The mixing operator, however, has remained softmax attention, inherited from sequence modeling rather than derived from denoising. For Gaussian noise, the optimal denoiser admits a canonical expansion in polynomial functionals of the observation, the Wiener–Volterra series. We prove that softmax attention, and any normalized mixing rule (including linear attention), cannot realize any polynomial map beyond an affine one. In contrast, Volterra layers realize this class directly and, on a data manifold, achieve approximation rates governed by the manifold dimension rather than the ambient pixel dimension. We therefore propose the architecture these results prescribe; raw image patches passed through a bottleneck embedding, followed by cascaded low-rank Volterra layers, which carry no activation function because each layer is nonlinear in its own right. To our knowledge this is the first pixel-space diffusion architecture whose mixing operators are Volterra layers throughout, with no attention, no convolution, no pretrained components, and no activation function on the residual path. On ImageNet 256×256 and 512×512, the resulting model is competitive with self-contained pixel-space diffusion at matched parameter and FLOP budgets.

1 INTRODUCTION

Diffusion models now dominate high-fidelity image generation, and most use the Transformer, a sequence operator whose fit to the geometry of natural images remains unclear. The strongest systems operate in the latent space of a pretrained tokenizer (Rombach et al., 2022; Peebles & Xie, 2023), which ties generation to external components trained with adversarial and perceptual objectives and, increasingly, aligned to self-supervised encoders (Oquab et al., 2024; Yu et al., 2025). On the other hand, pixel-space alternatives avoid the tokenizer but typically add specialized machinery of their own (Section 2). Li & He (2026) showed that a plain Transformer on raw image patches suffices, provided it predicts the clean image rather than the noise or velocity and compresses each patch through a low-dimensional bottleneck. They motivate this with the manifold assumption, that natural images occupy a low-dimensional subset of pixel space, which clean-image prediction exploits. This leaves the architectural question open. If denoising is intrinsically low-dimensional, which *function class* represents it most efficiently?

We find that the answer is classical. The minimum mean-squared-error denoiser for a signal in additive Gaussian noise admits a canonical expansion as a polynomial functional series, the Wiener–Volterra series (Volterra, 1887; Schetzen, 1980). We turn this characterization into two results for pixel-space diffusion. First, restricted to a $d_{\mathcal{M}}$-dimensional data manifold, the optimal denoiser is approximated by a polynomial in a low-dimensional embedding, with squared error decaying as $N^{-2s/(2d_{\mathcal{M}}+1)}$ in the number of coefficients N for smoothness s, a rate set by the intrinsic rather than the ambient dimension (Proposition 1). Second, attention lies structurally outside this class. A softmax attention

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.

layer that realizes a polynomial map at all realizes an affine one, whereas a low-rank Volterra stack represents the whole class exactly once its width clears an explicit threshold (Proposition 2).

Our architecture is the explicit construction these results call for. Each raw patch passes through a bottleneck embedding, followed by a stack of low-rank Volterra layers, each pairing a cross-token mixer over patch positions with a channel-wise mixer within a patch. Because each mixer is nonlinear in its own right, the residual path carries no activation function, and the model trains end-to-end on pixels with no tokenizer, no adversarial or perceptual loss, no representation alignment and no pretrained components. To our knowledge it is the first pixel-space diffusion architecture derived as the constructive realization of an approximation-theoretic separation rather than by analogy with sequence modeling.

On ImageNet 256×256 the resulting model outperforms JiT in FID at every scale at matched parameters and FLOPs, reaching FID 1.48 with no pretrained components, and it does so again at 512×512, where it reaches 1.55 (Section 5.3). In a controlled study on a known manifold, the necessity of x-prediction reported by Li & He (2026) carries over to our architecture (Section 5.1).

Contributions. We give an intrinsic-dimension approximation rate for the optimal denoiser on the data manifold, and a separation showing that softmax attention realizes no polynomial map beyond affine ones while the Volterra class attains the rate exactly (Section 3). From this theory we derive a self-contained, activation-free pixel-space diffusion backbone built from low-rank Volterra token and channel mixers (Section 4). It improves FID over JiT at every scale on ImageNet at matched compute (Section 5.3), and an ablation shows that each Volterra sublayer lowers FID, with the activation-free channel mixer outperforming the feed-forward network it replaces (Section 5.4).

2 Background and Related Work

2.1 Pixel-Space Diffusion

Diffusion models started in pixel space with convolutional U-Nets (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; Dhariwal & Nichol, 2021). Latent diffusion (Rombach et al., 2022) moved them to autoencoder latents and became standard for Transformer backbones such as DiT (Peebles & Xie, 2023) and SiT (Ma et al., 2024), at the cost of pretrained tokenizers and alignment to self-supervised encoders (Oquab et al., 2024; Yu et al., 2025). Pixel-space diffusion remains competitive via specialized designs, namely latent-token interfaces (Jabri et al., 2023), hourglass Transformers (Crowson et al., 2024), U-ViT hybrids (Hoogeboom et al., 2023; 2025), multi-scale cascades (Chen et al., 2025), dual-level pixel Transformers (Yu et al., 2026), aligned neural-field decoders (Wang et al., 2026a), or self-supervised pretraining (Lei et al., 2026). Li & He (2026) showed that a plain ViT on raw patches works if it predicts the clean image and aggressively bottlenecks each patch. We likewise use a self-contained, plain architecture, but replace attention with Volterra layers, motivated by approximation theory on the data manifold (Section 3).

Formulation. We use the flow-matching diffusion formulation (Lipman et al., 2023; Esser et al., 2024). Given data $x \sim \rho_{\text{data}}$ and noise $\epsilon \sim \mathcal{N}(0, I)$, we define the noised sample $z_t = t\,x + (1-t)\,\epsilon$ for $t \in [0, 1]$, with velocity $v = \partial_t z_t = x - \epsilon$. Following Li & He (2026), a network $\text{net}_\theta(z_t, t, c)$ (with class label c) predicts the clean image $\hat{x}_\theta$, and we form the velocity prediction as

$$v_\theta = \frac{\hat{x}_\theta - z_t}{1-t}, \tag{1}$$

with the denominator clipped near $t = 1$. We train with the squared velocity loss $\mathcal{L} = \mathbb{E}_{t,x,\epsilon} \left\| v_\theta - v \right\|^2$ (Salimans & Ho, 2022), sampling t from a logit-normal distribution (Esser et al., 2024). At inference we integrate the learned velocity field from $t = 0$ to $t = 1$ with an ODE solver. The network output is kept in x-space, for the reason given in Section 4.1.

2.2 The Manifold Assumption

The hypothesis that high-dimensional natural data lie near a low-dimensional manifold is longstanding (Chapelle et al., 2006; Carlsson, 2009), and manifold-learning methods (Tenenbaum et al., 2000;

Belkin & Niyogi, 2003) were designed to recover it. Denoising autoencoders (Vincent et al., 2008) made the link to denoising explicit, anticipating the objective of modern diffusion models.

Generative models leverage the manifold assumption differently. Latent diffusion (Rombach et al., 2022) learns it *implicitly* via a pretrained autoencoder, while Li & He (2026) keep it *ambient*, arguing that clean-image prediction uses low-dimensional structure without an explicit encoder. We build it into the architecture, where the bottleneck embedding and Volterra mixers directly parametrize smooth maps on it. Jackson-type polynomial rates on compact submanifolds are classical (Ragozin, 1970); our contribution is the observation that a linear bottleneck into $\mathbb{R}^{2d_{\mathcal{M}}+1}$ makes the rate depend only on intrinsic dimension, and we provide an architecture that attains it (Section 3). We work with image patches, so the ambient dimension is the per-patch pixel dimension $D := D_p = 3p^2$ (hundreds to low thousands). We assume ρ_{data} is supported on a compact smooth submanifold $\mathcal{M} \subset \mathbb{R}^D$, the per-patch data manifold, of intrinsic dimension $d_{\mathcal{M}} \ll D$.

Real image distributions concentrate near rather than exactly on such a manifold; we state results for exact support, which the proofs use through compactness and reach, and near-manifold data behave as this case at noise levels well above the data's thickness. That $d_{\mathcal{M}} \ll D$ is an empirical and conjectural premise, supported by the success of strong bottleneck embeddings in (Li & He, 2026) and by intrinsic-dimension estimates (Pope et al., 2021; Levina & Bickel, 2004).

2.3 Volterra Series and Polynomial Architectures

Polynomial and multiplicative interactions are longstanding alternatives to linear maps with pointwise nonlinearities, though largely justified empirically rather than by target-function structure (Lin et al., 2015; Chrysos et al., 2021). Among attention-free token mixers, MLP-Mixer (Tolstikhin et al., 2021) mixes tokens with a learned linear map, while gMLP (Liu et al., 2021) uses multiplicative gating. Its spatial gating unit is a special case of our token mixer with one identity projection.

The discrete Volterra series. For $u \in \mathbb{R}^d$, the order-K discrete Volterra expansion is expressed as,

$$y_c \;=\; \sum_{k=1}^{K} \Bigg[\sum_{j_1=1}^{d} \cdots \sum_{j_k=1}^{d} \Big[W^{(k)}_{c,\, j_1,\ldots,j_k} \prod_{m=1}^{k} u_{j_m} \Big] \Bigg], \tag{2}$$

where y_c is the c-th output and the kernel $W^{(k)} \in \mathbb{R}^{d_{\text{out}} \times d \times \cdots \times d}$, a $(k+1)$-way tensor, weights every degree-k monomial $u_{j_1} \cdots u_{j_k}$ of the input. Introduced as a functional series for smooth nonlinear systems (Volterra, 1887; Schetzen, 1980), the expansion is, by Stone–Weierstrass (Stone, 1948), dense in continuous functions on compact domains as $K \to \infty$.

Volterra neural networks. Cascaded low-rank Volterra filtering (Roheda & Krim, 2020; Roheda et al., 2024a) first made the discrete Volterra series a competitive deep backbone, reaching state-of-the-art action recognition with far fewer parameters than comparable convolutional networks. Volterra networks have since proved competitive at low parameter budgets in classification, multimodal fusion, restoration and signal processing (Roheda et al., 2026; Ghanem et al., 2023; Roheda et al., 2024b; Bai et al., 2025). Banerjee et al. (2020) define Volterra convolutions on *known* homogeneous manifolds such as the sphere, where the manifold is the signal domain; here it is the *unknown* support of the data. Jaiswal et al. (2026) add Volterra layers to an adapter of a pretrained latent diffusion model, whereas we use them as the entire backbone, on raw pixels with no pretrained component.

3 Theoretical Motivation

3.1 The Denoising Map on the Manifold

Throughout, $\sigma := 1 - t$ and $t \in (0, 1)$, so $\sigma > 0$; notation is summarized in Table 3. The optimal denoiser at time t is the conditional mean $f_t(z) = \mathbb{E}[x \mid z_t = z]$, which is both the minimum mean-squared-error (MMSE) estimate of the clean signal x from the noised observation $z_t = z$, and the population minimizer of the training objective. Its second role holds because, at each t, the velocity loss of Section 2.1 equals the x-prediction error multiplied by σ^{-2}, so both losses share the minimizer f_t (Appendix B.2). Asking *which architecture approximates f_t efficiently* is therefore asking which

architecture is matched to the denoising task. Because the forward process adds Gaussian noise, $z_t \mid x \sim \mathcal{N}(tx, \sigma^2 I)$, Bayes' rule turns this mean into an average over the manifold, each clean point x weighted by how closely tx matches z, and its Jacobian is the scaled posterior covariance,

$$f_t(z) \;=\; \frac{\int_{\mathcal{M}} x\, e^{-\|z-tx\|^2/2\sigma^2}\, d\rho_{\mathrm{data}}(x)}{\int_{\mathcal{M}} e^{-\|z-tx\|^2/2\sigma^2}\, d\rho_{\mathrm{data}}(x)}, \qquad \nabla_z f_t(z) \;=\; \frac{t}{\sigma^2}\,\mathrm{Cov}_{\,x\sim p(\cdot\mid z_t=z)}(x). \tag{3}$$

So $f_t(z)$ is a soft projection of the observation onto the data manifold. Two structural facts drive the analysis. First, because $\sigma > 0$ the Gaussian weight is strictly positive and real-analytic, so f_t is real-analytic on all of $\mathbb{R}^D$ (Appendix C) and its smoothness order s may be taken arbitrarily large. Second, the Jacobian of f_t is exactly the scaled posterior covariance (right of equation 3), a denoising analogue of Tweedie's formula (Robbins, 1956; Efron, 2011). As the posterior is supported on the $d_{\mathcal{M}}$-dimensional manifold, this covariance is concentrated on about $d_{\mathcal{M}}$ tangent directions, so f_t varies quickly along $\mathcal{M}$ and slowly transverse to it. The first fact governs the approximation *rate*, the second its *cost*.

Why a polynomial class. For a signal in Gaussian noise the MMSE estimator is classically a series of polynomial functionals of the observation, the Wiener–Volterra series (Wiener, 1958; Schetzen, 1980), and here that expansion can be written explicitly. Write $\tilde{\rho}$ for the data distribution ρ_{data} reweighted by $e^{-t^2\|x\|^2/2\sigma^2}$ and renormalized, a fixed measure that does not depend on the observation. The denoiser is the mean of $\tilde{\rho}$ under the exponential tilt $e^{\theta^\top x}$ with natural parameter $\theta := tz/\sigma^2$, and expanding that mean in θ about $\theta = 0$ gives

$$f_t(z) \;=\; \kappa_1 \;+\; \tfrac{t}{\sigma^2}\,\kappa_2\, z \;+\; \tfrac{1}{2}\big(\tfrac{t}{\sigma^2}\big)^2 \kappa_3[z,z] \;+\; \cdots, \tag{4}$$

a Volterra series whose order-n kernel is the $(n{+}1)$-st cumulant κ_{n+1} of $\tilde{\rho}$, scaled by $(t/\sigma^2)^n/n!$ (Appendix B.1). The series is local, converging only for $\|z\|$ below a radius set by $\tilde{\rho}$, so it identifies the polynomial class, and Proposition 1 supplies the approximation on all of $t\mathcal{M}$.

Proposition 1 (Polynomial approximation of the denoising map). *Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact C^∞ submanifold of intrinsic dimension $d_{\mathcal{M}}$, ρ_{data} a probability measure on $\mathcal{M}$, and f_t the optimal denoiser at $t \in (0,1)$.* (a) Regularity. *f_t is real-analytic on $\mathbb{R}^D$.* (b) Approximation. *There is a linear projection $\pi : \mathbb{R}^D \to \mathbb{R}^{d'}$, $d' = 2d_{\mathcal{M}} + 1$, whose restriction to the scaled manifold $t\mathcal{M}$, on which the noised inputs concentrate, is a smooth embedding for every t, such that for every $s > 0$ and degree K there is a polynomial $P_K : \mathbb{R}^{d'} \to \mathbb{R}^D$ with*

$$\sup_{z \in t\mathcal{M}} \big\| f_t(z) - P_K\big(\pi(z)\big)\big\|^2 \;\le\; C \cdot K^{-2s}.$$

Since P_K has $N = \binom{d'+K}{d'} \asymp K^{2d_{\mathcal{M}}+1}$ coefficients per output coordinate, the error is $O(N^{-2s/(2d_{\mathcal{M}}+1)})$, set by $d_{\mathcal{M}}$ rather than by D, whereas a polynomial on the raw patch would need $\asymp K^D$. The constant $C = C(s,t)$ depends on t, the reach and curvature of $t\mathcal{M}$ and $\|f_t\|_{C^s}$, but not on D.

The rate depends on the intrinsic dimension $d_{\mathcal{M}}$ and, by part (a), is unbounded in the smoothness s. On $t\mathcal{M}$ the four targets are affine reparametrizations of one another, so the rate holds for ϵ-, v- and score prediction too; off $t\mathcal{M}$ only x survives the bottleneck (Appendix B.2). Proposition 1 studies f_t on the scaled manifold $t\mathcal{M}$, but the network is evaluated at the noised input z_t in a radius-δ tube $\mathcal{T}_\delta$ around it. A compressing linear map π discards directions normal to $t\mathcal{M}$, and the Jacobian identity (right of equation 3) shows these directions contribute little. In a unit direction u, the sensitivity of f_t is $\frac{t}{\sigma^2}\, u^\top \mathrm{Cov}(x)\, u$ (the posterior variance along u), which is small for normal directions once the noise falls below the manifold reach τ. Lemma 1 (Appendix C.4) formalizes this and bounds the extra error from replacing the nonlinear retraction onto $t\mathcal{M}$ with the linear bottleneck π.

Combined with Proposition 1, the lemma gives the bound on $\mathcal{T}_\delta$ that the network needs,

$$\sup_{z \in \mathcal{T}_\delta} \big\| f_t(z) - P_K\big(\pi(z)\big)\big\|^2 \;\le\; \underbrace{C \cdot K^{-2s}}_{\text{polynomial approximation}} \;+\; \underbrace{C' \cdot \big(\eta(t,\sigma,\tau) + c_\pi/t\big)^2\, \delta^2}_{\text{compression cost}}. \tag{5}$$

The first term is the rate of Proposition 1. The second, the price of compression, has two sources. The factor $\eta(t,\sigma,\tau) = C_\eta \min\{1/t,\ \sigma^2/(t^3\tau^2)\}$ bounds how much f_t changes along the normal

directions that π discards, and is small once $\sigma \ll t\tau$. The constant c_π covers the tangential component in the kernel of π and the error of evaluating P_K slightly off the manifold, and it shrinks as the bottleneck widens, one reason the trained d' exceeds $2d_{\mathcal{M}} + 1$. Neither depends on σ, δ, or D. The normal component of the noise has length about $\sigma\sqrt{D - d_{\mathcal{M}}}$, so in the worst case the compression term is $O((D - d_{\mathcal{M}})\,\sigma^2/t^2)$, which vanishes late in sampling but is not free of D. Averaged over the noise, its slippage part depends on d' rather than D (Remark 6). All proofs are in Appendix C.

Behaviour across the sampling schedule. These results hold at a fixed t, and Appendix B.3 checks that the bound stays controlled as sampling runs t from 0 to 1. Early in sampling f_t is nearly affine, and late in sampling the posterior is concentrated and the compression cost vanishes. The hard case is intermediate t, where the posterior spreads over a curved patch whose tangent spaces together span more than $d_{\mathcal{M}}$ directions. The bottleneck must cover that span, a second reason d' exceeds $2d_{\mathcal{M}} + 1$. The required degree is also highest there but finite, and depth supplies it. None of this involves D.

Proposition 2 (Attainment, and the polynomial content of attention). *With the setting of Proposition 1, suppose ρ_{data} is not a point mass and fix $L \geq 1$. Write $K_L := 2^L$ and $\mathcal{P}_K$ for the polynomial maps $\mathbb{R}^{d'} \to \mathbb{R}^D$ of degree at most K.* (i) Attainment. *Let $\mathcal{V}_{L,\mathsf{P}}$ be the depth-L stacks of the low-rank Volterra mixers of Section 4 (degree 2 per layer, gated residual, affine embedding, linear head, no normalization) applied to $\pi(z)$, with at most P parameters per layer. For $\mathsf{P} \geq \mathsf{P}_0(L, d') \asymp \binom{d'+K_L}{d'}^2$,*

$$\left\{\, Q \circ \pi \;:\; Q \in \mathcal{P}_{K_L} \,\right\} \;\subseteq\; \mathcal{V}_{L,\mathsf{P}}, \qquad \textit{hence} \qquad \inf_{g \in \mathcal{V}_{L,\mathsf{P}}} \sup_{z \in t\mathcal{M}} \|f_t(z) - g(z)\|^2 \;\leq\; C \cdot K_L^{-2s},$$

independently of D, and on $\mathcal{T}_\delta$ up to the compression term of equation 5. (ii) Attention. *A softmax self-attention layer with residual connection (any heads and projections) that agrees with a polynomial map on a nonempty open set is affine, and f_t is not affine. So the class $\{Q \circ \pi : Q \in \mathcal{P}_{K_L}\}$, which $\mathcal{V}_{L,\mathsf{P}}$ represents exactly, meets the attention class only in its affine part, at every budget.*

Part (i) shows that the low-rank factorization loses nothing. Above the threshold the class equals $\mathcal{P}_{K_L}$, since it contains every polynomial of degree K_L and a depth-L stack of degree-2 layers has no higher degree. The proof (Appendix C.5) is constructive, since iterated squaring raises an affine form to the power 2^L in L layers and such powers span the polynomials. The threshold P_0 certifies representability and is not a practical width. For part (ii) (Appendix C.6), the infinite Taylor series of the softmax is not extra capacity. An analytic function whose power series does not terminate agrees with no polynomial, which puts attention outside the class rather than above it. That attention also attains a slower rate at matched budget is natural to expect, and we state it as Conjecture 1.

