Thermodynamic Illumination: Measuring the Volume of Structured Images Under Neural Network Priors

Matthew Kalenuik
Independent Researcher
British Columbia, Canada
matthew.kalenuik@icloud.com

February 11, 2026

Abstract

Inductive bias is often discussed qualitatively. We introduce *Thermodynamic Illumination*, a framework to quantify the inductive biases of neural network architectures by measuring the *density of structured images* (those with non-random patterns distinguishing them from noise) under their random-weight priors. Using nested sampling from statistical physics, we estimate prior volumes for images exceeding structure thresholds.

Using coordinate-conditioned architectures with matched inputs, we find CoordConvNet reaches $\tau = 0.1$ at 0.72 NS crossing bits ($B_{\text{NS}}^{\text{cross}}$) with a 64% initial live-point pass fraction, while CoordViT reaches $\tau = 0.1$ at 2.67 $B_{\text{NS}}^{\text{cross}}$ bits with a 3% initial live-point pass fraction. This is a $\sim$2-bit $B_{\text{NS}}^{\text{cross}}$ threshold-crossing gap and a separate $\sim 21\times$ live-point pass-fraction gap. This controlled comparison isolates locality with weight sharing (convolutions) versus global mixing (attention) as the dominant architectural source of bias. In an independent prior-comparison baseline, CPPN reaches $\tau = 0.1$ at $\sim$1.9 $B_{\text{NS}}^{\text{cross}}$ bits while uniform remains above the explored $B_{\text{NS}}^{\text{cross}} > 72$ bit floor.

We validate that thermodynamic volume predicts reconstruction quality (Spearman $\rho = 0.874$, $p = 0.0001$, $n = 13$ architectures) and optimization dynamics: low-$B_{\text{NS}}^{\text{cross}}$ architectures achieve 12dB better denoising in Deep Image Prior experiments. Across 13 architectures, we discover three thermodynamic regimes—*Shielded* (forces generalization), *Memorization* (fits without generalizing), and *Broken* (fails untrained reconstruction)—with structure predicting generalization gap (Spearman $\rho = 0.79$, permutation $p = 0.002$). The bits metric does *not* predict random-feature linear classification, confirming it captures generative priors specifically.

Our work transforms inductive bias from a qualitative notion to a measurable quantity, enabling principled architecture selection for generative tasks.

1 Introduction

Vision Transformers require substantially more training data than convolutional networks to achieve comparable performance on image tasks [14]. Why? The standard explanation invokes "inductive bias"—convolutional architectures encode assumptions about spatial locality and translation invariance that transformers must learn from data. But this explanation remains qualitative. *How much* more bias do ConvNets have? Can we measure it? Our framework quantifies architectural priors for image generation and reconstruction; while we show this metric does not directly predict classification performance (Section 5.7), the absence of generative priors in untrained ViTs is consistent with their need to learn such structure from data.

Consider the Deep Image Prior [16]: a randomly-initialized convolutional network, when optimized to reconstruct a corrupted image, produces clean natural images rather than noise—without any training data. Vision Transformers underperform substantially in this setup (about 10dB vs 25dB for ConvNets). This striking result demonstrates that CNN architecture alone encodes a powerful prior over images. But *how powerful*? And how does this compare to transformers?

The challenge is measurement. We cannot enumerate all possible outputs of a neural network, and naive sampling fails catastrophically for rare events. For architectures like uniform random pixels, where structured images occupy 10^{-20} of output space, random sampling would require cosmic timescales. However, for architectures with favorable inductive biases like CPPNs, the story is dramatically different.

Our contribution. We introduce *Thermodynamic Illumination*, a framework that quantifies inductive bias by measuring the *volume of structured images* under a network's random-weight prior. Borrowing nested sampling from statistical physics [12], we efficiently estimate probabilities as small as 10^{-22} (72 bits of prior volume), answering:

"What fraction of a network's output space consists of structured images?"

We report two related quantities. **Tail-surprisal bits** are

$$B_{\mathrm{MC}}(\tau) = -\log_2 p(\tau), \quad p(\tau) = \Pr[\mathrm{order}(x) \geq \tau].$$

This is a direct prior-mass quantity estimated by Monte Carlo. For nested-sampling trajectories we also report a **threshold-crossing cost**

$$B_{\mathrm{NS}}^{\mathrm{cross}}(\tau) = \frac{i_\tau}{N_{\mathrm{live}} \ln 2},$$

where i_τ is the first iteration whose running threshold exceeds τ. $B_{\mathrm{NS}}^{\mathrm{cross}}$ is a search-cost proxy; B_{MC} is a direct tail-mass estimate. We use "thermodynamic volume" to denote probability mass under the architecture-induced prior (the pushforward of the weight distribution through the network), not geometric volume in image space. Probability-ratio interpretations are made from B_{MC} only; $B_{\mathrm{NS}}^{\mathrm{cross}}$ is interpreted as an operational threshold-crossing cost. At matched calibration settings, $B_{\mathrm{NS}}^{\mathrm{cross}}$ shows a small positive bias in analytic tests (about 7% mean overestimation; Appendix E), so large NS crossing-cost separations are conservative under our fixed protocol.

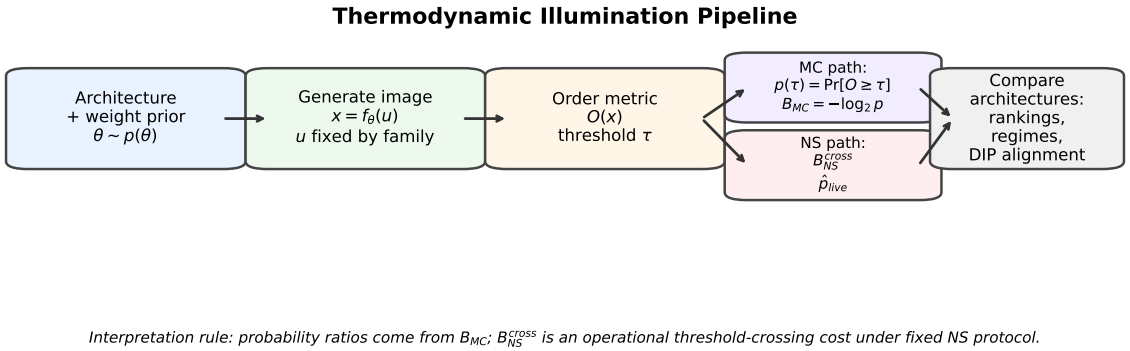

Figure 1: Thermodynamic Illumination schematic. An architecture with an explicit random-weight prior induces an image distribution $P(x)$. An order metric $O(x)$ and threshold τ define "structured" events. When Monte Carlo is feasible, we estimate tail mass $p(\tau)$ and tail-surprisal bits $B_{\mathrm{MC}}(\tau) = -\log_2 p(\tau)$. Nested sampling yields an operational threshold-crossing cost $B_{\mathrm{NS}}^{\mathrm{cross}}(\tau)$ and initial live-point pass fraction $\hat{p}_{\mathrm{live}}(\tau)$ under a fixed protocol.

Our contributions are:

1. **A novel measurement framework**: We adapt nested sampling to estimate the density of structured images under neural network priors, enabling quantitative comparison of architectures that was previously impossible.
2. **Massive differences revealed**: In a controlled coordinate-conditioned 32×32 comparison, $B_{\mathrm{NS}}^{\mathrm{cross}}(\tau{=}0.1)$ is 0.72/2.11/2.67 bits for CoordConvNet/CoordMLP/CoordViT with initial live-point pass fractions 64%/21%/3%. In independent prior-comparison baselines, CPPN reaches $\tau = 0.1$ at $\sim$1.9 $B_{\mathrm{NS}}^{\mathrm{cross}}$ bits while uniform remains above the explored $B_{\mathrm{NS}}^{\mathrm{cross}} > 72$ bit floor.

3. **Predictive validation**: We demonstrate that $B_{\text{NS}}^{\text{cross}}$ strongly predicts reconstruction quality (Spearman $\rho = 0.874$, $p = 0.0001$, $n = 13$), providing practical utility beyond theoretical interest.

4. **A Generative-Discriminative Trade-off**: We observe that architectures optimal for reconstruction (low $B_{\text{NS}}^{\text{cross}}$) are suboptimal for classification, and vice versa. This trade-off, demonstrated on two classification datasets, suggests why structure-based metrics predict reconstruction but not classification.

5. **Scale validation**: We report threshold-matched CPPN size-scaling exponents (Appendix J) and run additional scale-normalized convolutional sweeps up to 1024×1024. Scaling claims are restricted to matched thresholds and architecture families.

Key Results at a Glance:

- **The ConvNet-ViT Gap**: With matched coordinate inputs, CoordConvNet reaches $\tau = 0.1$ at $B_{\text{NS}}^{\text{cross}} = 0.72$ bits with 64% initial live-point pass fraction, while CoordViT reaches $B_{\text{NS}}^{\text{cross}} = 2.67$ bits with 3% initial live-point pass fraction—a ∼2-bit $B_{\text{NS}}^{\text{cross}}$ gap and a separate ∼21× live-point pass-fraction gap, isolating locality with weight sharing versus global mixing
- **Kinetic Validation**: Low-$B_{\text{NS}}^{\text{cross}}$ architectures achieve 12dB better denoising in Deep Image Prior experiments
- **Three Regimes**: Architectures divide into Shielded (forces generalization), Memorization (fits without generalizing), and Broken (fails untrained generative reconstruction)
- **Generative-Specific**: Predicts reconstruction ($\rho = 0.874$) but NOT random-feature classification—confirming the metric captures generative priors
- **Scale-Validated**: Threshold-matched CPPN scaling is explicit ($\beta \approx 0.80$ at $\tau = 0.1$, $\beta \approx 1.45$ at $\tau = 0.25$), and high-resolution sweeps extend to 1024×1024 with scope-limited comparisons

Initialization is part of the prior. All headline comparisons fix a single explicit weight prior within each comparison (no fan-in scaling). Alternative initializations (He/Xavier) define different priors and can materially change pass fractions for some architectures (notably CoordViT), so we treat initialization as part of the object of study and report sensitivity in Appendix K.3. Claims about "bias" are therefore scoped to the stated protocol and weight prior.