3.2 Beyond Softmax

Proposition 2(ii) is specific to the softmax, but the separation is not. Proposition 3 (Appendix C.7) extends its global form to any mixing layer whose weights are nonnegative and sum to one, however computed. Such a layer averages values that are linear in the input, so its output grows at most linearly, and a polynomial on the whole input space that does so is affine; for softmax, analyticity makes an open set enough. The result covers linear attention with a nonnegative feature map (Katharopoulos et al., 2020; Xie et al., 2025), kernel approximations of the softmax (Choromanski et al., 2021), and any other rule that averages. Appendix B.4 treats unnormalized attention, where it fails.

Where the nonlinearity comes from. A Transformer block splits the work between two components. The attention sublayer mixes tokens, but its polynomial content is only affine (Propositions 2(ii) and 3). The feed-forward sublayer supplies the nonlinearity but acts on each token separately. Applied to the attention output, its activation does produce cross-token polynomial terms, but not a free parametrization of them, since every slot draws on the same attention weights, the activation fixes the coefficients across orders, and a wider sublayer buys channel rank but no new token structure (Table 5, Appendix B.6). The Volterra token mixer mixes tokens and raises the degree in the same step by multiplying two tokens, and the channel mixer does the same within a token. The claim is therefore not that a Transformer cannot approximate f_t, which it can, but that its mixing operator contributes no polynomial content beyond affine, so the polynomial structure of the target must be recovered through a pointwise nonlinearity whose orders are coupled rather than free.

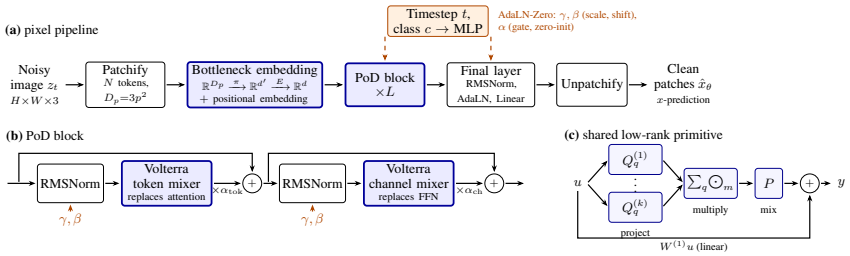

Figure 1: **The PoD architecture.** (a) pixel pipeline; (b) block, both Transformer sublayers replaced by Volterra mixers; (c) the primitive they share, equation 6.

4 Polynomial Denoisers on the Image Manifold

The architecture realizes the construction of Proposition 1. Its projection π is a bottleneck embedding (Section 4.1), and the degree-K polynomial P_K is a stack of low-rank Volterra layers alternating a cross-token and a channel-wise mixer (Section 4.3), as summarized in Figure 1.

4.1 Patch Tokenization and Manifold Embedding

Given an image of shape $H \times W \times 3$, we partition it into non-overlapping $p \times p$ patches, giving $N = HW/p^2$ tokens of ambient dimension $D_p = 3p^2$ ($N = 256$ in all our models).

The patch embedding is a low-rank linear map factored through a bottleneck, $\mathbb{R}^{D_p} \xrightarrow{\pi} \mathbb{R}^{d'} \xrightarrow{E} \mathbb{R}^d$, where $d' \ll D_p$ is the bottleneck dimension and $d \geq d'$ the hidden width of the mixers that follow. Sinusoidal positional embeddings (Vaswani et al., 2017) are added after the embedding. Such bottlenecks improve sample quality in pixel-space diffusion (Li & He, 2026), and Proposition 1 offers a theoretical explanation.

The bottleneck is also why the network predicts the clean image. The four prediction targets share one function class, but only the clean image lies on the manifold, where the compressed representation can produce it, whereas a noise or velocity output must span the ambient space (Appendix B.2).

4.2 The Low-Rank Volterra Primitive

Both mixers in this section are instances of a single low-rank Volterra primitive (Figure 1c). Rather than storing a dense order-k kernel, the primitive performs a *project–multiply–mix* operation. Write n for the size of the axis being mixed, $n = N$ for the cross-token mixer and $n = d$ for the channel mixer. The input $u \in \mathbb{R}^n$ is projected to a hidden width h by R groups of k independent factors $Q_q^{(1)}, \ldots, Q_q^{(k)} \in \mathbb{R}^{h \times n}$, the k projections of each group are multiplied elementwise, and the R products are summed and mixed back through a factor $P \in \mathbb{R}^{d_{\text{out}} \times h}$,

$$y = W^{(1)}u + P \sum_{q=1}^{R} \bigodot_{m=1}^{k} \left(Q_q^{(m)} u\right), \qquad W^{(k)}_{i,j_1,\ldots,j_k} = \sum_{c=1}^{h} \sum_{q=1}^{R} P_{ic} \prod_{m=1}^{k} Q^{(m)}_{q,\,cj_m}. \tag{6}$$

Each coordinate $(Q_q^{(m)}u)_c$ is a learned mixture of *all* inputs, so the product produces every degree-k monomial, cross-terms included. The product term is therefore the order-k Volterra term equation 2, and $W^{(k)}$ becomes its order-k kernel, a canonical-polyadic factorization of rank at most hR with an independent (untied) factor per slot, which constrains the kernel but not the set of interactions it represents (Appendix D). Evaluation costs $\mathcal{O}(k\,nhR)$ per kernel against $\mathcal{O}(n^{k+1})$ for a dense one. We use $k = 2$ throughout, since depth already supplies degree 2^L (Proposition 2(i)).

4.3 Cross-Token and Channel-Wise Mixing

The block instantiates the primitive twice, once along each axis. Write u_{ia} for channel a of token i, with $i = 1, \ldots, N$ and $a = 1, \ldots, d$. The token mixer takes the N positions as slots at fixed channel,

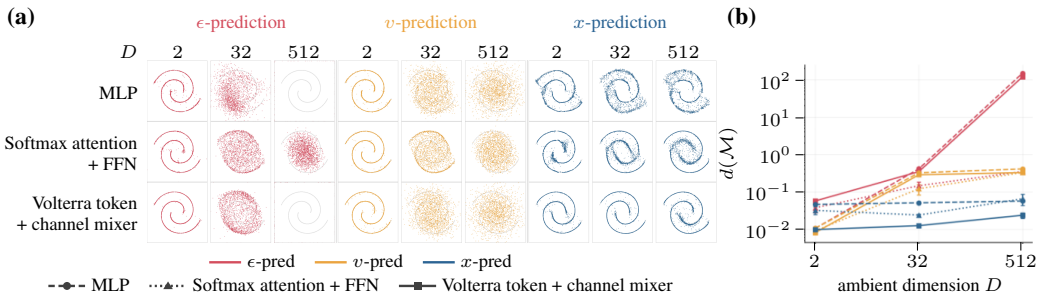

Figure 2: **The manifold picture in a controlled setting.** (a) Samples from a $d_{\mathcal{M}} = 1$ spiral in $\mathbb{R}^D$ (grey) for three targets and three architectures. (b) Mean distance to the manifold $d(\mathcal{M})$ against D.

and the channel mixer takes the d channels of one token as slots at fixed position,

$$y_{ia} = \sum_j W^{(1)}_{\text{tok},ij}\, u_{ja} + \sum_{j,j'} W^{(2)}_{\text{tok},ijj'}\, u_{ja} u_{j'a}, \qquad y_{ia} = \sum_b W^{(1)}_{\text{ch},ab}\, u_{ib} + \sum_{b,c} W^{(2)}_{\text{ch},abc}\, u_{ib} u_{ic}. \quad (7)$$

Each mixer shares its kernels along the axis it does not mix (W_{tok} carries no channel index, W_{ch} no position index), and each is an instance of equation 6 (Appendix D). The token mixer has $h = N$, $P = I$ and dense $Q_q^{(m)} \in \mathbb{R}^{N\times N}$, so it holds $(2R+1)N^2$ parameters per block for any width d, against $4d^2$ for self-attention, and costs $\mathcal{O}(RN^2 d)$ against $\mathcal{O}(N^2 d + Nd^2)$. The channel mixer has hidden width h and rank R', and costs $\mathcal{O}(dhR')$ per token.

4.4 Block Structure and Conditioning

A polynomial denoiser block applies, in order: RMSNorm (Zhang & Sennrich, 2019) with affine modulation, cross-token polynomial mixer, residual connection, RMSNorm with modulation, channel-wise polynomial mixer, residual connection (Figure 1b). The pre-norm residual structure follows standard practice for stably training deep residual networks.

Diffusion time t and class label c enter through unmodified AdaLN-Zero conditioning (Peebles & Xie, 2023). A small MLP of (t, c) produces, for each sublayer, a scale γ and shift β applied as $\gamma \odot \text{RMSNorm}(\cdot) + \beta$, and a zero-initialized gate α on the sublayer output, so that each block starts as the identity.

No activation function. The residual path contains no activation. The only nonlinearity is the degree-2 product, which takes the place of the expansion–activation–contraction of a feed-forward block, and GELU-type functions appear only in the conditioning MLP. This is required rather than stylistic, since a pointwise activation on the residual path would leave the polynomial class and void Proposition 2(i). In the ablation of Section 5.4, replacing the feed-forward channel sublayer with the activation-free Volterra mixer lowers FID.

The RMSNorm and the conditioning branch remain outside the polynomial class, but both are shared with the attention baseline and neither mixes tokens. At fixed (t, c) the conditioning acts as a diagonal affine map, and RMSNorm is a positive per-token rescaling, so Proposition 2(i) holds exactly for the normalization-free class and the rate of Proposition 1 is unaffected (Appendix B.5).

5 Experiments

5.1 The Manifold Picture in a Controlled Setting

Before ImageNet, we test the theory on a known manifold with adjustable ambient dimension. We embed a one-dimensional spiral ($d_{\mathcal{M}} = 1$) in $\mathbb{R}^D$, $D \in \{2, 32, 512\}$, and compare the MLP of Li & He (2026), a softmax attention block and a Volterra block at matched parameter count. We report the mean distance of generated points to the manifold, $d(\mathcal{M})$, which measures whether they lie on the manifold at all. Figure 2 reproduces the failure pattern of Li & He (2026). Under ϵ- and v-prediction, $d(\mathcal{M})$ grows by one to more than three orders of magnitude as D increases and the samples lose the



Figure 3: **Class-conditional samples** from PoD-G/16 on ImageNet 256×256.

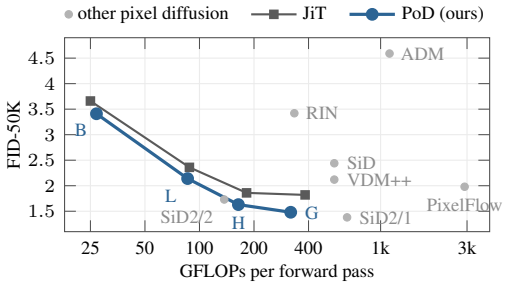

Figure 4: **Quality against compute on ImageNet** 256×256**.** PoD and JiT at scales B–G, which differ only in the mixing operator, and other pixel-space diffusion models.

spiral, whereas under x-prediction every denoiser stays within a factor of about three of its $D = 2$ value. Attention degrades rather than collapses under ϵ-prediction, its output being confined to the hull of its values, the boundedness that caps its polynomial content at affine (Proposition 3). Under x-prediction the Volterra block attains the lowest $d(\mathcal{M})$ at every D, by a factor of roughly two to five over the MLP and attention, across a 256-fold increase in D. The polynomial class is matched to the geometry of f_t, and what it buys is fidelity to the manifold.

5.2 Training Details

We train class-conditional models on ImageNet-1K at 256×256 and 512×512 with patch sizes 16 and 32, so $N = 256$ in both, and report FID and Inception Score on 50K samples against the ADM reference statistics (Dhariwal & Nichol, 2021). Four scales, B, L, H and G, match the parameters and FLOPs of Li & He (2026), with $d' = 128$ for B, L and 256 for H, G, token rank $R = 10$, and a channel mixer of hidden width $h = 2d$ and rank $R' = 1$ (Appendix D). The x-prediction head is trained under the velocity loss of Section 2.1 for 800 epochs with classifier-free guidance (Ho & Salimans, 2022) and sampled with a 50-step Heun solver (Heun, 1900; Karras et al., 2022); the time distribution, label dropout and sampler settings are in Appendix E. Each quadratic branch carries a learned per-channel scale initialized to 0.1, which can be absorbed into $W^{(2)}$ and leaves the function class unchanged. Our primary comparison is JiT (Li & He, 2026), which shares our recipe and differs only in the mixing operator, with other models given for context.

5.3 Main Results on ImageNet

PoD is ahead of JiT in FID at every scale at matched parameters and FLOPs, most at G (1.48 against 1.82; Table 1, Figure 4), and the ordering holds at 512×512, where PoD-G/32 reaches 1.55 against 1.78 (Table 6, Appendix F). Among pixel-space models without pretrained components, PoD-G/16 is the strongest at comparable compute. SiD2-UViT/1 reaches 1.38 but at 653 GFLOPs against our 319, and SiD2-UViT/2 lies on the PoD curve between L and H. The best latent-space model, RAE at 1.13 (Zheng et al., 2026), remains ahead but relies on a pretrained tokenizer, a perceptual loss and self-supervised features. Figure 3 shows samples from PoD-G/16. More samples at both resolutions are in Appendix G.

5.4 Architecture Ablation

Table 2 replaces the two Transformer sublayers one at a time. The Volterra token mixer in place of self-attention lowers FID from 3.66 to 3.54, and the activation-free Volterra channel mixer in place

Table 1: ImageNet 256×256, FID and IS on 50K samples. Pre-training columns list the external models each method needs, and parameters include the tokenizer decoder where used.

Method	pre-training			params	GFLOPs	FID ↓	IS ↑
	token	perc.	self-sup.				
Latent-space diffusion							
DiT-XL/2 (Peebles & Xie, 2023)	SD-VAE	VGG	–	675+49M	119	2.27	278.2
REPA, SiT-XL/2 (Yu et al., 2025)	SD-VAE	VGG	DINOv2	675+49M	119	1.42	305.7
LightningDiT-XL/2 (Yao et al., 2025)	VA-VAE	VGG	DINOv2	675+49M	119	1.35	295.3
DDT-XL/2 (Wang et al., 2026b)	SD-VAE	VGG	DINOv2	675+49M	119	1.26	310.6
RAE, DiT^{DH}-XL/2 (Zheng et al., 2026)	RAE	VGG	DINOv2	839+415M	146	1.13	262.6
Pixel-space, non-diffusion							
JetFormer (Tschannen et al., 2025)	–	–	–	2.8B	–	6.64	–
FractalMAR-H (Li et al., 2025)	–	–	–	848M	–	6.15	348.9
Pixel-space diffusion							
ADM-G (Dhariwal & Nichol, 2021)	–	–	–	554M	1120	4.59	186.7
RIN (Jabri et al., 2023)	–	–	–	410M	334	3.42	182.0
SiD, UViT/2 (Hoogeboom et al., 2023)	–	–	–	2B	555	2.44	256.3
VDM++, UViT/2 (Kingma & Gao, 2023)	–	–	–	2B	555	2.12	267.7
SiD2, UViT/2 (Hoogeboom et al., 2025)	–	–	–	N/A	137	1.73	–
SiD2, UViT/1 (Hoogeboom et al., 2025)	–	–	–	N/A	653	1.38	–
PixelFlow, XL/4 (Chen et al., 2025)	–	–	–	677M	2909	1.98	282.1
PixNerd, XL/16 (Wang et al., 2026a)	–	–	DINOv2	700M	134	2.15	297.0
JiT-B/16 (Li & He, 2026)	–	–	–	131M	25	3.66	275.1
JiT-L/16 (Li & He, 2026)	–	–	–	459M	88	2.36	298.5
JiT-H/16 (Li & He, 2026)	–	–	–	953M	182	1.86	303.4
JiT-G/16 (Li & He, 2026)	–	–	–	2B	383	1.82	292.6
Pixel-space diffusion, polynomial (ours)							
PoD-B/16	–	–	–	119M	27	3.41	282.1
PoD-L/16	–	–	–	391M	86	2.14	301.6
PoD-H/16	–	–	–	787M	164	1.63	307.9
PoD-G/16	–	–	–	1.6B	319	1.48	290.2

Table 2: Architecture ablation on ImageNet 256×256 at B/16 scale.

Configuration	Token mixer	Channel mixer	FID-50K ↓
JiT baseline	Self-attention	FFN	3.66
Polynomial token only	Volterra $k=2$	FFN	3.54
Full polynomial	Volterra $k=2$	Volterra $k=2$	3.41

of the feed-forward network lowers it further to 3.41, so the gain over JiT is not carried by a single component. The second step removes the only activation in the block and still helps, as Section 3.2 predicts. A Transformer block needs its activation because attention contributes no polynomial nonlinearity, whereas a Volterra mixer supplies its own.

6 Conclusion

The optimal denoiser is locally a Wiener–Volterra series, approximated on the data manifold by a polynomial of a low-dimensional linear projection at a rate set by the intrinsic dimension. Low-rank Volterra stacks represent that class exactly once their width clears an explicit threshold, whereas softmax attention, and any mixing rule that averages, meets it only in its affine part. The resulting architecture, a bottleneck embedding followed by activation-free Volterra mixers, trains on raw pixels without pretrained components and improves on JiT in FID at every scale at 256×256 and 512×512 at matched parameters and FLOPs. The analysis assumes an exact smooth manifold, its worst-case compression term can grow with the ambient dimension, and a rate gap for attention remains a conjecture. Natural extensions are polynomial conditioning, where the diffusion time gates the Volterra kernels, and video and scientific data, where pretrained tokenizers are hard to obtain.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, in revising the statements and steps of proofs developed by the authors. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results, and generating synthetic data sets, cleaning and reformatting datasets, qualitative and thematic data analysis, and translation are not applicable to this work. Additionally, we used generative AI tools to polish and aid the writing of the paper, including editing for readability, for the retrieval and discovery of related work, including searching for and identifying relevant literature, and to create or modify scientific figures. We have reviewed all AI-assisted work. Every reference surfaced with AI assistance was checked against the original source, and all AI-assisted text, proofs and figures were read and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated with their assumptions in Section 3, and complete proofs are given in Appendix C, with supporting derivations in Appendix B. The architecture is specified in Section 4, with the derivations behind the primitive and the mixers in Appendix D. The training objective, schedule and sampler are given in Section 5.2, with pseudocode in Appendix E. All experiments use the public ImageNet-1K dataset with the standard ADM evaluation protocol.

REFERENCES

Lu Bai, Siyuan Lian, Mengtong Li, Yiming He, Li Rao, Xiaofeng Zeng, Ruquan Sun, Kai Chen, and Jing Lu. WaveNet-Volterra neural network for active noise control: A fully causal approach. *Mechanical Systems and Signal Processing*, 241:113486, 2025.

Monami Banerjee, Rudrasis Chakraborty, Jose Bouza, and Baba C. Vemuri. VolterraNet: A higher order convolutional network with group equivariance for homogeneous manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):823–833, 2020.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. PixelFlow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021.