Reproducibility (Public Release).

- Code, figures, and artifact bundles: `https://github.com/kaneda2004/thermodynamic-illumination-public` (tag `arxiv-v1`)
- Quickstart: `uv sync` then `uv run python reproduce_results.py`
- Determinism: scripts set fixed seeds in their configs; see `REPRODUCE.md` for the exact seeds and run counts per table/figure
- Submission artifacts: the exact arXiv flat-source upload is included as `paper/ax.tar` and `paper/arxiv_flat_source/`

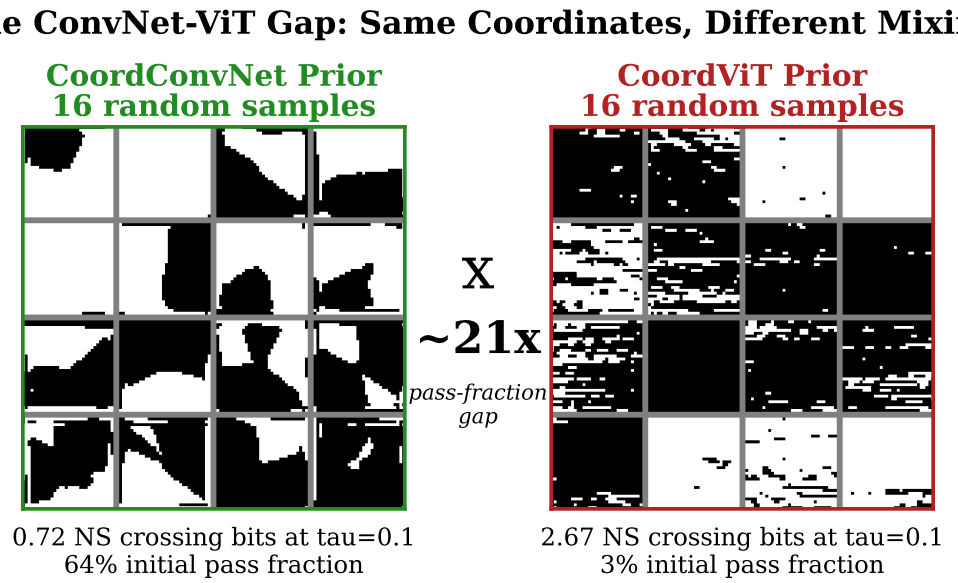

Figure 2: The ConvNet-ViT efficiency gap. **Left**: A CoordConvNet prior ($B_{\text{NS}}^{\text{cross}} = 0.72$ bits at $\tau = 0.1$, 64% initial live-point pass fraction) produces structured images—curves, edges, and geometric patterns emerge from 3×3 local convolutions. **Right**: A CoordViT prior ($B_{\text{NS}}^{\text{cross}} = 2.67$ bits at $\tau = 0.1$, 3% initial live-point pass fraction) produces noise-like outputs under this untrained prior. Same coordinate inputs and random weights, but higher NS crossing cost and lower initial live-point pass fraction for CoordViT. The shown bits are $B_{\text{NS}}^{\text{cross}}$ (crossing cost), not direct tail-surprisal B_{MC}.

Paper organization. Section 2 reviews related work on inductive biases and untrained networks. Section 3 provides background on nested sampling and order metrics. Section 4 presents our measurement framework. Section 5 validates the framework through prior comparison, reconstruction, and classification experiments. Section 6 discusses implications and limitations, and Section 7 concludes.

2 Related Work

Deep Image Prior and Untrained Networks. The discovery that untrained CNNs serve as powerful image priors [16] sparked interest in understanding architecture-induced biases. Subsequent work showed that network depth, skip connections, and activation functions all influence the prior [3]. However, these works demonstrate bias qualitatively through reconstruction examples. Our framework provides the first *quantitative* comparison, enabling statements like "CPPN priors require at least $10^{21}\times$ less threshold-crossing effort than uniform priors under a fixed nested-sampling protocol."

Weight Agnostic Neural Networks. Gaier & Ha [2] evolved network topologies that solve tasks with *any* random weight assignment, demonstrating that architecture alone can encode solutions. Their work asks "what tasks can an architecture solve?" while we ask "what outputs does an architecture prefer?" These are complementary: task performance requires both appropriate bias *and* sufficient capacity, while our metric isolates the bias component.

Zero-Shot Neural Architecture Search. Mellor et al. [8] introduced metrics to predict trained accuracy from untrained network statistics, enabling neural architecture search without training. Their metrics examine activation patterns and gradient flow. Our key finding is that such predictions are *task-dependent*: bits predicts reconstruction but shows no significant relationship with classification on our tested datasets. This reveals a trade-off between generative and discriminative optimization that activation-based metrics cannot detect.

Compositional Pattern Producing Networks. CPPNs [13] were introduced for evolutionary art, producing structured patterns through coordinate-based mappings with specialized activation functions. The Picbreeder system [11] demonstrated their creative potential through human-guided evolution. We provide the first quantitative measure of CPPN bias strength, explaining *why* they produce structure: they occupy an unusually dense region of image space.

Physics-Inspired Methods for Neural Networks. Wang-Landau sampling and related methods have been applied to map loss landscapes in weight space [7]. Random feature theory [17] provides theoretical grounding for untrained networks. Lee et al. [5] showed that infinitely wide deep networks are equivalent to Gaussian processes with architecture-dependent covariance functions, providing a theoretical lens on untrained network priors that complements our empirical sampling approach. Our work applies nested sampling [12] to *output* space, measuring how much of an architecture's output manifold contains structured images.

Positioning. Unlike prior work that examines specific architectures or tasks, we provide a general framework for measuring and comparing structural bias across any architecture. Our bits metric is architecture-agnostic and measurement-task-agnostic for untrained generative structure; downstream supervised performance remains task-dependent.

3 Background

Before presenting our method, we introduce two key concepts: nested sampling for rare event estimation, and order metrics for quantifying image structure.

3.1 Nested Sampling

Nested sampling [12] is a Monte Carlo algorithm originally developed for Bayesian evidence computation. Its key insight is that exploring probability space by *order statistic* rather than by region enables efficient estimation of exponentially small probabilities.

The algorithm maintains N "live points" sampled from a prior distribution. At each iteration, the lowest-likelihood point is removed and replaced with a new sample constrained to have higher likelihood. This process compresses the prior volume geometrically: after i iterations, the remaining volume is approximately $X_i \approx e^{-i/N}$.

This geometric shrinkage makes nested sampling ideally suited for estimating rare events. While naive rejection sampling requires $\mathcal{O}(1/p)$ samples to find an event with probability p, nested sampling requires only $\mathcal{O}(N \log(1/p))$ samples—exponentially more efficient for rare events.

3.2 Quantifying Image Structure

What makes an image "structured" versus "random"? We seek metrics that capture intuitive notions of order while remaining computationally tractable. A good order metric should:

1. Assign high values to images with recognizable patterns (symmetry, connectivity, regularity)
2. Assign low values to noise-like images
3. Be robust to single-pixel perturbations (not dominated by noise)
4. Avoid trivial solutions (all-black, all-white)

We use a multiplicative combination of four such properties (compressibility, symmetry, connectivity, and color balance), detailed in Section 4.3. The multiplicative form ensures that *all* properties must be present—a highly compressible but disconnected image scores low, as does a symmetric but noisy image.

4 Method

4.1 Problem Formulation

Let $P(x)$ be the distribution of images induced by a generative prior (e.g., a network with random weights). Let $O(x) : \mathcal{X} \to [0, 1]$ be an *order metric* measuring image structure (1 = highly ordered, 0 = random noise).

We seek to estimate the *density of structured images*:

$$V(\tau) = \Pr_{x \sim P}[O(x) \geq \tau] \tag{1}$$

This is analogous to the *density of states* in statistical mechanics: $V(\tau)$ measures the fraction of configuration space at or above energy level τ. The key difference is that we measure in *output space* (images) rather than weight space.

The theoretical tail-surprisal at threshold τ is:

$$B_{\text{MC}}(\tau) = -\log_2 V(\tau) \tag{2}$$

This has an information-theoretic interpretation: $B_{\text{MC}}(\tau)$ is the *surprisal* (negative log probability) of drawing a structured image from the prior—equivalently, the bits of prior volume required. Random sampling requires $\mathcal{O}(2^{B_{\text{MC}}(\tau)})$ samples to find one image with $O(x) \geq \tau$.

Tail-surprisal bits enable cross-architecture probability comparison: if one prior has $B_{\text{MC}}(\tau) = 2$ and another has $B_{\text{MC}}(\tau) > 72$ at the same threshold, then the corresponding tail probabilities differ by at least $2^{70} \approx 10^{21}$.

Interpreting measurements. We distinguish three types of quantities:

- **Probabilities from Monte Carlo**: Direct sampling provides empirical bounds. With 0/1,000,000 ViTDecoder samples exceeding $\tau = 0.1$, we have $P < 10^{-6}$ (one-sided bound). ConvDecoder's own architecture-induced prior achieves 51.2% at the same threshold—a >500,000× gap versus the ViTDecoder prior.
- **Mean order values**: Some architectures produce mean order scores $\sim 10^{-23}$ (Table 5). This is a metric magnitude, not a probability—it reflects how far typical outputs are from the structure threshold.
- **Bits from nested sampling**: NS reports an operational threshold-crossing cost $B_{\text{NS}}^{\text{cross}}(\tau) = i_\tau/(N_{\text{live}} \ln 2)$ (Section 5.14). CPPN reaches $\tau = 0.1$ at $\sim$2 $B_{\text{NS}}^{\text{cross}}$ bits; uniform pixels remain un-crossed past the explored $B_{\text{NS}}^{\text{cross}} > 72$ bit floor, giving a rigorous lower bound on threshold-crossing effort under this fixed protocol (not a calibrated B_{MC} probability ratio).

The key result is the *relative gap* between architectures, which is robust across metrics and thresholds.