Grigorios G. Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Jiankang Deng, Yannis Panagakis, and Stefanos Zafeiriou. Deep polynomial neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4021–4034, 2021.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Sally Ghanem, Siddharth Roheda, and Hamid Krim. Latent code-based fusion: A Volterra neural network approach. *Intelligent Systems with Applications*, 18:200210, 2023.

Karl Heun. Neue Methoden zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Zeitschrift für Mathematik und Physik*, 45:23–38, 1900.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning (ICML)*, pp. 13213–13232, 2023.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 FID on ImageNet512 with pixel-space diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18062–18071, 2025.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In *International Conference on Machine Learning (ICML)*, 2023.

Rohan Jaiswal, Aniruddha Bala, Siddharth Roheda, and Rohit Chowdhury. GalaxyEdit: Large scale image editing dataset with enhanced diffusion adapter. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11842–11846, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26565–26577, 2022.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, pp. 5156–5165, 2020.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 65484–65516, 2023.

Jiachen Lei, Keli Liu, Julius Berner, Haiming Yu, Hongkai Zheng, Jiahong Wu, and Xiangxiang Chu. There is no VAE: End-to-end pixel-space generative modeling via self-supervised pre-training. In *International Conference on Learning Representations (ICLR)*, 2026.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, 2004.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, 2026.

Tianhong Li, Qinyi Sun, Lijie Fan, and Kaiming He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1449–1457, 2015.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay attention to MLPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9204–9215, 2021.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, pp. 23–40. Springer, 2024.

Donald J. Newman. Rational approximation to $|x|$. *Michigan Mathematical Journal*, 11(1):11–14, 1964.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

David L Ragozin. Polynomial approximation on compact manifolds and homogeneous spaces. *Transactions of the American Mathematical Society*, 150(1):41–53, 1970.

David L Ragozin. Constructive polynomial approximation on spheres and projective spaces. *Transactions of the American Mathematical Society*, 162:157–170, 1971.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 157–163. University of California Press, 1956.

Siddharth Roheda and Hamid Krim. Conquering the CNN over-parameterization dilemma: A Volterra filtering approach for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11948–11956, 2020.

Siddharth Roheda, Hamid Krim, and Bo Jiang. Volterra neural networks (VNNs). *Journal of Machine Learning Research*, 25(182):1–29, 2024a.

Siddharth Roheda, Amit Unde, and Loay Rashid. MR-VNet: Media restoration using Volterra networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6098–6107, 2024b.

Siddharth Roheda, Aniruddha Bala, Rohit Chowdhury, and Rohan Jaiswal. VNODE: A piecewise continuous Volterra neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5896–5900, 2026.

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 (CVPR)*, pp. 10674–10685, 2022.

Edward B. Saff and Richard S. Varga (eds.). *Padé and Rational Approximation: Theory and Applications*. Academic Press, 1977.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Martin Schetzen. *The Volterra and Wiener Theories of Nonlinear Systems*. John Wiley & Sons, 1980.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015.

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

Marshall H. Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21 (5):237–254, 1948.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 24261–24272, 2021.

Michael Tschannen, André Susano Pinto, and Alexander Kolesnikov. JetFormer: An autoregressive generative model of raw images and text. In *International Conference on Learning Representations (ICLR)*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 1096–1103, 2008.

Vito Volterra. *Sopra le funzioni che dipendono da altre funzioni*. Tipografia della R. Accademia dei Lincei, 1887.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. PixNerd: Pixel neural field diffusion. In *International Conference on Learning Representations (ICLR)*, 2026a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40633–40642, 2026b.

Norbert Wiener. *Nonlinear Problems in Random Theory*. The MIT Press, 1958.

Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with ReLU in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14273–14282, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX

A NOTATION

Table 3: Summary of notation.

Symbol	Meaning
Dimensions and indices	
$D = D_p$	ambient dimension of the theory, the per-patch pixel dimension
$D_p = 3p^2$	ambient dimension of one $p \times p$ RGB patch ($D_p \in \{768, 3072\}$)
p	patch side length ($p = 16$ at 256×256, $p = 32$ at 512×512)
$d_{\mathcal{M}}$	intrinsic dimension of the data manifold $\mathcal{M}$, with $d_{\mathcal{M}} \ll D$
d'	per-token bottleneck dimension ($d' \in \{128, 256\}$); $2d_{\mathcal{M}} + 1$ suffices in Proposition 1, and the trained value is deliberately larger (Appendix C)
d	hidden width of the network, the working dimension of all token and channel mixers; $d \geq d'$
N	number of patch tokens per image ($N = 256$)
N	number of coefficients of a degree-K polynomial on $\mathbb{R}^{d'}$, per output coordinate; $N = \binom{d'+K}{d'} \asymp K^{2d_{\mathcal{M}}+1}$
$R,\ R'$	rank of the Volterra primitive equation 6 in the token and channel mixers, the number of products summed per hidden unit ($R = 10$, $R' = 1$)
n_{h}	number of attention heads (Appendices B.6 and C.6)
Diffusion and denoising	
$x,\ \epsilon$	clean image $x \sim \rho_{\mathrm{data}}$ and noise $\epsilon \sim \mathcal{N}(0, I)$
$t,\ \sigma$	flow time $t \in (0, 1)$ and noise level $\sigma := 1 - t$
z_t	noised observation, $z_t = tx + (1 - t)\epsilon$
ρ_{data}	data distribution; $d\rho_{\mathrm{data}}(x)$ denotes integration against it
$f_t(z)$	optimal (MMSE) denoiser, $f_t(z) = \mathbb{E}[x \mid z_t = z]$
$\hat{x}_\theta$	the network's prediction of the clean image x
$\tilde{\rho}$	ρ_{data} reweighted by $e^{-t^2\|x\|^2/2\sigma^2}$ and renormalized
θ	natural parameter $\theta := tz/\sigma^2$ of the tilt of $\tilde{\rho}$ (unrelated to the network parameters in $\hat{x}_\theta$)
Λ	cumulant generating function of $\tilde{\rho}$, $\Lambda(\theta) = \log \int e^{\theta^\top x}\, d\tilde{\rho}(x)$
κ_n	n-th cumulant tensor of $\tilde{\rho}$; the order-n Volterra kernel of f_t is κ_{n+1} scaled by $(t/\sigma^2)^n/n!$
Manifold and approximation	
$\mathcal{M} \subset \mathbb{R}^D$	data manifold; τ its reach
π	linear bottleneck projection $\mathbb{R}^D \to \mathbb{R}^{d'}$ (Proposition 1)
$\rho,\ \rho_{\min}$	density of ρ_{data} with respect to the volume measure of $\mathcal{M}$, and its lower bound (Lemma 1)
τ_π	reach of the projected manifold $\pi(\mathcal{M})$ (Lemma 1(b))
$\delta,\ \mathcal{T}_\delta$	tube radius, and the tube $\{z : \mathrm{dist}(z, t\mathcal{M}) \leq \delta\}$ about the scaled manifold
$r,\ \mathrm{np}$	nearest-point retraction onto $t\mathcal{M}$ in $\mathbb{R}^D$, and nearest-point projection onto $\pi(t\mathcal{M})$ in $\mathbb{R}^{d'}$
$\eta,\ C_\eta$	normal-sensitivity factor of Lemma 1, $\eta = C_\eta \min\{1/t,\ \sigma^2/(t^3\tau^2)\}$, and its constant
c_π	slippage constant of Lemma 1(b), set by π and $\mathcal{M}$
$s,\ K$	smoothness order of f_t; polynomial degree
P_K	degree-K polynomial approximant $\mathbb{R}^{d'} \to \mathbb{R}^D$ of Proposition 1 (distinct from the factor P)
$\mathcal{P}_K$	space of polynomial maps $\mathbb{R}^{d'} \to \mathbb{R}^D$ of total degree at most K
K_L	$K_L = 2^L$, the degree reached by a depth-L stack of degree-2 layers
Architecture	
L	number of stacked PoD blocks
k	Volterra order (degree) per layer, default $k = 2$
n	size of the axis a mixer acts on: $n = N$ for the token mixer, $n = d$ for the channel mixer
h	hidden width of the primitive ($h = N$ in the token mixer, $h = 2d$ in the channel mixer)
$P,\ Q_q^{(m)}$	factors of the primitive, $Q_q^{(m)} \in \mathbb{R}^{h \times n}$ projects input slot m of group q and $P \in \mathbb{R}^{d_{\mathrm{out}} \times h}$ mixes back (tied variant, $Q_q^{(m)}$ equal across m)
$W_{\mathrm{tok}}^{(k)},\ W_{\mathrm{ch}}^{(k)}$	order-k Volterra kernels of the token and channel mixers
$\alpha_{\mathrm{tok}}, \alpha_{\mathrm{ch}}$	per-channel, zero-initialized residual gates
$\gamma,\ \beta$	AdaLN scale and shift computed from the time t and class label c, applied as $\gamma \odot \mathrm{RMSNorm}(\cdot) + \beta$

Table 3, continued.

Symbol	Meaning
Architecture	
P	per-layer parameter budget in the function-class statements of Section 3 (distinct from the factor P)
$\mathcal{V}_{L,\mathsf{P}}$	depth-L stacks of low-rank Volterra mixers with at most P parameters per layer (Proposition 2); P_0 is the attainment threshold

B Extended Derivations and Discussion

This appendix holds material referenced from Section 3 in summary form: the derivation of the cumulant expansion and the condition under which it terminates, the all-order Tweedie identity, the prediction-target discussion, the covering argument across the sampling schedule, the extended comparison against attention-shaped operators, the role of normalization and conditioning in our block, and the feed-forward objection. It gives derivations and discussion; the proofs themselves are in Appendix C.

B.1 The cumulant expansion, in full

This subsection derives the expansion quoted as Equation equation 4 in Section 3, verifies it against the Jacobian identity, and extends it to every order of differentiation.

The three-term split. Expanding the square in the exponent of the denoiser formula (left of equation 3) separates the observation from the signal,

$$-\frac{\|z-tx\|^2}{2\sigma^2} \;=\; \underbrace{-\frac{\|z\|^2}{2\sigma^2}}_{z\text{ alone}} \;+\; \underbrace{\frac{t}{\sigma^2}\,z^\top x}_{\text{the coupling}} \;-\; \underbrace{\frac{t^2}{2\sigma^2}\,\|x\|^2}_{x\text{ alone}}, \tag{8}$$

and the three terms play different roles. The first depends on z alone, so it factors out of both integrals on the left of equation 3 and cancels between them. The third depends on x alone, so rather than being carried through the calculation it can be absorbed into the measure one integrates against: write $\tilde\rho$ for the data distribution reweighted by $e^{-t^2\|x\|^2/2\sigma^2}$ and renormalized, a fixed measure that mildly down-weights images of large norm and, crucially, does not depend on the observation. Only the middle term couples z to x, and it does so *linearly*, through the single vector $\theta := tz/\sigma^2$. Substituting into the left of equation 3 and cancelling the z-only factor and the normalizing constant of $\tilde\rho$ between numerator and denominator leaves

$$f_t(z) \;=\; \frac{\int x\, e^{\theta^\top x}\, d\tilde\rho(x)}{\int e^{\theta^\top x}\, d\tilde\rho(x)}.$$

The tilt is a gradient. The right-hand side is the mean of $\tilde\rho$ reweighted by $e^{\theta^\top x}$, an *exponential tilt* of $\tilde\rho$ with natural parameter θ. Such a mean is always a gradient. Setting $\Lambda(\theta) := \log \int e^{\theta^\top x}\, d\tilde\rho(x)$ and differentiating under the integral,

$$\nabla_\theta \Lambda(\theta) \;=\; \frac{\int x\, e^{\theta^\top x}\, d\tilde\rho(x)}{\int e^{\theta^\top x}\, d\tilde\rho(x)},$$

which is exactly the mean of $\tilde\rho$ under the tilt. Hence

$$f_t(z) \;=\; \nabla_\theta \Lambda(\theta)\big|_{\theta=\frac{t}{\sigma^2}z}, \qquad \Lambda(\theta) := \log \int_{\mathcal{M}} e^{\theta^\top x}\, d\tilde\rho(x), \qquad d\tilde\rho \;\propto\; e^{-\frac{t^2}{2\sigma^2}\|x\|^2}\, d\rho_{\text{data}}. \tag{9}$$

Taking the logarithm is what turns the unnormalized integral into a mean: with $M(\theta) := \int e^{\theta^\top x} d\tilde\rho$ we have $\nabla_\theta \Lambda = \nabla_\theta M / M$, and it is the denominator that normalizes the weights. By definition Λ is then the *cumulant generating function* of $\tilde\rho$, and its Taylor coefficients at $\theta = 0$ are the cumulant tensors κ_n. κ_1 is the mean, κ_2 the covariance, and the higher κ_n are n-way arrays, symmetric in their indices. Differentiating that Taylor expansion once and substituting $\theta = tz/\sigma^2$ gives equation 4, whose kernels are therefore named rather than left abstract.

The expansion is local. The Taylor series of Λ at $\theta = 0$, and with it equation 4, has a finite radius of convergence in general. Since $\tilde{\rho}$ has compact support, M extends to an entire function of $\theta \in \mathbb{C}^D$ that is positive on $\mathbb{R}^D$, so Λ is real-analytic on all of $\mathbb{R}^D$. But $\Lambda = \log M$ is singular wherever M vanishes, and M generally has complex zeros, so along each real direction the series in θ converges only up to the nearest zero on the complex line through that direction. For the uniform measure on a circle of radius a about the origin (a case in which $\tilde{\rho} = \rho_{\text{data}}$, since every point has the same norm), $M(\theta) = I_0(a\,\|\theta\|)$ for real θ in the plane of the circle, and its nearest complex zeros lie at $\|\theta\| = j_{0,1}/a$, with $j_{0,1} \approx 2.405$ the first zero of the Bessel function J_0. With $\theta = tz/\sigma^2$ and $\|z\| = ta$ on $t\mathcal{M}$, the series converges on $t\mathcal{M}$ only while $t^2a^2/\sigma^2 < j_{0,1}$, and diverges there late in sampling. The expansion therefore identifies the polynomial class and names its kernels, but it is not the approximant. Proposition 1 rests instead on the real-analyticity of f_t on all of $\mathbb{R}^D$, which holds at every t, and its polynomial P_K is not a truncation of the series (see the remark after its proof in Appendix C).

When the expansion terminates. The expansion terminates exactly when $\tilde{\rho}$ is Gaussian, since a Gaussian is characterized by the vanishing of all cumulants beyond the second. For $\tilde{\rho} = \mathcal{N}(\mu, S)$ one gets $\Lambda(\theta) = \theta^\top\mu + \frac{1}{2}\theta^\top S\theta$ and hence $f_t(z) = \mu + \frac{t}{\sigma^2}Sz$ with no truncation. Substituting the precision addition $S^{-1} = \Sigma_\rho^{-1} + \frac{t^2}{\sigma^2}I$ recovers the classical Wiener filter, the linear MMSE estimator that the Wiener–Volterra series generalizes. This case sits outside our setting, which requires compact support (Section 2.2), and the contrast is informative. With compact support, f_t is affine only if ρ_{data} is a point mass (Lemma 6), so within our setting the expansion never terminates, i.e. the optimal denoiser lies in the closure of the polynomial class rather than in the class itself, approximated to arbitrary accuracy by polynomials of sufficient degree but never equal to one. This is why Proposition 1 is an approximation result. What classical theory leaves open is the degree required, and this is what Proposition 1 settles. The real-analyticity of f_t makes the best degree-K approximation error decay faster than any fixed power of $1/K$, while the low intrinsic dimension of $\mathcal{M}$ makes the degree, and hence the parameter count, depend on $d_{\mathcal{M}}$ rather than the ambient D.

An all-order Tweedie identity. The cumulant form equation 9 sharpens the Jacobian identity (right of equation 3) to every order of differentiation. Since $f_t(z) = \nabla_\theta\Lambda(\theta)$ at $\theta = tz/\sigma^2$, each derivative in z brings down a factor t/σ^2 by the chain rule, and the higher derivatives of a cumulant generating function at parameter θ are, by definition, the cumulants of the measure tilted by $e^{\theta^\top x}$, which at $\theta = tz/\sigma^2$ is exactly the posterior $p(\cdot \mid z_t = z)$. Hence

$$\nabla_z^k f_t(z) \;=\; \Big(\tfrac{t}{\sigma^2}\Big)^k \,\kappa_{k+1}\big(p(\cdot \mid z_t = z)\big), \qquad k \geq 1, \tag{10}$$

of which the right of equation 3 is the case $k = 1$, the second cumulant being the covariance. Both identities are proved formally in Appendix C; the derivation here is the informal route. The identity makes the t-dependence of the approximation theory explicit rather than opaque: the norm $\|f_t\|_{C^s}$, which controls the constant $C(s,t)$ of Proposition 1(b), is by equation 10 a statement about the posterior cumulants up to order $s{+}1$, scaled by powers of t/σ^2, so uniform-in-t control of the constant is equivalent to uniform-in-t bounds on posterior cumulants. Both ends of the schedule are benign on this reading. Late in sampling the posterior concentrates at scale σ/t and near-Gaussianity leaves $\kappa_{k+1} = O\big((\sigma^2/t^2)^k\,\tau^{-(k-1)}\big)$, with τ the reach of $\mathcal{M}$ introduced with Lemma 1, so the powers of σ cancel exactly and $\nabla_z^k f_t = O\big(t^{-k}\tau^{-(k-1)}\big)$; early in sampling the posterior relaxes to ρ_{data}, whose cumulants are bounded on the compact $\mathcal{M}$, and $\nabla_z^k f_t = O(t^k)$. Any growth of $C(s,t)$ is therefore confined to intermediate t, consistent with the covering argument of Appendix B.3.

The denominator. The denominator on the left of equation 3 is the marginal density $p_{z_t}(z) = \int_{\mathcal{M}} p(z \mid x)\, d\rho_{\text{data}}(x)$, which normalizes the weights. The Gaussian prefactor $(2\pi\sigma^2)^{-D/2}$ is common to numerator and denominator and cancels, so the exponential written there is proportional to the likelihood rather than equal to it.

B.2 Prediction targets and learnability

Section 3 works throughout with the clean-image denoiser f_t, whereas a diffusion model may be trained against any of four interchangeable targets (clean image, score, noise, velocity), and the

choice among them is usually treated as consequential. Two questions follow, and they have opposite answers. The approximation theory does not depend on the choice: the four targets are affine reparametrizations of one another at fixed t, so they span the same polynomial class and Proposition 1 applies verbatim to each. The architecture does depend on it, but the equivalence locates where. It separates two decisions that are ordinarily conflated, namely which quantity the network's output layer emits and in which coordinates the training error is measured, and only the first is constrained by the manifold hypothesis. Section 2.1 accordingly settles them separately, with an x-prediction head and a v-space loss.

The four targets span one class. The score of the noised marginal is a rescaling of the denoiser. Since $p_t(z) \propto \int e^{-\|z-tx\|^2/2\sigma^2}\, d\rho_{\mathrm{data}}(x)$, differentiating the logarithm brings $(tx - z)/\sigma^2$ down inside the integral and divides by the integral itself, leaving a posterior expectation. This is the same differentiation under the integral sign that identifies f_t with $\nabla_\theta \Lambda$ in Appendix B.1:

$$\nabla_z \log p_t(z) \;=\; \frac{t\, f_t(z) - z}{\sigma^2}, \qquad \text{equivalently} \qquad f_t(z) \;=\; \frac{z + \sigma^2\, \nabla_z \log p_t(z)}{t}. \tag{11}$$

The noise and velocity targets follow from $z_t = tx + \sigma\epsilon$ and $v = x - \epsilon$, using $t + \sigma = 1$:

$$\mathbb{E}[\epsilon \mid z_t = z] \;=\; \frac{z - t f_t(z)}{\sigma}, \qquad\qquad \mathbb{E}[v \mid z_t = z] \;=\; \frac{f_t(z) - z}{\sigma}.$$

Each of the three is an invertible affine function of $f_t(z)$ and z whose coefficients depend on t alone. At fixed t such a map carries a degree-K polynomial in z to a degree-K polynomial in z, so the four targets generate the same class and Proposition 1 transfers to each with only the constant changed. The approximation theory of Section 3 is in this sense a statement about diffusion denoisers in general rather than about one parametrization.

The loss coordinates are a reweighting. Section 3.1 states that the flow-matching loss is the x-prediction error rescaled by σ^{-2}; here is the substitution. Putting $v_\theta = (\hat{x}_\theta - z_t)/\sigma$ and $v = x - \epsilon$ into the velocity loss of Section 2.1, and using $z_t = tx + \sigma\epsilon$ so that the noise cancels and $t + \sigma = 1$ so that the signal collapses as well,

$$v_\theta - v \;=\; \frac{\hat{x}_\theta - z_t}{\sigma} - (x - \epsilon) \;=\; \frac{\hat{x}_\theta - (t+\sigma)\, x}{\sigma} \;=\; \frac{\hat{x}_\theta - x}{\sigma}.$$

The factor σ^{-2} is positive and independent of θ, so minimizing the loss at each t minimizes $\mathbb{E}\Big[\|\hat{x}_\theta - x\|^2\Big]$, whose minimizer over measurable functions of z_t is the conditional mean f_t; training drives $\hat{x}_\theta(\cdot, t)$ toward it. The factor is not constant in t, so the t-averaged objective weights low-noise times heavily, which is why the denominator is clipped near $t = 1$ (Section 2.1). The same substitution applied to the noise target gives $\|\hat{\epsilon}_\theta - \epsilon\|^2 = (t^2/\sigma^2)\, \|\hat{x}_\theta - x\|^2$. Measuring the error in ϵ- or v-coordinates therefore alters the weight each noise level carries in the t-averaged objective and nothing else, and both up-weight the low-noise end. This is a choice about optimization, and it is made freely. It also cannot explain the failure of ϵ- and v-prediction in Figure 2. The weights depend on t alone and not on the ambient dimension D, whereas that failure grows with D at fixed $d_{\mathcal{M}}$.

The output space is not. What the final layer must emit is a different matter, and here the targets are not interchangeable. The patch embedding of Section 4.1 passes each patch through the rank-d' projection π before anything else, so everything the network computes depends on a patch z only through $\pi(z)$. The noised input lies at distance about $\sigma\sqrt{D - d_{\mathcal{M}}}$ from $t\mathcal{M}$, and the question for each head is how much of its target survives this compression. For the clean image, most of it does. Late in sampling the posterior concentrates (Step 2 of the proof of Lemma 1), so $f_t(z)$ lies within $O(\sigma^2/(t^2\tau))$ of $\mathcal{M}$, and by Lemma 1(b) it is a function of $\pi(z)$ up to the compression term of equation 5. Once the lemma applies to a typical input (Appendix B.3), that term is at most $O((D - d_{\mathcal{M}})\, \sigma^2/t^2)$ and vanishes as $\sigma \to 0$. For the noise, little of it does. Late in sampling the noise target $\mathbb{E}[\epsilon \mid z_t = z] = (z - t f_t(z))/\sigma$ is essentially the component of ϵ normal to the manifold, a vector of length about $\sqrt{D - d_{\mathcal{M}}}$ spread evenly over the normal directions. The part of ϵ in the $D - d'$ directions that π discards is independent of $\pi(z)$ and of the other patches, so no function of them recovers it, and an ϵ-head incurs a squared error of at least $D - d'$ at every noise level. The same holds for a head emitting $v = x - \epsilon$. The prediction is therefore about the ambient dimension.

At fixed $d_{\mathcal{M}}$, the error an ϵ- or v-head incurs through the bottleneck grows with D at every noise level, while that of an x-head is bounded, late in sampling, by a term that vanishes as $\sigma \to 0$.

A second and softer effect compounds the first. When σ is small the posterior concentrates, so $z \approx t f_t(z)$ and the noise target is a difference of two nearly equal quantities divided by σ; obtaining it to fixed relative accuracy demands relative accuracy of order σ in the manifold quantity. At matched loss weighting this makes no difference to the population optimum, which is why it is stated as a claim about capacity allocation under a finite budget rather than about function classes.

The empirical necessity of x-prediction reported by Li & He (2026), and reproduced in the controlled study of Section 5.1, is on this reading a statement about the output layer and not about function classes; the classes coincide. That study separates the two accounts by varying D at fixed $d_{\mathcal{M}}$, and the distinction it draws is between heads rather than losses: the runs labelled v-prediction there emit v directly, whereas the architecture of Section 4 emits $\hat{x}_\theta$ and forms v_θ from it afterwards by an explicit formula in z_t, and is therefore an x-head trained under a v-space loss, which is not the configuration that fails. The affine maps above are invertible and computable in closed form from z, so they can be applied outside the network, which leaves the bottleneck responsible for the manifold-valued factor alone.

B.3 Behaviour across the sampling schedule

Proposition 1, Lemma 1 and the combined bound equation 5 hold at a fixed t, while sampling runs t from 0 to 1. This subsection checks that the bound stays controlled along the whole path. What changes along the path is the noise-to-signal ratio $\sigma/t = (1-t)/t$, which falls from ∞ to 0. We compare it with the reach τ of the manifold, the largest distance from $\mathcal{M}$ within which every point has a unique nearest point on $\mathcal{M}$ (for a circle, the reach is its radius). Given a noised input, the posterior over clean images spreads by about σ/t along the manifold (Step 1 of the proof of Lemma 1), so comparing σ/t with τ asks whether the posterior sees a nearly flat piece of the manifold or a curved one. This gives three regimes, summarized in Table 4 and taken below in the order sampling visits them. A final paragraph explains where Lemma 1 itself applies, which is a separate question.

Table 4: The three regimes of the sampling schedule, in the order sampling visits them.

Regime	σ/t	Why the bound is controlled	Constrains
Early	$\gg \tau$	target nearly affine	—
Intermediate	$\sim \tau$	tangent span; analyticity and depth	width, degree
Late	$\ll \tau$	posterior concentrated; compression cost vanishes	—

Early, $\sigma/t \gg \tau$. The observation carries almost no information about x, so the posterior stays close to ρ_{data} and its covariance relaxes to $\Sigma_\rho = \mathrm{Cov}_{\rho_{\text{data}}}(x)$. By the Jacobian identity (right of equation 3), $\nabla_z f_t \approx \frac{t}{\sigma^2}\Sigma_\rho$. What matters is not that this Jacobian is small but that it is nearly the same at every z, and a map with a constant Jacobian is affine. Truncating the expansion equation 4 after its linear term gives $f_t(z) \approx \bar{x} + \frac{t}{\sigma^2}\Sigma_\rho\, z$, with $\bar{x}$ the data mean, and the higher-order terms carry higher powers of t. A degree-one polynomial therefore nearly suffices, and Lemma 1 is not needed.

Intermediate, $\sigma/t \sim \tau$. This is the hard regime. The posterior spreads over a patch of the manifold comparable in size to the reach, so the patch is curved and f_t is genuinely nonlinear. In the proof of Lemma 1 (Step 2), the spread of the posterior away from the manifold is smaller than its spread along it by the factor $(\sigma/t\tau)^2$, and here that factor is of order one, so the posterior no longer lies close to a single tangent plane. Two facts keep the bound controlled, one for each design parameter. For the *width*, the bottleneck must retain the span of the tangent spaces across the patch. This is larger than a single tangent space, but it is set by $d_{\mathcal{M}}$ and the curvature, not by D (Remark 4). For the *degree*, f_t is real-analytic at every t by Proposition 1(a), so the degree needed for a given accuracy is largest here but finite, and depth supplies it, since an L-layer stack reaches degree 2^L. Neither fact involves D.

Late, $\sigma/t \ll \tau$. The posterior is concentrated near a single point of the manifold, on a patch small enough to be nearly flat, and the target is as simple as it gets. Once Lemma 1 applies to a typical input (see below), the compression term of equation 5 is $O((D - d_{\mathcal{M}})\,\sigma^2/t^2)$ and vanishes as $\sigma \to 0$ at

any fixed D. The shrinking distance does the work, since $\eta \to 0$ but $c_\pi/t \to c_\pi$ does not. Averaged over the noise, the slippage part of the term scales with d' instead of D (Remark 6).

Where Lemma 1 applies. The lemma bounds the worst case over all inputs within distance $\delta < t\tau$ of the scaled manifold $t\mathcal{M}$. A typical noised input is farther away than σ suggests. Write $z = tx + \sigma\epsilon$ with $x \in \mathcal{M}$ and $\epsilon \sim \mathcal{N}(0, I_D)$. Of the D directions of ϵ, $d_\mathcal{M}$ run along the manifold at x and $D - d_\mathcal{M}$ are perpendicular to it. The perpendicular part has length about $\sqrt{D - d_\mathcal{M}}$, so z lies at distance about $\sigma\sqrt{D - d_\mathcal{M}}$ from $t\mathcal{M}$. The lemma therefore applies to a typical input only once $\sigma\sqrt{D - d_\mathcal{M}} < t\tau$, that is $\sigma/t < \tau/\sqrt{D - d_\mathcal{M}}$, a threshold well below τ when D is large (about 27 times smaller for a 16×16 RGB patch). Part (b) needs in addition $\|\pi\|\, \sigma\sqrt{D - d_\mathcal{M}} \le t\tau_\pi/2$, the same condition with τ replaced by $\tau_\pi/(2\,\|\pi\|)$. This factor $\sqrt{D - d_\mathcal{M}}$ measures how far the *input* z lies from $t\mathcal{M}$, not the target $f_t(z)$, which late in sampling stays within $O(\sigma^2/(t^2\tau))$ of $\mathcal{M}$ (Appendix B.2). It limits where the lemma's worst-case argument reaches, not how hard f_t is to approximate.

This does not make the posterior less concentrated between $\tau/\sqrt{D - d_\mathcal{M}}$ and τ. By the left of equation 3, the posterior gives each candidate clean image $x' \in \mathcal{M}$ the weight $e^{-\|z - tx'\|^2/2\sigma^2}\, d\rho_{\mathrm{data}}(x')$, so comparing two candidates x' and x depends only on the difference of their exponents and on the data density at the two points. Substituting $z = tx + \sigma\epsilon$,

$$\|z - tx'\|^2 - \|z - tx\|^2 \;=\; \|t(x - x') + \sigma\epsilon\|^2 - \|\sigma\epsilon\|^2 \;=\; t^2\,\|x - x'\|^2 + 2t\sigma\,\epsilon^\top (x - x').$$

The term $\sigma^2\,\|\epsilon\|^2$ is the squared distance of z from the true point tx. It is the same for every candidate, so it cancels, and with it the large distance $\sigma\sqrt{D - d_\mathcal{M}}$ of z from the manifold drops out. What remains depends on the noise only through $\epsilon^\top(x - x')$, which is $\|x - x'\|$ times the component of ϵ along the direction from x' to x. That component is a single standard Gaussian number, not a sum over D coordinates. As x' ranges over the manifold, these directions form a family of dimension $d_\mathcal{M}$, so the largest such component grows like $\sqrt{d_\mathcal{M}}$ rather than $\sqrt{D}$. The posterior therefore concentrates at scale about $\sigma\sqrt{d_\mathcal{M}}/t$, as if the noise acted only along the manifold, and it is sharp once σ/t is well below τ up to factors of $d_\mathcal{M}$. Between the two thresholds the target thus behaves as in the late regime, and only the worst-case bound of the lemma does not yet apply. We give this argument as a sketch for a typical input rather than as a bound over the tube.

Summary. Only the intermediate regime constrains the architecture. The two ends are easy for opposite reasons. Early in sampling the target is nearly affine even though the input is nowhere near the manifold, and late in sampling the geometry is clean. The ambient dimension does not enter how hard the target is in any regime. It enters only the worst-case bound, through where Lemma 1 begins to apply and through the size of the compression term late in sampling. Finally, Proposition 2 and Conjecture 1 compare Volterra and attention stacks on the same f_t, so the comparison does not depend on how hard any individual t is.

B.4 The unnormalized boundary, and what distinguishes the Volterra parametrization

Propositions 2(ii) and 3 are complementary rather than redundant. The first is sharper in hypothesis, needing agreement only on an open set, and is available for softmax because analyticity propagates that agreement globally; the second is far broader in reach, applying to weightings that are not analytic at all. For real-analytic weightings the two coincide, since agreement on an open set extends to $\mathbb{R}^{N\times d}$.

Proposition 3 is also sharp: normalization is not a technical convenience in the proof but the whole of it. Remove it and the conclusion genuinely fails. Unnormalized linear attention with the identity feature map,

$$h_i(X) \;=\; x_i \;+\; \sum_{j=1}^{N} \left(x_i^\top W_Q^\top W_K x_j\right) W_O W_V x_j,$$

is a cubic polynomial map of X, and a stack of such layers lies inside the polynomial class. This is the one attention-shaped operator our results do not exclude, and we regard it as a corroboration rather than a counterexample: the claim of this section is about the function class the denoising

task inhabits, not about a proprietary operator, so a polynomial attention variant performing well on pixel diffusion is what the theory predicts. What the operator forfeits is the property that motivated attention in the first place. The convex-combination bound that Proposition 3 turns against normalized attention is also what keeps its outputs inside the hull of the values; dropping the denominator to gain polynomiality gives up that guarantee, which is the known stability difficulty with unnormalized attention.

That difficulty follows from leaving the normalized class *without replacing the bound*, not from leaving it. The softmax discharges two roles at once: it makes the mixing weights a convex combination, which is the hypothesis Proposition 3 turns against it, and it confines the output to the hull of the values, which is what keeps activations from growing with sequence length. The two are separable, and the block of Section 4.4 keeps the second while discarding the first, by moving the bound off the mixing weights and onto the mixer's input. The RMSNorm preceding each mixer fixes the norm of its output at $\sqrt{d}$, and the modulation $u = \gamma \odot \mathrm{RMSNorm}(\cdot) + \beta$ that follows keeps $\|u\| \le b := \|\gamma\|_\infty \sqrt{d} + \|\beta\|$, where γ and β depend on (t, c) alone. The bilinear term therefore obeys $\left\|W^{(2)}(u,u)\right\| \le b^2 \left\|W^{(2)}\right\|$, a bound in the parameters and the conditioning with no dependence on the data or on N. Unnormalized attention discards both roles together: its weights stay data-dependent and are no longer confined, so nothing bounds the layer, and the output grows with the N unnormalized terms in its sum.

The derivatives separate the two as well. The Jacobian of a polynomial mixer, $W^{(1)} + W^{(2)}(u,\cdot) + W^{(2)}(\cdot,u)$, which is $W^{(1)} + 2W^{(2)}(u,\cdot)$ when $W^{(2)}$ is symmetric, is affine in u, with no saturating and no diverging regime; the mixer contains neither an exponential nor a reciprocal, so neither the gradient growth that accompanies large logits in exponential attention nor the vanishing gradient of a saturated softmax has an analogue here. The one unnormalized attention variant trained at scale corroborates the separation: the ReLU attention of Wortsman et al. (2023) reintroduces an explicit factor $1/N$ in front of the scores. That factor is data-independent, so the weights still do not sum to one and Proposition 3 still does not apply; what it restores is boundedness in sequence length, not convexity. It is the same split we make, arrived at from the other side.

Being inside the polynomial class is necessary, not sufficient, and four things separate the architecture of Section 4 from unnormalized attention within that class. First, its kernels are the ones the theory names: by equation 4 the order-n Volterra kernel of the optimal denoiser is the $(n{+}1)$-st cumulant of the reweighted data measure $\tilde{\rho}$, and $W^{(2)}_{\mathrm{tok}}$ is a free low-rank kernel of exactly that order and type, learned rather than set to the cumulant value, whereas the cubic tensor of unnormalized attention is constrained to the query–key–value form and realizes a structured slice of it. Second, degree is a per-layer choice: the project–multiply–mix primitive delivers order k from a single layer through the k-fold product, so degree and depth are independent knobs, while attention's degree is pinned at three by its own factorization. Third, the mixing kernels $W^{(1)}_{\mathrm{tok}}$ and $W^{(2)}_{\mathrm{tok}}$ are indexed by absolute token position rather than derived from content, so spatial structure on a fixed image grid is native to the operator rather than supplied only by positional encodings or by convolutional components reintroduced alongside it. Finally, Proposition 2(i) certifies that our class contains $\mathcal{P}_{K_L}$ exactly at finite budget; the corresponding question for unnormalized attention is open, and we claim no answer to it.

B.5 Normalization and conditioning in the PoD block

Our own use of normalization is of a different kind, and bears on Proposition 2(i) in a way worth stating precisely. The RMSNorm preceding each mixer rescales a token by a positive scalar, $u \mapsto c(u)\,u$. Were c constant the kernels would absorb it outright: $W^{(1)}$ and $W^{(2)}$ are learned independently, so the linear term takes c and the quadratic c^2 at no cost to the class. Being input-dependent, c instead divides each homogeneous component by the matching power of the token norm, so what the trained network realizes is a degree-graded reweighting of the polynomial rather than an unstructured departure from it. We do not claim that the reweighted class still contains $\mathcal{P}_{K_L}$: recovering the discarded radial dependence from the residual stream is plausible, and we have not shown it. The rate of Proposition 1 is untouched either way, being a statement about the target denoiser rather than about the network that approximates it.

Table 5: **What each operator can parametrize at order k.** The feed-forward sublayer does produce cross-token polynomial terms, but as a k-fold power of the attention matrix, scaled by a coefficient the architecture cannot choose. Shown for one head. With n_{h} heads each slot draws from the same n_{h} matrices.

At order k	Transformer block (one head), smooth φ	Low-rank Volterra layer
Kernel	$c_k\big[\sum_a (W_2)_{ca}(W_1)_{a\cdot}^{\otimes k}\big] \otimes A_{i\cdot}^{\otimes k}$	$W^{(k)}$ of equation 6, rank at most hR
Slots	the same A in every slot; token rank one ($\le n_{\mathrm{h}}$ with n_{h} heads)	independent $Q_q^{(m)}$ per slot
Across orders	one shared W_1A; c_k fixed by φ	independent kernels
Wider sublayer	buys channel rank, never token rank	—

The conditioning branch is the block's other non-polynomial component, and it is harmless for a different reason. At fixed (t, c) it emits constants, so modulation acts on the data path as a diagonal affine map and the branch's own nonlinearities never reach the data; this is the regime in which Propositions 1 and 2 are stated, so the fixed-t form of the theory and the harmlessness of the branch are one fact rather than two concessions. That the kernels vary with t at all is required rather than tolerated, since by equation 4 they are cumulants of a t-dependent tilt. A conditioning branch that is itself polynomial in t is available and we do not use one. Nothing makes the dependence on t polynomial, and Lemma 1 points the other way, its binding quantities behaving like $1/t$ and $\sigma^2/(t^3\tau^2)$, which a low-degree polynomial in t models poorly where the schedule spends its resolution. This is distinct from letting t gate the Volterra kernels directly, which changes how the conditioning acts on the data path rather than how it is computed, and which Section 6 lists as a next step. Keeping the branch identical to the baseline's is also what makes every comparison we report isolate the mixing operator.