4.2 Nested Sampling

We use nested sampling [12] to estimate $V(\tau)$ efficiently. The algorithm maintains N "live points" from the prior and progressively restricts to higher-order regions:

Algorithm 1 Nested Sampling for Structure Density

```
Initialize N samples from prior P(x)
for i = 1 to M iterations do
    Find lowest-order sample x_min
    Record (x_min, O(x_min), log X_i) as "dead point"
    Replace x_min with new sample having higher order
    log X_i ← −(i + 1)/N (volume shrinkage)
end for
return dead points with volume estimates
```

The key property is *geometric volume shrinkage*: after i iterations, the remaining prior volume is $X_i \approx e^{-i/N}$. This enables exploration of exponentially rare regions with only linear cost in iteration count.

Reported NS quantities. For threshold τ, let i_τ denote the first iteration where the running best threshold exceeds τ. We report

$$B_{\mathrm{NS}}^{\mathrm{cross}}(\tau) = \frac{i_\tau}{N_{\mathrm{live}} \ln 2},$$

which captures threshold-crossing search cost under a fixed NS protocol. We also report the initial live-point pass fraction

$$\hat{p}_{\mathrm{live}}(\tau) = \frac{1}{N_{\mathrm{live}}} \sum_{j=1}^{N_{\mathrm{live}}} \mathbf{1}[O(x_j^{(0)}) \geq \tau].$$

$\hat{p}_{\mathrm{live}}$ is an initial Monte Carlo estimate from live points; $B_{\mathrm{NS}}^{\mathrm{cross}}$ is a trajectory cost. They are related but not numerically identical at finite sample sizes.

Prior-Preserving Sampling. The critical step is replacing $x_{\min}$ with a new sample that (1) exceeds the threshold $O(x) > \tau$ and (2) is distributed according to the prior P. For neural network priors parameterized by weights or latent codes drawn from $\mathcal{N}(0, I)$ (CPPN network parameters; MLP latent codes), we use Elliptical Slice Sampling (ESS) [9]:

1. Sample auxiliary variable $\nu \sim \mathcal{N}(0, I)$
2. Choose angle uniformly: $\phi \sim \mathrm{Uniform}[0, 2\pi]$
3. Propose $w' = w \cos\phi + \nu \sin\phi$
4. Accept if $O(\mathrm{decode}(w')) > \tau$, else shrink bracket and repeat

ESS is exact in the idealized algorithm (no Metropolis-Hastings acceptance ratio), ensuring Gaussian-prior preservation. Our implementation adds finite contraction/restart caps for robustness; diagnostics in Appendix H show stable mixing and no evidence these limits drive reported results. Importantly, our acceptance test is *order-threshold based* (accept if $O(\mathrm{decode}(w')) > \tau$), not likelihood-based—the likelihood threshold constraint is measurable and the decoding function is deterministic, satisfying ESS correctness conditions.

Prior Specification. We formalize the generative process as $x = f_\theta(u)$ where $\theta \sim \mathcal{N}(0, \sigma^2 I)$ are network weights and u is an architecture-dependent input:

- **Coordinate-based** (CPPN, Fourier Features): u is a fixed coordinate grid $\{(i/N, j/N)\}_{i,j=1}^{N}$
- **Latent-decoded** (VAE, GAN decoders): $u \sim \mathcal{N}(0, I)$ is a random latent vector
- **Direct-decoder** (ConvNet, ViT, MLP decoders): u is fixed seed noise or learned embedding

Table 1: Prior contract by architecture family

Family	Random variables	Held fixed	Induced image prior
Coordinate-conditioned	$\theta \sim \mathcal{N}(0, \sigma^2 I)$	coordinate grid (x, y) or (x, y, r)	$x = f_\theta(u_{\mathrm{coord}})$
Latent-decoded	$\theta \sim \mathcal{N}(0, \sigma^2 I)$, $z \sim \mathcal{N}(0, I)$	decoder architecture	$x = f_\theta(z)$
Direct-decoder	$\theta \sim \mathcal{N}(0, \sigma^2 I)$	seed tensor u_0 / embeddings	$x = f_\theta(u_0)$
Procedural/statistical	method state ξ from method-specific prior	generation rules/hyperparameters	$x = g_\xi$

Our "prior" is thus the distribution over images induced by randomness in θ (and u where applicable). We sample weights from $\mathcal{N}(0, \sigma^2 I)$ with σ fixed per experiment ($\sigma = 0.3$ for the 32×32 coordinate-input hierarchy; $\sigma = 1.0$ for the 64×64 RGB spectrum and other experiments unless noted). We deliberately avoid fan-in scaling (He/Xavier initialization) because our goal is to characterize *architecture-as-prior*, not

trained-network behavior. BatchNorm running statistics are kept at PyTorch eval defaults (running mean = 0, running variance = 1), and BatchNorm affine parameters are sampled with the rest of θ unless otherwise noted. Because initialization can affect pass fractions for some architectures (Appendix K.3), we hold σ and initialization fixed within each comparison. Appendix G further confirms that architecture rankings are preserved across different order metric formulations.

4.3 Order Metrics

We use a **multiplicative metric** combining four factors, each normalized to $[0, 1]$:

$$O(x) = O_{\text{compress}}(x) \times O_{\text{symmetry}}(x) \times O_{\text{connectivity}}(x) \times O_{\text{balance}}(x) \tag{3}$$

Compressibility measures pattern regularity via compression ratio:

$$O_{\text{compress}}(x) = 1 - \frac{\text{compressed\_size}(x)}{\text{raw\_size}(x)} \tag{4}$$

We use zlib compression on the binary image. Random noise is incompressible ($O_{\text{compress}} \approx 0$); regular patterns compress well ($O_{\text{compress}} \to 1$).

Symmetry measures bilateral reflection similarity:

$$O_{\text{symmetry}}(x) = 1 - \frac{\|x - \text{flip}(x)\|_1}{n_{\text{pixels}}} \tag{5}$$

This captures a basic structural property common in natural and designed images.

Connectivity measures whether foreground forms a single connected component:

$$O_{\text{connectivity}}(x) = \frac{\text{largest\_component\_size}}{\text{total\_foreground\_pixels}} \tag{6}$$

This penalizes scattered, fragmented images in favor of coherent shapes. For images with zero foreground pixels, connectivity is defined as 0.

Color Balance penalizes trivial all-black or all-white images:

$$O_{\text{balance}}(x) = 4 \cdot p \cdot (1 - p) \tag{7}$$

where p is the fraction of white pixels. This peaks at $p = 0.5$ and goes to zero at extremes.

The multiplicative form ensures *all* properties must be present—a highly compressible but disconnected image scores low, as does a symmetric but imbalanced image.

RGB Order Metric Used in Sections 5.1, 5.2, 5.3, 5.10, and 5.11. For RGB experiments we use

$$O_{\text{RGB}}(x) = C_{\text{JPEG}}(x) \times S_{\text{TV}}(x), \tag{8}$$

with

$$C_{\text{JPEG}}(x) = \max\left(0,\ 1 - \frac{b_{\text{JPEG}}(x; Q = 85)}{b_{\text{raw}}(x)}\right), \tag{9}$$

$$\text{TV}(x) = \frac{1}{N} \sum |x_{i+1,j,c} - x_{i,j,c}| + \frac{1}{N} \sum |x_{i,j+1,c} - x_{i,j,c}|, \tag{10}$$

$$S_{\text{TV}}(x) = \exp(-10 \cdot \text{TV}(x)). \tag{11}$$

Here $x \in [0, 1]^{H \times W \times C}$ before encoding; b_{JPEG} uses 8-bit JPEG at quality 85.

Why Gates Are Necessary: Excluding Degenerate Solutions. Simple structure metrics—compression ratio, autocorrelation, spectral entropy—can be maximized by *degenerate* outputs that are technically "structured" but visually uninteresting:

- **Constant images** (all black or all white): perfectly compressible, perfect autocorrelation, zero entropy
- **Uniform gray**: high local variance, balanced density
- **Checkerboard patterns**: high compressibility, regular edges

Without explicit gates, architectures producing such degenerate outputs would score highest on simple metrics. For example, an MLP with random weights often produces nearly constant images (due to saturated activations), achieving near-perfect compression and autocorrelation scores. Our multiplicative design addresses this by requiring *multiple necessary conditions*: non-trivial density (Color Balance gate), presence of edges (Connectivity gate), moderate compressibility (not trivial patterns), and spatial coherence (Symmetry gate). Only images satisfying *all* conditions—excluding both random noise *and* degenerate constants—score highly.

We validated this design empirically: our initial tests on 8 simple metrics (compression, autocorrelation, entropy, etc.) found only 2 produced consistent rankings when comparing binary image priors (CPPN, MLP, Walk, Uniform)—a failure mode we initially attributed to MLP degeneracy. However, extended validation across 19 metrics and 5 architectures (below) revealed a more nuanced picture: most standard metrics *do* rank structured architectures highest when CPPN (which produces smooth, non-degenerate outputs) is included. The key insight is that our multiplicative gates provide *one* valid approach to measuring structure, but are not uniquely necessary—multiple principled approaches converge on the same rankings. We use *percentile-based thresholds* for comparison (pooling across architectures to define fair cut points), ensuring rankings are not artifacts of scale differences between metrics.

Robustness Across Standard Metrics. To address concerns about metric arbitrariness, we tested 19 diverse metrics from the image quality, texture analysis, and signal processing literatures—including compression ratio, autocorrelation, GLCM features, LBP entropy, Betti numbers (persistent homology), NIQE, BRISQUE, spectral centroid, spectral slope, and high-frequency energy ratio. **Result: 15/19 metrics (79%) correctly rank structured architectures (CPPN, ConvNet, ResNet) highest** when accounting for metric direction (e.g., lower HF-ratio = more structured). Only 1/19 metrics falls into a potential degeneracy trap (LBP entropy ranks MLP highest), while 3/19 measure orthogonal properties like edge density where uniform noise wins trivially. The spectral slope provides particularly strong validation: CPPN achieves $\beta = -2.75$ (close to the $1/f^2$ natural image law), while MLP shows $\beta = +0.06$ (flat spectrum). Our gated metric correlates strongly (Spearman $\rho = -0.91$) with this standard, non-heuristic measure from signal processing. See Appendix G, Table 26 for full results.

Note that multiplicative metrics impose implicit AND logic, which can create sharp transitions as thresholds approach 1. However, the architecture-dependent *location* of these transitions is robust (Appendix G shows consistent rankings across metrics), confirming our findings reflect genuine architectural differences rather than metric construction. **Key robustness claim:** the ranking of architectures by thermodynamic volume is preserved across all reasonable structure metrics we tested—the specific metric choice affects absolute values but not the ordering.

What "Structure" Means. We emphasize that "structure" in this framework refers specifically to geometric regularity: spatial coherence, low-frequency dominance, and compressibility—properties ubiquitous in natural image statistics ($1/f$ spectra). This is a narrower notion than semantic structure, but it is precisely this geometric prior that distinguishes convolutional from attention-based architectures at initialization. Our goal is not to capture all notions of structure, but to quantify the specific inductive bias that predicts downstream task performance. Independent spectral analysis confirms this: high-order architectures exhibit $1/f$ frequency scaling characteristic of natural images, while low-order architectures show flat spectra (Appendix F).

Scale Normalization. The edge density component (measured via connectivity) scales as $O(1/N)$ for smooth CPPN patterns, requiring resolution-dependent gate calibration. We normalize the edge gate center by resolution: $\text{center}(N) = 0.15 \times (32/N)$, ensuring that the same CPPN produces consistent order scores

when sampled at different resolutions. This preserves the metric's discriminative power while enabling principled cross-resolution comparison.

4.4 Architectures Tested

We test architectures across two experimental setups with different image formats and order metrics (see Table 2).

4.4.1 Baseline Priors (32×32 Binary)

For foundational validation (Section 5.5), we test four priors spanning extreme structural biases:

CPPN (Compositional Pattern Producing Network). CPPNs [13] map spatial coordinates to pixel values: $f : (x, y, r) \to [0, 1]$, where $r = \sqrt{x^2 + y^2}$ provides radial information. We use tanh activations throughout, which produce smooth, bounded outputs. The coordinate-based parameterization induces strong spatial coherence—nearby pixels receive similar inputs and thus similar outputs. Architecture: 2 hidden layers of 32 units (∼3,600 parameters).

MLP (Multi-Layer Perceptron). The MLP maps a latent vector to a flattened image: $f : \mathbb{R}^d \to \mathbb{R}^{n \times n}$. Unlike CPPNs, there is no explicit spatial structure—pixels are independent output dimensions. This acts as a random projection [4], which preserves distances but destroys topology. Architecture: 2 hidden layers of 128 units with ReLU (∼180,000 parameters).

Walk Prior. The walk prior generates images via random walks on a grid, producing connected line-like structures. Starting from a random position, the walk takes steps in random directions, marking visited pixels. This encodes connectivity by construction but lacks other structure.

Uniform Prior (Baseline). The maximum entropy baseline: each pixel is independently sampled from Bernoulli(0.5). This represents "no prior" and serves as the null hypothesis.

4.4.2 Neural Network Architectures (64×64 RGB)

For the main experiments (Sections 5.1–5.3), we test 13 architectures spanning convolutional, attention-based, coordinate-based, and fully-connected families:

Convolutional Networks (6 variants). *ResNet* variants with 2, 4, 6, and 9×9 kernel sizes use residual blocks with BatchNorm, producing strong locality bias. *U-Net* adds skip connections between encoder and decoder. *Depthwise Separable Conv* factorizes spatial and channel mixing. Across these variants, parameter counts span 23K–2.1M at 64×64 RGB.

Attention-Based Networks (4 variants). *Vision Transformer (ViT)* uses 4 transformer layers with 4 attention heads, patch size 8×8, embedding dimension 128, and learnable positional embeddings (563K parameters). *Windowed ViT* restricts attention to local windows. *Local Attention* uses fixed local receptive fields. *Hybrid ViT* combines convolutional stems with transformer blocks.

Coordinate-Based Networks (2 variants). *CPPN* as described above, scaled to 64×64 RGB output. *Fourier Features* [15] use sinusoidal positional encodings before an MLP, enabling high-frequency pattern generation.

Fully-Connected (1 variant). *MLP* maps a latent vector directly to all pixels without spatial structure (6.5M parameters at 64×64).

All weights are sampled from $\mathcal{N}(0, 1)$ without fan-in scaling, as our goal is to characterize architecture-as-prior. See Appendix C for detailed specifications.

4.5 Validation Against Ground Truth

To validate our nested sampling implementation, we test against metrics with *known* analytic probabilities. For example, the probability of a random $n \times n$ binary image having mean pixel value $\geq T$ follows the binomial distribution exactly. At experimental resolutions (32×32, 64×64), our estimates show systematic overestimation of 6–17% with mean 9%. This ∼7% bias floor persists regardless of live points—additional sampling reduces variance but not systematic bias. Relative orderings remain valid; a 7% correction on a 70-bit gap still yields $> 10^{19}$-fold efficiency differences. See Appendix E for extended calibration.

5 Experiments

We conduct three experiments: (1) comparing prior volumes across architectures, (2) validating correlation with reconstruction quality, and (3) testing correlation with classification accuracy. All experiments use 32×32 binary images unless otherwise noted.

Order Metrics by Experiment. For clarity, Table 2 specifies which order metric is used in each experiment section, since different image types (binary vs RGB) require different metrics. **Important:** Bits are not compared numerically across metric families; only within-experiment rankings are interpreted.

Table 2: Order metrics used in each experiment section

Section	Image Type	Order Metric
5.1 ConvNets vs Transformers	64×64 RGB	Compression + TV
5.2 Deep Image Prior	64×64 RGB	Compression + TV
5.3 Structure-Generalization	64×64 RGB	Compression + TV
5.4 Monte Carlo Validation	64×64 binary	Multiplicative (Eq. 3)
5.5 Prior Comparison	32×32 binary	Multiplicative (Eq. 3)
5.6 Reconstruction	28×28 grayscale[‡]	Multiplicative
5.7 Classification	28×28 grayscale[‡]	Multiplicative
5.10 RGB Scaling	32×32 RGB	Compression + TV
5.11 Thermodynamic Alignment	64×64 RGB	Compression + TV

[‡]Grayscale images binarized at threshold 0.5 before applying multiplicative metric.

Summary of Claims and Evidence. Table 3 summarizes our key quantitative claims with their measurement methods and uncertainty bounds.

Table 3: Summary of key claims and supporting evidence

Claim	Method	Samples	Result	Uncertainty
ConvNet $P(\tau{=}0.1) = 51\%$	MC	10,000	51.2%	±3% (bootstrap)
ViT $P(\tau{=}0.1) < 10^{-6}$	MC	1,000,000	0/1M	One-sided bound
CPPN $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1) \approx 2$	NS	500 live	2.1 bits	±7% (App. E)
Uniform $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1) > 72$	NS	500 live	>72 bits	Lower bound
$B_{\text{NS}}^{\text{cross}}$–reconstruction ρ	Spearman	13 arch.	0.874	$p = 0.0001$
$B_{\text{NS}}^{\text{cross}}$–classification ρ	Spearman	24 arch.	-0.24	$p = 0.27$ (n.s.)

5.1 Inductive Bias Spectrum: ConvNets vs Transformers

The debate between convolutional networks and Vision Transformers (ViT) typically focuses on trained performance. Our framework enables comparison of their *untrained* priors—revealing surprising differences.

Setup. We compare three 64×64 RGB architectures:

- **ResNet** (245K params): 4-layer convolutional decoder with BatchNorm
- **ViT** (563K params): 4-layer transformer with learnable positional embeddings
- **MLP** (6.5M params): 3-layer fully-connected network

Hypothesis. We expected ViT to show *medium* inductive bias—between ResNet (strong) and MLP (weak)—since positional embeddings encode spatial information.

Architecture Details. The ViT uses 4 transformer layers with 4 attention heads, patch size 8×8 (8 patches per row), embedding dimension 128, MLP ratio 2:1, LayerNorm, and learnable positional embeddings (563K parameters total). The ResNet uses 4 residual blocks with 3×3 convolutions and BatchNorm (245K parameters). The MLP is a 3-layer fully-connected network with ReLU activations (6.5M parameters). All

weights are sampled from $\mathcal{N}(0, 1)$ without fan-in scaling. For BatchNorm layers, we evaluate in inference mode with default running statistics, and BN affine parameters are sampled with the same Gaussian prior unless explicitly ablated. We emphasize that this is a probe of *untrained generator* behavior in a generative setup, not a statement about trained ViT performance on discriminative tasks (e.g., ImageNet classification) where ViTs excel. We treat initialization as part of the prior and hold it fixed within each comparison (Appendix K.3).

Results. Figure 3 reveals a surprising finding: **in our untrained image-generation setup, Global ViT is thermodynamically close to MLP**. Both flatline at score ≈ 0.0001, while ResNet reaches 0.84 after exploring 11 bits of prior volume ($-\log_2 X$), crossing $\tau = 0.1$ at $B_{\text{NS}}^{\text{cross}} = 0.96$ bits. This does not imply ViTs lack inductive bias when trained on real data—only that their *untrained output distribution* shows no structural preference. (Windowed ViT variants show similar zero-structure behavior but differ in optimization dynamics; see Section 5.3.)