B.6 The feed-forward objection

A Transformer block is not attention alone. Its feed-forward sublayer applies a pointwise nonlinearity, and a block so equipped is a universal approximator that certainly approximates f_t. We first take a single attention head with attention matrix A, held fixed at the point of expansion, and return to several heads below. Expanding the nonlinearity does produce cross-token polynomial terms, since the k-th power of $W_1 \sum_j A_{ij}x_j$ contains every degree-k monomial across tokens.[1] What it does not produce is a *free* parametrization of them, and Table 5 sets the two operators side by side. The constraint worth drawing out is that width does not relax the others: the token factor $A_{i\cdot}^{\otimes k}$ carries no hidden-unit index, so it survives the sum over hidden units untouched and extra width buys rank only in the channel factor. Only depth, composing blocks with different A, changes the token-mixing structure at all. With n_{h} heads the argument of the activation becomes $\sum_{e=1}^{n_{\mathrm{h}}} W_1 W_O^e W_V^e \sum_j A_{ij}^e x_j$, so each slot of the order-k term draws its token factor from the same n_{h} matrices $A^1, \dots, A^{n_{\mathrm{h}}}$. The token rank per slot is then at most n_{h} rather than one, the coefficients across orders are still fixed by the activation, and width still adds no token structure, which stays capped by the number of heads. A further constraint is recorded in Remark 1. The claim is therefore not that a Transformer cannot approximate f_t, which it can, but that its *mixing* operator contributes no polynomial content beyond affine, so the polynomial structure of the target must be recovered through the pointwise nonlinearity, where the orders are coupled rather than free.

Remark 1 (A further constraint on the feed-forward kernel). The effective order-k kernel of a feed-forward sublayer after a single head, $c_k\big[\sum_a (W_2)_{ca}(W_1)_{a\cdot}^{\otimes k}\big] \otimes A_{i\cdot}^{\otimes k}$, retains one genuine freedom beyond those noted above, and it is worth pricing because it is the strongest version of the objection. Rescaling $W_1 \mapsto \lambda W_1$ sends the order-k term to $c_k\lambda^k$, so the network can shift weight between orders after all. But it does so along the fixed one-parameter curve $\{(c_k\lambda^k)_k : \lambda > 0\}$, whose shape is determined by the activation. A degree-K Volterra layer chooses K kernels independently. The comparison is one degree of freedom against K, not none against K.

[1] The expansion presumes an activation with a nonvanishing Taylor series, which covers GELU and SiLU. For ReLU there is none, and the conclusion is more direct: the sublayer is then piecewise affine and realizes no polynomial of degree ≥ 2 exactly on any open set.

C PROOFS FOR SECTION 3

Throughout this appendix, $t \in (0,1)$ is the flow time and $\sigma := 1 - t \in (0,1)$ the corresponding noise level; $\mathcal{M} \subset \mathbb{R}^D$ is the compact data manifold inside the ambient pixel space of dimension D, and $\rho := \rho_{\text{data}}$ is the data distribution, a probability measure supported on $\mathcal{M}$. For a noised observation $z \in \mathbb{R}^D$ and a clean image $x \in \mathcal{M}$, the *Gaussian likelihood weight*

$$\psi(z,x) := \exp\big(-\|z - tx\|^2 / 2\sigma^2\big)$$

is the relative weight the observation z assigns to the clean point x at noise level σ. The optimal denoiser is the posterior-weighted average of x,

$$f_t(z) \;=\; \frac{Y(z)}{Z(z)}, \qquad Y(z) := \int_{\mathcal{M}} x\,\psi(z,x)\,d\rho(x), \qquad Z(z) := \int_{\mathcal{M}} \psi(z,x)\,d\rho(x),$$

where $Z(z)$ is the normalizer (partition function) and $Y(z)$ the weight-averaged numerator. We write $p(x \mid z) \propto \psi(z,x)\,\rho(x)$ for the posterior distribution over clean images given z, and $\mathbb{E}_z[\cdot]$ and $\mathrm{Cov}_z(\cdot)$ for its mean and covariance, so that $f_t(z) = \mathbb{E}_z[x]$.

C.1 REGULARITY (PROPOSITION 1(A))

We prove that f_t is *real-analytic* on $\mathbb{R}^D$, that is, locally equal to a convergent power series, and in particular C^∞, with derivatives of every order. The strategy is standard: a function that extends to a holomorphic (complex-differentiable) function of a complex argument $z \in \mathbb{C}^D$ is real-analytic on the real slice $\mathbb{R}^D$, so it suffices to show that the numerator Y and the normalizer Z each admit such an extension, and then to divide.

Proof. Step 1 (the normalizer is strictly positive). For every z the integrand $\psi(z,x)$ is strictly positive, an exponential never vanishing, and ρ is a probability measure, so $Z(z) = \int_{\mathcal{M}} \psi(z,x)\,d\rho(x) > 0$. The ratio $f_t = Y/Z$ is therefore well-defined for every z.

Step 2 (each weight extends to an entire function). Fix $x \in \mathcal{M}$. Viewed as a function of z, the weight $\psi(z,x) = \exp(-\|z - tx\|^2/2\sigma^2)$ is the exponential of a polynomial in z, so it extends from $\mathbb{R}^D$ to an entire function of $z \in \mathbb{C}^D$, one holomorphic on all of $\mathbb{C}^D$.

Step 3 (bounds uniform in x). For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_D) \in \mathbb{Z}^D_{\geq 0}$, a tuple of non-negative integers recording how many times to differentiate in each coordinate, write $\partial_z^\alpha := \partial_{z_1}^{\alpha_1} \cdots \partial_{z_D}^{\alpha_D}$ for the corresponding mixed partial derivative in z, of total order $|\alpha| := \alpha_1 + \cdots + \alpha_D$. On any compact set $\Omega \subset \mathbb{C}^D$, both $|\psi(z,x)|$ and every derivative $|\partial_z^\alpha \psi(z,x)|$ are bounded by a constant that does not depend on x. Indeed, differentiating ψ in z brings down, by the chain rule, a fixed polynomial in $(z - tx)/\sigma$ multiplying ψ itself; as x ranges over the compact manifold $\mathcal{M}$ and z over the compact set Ω, both that polynomial factor and ψ remain bounded, uniformly in x.

Step 4 (holomorphy passes through the integral). The integrands are holomorphic in z with the locally uniform, x-independent bounds of Step 3, so dominated convergence justifies differentiating under the integral sign: the difference quotients of Y and Z converge to the integrals of the corresponding z-derivatives of ψ. Hence Y and Z are holomorphic on $\mathbb{C}^D$, and therefore real-analytic on $\mathbb{R}^D$.

Step 5 (the quotient). Since Z is real-analytic and nowhere zero (Step 1), the quotient $f_t = Y/Z$ is real-analytic on $\mathbb{R}^D$. In particular f_t is C^∞, with derivatives of every order, so the smoothness order s in Proposition 1(b) may be taken arbitrarily large. □

Remark 2. The smoothness comes from the noise, not from the data: adding noise at level $\sigma > 0$ convolves the data distribution with a Gaussian, an infinitely smoothing operation, so f_t is real-analytic for every $t \in (0,1)$ however irregular ρ_{data} may be. The regularity of f_t is thus never the binding constraint in the rate of Proposition 1(b); the polynomial degree one is willing to spend is.

C.2 THE JACOBIAN IDENTITY (RIGHT OF EQUATION 3)

Proof. From $\nabla_z \psi = -\sigma^{-2}(z - tx)\,\psi$,

$$\nabla_z Z = -\sigma^{-2}\big(zZ - tY\big),$$
$$\nabla_z Y = -\sigma^{-2}\int_{\mathcal{M}} x\,(z-tx)^\top \psi\, d\rho = -\sigma^{-2}\big(\mathbb{E}_z[x]\,z^\top - t\,\mathbb{E}_z[xx^\top]\big)Z.$$

Using $\nabla_z f_t = \nabla_z Y/Z - (Y/Z)\,(\nabla_z Z/Z)^\top$ and $Y/Z = \mathbb{E}_z[x]$, the two terms are

$$\nabla_z Y/Z = -\sigma^{-2}\big(\mathbb{E}_z[x]\,z^\top - t\,\mathbb{E}_z[xx^\top]\big),$$
$$(Y/Z)(\nabla_z Z/Z)^\top = -\sigma^{-2}\big(\mathbb{E}_z[x]\,z^\top - t\,\mathbb{E}_z[x]\mathbb{E}_z[x]^\top\big).$$

Subtracting, the $\mathbb{E}_z[x]\,z^\top$ terms cancel and

$$\nabla_z f_t(z) = \frac{t}{\sigma^2}\big(\mathbb{E}_z[xx^\top] - \mathbb{E}_z[x]\mathbb{E}_z[x]^\top\big) = \frac{t}{\sigma^2}\,\mathrm{Cov}_z(x). \qquad \square$$

The all-order identity equation 10. By equation 9, $f_t(z) = \nabla_\theta \Lambda(\theta)$ at $\theta = tz/\sigma^2$, so the chain rule gives

$$\nabla_z^k f_t(z) \;=\; \Big(\tfrac{t}{\sigma^2}\Big)^k \nabla_\theta^{k+1}\Lambda(\theta)\Big|_{\theta = tz/\sigma^2}.$$

The derivatives of a cumulant generating function at parameter θ are the cumulants of the measure tilted by $e^{\theta^\top x}$, and tilting $\tilde{\rho}$ by $e^{\theta^\top x}$ with $\theta = tz/\sigma^2$ reconstitutes exactly the posterior: by the derivation of equation 9, the posterior density is proportional to $e^{\theta^\top x}\,d\tilde{\rho}(x)$, the z-only Gaussian factor having cancelled between numerator and denominator. Hence $\nabla_z^k f_t(z) = (t/\sigma^2)^k\,\kappa_{k+1}(p(\cdot \mid z_t = z))$, which is equation 10; at $k = 1$ the second cumulant is the covariance, recovering the computation above. $\square$

C.3 APPROXIMATION (PROPOSITION 1(B))

We prove that the best polynomial of total degree at most K approximates f_t on $t\mathcal{M}$ with squared error of order K^{-2s}, equivalently $N^{-2s/(2d_{\mathcal{M}}+1)}$ in the number N of coefficients per output coordinate, with every constant free of the ambient dimension D. Here

$$\|h\|_{C^s(\Omega)} \;:=\; \max_{|\alpha|\le s}\,\sup_{w\in\Omega}\,\big|\partial^\alpha h(w)\big|$$

is the C^s norm, the largest size attained by a function h and its partial derivatives up to order s over a domain Ω (with α a multi-index, as in the regularity proof above). The argument has three stages: embed $t\mathcal{M}$ into a space whose dimension is set by $d_{\mathcal{M}}$, run a classical polynomial approximation rate there, and convert the degree K into a parameter count.

Proof. Step 1 (linear embedding into the intrinsic dimension). Since $t\mathcal{M}$ is a compact C^∞ manifold of dimension $d_{\mathcal{M}}$, Whitney's projection theorem provides a residual (hence dense) set of linear maps $\mathbb{R}^D \to \mathbb{R}^{d'}$ with $d' = 2d_{\mathcal{M}} + 1$ whose restriction to $t\mathcal{M}$ is an injective immersion, and therefore, by compactness, a smooth embedding; the value $2d_{\mathcal{M}} + 1$ is the least dimension at which a generic linear projection avoids the double points of $t\mathcal{M}$, whose offending directions form a set of dimension $2d_{\mathcal{M}} + 1$. Fix such a map π (the bottleneck of Section 4). Its image $\mathcal{M}_\pi := \pi(t\mathcal{M})$ is a compact C^∞ submanifold of $\mathbb{R}^{d'}$ of dimension $d_{\mathcal{M}}$, and $\pi|_{t\mathcal{M}}$ is a diffeomorphism onto $\mathcal{M}_\pi$ that is *bi-Lipschitz*: for some $L_\pi \ge 1$,

$$L_\pi^{-1}\,\|z - z'\| \;\le\; \|\pi(z) - \pi(z')\| \;\le\; L_\pi\,\|z - z'\|, \qquad z, z' \in t\mathcal{M},$$

with L_π, like every constant below, set by the reach and curvature of $t\mathcal{M}$ and by π, not by D. Because $\pi|_{t\mathcal{M}}$ is injective, each $w \in \mathcal{M}_\pi$ has a unique preimage, so

$$g \;:=\; f_t \circ (\pi|_{t\mathcal{M}})^{-1} \;:\; \mathcal{M}_\pi \to \mathbb{R}^D$$

is well-defined; it is the target the polynomial must match on $\mathcal{M}_\pi$, the clean answer $f_t(z)$ indexed by the projected point $\pi(z)$. As a composition of the real-analytic f_t (part (a)) with the smooth inverse $(\pi|_{t\mathcal{M}})^{-1}$, g is C^∞, and the chain rule through this fixed smooth map on a compact set gives

$$\|g\|_{C^s(\mathcal{M}_\pi)} \;\le\; c_1\,\|f_t\|_{C^s(t\mathcal{M})},$$

the constant c_1 depending on the derivatives of $(\pi|_{t\mathcal{M}})^{-1}$, hence on L_π and the curvature of $t\mathcal{M}$, not on D. Approximating f_t on $t\mathcal{M}$ is thereby reduced to approximating the equally smooth g on the $d_{\mathcal{M}}$-dimensional $\mathcal{M}_\pi$.

Step 2 (polynomial rate by Jackson's theorem). A Jackson-type theorem on a compact submanifold of Euclidean space is not new: Ragozin (1970) relates the smoothness of a function on such a submanifold to its rate of best polynomial approximation, with a constructive companion for spheres and projective spaces (Ragozin, 1971). We give the argument here in the form we need, via Whitney extension to a cube, because the textbook statement of Jackson's theorem is on a solid domain and the extension route keeps every constant explicitly attached to d'; the step should be read as a rederivation for self-containedness, not as a new result. What it does *not* make redundant is Step 1: the constants in any such theorem track the dimension of the ambient space in which the submanifold sits, so applying one directly to $t\mathcal{M} \subset \mathbb{R}^D$ would reintroduce D into the bound. The projection into $\mathbb{R}^{2d_{\mathcal{M}}+1}$ is what makes a Jackson-type rate usable here, and it is the load-bearing step of the proof.

Treat g one output coordinate at a time; fix $a \in \{1, \dots, D\}$, a scalar g^a on $\mathcal{M}_\pi$. The Whitney extension theorem produces a compactly supported $G^a \in C^s(\mathbb{R}^{d'})$ with $G^a|_{\mathcal{M}_\pi} = g^a$ and $\|G^a\|_{C^s(\mathbb{R}^{d'})} \le c_2 \, \|g^a\|_{C^s(\mathcal{M}_\pi)}$, extending the target from the curved $\mathcal{M}_\pi$ to the flat space on which a polynomial is defined. The multivariate Jackson theorem on a cube $\Omega \supset \operatorname{supp} G^a$ then yields a polynomial P_K^a of total degree at most K with

$$\sup_{w \in \Omega} \left| G^a - P_K^a \right| \;\le\; c_3(d') \, K^{-s} \, \|G^a\|_{C^s} \,,$$

the classical statement that a C^s function on a d'-dimensional cube is approximated to order K^{-s} by polynomials of degree K, the constant depending on the cube dimension d', not on D. Assembling $P_K = (P_K^1, \dots, P_K^D)$ and restricting to $\mathcal{M}_\pi$,

$$\sup_{w \in \mathcal{M}_\pi} \|g(w) - P_K(w)\|^2 \;\le\; \sum_{a=1}^{D} \left(c_3 K^{-s} \, \|G^a\|_{C^s} \right)^2 \;\le\; c_4 \, K^{-2s} \, \|g\|^2_{C^s(\mathcal{M}_\pi, \mathbb{R}^D)} \,,$$

where squaring K^{-s} reflects the squared error and the output dimension D enters *only* through $\|g\|^2_{C^s} = \sum_a \|g^a\|^2_{C^s}$, a fixed quantity for fixed $(t\mathcal{M}, f_t)$ that touches neither the exponent nor the rate.

Step 3 (assembly, and the parameter form of the rate). For $z \in t\mathcal{M}$ we have $(\pi|_{t\mathcal{M}})^{-1}(\pi(z)) = z$, so $f_t(z) = g(\pi(z))$ and $P_K(\pi(z)) = (P_K \circ \pi)(z)$; the inverse cancels and is never evaluated. Hence

$$\sup_{z \in t\mathcal{M}} \|f_t(z) - P_K(\pi(z))\|^2 \;=\; \sup_{w \in \mathcal{M}_\pi} \|g(w) - P_K(w)\|^2 \;\le\; C \, K^{-2s}, \qquad C = c_4 \, \|g\|^2_{C^s} \,,$$

the degree form of the rate. For the parameter form, a polynomial of total degree at most K on $\mathbb{R}^{d'}$ has $N = \binom{d'+K}{d'} \asymp K^{d'}$ coefficients per output coordinate, so with $d' = 2d_{\mathcal{M}} + 1$, substituting $K \asymp N^{1/(2d_{\mathcal{M}}+1)}$ gives

$$K^{-2s} \;\asymp\; N^{-2s/(2d_{\mathcal{M}}+1)}.$$

The exponent is set by $d_{\mathcal{M}}$ and not by D, whereas a polynomial of the raw patch would need $\binom{D+K}{D} \asymp K^D$ coefficients. The dimension of the restrictions $P|_{\mathcal{M}_\pi}$ can be smaller than N when $\mathcal{M}_\pi$ lies on a lower-dimensional algebraic variety, but for a generic smooth manifold no nonzero polynomial vanishes on $\mathcal{M}_\pi$ and it equals N, so the count above is the one we use. By part (a) the smoothness order s is unbounded, so the rate exceeds any fixed polynomial order in K. Every constant ($c_1, \dots, c_4$ and C) tracks $d' = 2d_{\mathcal{M}} + 1$, the reach and curvature of $t\mathcal{M}$, and $\|f_t\|_{C^s}$; none carries an explicit D, which is the precise sense in which the rate escapes the ambient dimension. □

Remark 3. The polynomial P_K above is produced by Whitney extension and approximation on a cube, not by truncating the cumulant expansion of equation 4; the proposition does not assert that the truncation attains the rate, and the network never computes it. Wiener's orthogonalization, which does make the degree-K truncation optimal, is constructed against a Gaussian input measure, whereas z_t here is a Gaussian smoothing of a manifold-supported law. Neither property is needed: what is claimed is the existence of *some* degree-K polynomial attaining squared error $O(K^{-2s})$ uniformly on $t\mathcal{M}$, a sup-norm statement rather than an L^2 one, and Proposition 2(i) makes the existential form sufficient by showing the architecture contains the whole class.

Remark 4. The embedding dimension $d' = 2d_{\mathcal{M}} + 1$ is the least value at which a generic linear projection is guaranteed faithful, and is the dimension used in the bound, not a prescription for the trained width. Two points relate it to the architecture, and they have different status. First, the trained bottleneck should in fact exceed $2d_{\mathcal{M}} + 1$: at intermediate noise levels the posterior $p(x \mid z)$ spreads over a curved patch of $\mathcal{M}$, so the embedding must cover the span of tangent spaces across that patch rather than the manifold at a single point (Appendix B.3), and this surplus is still governed by $d_{\mathcal{M}}$ and the curvature of $\mathcal{M}$, never by D. Second, the practical width $d' \in \{128, 256\}$ lies below the ambient dimension by construction (here $D = D_p \in \{768, 3072\}$, a factor of six to twenty-four). Whether it also exceeds $2d_{\mathcal{M}} + 1$ depends on the intrinsic dimension $d_{\mathcal{M}}$, which we do not measure; under the manifold hypothesis, intrinsic-dimension estimates place natural images at a few dozen dimensions (Pope et al., 2021; Levina & Bickel, 2004), comfortably within the chosen d', but this is an empirical premise rather than a quantity established here.

C.4 Low-dimensional dependence (Lemma 1)

The lemma uses two pieces of geometry. The *reach* $\tau > 0$ of $\mathcal{M}$ is the largest radius such that every point within distance τ of $\mathcal{M}$ has a unique nearest point on $\mathcal{M}$; it bounds the curvature, the second fundamental form satisfying $\|\mathrm{II}\| \le 1/\tau$. The scaled manifold $t\mathcal{M}$ then has reach $t\tau$.