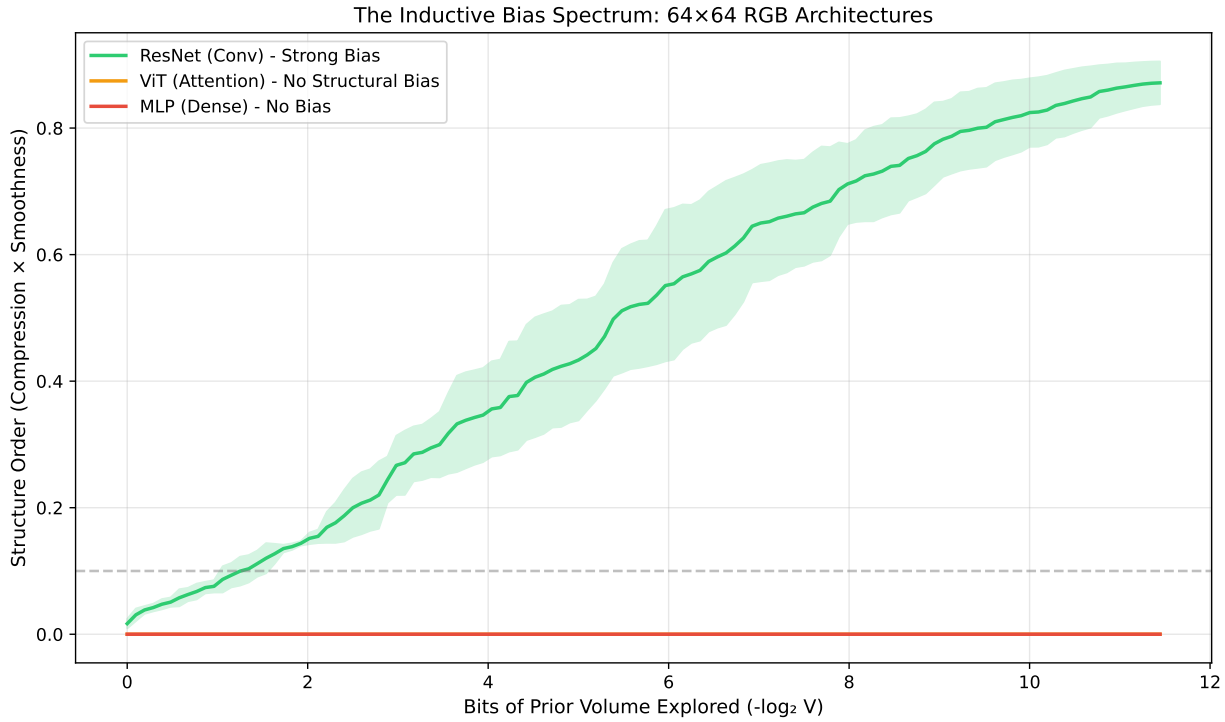

Figure 3: Inductive Bias Spectrum at 64×64: ResNet (green) shows strong structural bias, reaching high scores immediately. ViT (purple) and MLP (red) both flatline near zero—untrained ViT has *no* structural advantage over MLP despite positional embeddings.

Interpretation. Positional embeddings tell attention *where* patches are but don't constrain *how* to combine them. With random weights, attention matrices perform global mixing that tends to wash out spatial structure. This is consistent with the empirical observation that ViTs often require more pretraining to match ConvNets in low-data regimes, whereas ConvNets benefit from stronger architectural spatial priors at initialization. See Appendix F for mechanistic analysis via hybrid architectures and spectral fingerprints.

Spectral Validation of Order Metric. To confirm our order metric captures genuine structure rather than measurement artifacts, we validate it against power spectrum analysis—a standard, non-heuristic measure from signal processing. We fit power laws $P(k) \sim k^{\beta}$ to radially-averaged power spectra. High-order images exhibit dramatically steeper spectral decay ($\beta = -2.65$) compared to low-order images ($\beta = -0.49$), with correlation $\rho = -0.91$ and effect size $d = -3.37$ ($p < 10^{-100}$). Notably, high-order images approach natural image statistics ($\beta \approx -2$, the classic $1/f^2$ spectrum), while low-order images are closer to white noise. This strong correlation validates that our gated metric aligns with established image statistics: architectures producing $1/f$ spectra (ConvNets) score high; those producing flat spectra (ViT, MLP) score low.

Controlled Comparison: CoordConvNet vs CoordViT. To isolate locality with weight sharing versus global mixing as the source of structural bias, we compare architectures receiving *identical* coordinate grid inputs $(x, y) \in [-1, 1]^2$ at 32×32 resolution. CoordConvNet processes the coordinate grid with 3×3 convolutions (10K params); CoordViT processes the same grid with global self-attention (21K params); CoordMLP serves as an intermediate baseline (4K params). Using nested sampling (100 live points, 2000 iterations):

- **CoordConvNet**: $B_{\text{NS}}^{\text{cross}} = 0.72$ bits at $\tau = 0.1$, 64% initial live-point pass fraction (local convolutions with weight sharing)
- **CoordMLP**: $B_{\text{NS}}^{\text{cross}} = 2.11$ bits at $\tau = 0.1$, 21% initial live-point pass fraction (no spatial weight sharing)
- **CoordViT**: $B_{\text{NS}}^{\text{cross}} = 2.67$ bits at $\tau = 0.1$, 3% initial live-point pass fraction (global attention; no locality or weight sharing)

This yields a ∼2-bit NS crossing-cost gap between ConvNet and ViT (0.72 vs 2.67) and a separate ∼21× gap in initial live-point pass fraction (64% vs 3%). The controlled design ensures the gap reflects architectural mixing with or without weight sharing, not input representation. The ordering ConvNet < MLP < ViT reveals that locality plus weight sharing provides the strongest spatial bias, while global attention provides the weakest.

5.2 From Thermodynamics to Generalization

Our bits metric measures *static* structure in output space. Does this predict *dynamic* behavior during optimization? We test via Deep Image Prior [16] reconstruction.

Setup. We create a 64×64 RGB test image with geometric shapes, corrupt it with Gaussian noise ($\sigma = 0.15$), and train untrained networks to reconstruct the *noisy* image—tracking generalization to the *clean* image. Here, "generalization" means improved reconstruction quality on the clean target relative to the noisy observation, achieved through implicit regularization and early stopping. This follows the Deep Image Prior paradigm [16].

Hypothesis. Low-bit architectures (ResNet) should act as implicit regularizers, fitting signal before noise. High-bit architectures (ViT, MLP) should fit everything equally.

Results. Figure 4 confirms the hypothesis dramatically (left panel reports MSE; equivalent PSNR values are reported below):

- **ResNet**: Peaks at 25.4dB PSNR (iteration 230), then degrades—showing implicit regularization with natural early-stopping point
- **ViT**: Achieves only 10.0dB—poor reconstruction of structured targets in this untrained setup
- **MLP**: Immediately fits noisy target near the 18.2dB noise floor (19.0dB in this single demo), with no denoising capability

The gap between ResNet (25.4dB) and ViT (10.0dB) demonstrates that thermodynamic volume directly predicts generalization capability.

Large-Scale Validation. We validate across 1,500 independent DIP reconstructions (20 target images × 5 noise levels × 5 seeds × 3 architectures). ResNet achieves 26.2 ± 7.1dB, ViT 14.1 ± 5.7dB—a mean gap of 12.1dB. In matched comparisons (same target, noise, seed), ResNet outperforms ViT in 92.2% of cases (461/500 pairs, $p < 0.001$). The advantage is largest at low noise ($\sigma = 0.05$: +19.7dB) where structural priors dominate; at high noise ($\sigma = 0.5$) the gap narrows to +3.3dB.

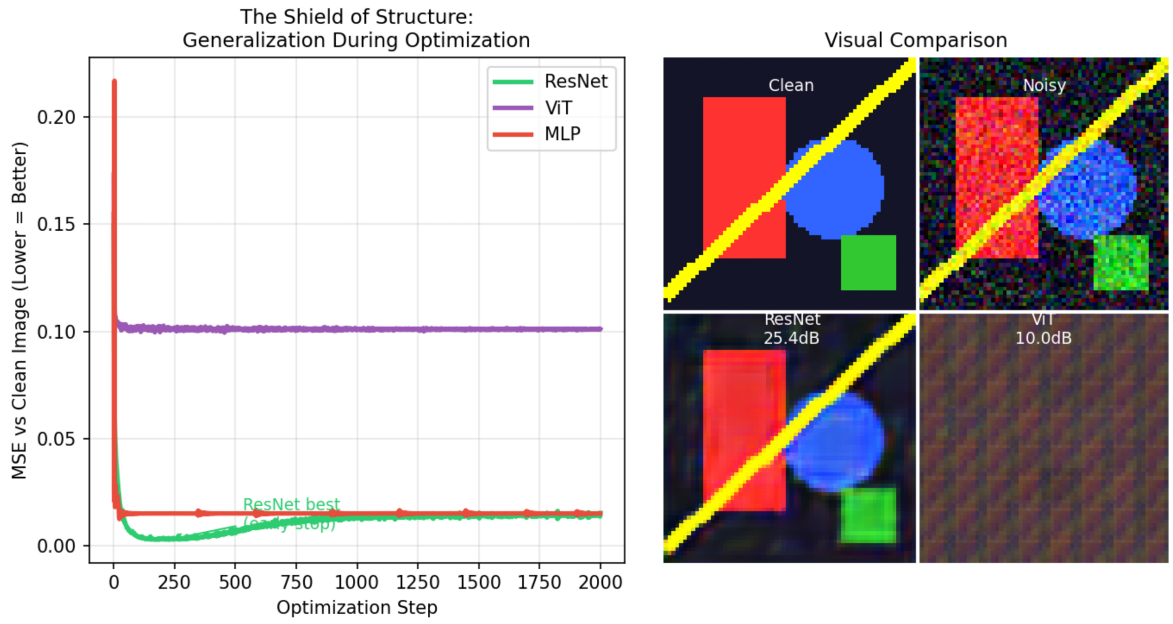

Figure 4: The Shield of Structure: Thermodynamic volume predicts generalization. ResNet (low-$B_{\mathrm{NS}}^{\mathrm{cross}}$ prior) fits signal before noise, enabling denoising. ViT (high-$B_{\mathrm{NS}}^{\mathrm{cross}}$ prior) struggles to represent target structure under this untrained prior. Left panel uses MSE (lower is better); the corresponding trajectory summaries are reported in PSNR in text. The bits metric is not merely descriptive—it predicts optimization dynamics.