Lemma 1 (Low-dimensional dependence of the denoiser). *Let $\mathcal{M}$ have reach $\tau > 0$, let ρ_{data} have a density ρ with respect to the volume measure of $\mathcal{M}$ that is C^3 and bounded below by some $\rho_{\min} > 0$, and fix $t \in (0, 1)$, $\sigma = 1 - t$. On the tube $\mathcal{T}_\delta = \{z : \mathrm{dist}(z, t\mathcal{M}) \le \delta\}$ with $\delta < t\tau$, let $r(z) := tx^\star(z)$ denote the nearest-point retraction onto the scaled manifold $t\mathcal{M}$.*

(a) (Normal insensitivity.)

$$\sup_{z \in \mathcal{T}_\delta} \big\| f_t(z) - f_t\big(r(z)\big) \big\| \;\le\; \eta(t, \sigma, \tau)\, \delta, \qquad \eta(t, \sigma, \tau) \;=\; C_\eta \min\Big\{ \underbrace{\tfrac{1}{t}}_{\textit{uniform}}, \; \underbrace{\tfrac{\sigma^2}{t^3\tau^2}}_{\textit{small-noise}} \Big\},$$

valid for every $t \in (0, 1)$, with C_η depending on $\mathcal{M}$, $\rho_{\min}$ and the C^3 norm of $\log \rho$, but not on σ, δ, or D; the small-noise branch is the active one once $\sigma \ll t\tau$, where $\eta = O(\sigma^2/(t^3\tau^2))$.

(b) (Linear compression.) *Let π be the projection of Proposition 1, let $\tau_\pi > 0$ be the reach of $\pi(\mathcal{M})$, and suppose in addition that $\|\pi\|\ \delta \le t\tau_\pi/2$. Then there is a map $F : \mathbb{R}^{d'} \to \mathbb{R}^D$ with*

$$\sup_{z \in \mathcal{T}_\delta} \big\| f_t(z) - F\big(\pi(z)\big) \big\| \;\le\; \Big(\eta(t, \sigma, \tau) + \tfrac{c_\pi}{t}\Big)\, \delta,$$

where $c_\pi \ge 0$ is a constant of the pair $(\pi, \mathcal{M})$, independent of σ, δ, and D. The retraction r is not a function of the linear image $\pi(z)$, and the c_π/t term prices the exchange: it reflects the tangential component that vectors of $\ker \pi$ carry, vanishes when $\ker \pi$ is orthogonal to every tangent space of $t\mathcal{M}$, and cannot be removed by any choice of F (Remark 5).

Intuitively (Figure 5), the clean signals consistent with a noised z differ by sliding along $\mathcal{M}$ rather than by leaving it, so f_t reads the along-manifold part of its input and is insensitive to the normal part. Discarding the normal directions is precisely what the bottleneck does, so it costs little. The resulting error is the product of two quantities: how far off the manifold z sits (δ), and how sensitive f_t is in the discarded directions (η). The second is small once the noise drops below the reach. A linear π cannot discard the normal directions exactly, since its kernel generically carries a tangential component along which f_t varies at the ordinary rate $O(1/t)$; part (b)'s c_π/t term prices this residual slippage, and it shrinks as the bottleneck widens. This is one reason the trained d' is set above the minimal $2d_{\mathcal{M}} + 1$; the second, posterior coverage, is discussed in Appendix B.3.

The proof makes this precise through a tangent–normal split of the posterior covariance. Along the manifold the posterior carries the ordinary spread of order σ^2/t^2, whereas reaching off the manifold costs an extra factor of curvature and shrinks the normal spread by $(\sigma/t\tau)^2$; the Jacobian identity (right of equation 3) then transfers this gap to $\nabla_z f_t$. Part (a) bounds the change of f_t under the nearest-point retraction, and part (b) prices the exchange of that nonlinear retraction for the linear bottleneck projection π.

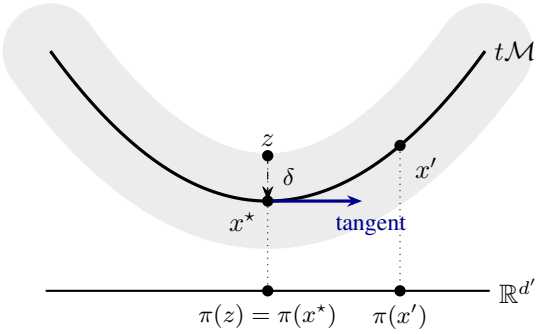

Figure 5: **Why a low-dimensional projection suffices** (schematic). Noised inputs z lie in a tube of radius δ about the scaled data manifold $t\mathcal{M}$. The clean signals consistent with z differ by moving *along* the manifold (tangent direction), not by leaving it, so the denoiser f_t varies along the tangent direction and is nearly flat along the normal one. The bottleneck π keeps a d'-dimensional image of the input. It sends z and $x^\star$ to the same point; this costs little, since f_t takes nearly the same value at both. Further, it keeps $x^\star$ and x' apart; this is essential, since f_t takes different values at them.

Proof. Step 0 (nearest point and unimodality). Fix $z \in \mathcal{T}_\delta$ with $\delta < t\tau$. Since $\operatorname{dist}(z, t\mathcal{M}) \le \delta < t\tau$ and $t\mathcal{M}$ has reach $t\tau$, the point z has a unique nearest point $tx^\star$ on $t\mathcal{M}$, and the residual $w := z - tx^\star$ is *normal* to $\mathcal{M}$ at $x^\star$ (perpendicular to the surface), with $\|w\| \le \delta$. Any competing posterior mode is separated from $x^\star$ by distance $\ge c\,\tau$ and so carries mass $O(e^{-c^2 t^2 \tau^2 / 2\sigma^2})$; for $\sigma \ll t\tau$ the posterior is effectively unimodal about $x^\star$ and these tails are absorbed below.

Step 1 (local model and the Gaussian posterior). Parametrize $\mathcal{M}$ near $x^\star$ in *tangent coordinates* $\xi \in \mathbb{R}^{d_\mathcal{M}}$, the displacement within the tangent plane $T_{x^\star}\mathcal{M}$ recorded against an orthonormal tangent frame $B \in \mathbb{R}^{D \times d_\mathcal{M}}$ ($B^\top B = I$):

$$x(\xi) = x^\star + B\xi + \tfrac{1}{2}\mathrm{II}(\xi,\xi) + O(\|\xi\|^3).$$

The linear term $B\xi$ moves in the tangent plane, and the *second fundamental form* II, the symmetric, normal-valued quadratic form measuring how $\mathcal{M}$ curves away from that plane ($\|\mathrm{II}\| \le 1/\tau$), bends the point back onto the surface. The posterior is $p(x \mid z) \propto e^{-\|z-tx\|^2/2\sigma^2}\,\rho(x)$, so its negative logarithm, restricted to the surface and dropping the ξ-independent normalizer, is $\Phi(\xi) := \frac{1}{2\sigma^2}\,\|z - tx(\xi)\|^2 - \log \rho(x(\xi))$. Writing $z - tx(\xi) = w - tB\xi - \frac{t}{2}\mathrm{II}(\xi,\xi) + O(\|\xi\|^3)$ and squaring, the cross term $-2t\,w^\top B\xi$ vanishes because w is normal and $B\xi$ tangent ($w^\top B\xi = 0$), so $\xi = 0$ is a critical point and the surviving quadratic is

$$\Phi(\xi) = \text{const} + \tfrac{1}{2\sigma^2}\,\xi^\top\big(\underbrace{t^2 I}_{\text{tangent stiffness}} - \underbrace{t\langle w, \mathrm{II}\rangle}_{\text{curvature} \times \text{offset}}\big)\xi + O(\|\xi\|^3),$$

where $\langle w, \mathrm{II}\rangle$ is the *shape operator* along w, the symmetric form with $\xi^\top \langle w, \mathrm{II}\rangle\, \xi = w^\top \mathrm{II}(\xi,\xi)$, satisfying $\|t\langle w, \mathrm{II}\rangle\| \le t\,\|w\|\,/\tau \le t\delta/\tau < t^2$. The Hessian $\frac{1}{\sigma^2}(t^2 I - t\langle w, \mathrm{II}\rangle)$ is thus positive definite, of size $\sim t^2/\sigma^2$. Since this Hessian is of order σ^{-2} while the third derivatives of $x(\xi)$ and $\log \rho$ are $O(1)$, the cubic remainder is smaller than the quadratic by a factor $\sim \sigma/(t\tau)$ on the bulk of the posterior (where $\|\xi\| \sim \sigma/t$), so by the Laplace approximation the posterior over ξ is Gaussian to leading order,

$$\xi \sim \mathcal{N}(0, \Sigma_\xi), \qquad \Sigma_\xi = \tfrac{\sigma^2}{t^2}\big(I + O(\delta/(t\tau))\big),$$

the prior $-\log \rho$ contributing only at the subleading order σ^2 absorbed into the correction.

Step 2 (tangent versus normal covariance). Decompose

$$x - \mathbb{E}_z[x] = B(\xi - \bar{\xi}) + \tfrac{1}{2}\big(\mathrm{II}(\xi,\xi) - \mathbb{E}\,\mathrm{II}(\xi,\xi)\big) + \cdots$$

into its tangent (linear in ξ) and normal (quadratic in ξ) parts. The tangent part has covariance $B\Sigma_\xi B^\top$, with eigenvalues $\frac{\sigma^2}{t^2}(1 + O(\delta/(t\tau)))$ on the tangent space $T_{x^\star}\mathcal{M}$. The normal part, being quadratic in ξ, has covariance $O\big(\|\mathrm{II}\|^2 \operatorname{Var}\|\xi\|^2\big) = O\big(\sigma^4/(t^4\tau^2)\big)$ on the normal space. The

posterior covariance is therefore of order σ^2/t^2 along the manifold but smaller by the factor $(\sigma/t\tau)^2$ across it: the largest eigenvalue of $\mathrm{Cov}_z(x)$ on directions normal to $\mathcal{M}$ is $O(\sigma^4/(t^4\tau^2))$.

Step 3 (transfer to the Jacobian; part (a)). By the Jacobian identity (right of equation 3), $\nabla_z f_t = \frac{t}{\sigma^2}\mathrm{Cov}_z(x)$, so its operator norm on directions normal to $t\mathcal{M}$ is $\frac{t}{\sigma^2}\cdot O(\sigma^4/(t^4\tau^2)) = O(\sigma^2/(t^3\tau^2))$, and $O(1/t)$ on tangent directions. The segment from z to its retraction $r(z) = tx^\star$ has length $\|w\| \le \delta$ and lies in the normal direction, so

$$\big\|f_t(z) - f_t\big(r(z)\big)\big\| \le \delta \cdot \sup_{z' \in \mathcal{T}_\delta} \big\|\nabla_z f_t(z')\big|_{T^\perp}\big\| = O\Big(\tfrac{\sigma^2}{t^3\tau^2}\Big)\,\delta,$$

which is the small-noise branch of part (a), $\eta = O(\sigma^2/(t^3\tau^2))$ for $\sigma \ll t\tau$.

Step 3′ (the uniform branch). The estimate of Steps 1–3 is a small-noise expansion; a bound valid for every $t \in (0,1)$ with $\delta < t\tau$ follows by localizing the posterior at scale σ/t without any Gaussian expansion. For a manifold point x' at chordal distance $d = \|x' - x^\star\|$, the reach inequality $\mathrm{dist}(x', T_{x^\star}\mathcal{M}) \le d^2/(2\tau)$ (Federer, 1959) and the normality of w give $|w^\top(x' - x^\star)| \le \delta\, d^2/(2\tau)$, so

$$\|z - tx'\|^2 - \|z - tx^\star\|^2 \;=\; t^2d^2 - 2t\, w^\top(x' - x^\star) \;\ge\; t^2d^2\Big(1 - \tfrac{\delta}{t\tau}\Big),$$

and for $\delta \le (1 - c_0)\, t\tau$ the posterior weight at distance d is at most $e^{-c_0 t^2 d^2/2\sigma^2}$ relative to the peak. The posterior second moment about $x^\star$ is therefore $O((\sigma/t)^2)$ (the function $d^2 e^{-c_0 t^2 d^2/2\sigma^2}$ is maximized at $d \asymp \sigma/t$), with constants depending on the density bounds of ρ_{data} and the geometry of $\mathcal{M}$, and since the covariance about the mean is dominated by the second moment about any fixed point, $\mathrm{Cov}_z(x) = O((\sigma/t)^2)$ on the tube. By the Jacobian identity (right of equation 3), $\|\nabla_z f_t\| \le \frac{t}{\sigma^2}\cdot O\big((\sigma/t)^2\big) = O(1/t)$ in every direction, for every t. Combining with Step 3 yields the form stated in the lemma, $\eta(t,\sigma,\tau) = C\min\{1/t,\ \sigma^2/(t^3\tau^2)\}$.

Step 4 (linear compression; part (b)). The retraction r is not a function of $\pi(z)$: π is linear with kernel of dimension $D - d'$, and two tube points on the same fiber $z + \ker\pi$ generally have different nearest points, so the assignment $\pi(z) \mapsto r(z)$ is ill-defined. Construct F instead through the projected manifold. Take π as in Proposition 1, bi-Lipschitz on $t\mathcal{M}$ with constant L_π. Its image $\pi(t\mathcal{M}) = t\,\pi(\mathcal{M})$ is a compact embedded submanifold of $\mathbb{R}^{d'}$ with reach $t\tau_\pi > 0$. A point of $\pi(\mathcal{T}_\delta)$ lies within $\|\pi\|\ \delta \le t\tau_\pi/2$ of it, so the nearest-point projection $\mathrm{np}(v) := \arg\min_{y \in \pi(t\mathcal{M})}\|v - y\|$, the point of $\pi(t\mathcal{M})$ closest to v, is well-defined on $\pi(\mathcal{T}_\delta)$ and Lipschitz there with constant at most 2 (Federer, 1959). Set

$$F(v) \;:=\; f_t\Big(\big(\pi|_{t\mathcal{M}}\big)^{-1}\big(\mathrm{np}(v)\big)\Big), \qquad v \in \pi(\mathcal{T}_\delta).$$

For $z \in \mathcal{T}_\delta$ write $\pi(z) = \pi(r(z)) + \pi(w)$ with $\|\pi(w)\| \le \|\pi\|\ \delta$; then $\mathrm{np}(\pi(z))$ differs from $\pi(r(z))$ by $O(\delta)$, and its bi-Lipschitz lift $t\tilde{x} := (\pi|_{t\mathcal{M}})^{-1}(\mathrm{np}(\pi(z)))$ differs from $r(z)$ by at most $c_\pi\,\delta$ along $t\mathcal{M}$, where c_π collects $\|\pi\|$, L_π, and the nearest-point Lipschitz constant, all independent of σ, δ, and D. Hence, using part (a) and the tangential operator norm $O(1/t)$ from Step 3,

$$\|f_t(z) - F(\pi(z))\| \;\le\; \big\|f_t(z) - f_t\big(r(z)\big)\big\| + \big\|f_t\big(r(z)\big) - f_t(t\tilde{x})\big\| \;\le\; \eta\,\delta + O\Big(\tfrac{1}{t}\Big)\, c_\pi\,\delta,$$

which is part (b) after absorbing constants into c_π. When $\ker\pi$ is orthogonal to every tangent space of $t\mathcal{M}$, the fiber and the slippage are purely normal and the tangential term vanishes. □

Remark 5. The c_π/t term of part (b) is not an artifact of this construction. Any F is constant on the fibers of π; a fiber segment of length $\sim \delta$ inside the tube generically carries a tangential component (vectors of $\ker\pi$ meet the tangent spaces at a positive angle, but not orthogonally), and along that component f_t varies at the ordinary tangential rate $O(1/t)$, independently of σ. Hence $\sup_{\mathcal{T}_\delta}\|f_t - F\circ\pi\| \gtrsim \delta/t$ for every F unless $\ker\pi$ lies in the common normal space of $t\mathcal{M}$, which requires d' at least the dimension of the span of the tangent spaces. Part (a) is therefore the sharp statement about the denoiser itself, and part (b) the honest statement about linear compression; the gap between them shrinks as the bottleneck widens, since a larger d' leaves a smaller kernel with less tangential content.

Remark 6 (Worst case and typical case). Part (b) is a supremum over the tube. For a noised input $z = tx + \sigma\epsilon$ the tube radius is set by the component of $\sigma\epsilon$ normal to $t\mathcal{M}$, whose length concentrates at $\sigma\sqrt{D - d_{\mathcal{M}}}$. Read at that radius, the compression term of equation 5 is $O((D - d_{\mathcal{M}})\,\sigma^2/t^2)$, and

the tube hypothesis $\delta < t\tau$ requires $\sigma\sqrt{D - d_{\mathcal{M}}} < t\tau$. The supremum is pessimistic for the slippage term, however. Step 4 uses the normal displacement $w = z - r(z)$ only through $\pi(w)$, bounding $\|\pi(w)\| \le \|\pi\|\,\|w\|$. To leading order in σ, $w = \sigma P_N \epsilon$, with P_N the orthogonal projection onto the normal space of $t\mathcal{M}$ at tx, so for isotropic noise

$$\mathbb{E}\,\|\pi(w)\|^2 \;=\; \sigma^2\,\mathrm{tr}\left(\pi P_N \pi^\top\right) \;\le\; \sigma^2\,\|\pi\|_F^2 \;\le\; d'\,\|\pi\|^2\,\sigma^2,$$

using that π has rank at most d'. Since $\epsilon \mapsto \|\pi P_N \epsilon\|$ is $\|\pi\|$-Lipschitz, Gaussian concentration gives $\|\pi(w)\| \le \|\pi\|\,\sigma\,(\sqrt{d'} + u)$ with probability at least $1 - e^{-u^2/2}$. For a typical noised input the slippage part of the compression cost is therefore of order $c_\pi^2\, d'\,\sigma^2/t^2$, set by the bottleneck width rather than by the ambient dimension. The normal-sensitivity part $\eta\,\|w\|$ is a property of f_t itself rather than of π, and this argument does not reduce it. We expect it to depend on the curvature of $\mathcal{M}$ rather than on D, since the posterior spreads in the normal directions only through curvature, but we do not prove this.

Remark 7 (Where each branch applies). Step 0, the Gaussian posterior of Step 1 and the small-noise branch of η are expansions in $\sigma/(t\tau)$ and hold only late in sampling, once $\sigma \ll t\tau$. The uniform branch of Step 3′ needs no expansion and holds at every t, but only on the tube $\delta < t\tau$. At intermediate t, where σ/t is comparable to τ, the two branches are both of order $1/t$, so for an input inside the tube the compression term of equation 5 is at most of order τ^2, bounded but not small. A typical noised input, however, lies at distance about $\sigma\sqrt{D - d_{\mathcal{M}}}$ from $t\mathcal{M}$ (Remark 6), which in this regime exceeds $t\tau$ by a factor of about $\sqrt{D - d_{\mathcal{M}}}$, so the lemma does not reach it. Appendix B.3 therefore treats intermediate t without the lemma, through the width the bottleneck needs (Remark 4) and the analyticity of f_t.

Combined bound equation 5. For $z \in \mathcal{T}_\delta$, the triangle inequality gives $\|f_t(z) - P_K(\pi(z))\| \le \|f_t(z) - F(\pi(z))\| + \|F(\pi(z)) - P_K(\pi(z))\|$. The first term is $\le (\eta + c_\pi/t)\,\delta$ by Lemma 1(b). For the second, $F(\pi(z)) = g(\mathrm{np}(\pi(z)))$ with $\mathrm{np}(\pi(z)) \in \pi(t\mathcal{M})$, and P_K approximates g on $\pi(t\mathcal{M})$ to $O(K^{-s})$ by Proposition 1(b); since P_K is a fixed polynomial, hence Lipschitz on the compact $\pi(\mathcal{T}_\delta)$, evaluating at $\pi(z)$ rather than at $\mathrm{np}(\pi(z))$ adds an $O(\delta)$ term whose σ-free constant joins the slippage inside the second bracket of equation 5, enlarging c_π. Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ yields equation 5. □

C.5 Attainment (Proposition 2(i))

Throughout this subsection set $K := K_L = 2^L$, write $\Pi_K(\mathbb{R}^{d'};\mathbb{R})$ for the scalar polynomials of total degree at most K, and let $H := \dim \Pi_K(\mathbb{R}^{d'};\mathbb{R}) = \binom{d'+K}{d'}$.

Lemma 2 (Powers of affine forms span). *For every $K \ge 0$,*

$$\mathrm{span}\left\{\,(a^\top u + b)^K \;:\; a \in \mathbb{R}^{d'},\; b \in \mathbb{R}\,\right\} \;=\; \Pi_K(\mathbb{R}^{d'};\mathbb{R}).$$

Proof. The inclusion $\subseteq$ is immediate. For $\supseteq$, fix $a \in \mathbb{R}^{d'}$ and consider, for $\lambda \in \mathbb{R}$,

$$\varphi_\lambda(u) \;:=\; (\lambda\, a^\top u + 1)^K \;=\; \sum_{j=0}^{K} \binom{K}{j} \lambda^j\,(a^\top u)^j.$$

Evaluating at $K+1$ distinct values $\lambda_0, \dots, \lambda_K$ produces a linear system for the functions $\left(\binom{K}{j}(a^\top u)^j\right)_{j=0}^{K}$ whose coefficient matrix $(\lambda_i^j)_{i,j}$ is Vandermonde, hence invertible. Every pure power $(a^\top u)^j$ with $j \le K$ therefore lies in the span of the family. It remains to verify that $\{(a^\top u)^j : a \in \mathbb{R}^{d'}\}$ spans the homogeneous polynomials of degree j, which is the classical polarization identity: for any indices $i_1, \dots, i_j \in \{1, \dots, m\}$,

$$u_{i_1} u_{i_2} \cdots u_{i_j} \;=\; \frac{1}{2^j\, j!} \sum_{\varepsilon \in \{\pm 1\}^j} \varepsilon_1 \cdots \varepsilon_j \left(\varepsilon_1 u_{i_1} + \cdots + \varepsilon_j u_{i_j}\right)^j,$$

expressing an arbitrary degree-j monomial as a combination of j-th powers of linear forms. Summing over $j \le K$ completes the proof. □

Lemma 3 (Exact realization by iterated multiplication). *Fix $R^* \geq 1$, affine forms $\ell_r(u) = q_r^\top u + c_r$ on $\mathbb{R}^{d'}$, and a matrix $C \in \mathbb{R}^{D \times R^*}$. A depth-$L$ stack of the mixers of Section 4, of stream width R^*, channel-mixer hidden width $h = R^*$ and rank $R' = 1$, with gated residual connections, affine embedding, and linear head, computes the map*

$$u \;\longmapsto\; \sum_{r=1}^{R^*} C_{\cdot r}\, \ell_r(u)^{2^L}$$

exactly, with at most $c_0(R^*)^2$ parameters per layer for an absolute constant c_0.

Proof. The construction acts on each patch separately and uses only the channel mixers. Every token mixer is switched off by setting its gate $\alpha_{\mathrm{tok}} = 0$, an admissible parameter setting, so each block reduces to its channel sublayer. The affine embedding writes $w_r^{(0)} := \ell_r(u)$ into channel r of the stream, $r = 1, \ldots, R^*$ (the constants c_r are supplied by the embedding bias). Each subsequent channel mixer consists of a linear branch and the quadratic primitive of equation 6, wrapped in a gated residual connection: on stream state w its output is

$$w \;+\; \mathrm{diag}(\alpha)\Big(W^{(1)} w + P\,\big[(Q_1^{(1)} w) \odot (Q_1^{(2)} w)\big]\Big).$$

Choose $Q_1^{(1)} = Q_1^{(2)} = P = I_{R^*}$ (the tied setting, which the untied class contains), $W^{(1)} = -I_{R^*}$, and $\alpha = \mathbf{1}$; the gates are zero-initialized but learnable, so this is an admissible parameter setting. The output is then $w + (-w + w \odot w) = w \odot w$: the linear branch cancels the residual and the layer squares the stream elementwise. After L such layers channel r holds $\ell_r(u)^{2^L}$, and the linear head applies C. The per-layer parameter count is that of $P, Q_1^{(1)}, Q_1^{(2)}, W^{(1)} \in \mathbb{R}^{R^* \times R^*}$ and the gates, in total $O\big((R^*)^2\big)$. □

Proof of Proposition 2(i). By Lemma 2 at degree K_L, the family $\{\ell^{K_L} : \ell \text{ affine}\}$ spans $\mathcal{P}_{K_L}(\mathbb{R}^{d'}; \mathbb{R})$ and therefore contains a basis $\ell_1^{K_L}, \ldots, \ell_H^{K_L}$ with $H = \binom{d'+K_L}{d'}$. Any $Q \in \mathcal{P}_{K_L}(\mathbb{R}^{d'}; \mathbb{R}^D)$ thus has each output coordinate a linear combination of the $\ell_r^{K_L}$, that is, $Q(u) = C\,\big(\ell_r(u)^{K_L}\big)_{r=1}^{H}$ for some $C \in \mathbb{R}^{D \times H}$. Lemma 3 with $R^* = H$ realizes this map exactly with at most $c_0 H^2 =: \mathsf{P}_0(L, d')$ parameters per layer, applied to the input $\pi(z)$ supplied by the bottleneck embedding. This proves the exact containment $\{Q \circ \pi : Q \in \mathcal{P}_{K_L}\} \subseteq \mathcal{V}_{L,\mathsf{P}}$ for $\mathsf{P} \geq \mathsf{P}_0$. The approximation bound follows by instantiating $g = P_{K_L} \circ \pi$ with P_{K_L} the polynomial of Proposition 1(b) at degree K_L, whose error on $t\mathcal{M}$ is at most $C_1 K_L^{-2s}$; the tube extension is equation 5. □

Remark 8. The statement is representational rather than practical. The threshold $\mathsf{P}_0 \asymp H^2$ is astronomically large at the depths used in practice, because the dense factor matrices charge H parameters per column to operations (carrying a channel, squaring a channel) that touch one coordinate each, and no attempt has been made to optimize the constant. A further reason for its size is that the construction realizes the *full* polynomial space $\mathcal{P}_{K_L}$ on $\mathbb{R}^{d'}$, of dimension $H = \binom{d'+K_L}{d'} \asymp K_L^{d'}$ (exponent $d' = 2d_{\mathcal{M}} + 1$), the same space in which Proposition 1 finds its approximants, and it charges dense factor matrices for every monomial of that space. Its role is to certify that restricting from unconstrained polynomials to the architecture's low-rank class loses nothing at the level of the rate: the degree form K_L^{-2s} of Proposition 1 is attained within the architecture's own function class. The construction realizes the target at the tied point $Q_q^{(1)} = Q_q^{(2)}$ of the primitive; untying the factors, as in the trained architecture, enlarges the parameterization without changing what is representable, so the threshold P_0 is unaffected. Normalization layers are excluded from $\mathcal{V}_{L,\mathsf{P}}$, as RMSNorm is not a polynomial map. What it changes in the trained network, a degree-graded reweighting of the polynomial, and why we do not claim the containment for the normalized class, is set out in Appendix B.5.

C.6 Polynomial content of softmax attention (Proposition 2(ii))

Lemma 4 (Independence of exponentials over polynomials). *Let $q_1, \ldots, q_n$ be real polynomials on $\mathbb{R}^\nu$ such that $q_i - q_j$ is nonconstant for every $i \neq j$, and let $p_1, \ldots, p_n$ be polynomials. If $\sum_{i=1}^n p_i(x)\, e^{q_i(x)} = 0$ for all x in a nonempty open set $U \subseteq \mathbb{R}^\nu$, then $p_i \equiv 0$ for every i.*

Proof. Both sides are real-analytic, so the identity extends from U to all of $\mathbb{R}^\nu$. Choose a unit vector v such that the top-degree homogeneous part of each difference $q_i - q_j$ ($i \neq j$) does not vanish at v; this is possible because finitely many nonzero homogeneous polynomials vanish only on a closed set of measure zero on the sphere. Then for every base point x_0, the univariate restriction $\lambda \mapsto (q_i - q_j)(x_0 + \lambda v)$ is a polynomial of the same degree as $q_i - q_j$ with leading coefficient determined by v alone; in particular it is nonconstant and tends to $+\infty$ or $-\infty$ as $\lambda \to +\infty$, with the sign independent of x_0. The relation $i \succ j$ iff $(q_i - q_j)(x_0 + \lambda v) \to +\infty$ is therefore a strict total order on $\{1, \dots, n\}$, independent of x_0; let i^* be its maximum. Dividing the identity restricted to the line by $e^{q_{i^*}(x_0+\lambda v)}$ gives

$$p_{i^*}(x_0 + \lambda v) \;=\; -\sum_{i \neq i^*} p_i(x_0 + \lambda v)\, e^{(q_i - q_{i^*})(x_0+\lambda v)},$$

and each term on the right tends to 0 as $\lambda \to +\infty$, since the exponent is a polynomial tending to $-\infty$ and the exponential decay dominates polynomial growth. A univariate polynomial with limit 0 at $+\infty$ vanishes identically, so $p_{i^*}(x_0 + \lambda v) \equiv 0$ for every x_0, hence $p_{i^*} \equiv 0$ on $\mathbb{R}^\nu$. Removing the i^* term and iterating completes the proof. □

Lemma 5 (Affine polynomial content of a softmax attention layer). *For a token matrix $X = (x_1, \dots, x_N) \in \mathbb{R}^{N \times d}$, let*

$$h_i(X) \;=\; x_i + \sum_{e=1}^{n_\mathrm{h}} W_O^e \, \frac{\sum_{j=1}^N \exp\big(s_{ij}^e(X)\big)\, W_V^e\, x_j}{\sum_{j=1}^N \exp\big(s_{ij}^e(X)\big)}, \qquad s_{ij}^e(X) \;=\; \tfrac{1}{\sqrt{d_h}}\,(W_Q^e x_i)^\top (W_K^e x_j),$$

be a softmax self-attention layer with n_h heads and residual connection. If h agrees with a polynomial map on a nonempty open set of inputs, then h is affine.

Proof. Both h and the candidate polynomial are real-analytic, so agreement on an open set implies agreement everywhere; write $g_i := h_i - x_i$, a polynomial map, and fix the output token i.

Consider first a single head ($n_\mathrm{h} = 1$) and drop the head index. Multiplying through by the (strictly positive) normalizer gives the global identity

$$\sum_{j=1}^N e^{s_{ij}(X)} \big(W_O W_V\, x_j - g_i(X)\big) \;\equiv\; 0.$$

Partition $\{1, \dots, N\}$ into classes by the equivalence $j \sim j'$ iff $s_{ij} - s_{ij'}$ is constant as a polynomial in X, and on each class $\mathcal{C}$ write $s_{ij} = s_{\mathcal{C}} + \kappa_j$ with $\kappa_j \in \mathbb{R}$. Representatives of distinct classes have nonconstant differences, so applying Lemma 4 coordinatewise (with polynomial coefficients $\sum_{j \in \mathcal{C}} e^{\kappa_j}(W_O W_V x_j - g_i)$) forces, for every class $\mathcal{C}$,

$$\sum_{j \in \mathcal{C}} e^{\kappa_j} \big(W_O W_V\, x_j - g_i(X)\big) \;\equiv\; 0, \quad \text{i.e.} \quad g_i(X) \;=\; \sum_{j \in \mathcal{C}} \omega_j\, W_O W_V\, x_j, \quad \omega_j = \frac{e^{\kappa_j}}{\sum_{j' \in \mathcal{C}} e^{\kappa_{j'}}},$$

since the weights e^{κ_j} are positive and sum to a nonzero constant. The right-hand side is affine in X, so g_i is affine and $h_i = x_i + g_i$ is affine.

For $n_\mathrm{h} \geq 2$ heads, multiply through by the product of all normalizers $D_i^e(X) := \sum_j e^{s_{ij}^e(X)}$:

$$g_i(X) \prod_{e=1}^{n_\mathrm{h}} D_i^e(X) \;-\; \sum_{e=1}^{n_\mathrm{h}} \Big(\sum_j e^{s_{ij}^e(X)}\, W_O^e W_V^e\, x_j \Big) \prod_{e' \neq e} D_i^{e'}(X) \;\equiv\; 0.$$

Expanding all products writes the left side as a finite sum $\sum_\alpha p_\alpha(X)\, e^{\bar{s}_\alpha(X)}$ in which every exponent is a sum $\bar{s}_\alpha = \sum_e s_{ij_e}^e$ over one score per head, and every coefficient p_α is either g_i (from the first product, with positive constant weight) or an affine map $W_O^e W_V^e x_j$ (from the second, with negative constant weight). Group the exponents into classes modulo additive constants and apply Lemma 4 as before: each class identity has the form $c\, g_i(X) =$ (affine map) where c is a sum of the positive constants attached to the g_i-terms in that class. Every term of the expansion of $g_i \prod_e D_i^e$ lies in some class, and within its class the g_i-coefficient c is a nonempty sum of positive numbers, hence nonzero. Therefore g_i is affine, and so is h. □

Lemma 6 (The optimal denoiser is not affine). *If ρ_{data} has compact support and is not a single point mass, then f_t is not affine, for any $t \in (0,1)$.*

Proof. Suppose f_t is affine. By the Jacobian identity (right of equation 3), $\mathrm{Cov}(x \mid z_t = z) = (\sigma^2/t)\,\nabla_z f_t(z)$ is then a constant matrix Σ_0. We claim $\Sigma_0 = 0$. The posterior of x given $z_t = \lambda v$ has density proportional to $\exp\left((\lambda t\, v^\top x - t^2 \|x\|^2/2)/\sigma^2\right)$ with respect to ρ_{data}. Since the support function of the convex hull of $\operatorname{supp}\rho_{\mathrm{data}}$ is differentiable at almost every direction, for almost every unit vector v the linear functional $x \mapsto v^\top x$ has a unique maximizer $x^\star$ over the hull; the maximizer is then an exposed point of the hull and so belongs to $\operatorname{supp}\rho_{\mathrm{data}}$. Fix such a v. For any $\zeta > 0$, the posterior weight of $\{x : v^\top x \le v^\top x^\star - \zeta\}$ relative to any neighbourhood of $x^\star$ of positive ρ_{data}-mass decays like $e^{-\lambda t \zeta'/\sigma^2}$ as $\lambda \to +\infty$, so the posterior converges weakly to $\delta_{x^\star}$, and by boundedness of the support its second moments converge as well; hence $\mathrm{Cov}(x \mid z_t = \lambda v) \to 0$ and $\Sigma_0 = 0$. Zero posterior covariance means the posterior is a Dirac mass for every z. But for every z the posterior is mutually absolutely continuous with ρ_{data}, the Gaussian weight being strictly positive, so it is a Dirac mass only if ρ_{data} is one, contradicting the assumption. □

Proof of Proposition 2(ii). Domain. Propositions 1 and 2(i) are stated for a single patch $z \in \mathbb{R}^D$. Attention mixes patches, and on a single token it is linear, since its softmax weight is 1, so part (ii) has content only at the level of an image. We therefore take the input to be an image $Z = (z_1, \dots, z_N) \in \mathbb{R}^{ND}$, whose tokens x_j are obtained from $\pi(z_j)$ by the affine embedding of Section 4.1, so that the token matrix $X = X(Z)$ is an affine function of Z. The per-patch class is compared with this one by applying $Q \circ \pi$ to each patch, which is a polynomial map of Z of the same degree.

Argument. The scores $s^e_{ij}(X(Z))$ are polynomials in Z and the values $W^e_V x_j$ are affine in Z, so clearing the normalizers as in the proof of Lemma 5 again yields an identity $\sum_\alpha p_\alpha(Z)\, e^{\bar{s}_\alpha(Z)} \equiv 0$ with polynomial coefficients and exponents. Lemma 4 applies in the Z variables, with the exponents grouped into classes modulo constants as polynomials in Z. This also handles the fact that $X(Z)$ sweeps an affine subspace of token matrices rather than an open set. Agreement of $h \circ X$ with a polynomial map of Z on a nonempty open set of images therefore forces $h \circ X$ to be affine in Z. Since f_t is not affine (Lemma 6) while $\mathcal{V}_{L,\mathsf{P}}$ contains $\{Q \circ \pi : Q \in \mathcal{P}_{K_L}\}$ exactly by part (i), the two classes meet only in affine maps. □

Remark 9. Two boundaries of the argument are worth recording. First, Lemma 5 concerns a single layer; a composition of attention layers could in principle agree with a polynomial map without any individual layer doing so, and extending the transcendence argument through compositions is the open depth-L part of Conjecture 1. Second, the transcendence argument itself is specific to the softmax, since it is the exponential form of the weights that Lemma 4 exploits. This does not, however, confine the affine-content property to softmax attention: Proposition 3 reaches the same conclusion for every nonnegative normalized weighting, by an argument that uses no algebraic structure at all (Appendix C.7). Linear attention under a nonnegative feature map is therefore covered as well, and the only attention-shaped operator left inside the polynomial class is the unnormalized one, for which the convexity bound fails. Pointwise MLP nonlinearities such as GELU are likewise non-polynomial maps, and their inclusion does not place a Transformer block inside the polynomial class, but we do not pursue a full block-level characterization here. The feed-forward sublayer is discussed in Appendix B.6.

The quantitative form of the separation, described informally after Proposition 2 in Section 3, is the following.

Conjecture 1 (Quantitative rate gap at matched budget). *In the setting of Proposition 2, let $\mathcal{A}_{L,\mathsf{P}}$ denote the class of depth-L stacks of softmax self-attention blocks with residual connections and at most P parameters per layer. For matched depth L and every $\mathsf{P} \ge \mathsf{P}_0(L, d')$,*

$$\inf_{h \in \mathcal{A}_{L,\mathsf{P}}} \sup_{z \in t\mathcal{M}} \|f_t(z) - h(z)\|^2 \;\ge\; C_2 \cdot K_L^{-2s} \cdot \big(1 + \Delta(s, L)\big)$$

for a constant $C_2 > 0$ and a gap $\Delta(s, L) > 0$ nondecreasing in the smoothness order s.

It is open for a specific reason: classical rational approximation theory bounds the *reverse* direction, with free-pole rationals approximating at least as well as polynomials of matched degree and some-

times strictly better (Newman, 1964; Saff & Varga, 1977), so no generic degree-counting argument will do. A proof must exploit the structural constraints of attention, positive normalized weights with input-tied denominators, for which quantitative lower bounds are not currently available. Extending Proposition 2(ii) itself from a single layer to depth-L stacks is likewise open, since a composition of non-polynomial layers can in principle be polynomial, though we expect the same transcendence mechanism to apply.

C.7 Affine content of normalized attention (Proposition 3)

Proposition 3 (Affine content of normalized attention). *Let h be a mixing layer with residual connection of the form*

$$h_i(X) \;=\; x_i \;+\; \sum_{j=1}^{N} \omega_{ij}(X)\, g(x_j), \qquad i = 1, \dots, N, \tag{12}$$

where $g : \mathbb{R}^d \to \mathbb{R}^d$ is affine and the weights satisfy, at every $X \in \mathbb{R}^{N\times d}$,

$$\omega_{ij}(X) \;\geq\; 0 \qquad \textit{and} \qquad \sum_{j=1}^{N} \omega_{ij}(X) \;=\; 1.$$

If h coincides with a polynomial map on $\mathbb{R}^{N\times d}$, then h is affine. No regularity of any kind is assumed of ω: the conclusion holds whether the weights are exponential, rational, piecewise linear, or discontinuous.