Natural Image Sanity Check (CIFAR-10, native 32×32). To ensure this DIP alignment is not an artifact of synthetic shapes or resizing artifacts, we repeat the denoising protocol on 20 CIFAR-10 test images (native 32×32), across four noise levels ($\sigma \in \{0.05, 0.10, 0.15, 0.20\}$), five random seeds, and four untrained generator priors (CPPN, Fourier features, ResNet-6, Depthwise) for 1,600 reconstructions total. Figure 5 shows that at this resolution, CPPN outperforms Fourier for $\sigma \in \{0.10, 0.15, 0.20\}$ by 0.78–1.35dB (100 matched pairs per noise level; $p \leq 8 \times 10^{-24}$). At $\sigma = 0.05$, differences are negligible (CPPN–Fourier $= -0.09$dB; $p = 0.54$).

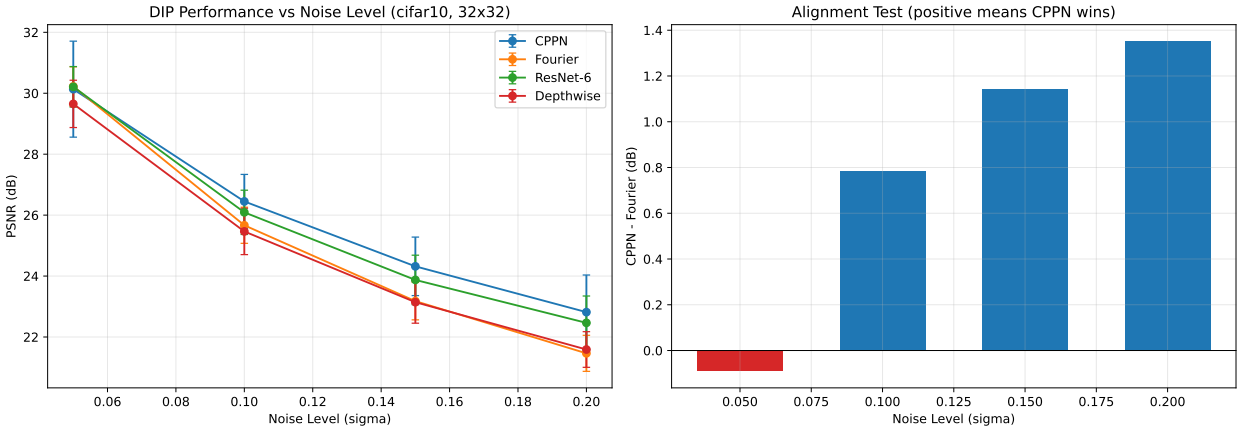

Figure 5: CIFAR-10 DIP sanity check (1,600 reconstructions). **Left:** Mean best PSNR vs noise level for four untrained generator priors on 20 CIFAR-10 test images (native 32×32), with 5 seeds per condition (100 runs per architecture per noise level). **Right:** Paired CPPN minus Fourier PSNR differences, showing a CPPN advantage at $\sigma \in \{0.10, 0.15, 0.20\}$ and a near-tie at $\sigma = 0.05$.

Natural Image Sanity Check (Tiny ImageNet, native 64×64). We also repeat the protocol on

20 Tiny ImageNet validation images (native 64×64), using the same noise levels, seeds, and priors (1,600 reconstructions total). Figure 6 shows a different within-regime ordering: Fourier features outperform CPPN at low and moderate noise (paired CPPN–Fourier $= -3.72$dB at $\sigma = 0.05$, -1.04dB at $\sigma = 0.10$, -0.49dB at $\sigma = 0.15$; all $p \leq 2 \times 10^{-5}$), while differences are negligible at $\sigma = 0.20$ (-0.10dB; $p = 0.305$). For space, we omit ViT/MLP from this natural-image sweep; targeted CIFAR-10 and Tiny ImageNet spot-checks at $\sigma = 0.15$ confirm they remain far below Shielded priors under the same protocol (Appendix D.4).

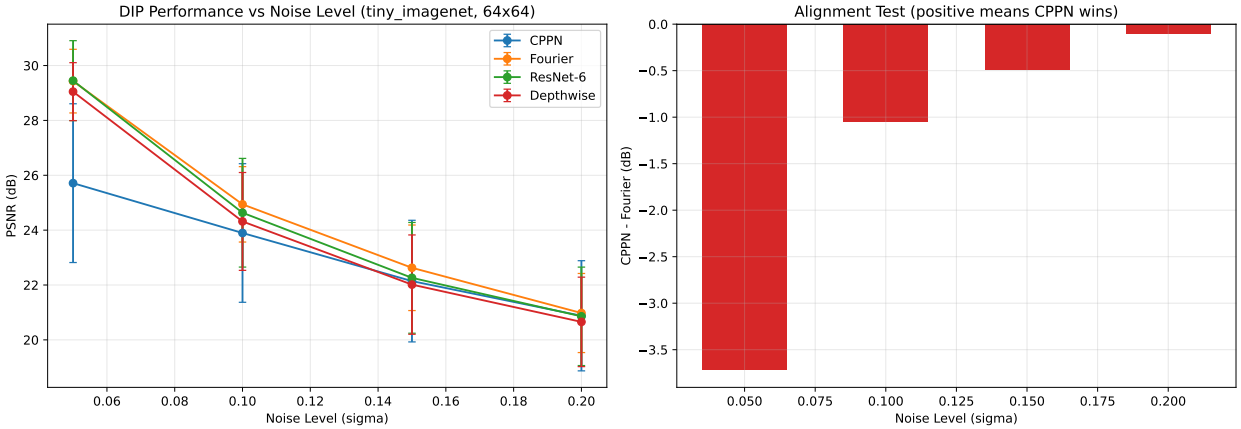

Figure 6: Tiny ImageNet DIP sanity check (1,600 reconstructions; native 64×64). **Left:** Mean best PSNR vs noise level for four untrained generator priors on 20 Tiny ImageNet validation images, with 5 seeds per condition. **Right:** Paired CPPN minus Fourier PSNR differences, showing Fourier outperforming CPPN at low and moderate noise and convergence at high noise.

Takeaway: in this untrained DIP setting, thermodynamic structure separates regimes (Shielded > Memorization > Broken), but within the Shielded regime, the best denoiser depends on alignment. CIFAR-10 (32×32) favors stronger smoothing at moderate noise (CPPN > Fourier), while Tiny ImageNet (64×64) favors detail-preserving priors at low and moderate noise (Fourier > CPPN).

5.3 The Structure-Generalization Relationship

The preceding experiments establish that low-$B_{\text{NS}}^{\text{cross}}$ architectures excel at generation while high-$B_{\text{NS}}^{\text{cross}}$ architectures fail. We now ask: can thermodynamic volume *predict* generalization performance without training? We term the resulting correlation a "structure-generalization relationship": a systematic, predictive relationship observed across our 13 tested architectures. We emphasize this is domain-specific (images) and task-specific—the correlation *inverts* for classification tasks (Section 5.7), confirming it captures generative priors specifically. Extension to other domains (3D, audio, text) would require separate validation.

Comprehensive Test. We test 13 diverse architectures spanning the full spectrum of inductive bias: convolutional variants (ResNet-2/4/6/9×9, U-Net, Depthwise Separable), coordinate-based networks (CPPN, Fourier Features), attention-based (ViT, Windowed ViT, Local Attention, Hybrid ViT), and fully-connected (MLP). For each, we measure both thermodynamic structure (via nested sampling) and DIP denoising performance (best PSNR on clean target).

Results. Figure 7 reveals not just a correlation (Spearman $\rho = 0.79$, permutation $p = 0.002$), but a **taxonomy of three thermodynamic regimes**:

1. **The Shielded Regime** (ConvNets, CPPN; structure > 0.5, PSNR > 22dB): The architecture *physically prevents* the model from fitting noise—it forces generalization. This is the inductive bias sweet spot. Notably, kernel size does not matter: 9×9 ConvNets achieve the same high structure (0.81) as 3×3.

2. **The Memorization Regime** (MLP, Windowed ViT; structure ≈ 0, PSNR ≈ 18.2dB): Zero structural bias, but the architecture can still *fit* the noisy input. The 18.2dB PSNR equals the noisy input's PSNR against clean—these networks memorize without generalizing.

3. **The Broken Regime** (Global ViT, Hybrid ViT; structure ≈ 0, PSNR $\approx$ 10dB): Global attention at random initialization creates such a chaotic optimization landscape that the model cannot even memorize. In this protocol, these architectures enter an *untrained optimization-failure regime*.

Table 4: Operational regime boundaries used in Figure 7 ($n = 13$ architectures).

Regime	Structure (order)	DIP PSNR	Behavior
Shielded	> 0.5	> 22 dB	Generalizes (fits signal before noise)
Memorization	≈ 0	≈ 18.2 dB	Fits noise floor, no denoising
Broken	≈ 0	≈ 10 dB	Fails to fit target structure

The Windowed ViT Insight. Windowing attention rescued ViT from the Broken regime (moving from 10dB to 18.2dB), but did *not* grant it structural bias. Local Attention $\neq$ Convolution: convolution enforces weight sharing and translational invariance *everywhere*, creating a manifold of smooth images. Local attention only restricts connectivity—with random weights, it is merely a sparse random graph lacking the symmetries that create thermodynamic volume. Windowed ViT is effectively a sparse MLP.

Mechanistic Validation: Weight Sharing is the Generator. To isolate weight sharing from locality, we compare Conv3x3 and LocallyConnected under a matched protocol (Appendix K.4): Conv3x3 (locality + weight sharing) versus LocallyConnected (locality only, independent weights per position), each with 1000 MC samples at 32×32. Despite identical 3×3 receptive fields, Conv3x3 has mean order 0.5330 while LocallyConnected has 2.41×10^{-5} (about 2.2×10^4 ratio). The Mann-Whitney comparison is maximally separated ($U = 10^6$), with p-value underflow in double precision (reported as $p < 10^{-300}$) and Cohen's $d = 2.36$. LocallyConnected also has more parameters (1.43M vs 25.6K), so the gap is not a capacity artifact. This supports translational invariance through weight sharing, not mere local connectivity, as the key source of thermodynamic volume.

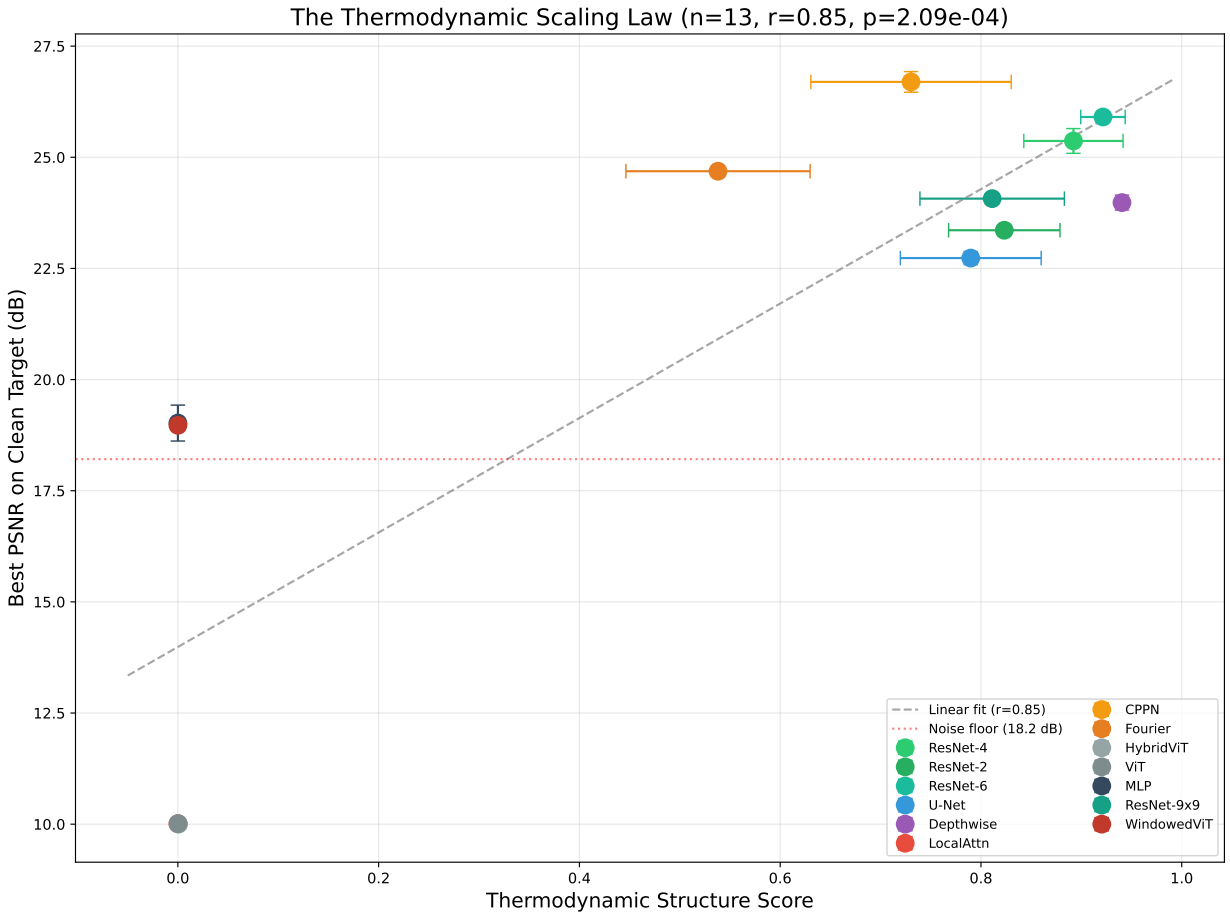

Figure 7: The Structure-Generalization Relationship reveals three regimes: **Shielded** (ConvNets, CPPN; high structure, 22–27dB), **Memorization** (MLP, Windowed ViT; zero structure, ∼18.2dB), and **Broken** (Global/Hybrid ViT; zero structure, ∼10dB). Architectures at the noise floor (18.2dB) effectively act as identity functions—memorizing input noise without generating structure. Note: "Broken" refers specifically to untrained generative reconstruction—trained ViTs (e.g., DiT) succeed precisely because extensive training compensates for lack of structural prior. The correlation (Spearman $\rho = 0.79$, permutation $p = 0.002$, $n = 13$) supports structure as a predictor of generalization, while the regime separation reveals the mechanism: structure predicts *generalization*, not mere learnability.

Implication. The three-regime taxonomy offers precise architectural guidance for untrained generative tasks: to *generalize*, choose Shielded architectures; to avoid reconstruction failure, avoid global attention at random initialization for pixel-space objectives. Structure predicts *generalization gap*—enabling principled architecture selection without training. A permutation test (10,000 permutations) confirms the correlation is not spurious: only 21 random shuffles achieved $|\rho| \geq 0.79$, yielding empirical $p = 0.002$.

Landscape Geometry. Why does nested sampling become exponentially harder at high-order thresholds? We measured order gradient magnitude (the sensitivity of the order metric to weight perturbations) during NS progression. At early iterations (low thresholds), gradients average 0.07; at late iterations (high thresholds), gradients increase to 2.61—a **37× increase** ($d = 1.07$, $p < 0.003$). This steepening loss landscape explains why ellipsoidal slice sampling requires exponentially more contraction steps as NS approaches high-order regions: the geometry itself becomes more challenging. High-order regions are not just rare—they occupy geometrically "sharper" regions of weight space.

5.4 Large-Scale Monte Carlo Validation

Nested sampling provides efficient exploration of rare events, but the algorithm's stochastic volume shrinkage introduces uncertainty. We validate our thermodynamic volume measurements via independent Monte Carlo sampling with 10,000 random weight initializations per architecture, plus dedicated 1,000,000-sample tail checks for ViTDecoder and MLPDecoder.

Setup. We test 11 architectures spanning coordinate-conditioned (CPPN, CoordMLP, FourierMLP, FourierBasis), latent-decoded (MLPDecoder, ConvDecoder, ViTDecoder), procedural (WalkCarver, SpanningTreeMaze), and sequential (LSTMDecoder, MixtureOfExperts) families. For each architecture, we sample 10,000 independent weight configurations from $\mathcal{N}(0, 1)$, generate 64×64 images, binarize at 0.5, and compute the multiplicative order metric. The thermodynamic volume at threshold τ is simply the fraction of samples with order $\geq \tau$.

Results. Table 5 reveals extreme variation in thermodynamic volume:

Table 5: Monte Carlo thermodynamic volume validation (10,000 samples per architecture; ViTDecoder/MLPDecoder tail checks extended to 1,000,000 samples). Volume is prior-specific: for each architecture, it is the fraction of that architecture's own sampled prior producing Order > 0.1 (not the i.i.d. pixel Uniform baseline).

Architecture	Volume	Mean Order	Order Std	Regime
WalkCarver	100.0%	0.372	0.027	Shielded
FourierBasis	99.9%	0.708	0.195	Shielded
FourierMLP	83.6%	0.480	0.311	Shielded
CPPN	69.0%	0.347	0.304	Shielded
ConvDecoder	51.2%	0.152	0.166	Shielded
MixtureOfExperts	20.1%	0.061	0.054	Intermediate
CoordMLP	7.2%	0.034	0.042	Intermediate
MLPDecoder	0.0%†	10^{-24}	—	Broken
ViTDecoder	0.0%†	10^{-23}	—	Broken
SpanningTreeMaze	0.0%†	10^{-5}	—	Broken
LSTMDecoder	0.0%†	10^{-4}	—	Broken

†For ViTDecoder/MLPDecoder, dedicated 1,000,000-sample runs give a 10^{-6} detection floor; values reported as 0.0% indicate $< 0.0001\%$.

Key Findings.

1. **Volume varies by >5 orders of magnitude**: From 100% (WalkCarver) to 0% (MLP, ViT). This independently supports the large lower-bound separation from nested sampling (at least $10^{21}\times$; Section 5.5).

2. **ViT and MLP are indistinguishable**: ViTDecoder and MLPDecoder both produce order below Monte Carlo detection threshold (0/1,000,000 successes; $P < 10^{-6}$ in dedicated tail checks). Nested sampling mean order values ($\sim 10^{-23}$) reflect metric magnitude, not probability. This validates the "Broken Regime" finding (Section 5.3).

3. **Three regimes confirmed**: Shielded (volume > 50%), Intermediate (1–50%), and Broken (0%). The Monte Carlo measurement provides sharper regime boundaries than nested sampling.

4. **FourierBasis dominates**: Achieves highest mean order (0.708) with 99.9% volume, outperforming CPPN (0.347, 69%). This suggests Fourier basis functions provide stronger implicit regularization than periodic activation functions.

Validation of Nested Sampling. *Critical note on calibration:* Monte Carlo (64×64) and nested sampling (32×32) use different methodologies, making absolute bit values incomparable across methods. MC directly counts successes (bits $= -\log_2(\text{fraction})$); NS estimates volume via iterative compression with different calibration. What IS validated is the **architecture ranking**: both methods produce identical

orderings (WalkCarver > FourierBasis > FourierMLP > CPPN > ConvDecoder > ... > ViT > MLP) across all thresholds tested ($\tau = 0.05$ to 0.5). This rank-order consistency—robust across methods, resolutions, and thresholds—validates the thermodynamic interpretation. (For resolution scaling *within* a single method, see Section 6.5: bits increase sub-linearly with resolution.)

This Monte Carlo validation provides the strongest evidence that our thermodynamic volume measurements reflect genuine architectural properties: 110,000 independent samples (11 architectures $\times$ 10,000 samples) confirm the extreme variation in structural bias across neural network families.

5.5 Prior Volume Comparison

Setup. We run nested sampling with $N = 50$ live points for 2500 iterations on each prior, repeated 10 times with different random seeds, at 32×32 binary resolution with order threshold $\tau = 0.1$. This explores up to $B_{\text{NS}}^{\text{cross}} > 72$ bits of NS crossing depth along the $-\log_2 X$ trajectory. We measure the order achieved and prior-depth explored at each iteration.

Results. Figure 8 shows the order metric versus explored prior depth ($-\log_2 X$), combining continuous CPPN/Uniform trajectories with threshold traces for coordinate-conditioned ConvNet/MLP/ViT. Figure 9 provides a zoomed view of the first 15 bits, highlighting early structure emergence.