The scope is worth stating plainly, because it covers essentially every attention mechanism in current use. Softmax weights are nonnegative and normalized. So are the weights of linear attention under any nonnegative feature map, including the ReLU variant used in high-resolution diffusion backbones (Katharopoulos et al., 2020; Xie et al., 2025), kernel-feature approximations of the softmax (Choromanski et al., 2021), and any convex mixing rule whatever. None of these lies in the polynomial class except in its affine part. Appendix B.4 treats the unnormalized boundary where the conclusion fails, and how this proposition and Proposition 2(ii) divide the work.

Throughout, $\|X\|$ denotes the Frobenius norm of $X = (x_1, \dots, x_N) \in \mathbb{R}^{N\times d}$, so that $\|x_j\| \leq \|X\|$ for every j.

Proof. Write $g(x) = Gx + c$ with $G \in \mathbb{R}^{d\times d}$ and $c \in \mathbb{R}^d$, and set

$$S_i(X) \;:=\; h_i(X) - x_i \;=\; \sum_{j=1}^{N} \omega_{ij}(X)\, g(x_j).$$

Step 1 (the convexity bound). Because the weights are nonnegative and sum to one, $S_i(X)$ is a convex combination of the points $g(x_1), \dots, g(x_N)$, hence lies in their convex hull. The norm of a convex combination is at most the largest norm among the points combined, so for every X,

$$\|S_i(X)\| \;\leq\; \max_{1\leq j\leq N} \|g(x_j)\| \;\leq\; \|G\|_{\mathrm{op}} \max_j \|x_j\| + \|c\| \;\leq\; a\,\|X\| + b, \tag{13}$$

with $a := \|G\|_{\mathrm{op}}$ and $b := \|c\|$ depending only on the layer's parameters. This is the only place the hypotheses on ω are used, and they are used only through nonnegativity and normalization; no continuity, boundedness, or algebraic form is required.

Step 2 (affine growth forces degree one). Suppose h coincides with a polynomial map on $\mathbb{R}^{N\times d}$. Then each S_i is polynomial, since x_i is. Decompose S_i into homogeneous parts,

$$S_i \;=\; \sum_{j=0}^{J} S_i^{[j]}, \qquad S_i^{[j]}(\lambda X) = \lambda^j S_i^{[j]}(X), \qquad S_i^{[J]} \not\equiv 0,$$

and suppose for contradiction that $J \geq 2$. Fix any $X_0 \in \mathbb{R}^{N\times d}$ and evaluate along the ray λX_0, $\lambda > 0$:

$$\frac{S_i(\lambda X_0)}{\lambda^J} \;=\; S_i^{[J]}(X_0) \;+\; \sum_{j<J} \lambda^{j-J} S_i^{[j]}(X_0) \;\xrightarrow[\lambda\to\infty]{}\; S_i^{[J]}(X_0),$$

while equation 13 gives

$$\frac{\|S_i(\lambda X_0)\|}{\lambda^J} \;\le\; \frac{a\,\lambda\,\|X_0\| + b}{\lambda^J} \;\xrightarrow[\lambda\to\infty]{}\; 0 \qquad \text{since } J \ge 2.$$

Hence $S_i^{[J]}(X_0) = 0$. As X_0 was arbitrary, $S_i^{[J]} \equiv 0$, contradicting the choice of J. Therefore $J \le 1$, so S_i is affine and $h_i(X) = x_i + S_i(X)$ is affine. □

Remark 10. The argument is insensitive to how the weights are produced, which is what gives the proposition its reach: softmax, any nonnegative feature map with normalization (including the piecewise-linear ReLU map, for which no analytic continuation is available), and any convex mixing rule are all covered by the same two steps. Multiple heads are covered as well: a sum of n_{h} heads, each of the form equation 12, obeys equation 13 with a replaced by $\sum_e \|G_e\|_{\mathrm{op}}$, and Step 2 is unchanged.

Remark 11 (Sharpness). Normalization cannot be dropped. For unnormalized linear attention with the identity feature map (Appendix B.4), the weights $x_i^\top W_Q^\top W_K x_j$ are neither nonnegative nor normalized and grow linearly in $\|X\|$, so equation 13 fails, the layer is a cubic polynomial map, and the conclusion of the proposition is false for it. The dichotomy is therefore exact. An attention-shaped layer with a data-dependent denominator has affine polynomial content; one without a denominator can be polynomial, at the cost of the convex-hull property that the same bound expresses.

D Architecture Details

This appendix gives the derivations behind Sections 4.2 and 4.3 in full.

The primitive as a Volterra term. Each projected coordinate $(Q_q^{(m)}u)_c = \sum_{j=1}^n Q_{q,cj}^{(m)}\,u_j$ is a learned mixture of *all* inputs, so multiplying the k projections of a group produces every degree-k monomial, cross-terms included: $\prod_m (Q_q^{(m)}u)_c = \sum_{j_1,\dots,j_k=1}^n \big(\prod_m Q_{q,cj_m}^{(m)}\big)\,u_{j_1}\cdots u_{j_k}$. Summing over the R groups and mixing through P collects these into the order-k Volterra term equation 2, $y_i = \sum_{j_1,\dots,j_k=1}^n W_{i,\,j_1,\dots,j_k}^{(k)}\,u_{j_1}\cdots u_{j_k}$, whose kernel is supplied directly in factored form,

$$W_{i,\,j_1,\dots,j_k}^{(k)} \;=\; \sum_{c=1}^{h}\sum_{q=1}^{R} P_{ic}\prod_{m=1}^{k} Q_{q,\,cj_m}^{(m)}. \tag{14}$$

This is a canonical-polyadic factorization with one rank-one term per hidden unit and group (c, q), so of rank at most hR, with an independent factor per slot; the factorization constrains the kernel, not the set of interactions it represents. We parameterize the factors independently (untied), at $k = 2$.

Three properties of the primitive matter here. First, only the symmetric part of the kernel is identifiable, since $u_{j_1}\cdots u_{j_k}$ is unchanged by permuting its factors; untied factors therefore cost nothing in what is represented, while giving the optimizer a larger space to move in. Second, evaluation costs $\mathcal{O}(k\,nhR)$ per kernel against $\mathcal{O}(n^{k+1})$ for a dense one. Third, a single instance realizes an order-k kernel through the k-fold product, so polynomial order is a per-layer choice independent of depth. We use $k = 2$ throughout, since depth already supplies degree 2^L (Proposition 2(i)).

Token mixer. The token mixer is the primitive over the N token positions with $h = N$ and $P = I$, applied to each channel column $u_{:,a}$ with the same maps on every channel,

$$y_{ia} = \sum_{j=1}^{N} W_{\mathrm{tok},ij}^{(1)}\,u_{ja} + \sum_{q=1}^{R}\big(Q_q^{(1)}u_{:,a}\big)_i\big(Q_q^{(2)}u_{:,a}\big)_i, \qquad Q_q^{(1)}, Q_q^{(2)} \in \mathbb{R}^{N\times N}, \tag{15}$$

so that $W_{\mathrm{tok},ijj'}^{(2)} = \sum_{q=1}^R Q_{q,ij}^{(1)}\,Q_{q,ij'}^{(2)}$, and for each output position i the $N \times N$ slice of the kernel has rank at most R. All $2R + 1$ maps are dense $N \times N$ matrices, so the mixer holds $(2R + 1)N^2$ parameters per block, independent of the width d, against N^3 for a dense $W_{\mathrm{tok}}^{(2)}$ alone. The maps carry no channel index, so in implementation all $2R$ quadratic maps are applied to the $N \times d$ token matrix at once and their outputs multiplied elementwise. Evaluation costs $(2R + 1)N^2 d$ multiply-adds for the maps and about $2RNd$ for the products and their sum, so $\mathcal{O}(RN^2 d)$ in total. Self-attention costs $\mathcal{O}(N^2 d + Nd^2)$, the second term for its four $d \times d$ projections, which the token mixer does not have.

Channel mixer. The channel mixer is the primitive over the d channels of one token, with hidden width h, R' pairs $Q_q^{(1)}, Q_q^{(2)} \in \mathbb{R}^{h\times d}$ and $P \in \mathbb{R}^{d\times h}$, applied at every position with the same weights. Its linear branch also passes through the hidden layer, $W_{\text{ch}}^{(1)} = PQ^{(0)}$ with $Q^{(0)} \in \mathbb{R}^{h\times d}$, so the mixer holds $(2R'+2)dh$ parameters and costs as many multiply-adds per token. At $h = 2d$ and $R' = 1$ (Section 5.2) this is $8d^2$, the parameter count of the gated feed-forward block of Li & He (2026) at the same width.

E Implementation Details

Algorithms 1 and 2 give the training step and the sampler. They follow Algorithms 1 and 2 of Li & He (2026), since PoD changes only the network net_θ, and use the settings of Section 5.2. Images are scaled to $[-1, 1]$, and $\varnothing$ denotes the null class label used for classifier-free guidance (Ho & Salimans, 2022).

Algorithm 1 Training step.

Input: clean image x, class label c, network net_θ
with probability 0.1, set $c \leftarrow \varnothing$ ▷ label dropout
$t \sim \text{logit-normal}(-0.8,\, 0.8)$ ▷ location -0.8, scale 0.8
$\epsilon \sim \mathcal{N}(0, I)$
$z_t \leftarrow t\, x + (1-t)\, \epsilon$
$\hat{x}_\theta \leftarrow \text{net}_\theta(z_t, t, c)$ ▷ x-prediction
$v \leftarrow (x - z_t)/\max(1-t,\, 0.05)$ ▷ $= x - \epsilon$ for $t \le 0.95$
$v_\theta \leftarrow (\hat{x}_\theta - z_t)/\max(1-t,\, 0.05)$
$\mathcal{L} \leftarrow \|v_\theta - v\|^2$
take a gradient step on $\mathcal{L}$ and update the EMA weights

Algorithm 2 Sampling with classifier-free guidance, Heun solver.

Input: class label c, guidance scale ω, guidance interval $(t_{\text{lo}}, t_{\text{hi}})$, number of steps $S = 50$; net_θ uses the EMA weights
Guided velocity $\bar{v}(z, t)$:
  $v_c \leftarrow (\text{net}_\theta(z, t, c) - z)/\max(1-t,\, 0.05)$
  $v_\varnothing \leftarrow (\text{net}_\theta(z, t, \varnothing) - z)/\max(1-t,\, 0.05)$
  $w \leftarrow \omega$ if $t_{\text{lo}} < t < t_{\text{hi}}$, else $w \leftarrow 1$ ▷ guidance interval
  $\bar{v}(z, t) \leftarrow v_\varnothing + w\,(v_c - v_\varnothing)$
Integration from $t = 0$ (noise) to $t = 1$ (data):
$z \sim \mathcal{N}(0, I),\ t_i \leftarrow i/S$ for $i = 0, \ldots, S$ ▷ linear time grid
for $i = 0, \ldots, S-2$ **do**
  $h \leftarrow t_{i+1} - t_i$
  $k_1 \leftarrow \bar{v}(z, t_i),\ k_2 \leftarrow \bar{v}(z + h\,k_1,\, t_{i+1})$
  $z \leftarrow z + \frac{h}{2}\,(k_1 + k_2)$ ▷ Heun step
end for
$z \leftarrow z + (t_S - t_{S-1})\,\bar{v}(z, t_{S-1})$ ▷ final step in Euler form
return z

F Results at 512×512

Table 6 reports ImageNet 512×512 with patch size 32, so $N = 256$ as at 256×256 and the recipe is otherwise unchanged. The patch dimension is now $D_p = 3072$ and the bottleneck compresses each patch by a much larger factor, which puts the manifold argument of Section 3 under the most pressure. The matched-scale ordering against JiT survives. PoD lowers FID at every scale at matched parameters and FLOPs, from 4.02 to 3.73 at B, 2.53 to 2.31 at L, 1.94 to 1.75 at H and 1.78 to 1.55 at G. These margins, 0.19 to 0.29, are close to those at 256×256 (0.22 to 0.34), but the largest gain is now at B rather than G. Inception Score is higher than JiT at B and L and slightly lower at H and G. Among pixel-space models without pretrained components, only SiD2-UViT/2 reaches a

Table 6: Reference results on ImageNet 512×512. Notation as in Table 1. At this resolution the patch dimension is $D_p = 3072$, exceeding the hidden width of every model listed, so the bottleneck embedding is compressing by a larger factor than at 256. Baseline numbers are as reported by Li & He (2026).

	pre-training						
Method	token	perc.	self-sup.	params	GFLOPs	FID ↓	IS ↑
Latent-space diffusion							
DiT-XL/2 (Peebles & Xie, 2023)	SD-VAE	VGG	–	675+49M	525	3.04	240.8
SiT-XL/2 (Ma et al., 2024)	SD-VAE	VGG	–	675+49M	525	2.62	252.2
REPA, SiT-XL/2 (Yu et al., 2025)	SD-VAE	VGG	DINOv2	675+49M	525	2.08	274.6
DDT-XL/2 (Wang et al., 2026b)	SD-VAE	VGG	DINOv2	675+49M	525	1.28	305.1
RAE, DiT^{DH}-XL/2 (Zheng et al., 2026)	RAE	VGG	DINOv2	839+415M	642	1.13	259.6
Pixel-space diffusion							
ADM-G (Dhariwal & Nichol, 2021)	–	–	–	559M	1983	7.72	172.7
RIN (Jabri et al., 2023)	–	–	–	320M	415	3.95	216.0
SiD, UViT/4 (Hoogeboom et al., 2023)	–	–	–	2B	555	3.02	248.7
VDM++, UViT/4 (Kingma & Gao, 2023)	–	–	–	2B	555	2.65	278.1
SiD2, UViT/4 (Hoogeboom et al., 2025)	–	–	–	N/A	137	2.19	–
SiD2, UViT/2 (Hoogeboom et al., 2025)	–	–	–	N/A	653	1.48	–
PixNerd, XL/16 (Wang et al., 2026a)	–	–	DINOv2	700M	583	2.84	245.6
JiT-B/32 (Li & He, 2026)	–	–	–	133M	26	4.02	271.0
JiT-L/32 (Li & He, 2026)	–	–	–	462M	89	2.53	299.9
JiT-H/32 (Li & He, 2026)	–	–	–	956M	183	1.94	309.1
JiT-G/32 (Li & He, 2026)	–	–	–	2B	384	1.78	306.8
Pixel-space diffusion, polynomial (ours)							
PoD-B/32	–	–	–	121M	28	3.73	279.1
PoD-L/32	–	–	–	394M	87	2.31	302.4
PoD-H/32	–	–	–	790M	165	1.75	307.3
PoD-G/32	–	–	–	1.6B	320	1.55	303.9

lower FID (1.48), at 653 GFLOPs against our 320, and PoD-H/32 (1.75 at 165 GFLOPs) improves on SiD2-UViT/4 (2.19 at 137 GFLOPs).

G Additional Qualitative Samples

Figures 6–8 show class-conditional samples from PoD-G/16 on ImageNet 256×256, generated with classifier-free guidance and the 50-step Heun sampler of Appendix E. Figures 9 and 10 show samples from PoD-G/32 on ImageNet 512×512 with the same sampler.



Figure 6: **Uncurated samples from PoD-G/16 on ImageNet** 256×256**, mammals.** Each row shows two classes, four samples each, with the ImageNet class index in parentheses. From the top, the rows show Chihuahua (151) and beagle (162); golden retriever (207) and cocker spaniel (219); Old English sheepdog (229) and boxer (242); Siberian husky (250) and Pomeranian (259); tabby (281) and Persian cat (283); Egyptian cat (285) and lion (291); tiger (292) and hyena (276); hare (331) and zebra (340).



Figure 7: **Uncurated samples from PoD-G/16 on ImageNet** 256×256**, birds, reptiles and other animals.** Each row shows two classes, four samples each, with the ImageNet class index in parentheses. From the top, the rows show hen (8) and vulture (23); peacock (84) and macaw (88); goose (99) and black swan (100); flamingo (130) and pelican (144); vine snake (59) and boa constrictor (61); Indian cobra (63) and triceratops (51); jellyfish (107) and snail (113); king crab (121) and coral reef (973).



Figure 8: **Uncurated samples from PoD-G/16 on ImageNet** 256×256**, objects, food and nature.** Each row shows two classes, four samples each, with the ImageNet class index in parentheses. From the top, the rows show castle (483) and dome (538); fountain (562) and space shuttle (812); cannon (471) and tractor (866); pizza (963) and bell pepper (945); espresso (967) and potpie (964); cliff (972) and rapeseed (984); yellow lady's slipper (986) and mushroom (947); earthstar (995) and acorn (988).



Figure 9: **Uncurated samples from PoD-G/32 on ImageNet** 512×512**, animals.** Each row shows two classes, four samples each, with the ImageNet class index in parentheses. From the top, the rows show goldfish (1) and cock (7); ostrich (9) and kite (21); vulture (23) and king snake (56); macaw (88) and flamingo (130); Sussex spaniel (220) and miniature poodle (266); jaguar (290) and cheetah (293); American black bear (295) and ice bear (296); hare (331) and marmoset (377).

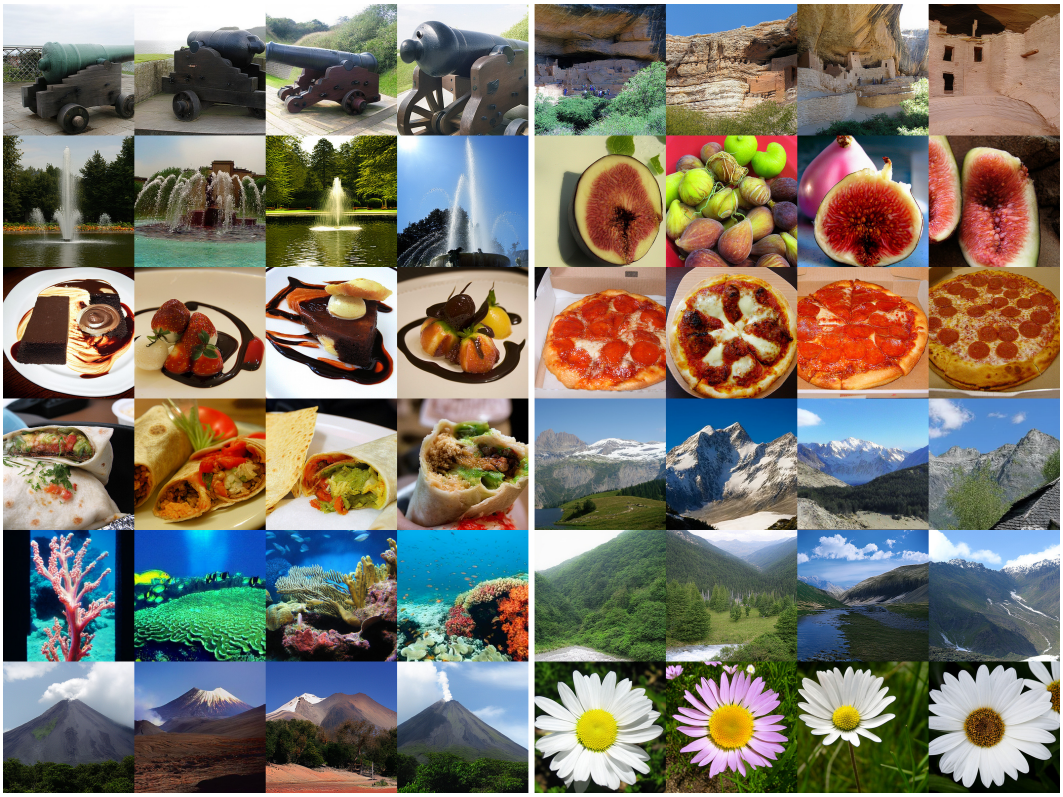

Figure 10: **Uncurated samples from PoD-G/32 on ImageNet** 512×512**, objects, food and scenes.** Each row shows two classes, four samples each, with the ImageNet class index in parentheses. From the top, the rows show cannon (471) and cliff dwelling (500); fountain (562) and fig (952); chocolate sauce (960) and pizza (963); burrito (965) and alp (970); coral reef (973) and valley (979); volcano (980) and daisy (985).