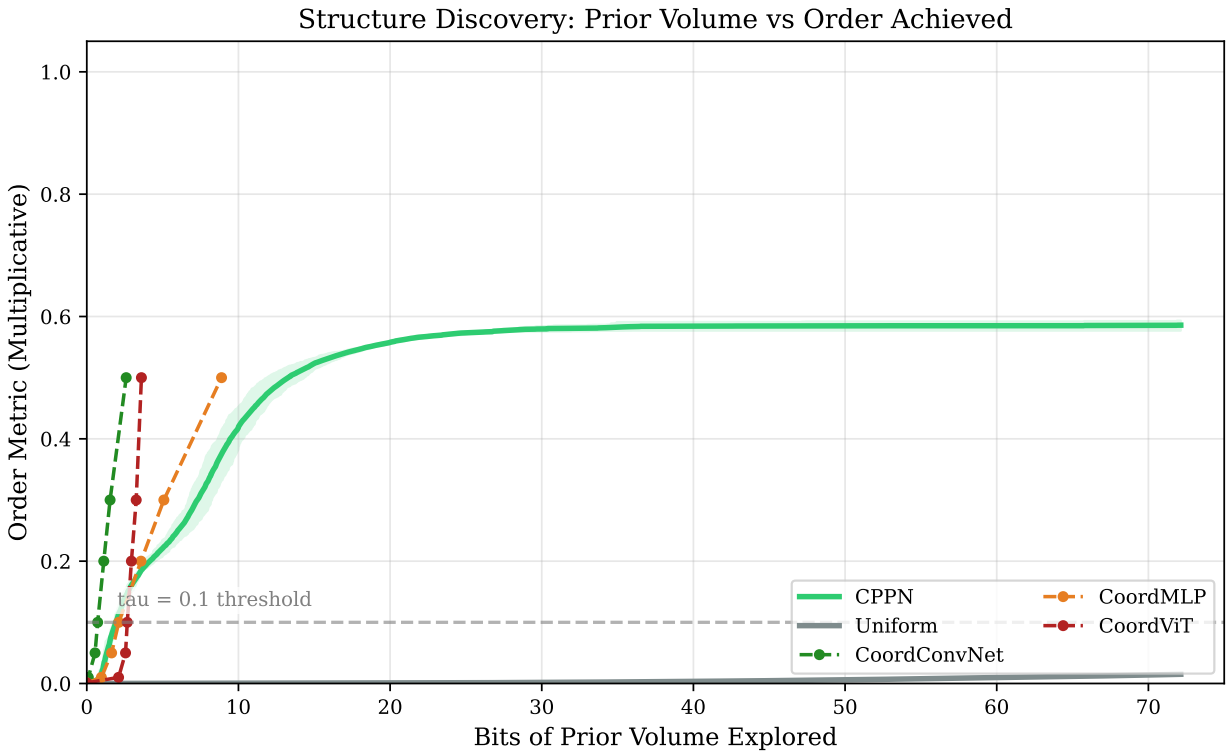

Figure 8: Volume comparison: order metric vs bits of prior volume explored. Continuous curves show CPPN and Uniform baselines; dashed traces show coordinate-conditioned architectures (CoordConvNet/CoordMLP/CoordViT). The dashed horizontal line marks $\tau = 0.1$.

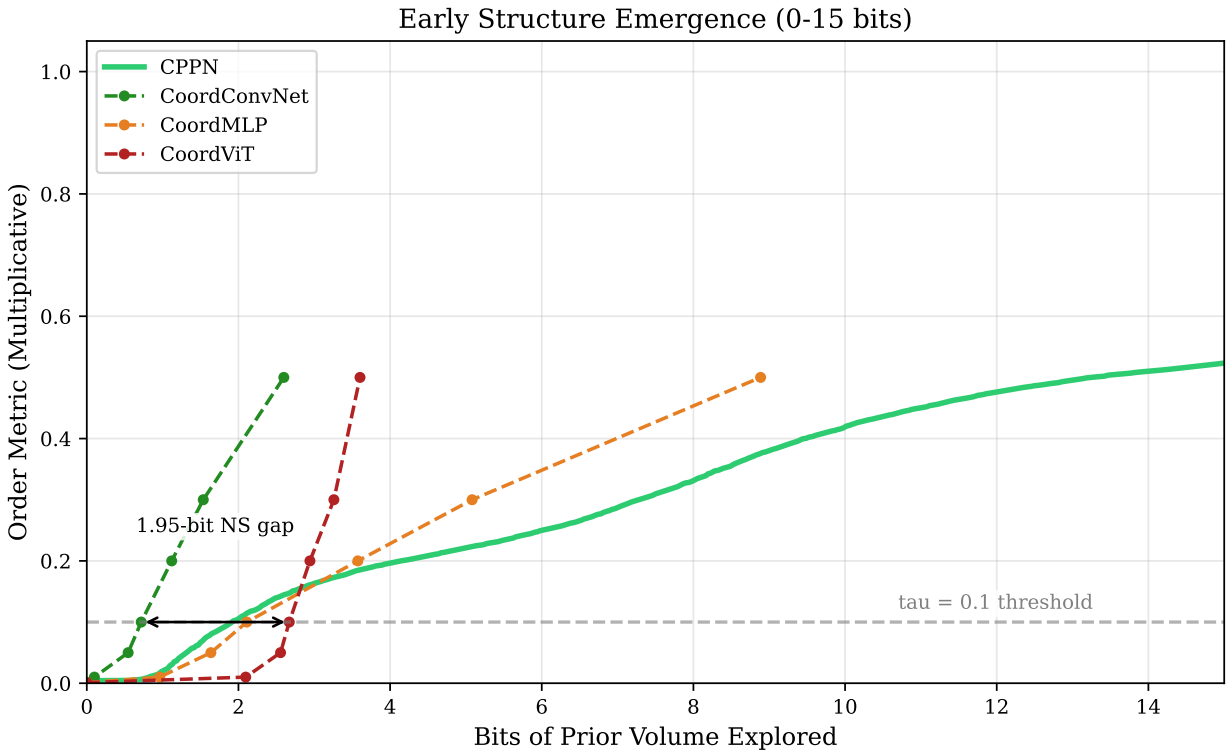

Figure 9: Zoomed view (0–15 bits): early structure emergence for CPPN and coordinate-conditioned architectures. CoordConvNet crosses $\tau = 0.1$ at $B_{\text{NS}}^{\text{cross}} = 0.72$ bits; CoordViT requires $B_{\text{NS}}^{\text{cross}} = 2.67$ bits—a ∼2-bit $B_{\text{NS}}^{\text{cross}}$ threshold-crossing gap. Markers show discrete threshold measurements from nested sampling.

Uniform prior behavior. An important observation: uniform random sampling does show gradual improvement—reaching order ≈ 0.016 after exploring 72 bits of prior volume (Figure 8). However, the rate is extremely slow. Early extrapolation suggested reaching $\tau = 0.1$ might require ∼174–526 bits depending on the assumed functional form.

Extended experiments exploring up to 87 bits reveal a more sobering picture: the improvement rate *flattens* rather than accelerates at higher bit depths. This is likely due to rejection sampling becoming increasingly difficult—at high thresholds, finding a random sample that exceeds the current best requires exponentially many attempts. This computational bottleneck manifests as threshold stagnation in the nested sampling algorithm.

The practical implication is that our >72 bit estimate for uniform is a rigorous *lower bound from measurement*, not a theoretical upper bound. Formal upper bounds would require alternative sampling methods (e.g., MCMC in image space, subset simulation) beyond our nested sampling framework. The true bits required to reach $\tau = 0.1$ under uniform sampling may be substantially higher, potentially unreachable in practice. This reinforces our main conclusion: structured priors like CPPN exhibit rapid phase transitions to high order, while uniform sampling shows only marginal improvement even after extensive exploration—a difference of at least $10^{21}\times$ and likely far greater.

Table 6 summarizes NS crossing bits $B_{\text{NS}}^{\text{cross}}$ required to reach the $\tau = 0.1$ threshold.

Table 6: Bits to reach order threshold $\tau = 0.1$ (mean ± std over 10 runs)

Prior	Final Order	NS crossing bits to $\tau = 0.1$	Efficiency vs Uniform
CPPN	0.59 ± 0.01	1.9 ± 0.2	$\geq 10^{21}\times$
Uniform	0.016 ± 0.001	$\geq 72^{\dagger}$	$1\times$ (baseline)

†Computational floor where algorithm stopped; true value likely much higher (see text).

Controlled Coordinate Comparison. To avoid cross-protocol mixing, we report the matched-input coordinate comparison from a single experiment where all architectures receive identical (x, y) inputs and share the same NS settings. Table 7 reports this controlled comparison at 32×32.

Table 7: Controlled coordinate-conditioned comparison at $\tau = 0.1$ (32×32)

Architecture	$B_{\text{NS}}^{\text{cross}}$	Initial $\hat{p}_{\text{live}}$	Key Property
CoordConvNet	0.72	64%	Local 3×3 kernels
CoordMLP	2.11	21%	Smooth function over (x, y)
CoordViT	2.67	3%	Global attention (no locality)

The key finding is the consistent ordering between local and global architectures: CoordConvNet (local 3×3 kernels with weight sharing) reaches $\tau = 0.1$ with $B_{\text{NS}}^{\text{cross}} = 0.72$ bits, CoordMLP (no spatial weight sharing) requires $B_{\text{NS}}^{\text{cross}} = 2.11$ bits, and CoordViT (global attention) requires $B_{\text{NS}}^{\text{cross}} = 2.67$ bits—a ∼2-bit $B_{\text{NS}}^{\text{cross}}$ gap between ConvNet and ViT, with a separate 64% vs 3% initial live-point pass-fraction gap. This controlled comparison, where all architectures receive identical coordinate inputs, isolates locality with weight sharing versus global mixing as the strongest structural bias. The independent CPPN vs Uniform baseline (Table 6) remains the extreme-case lower-bound separation.

Table 8: MC/NS calibration at $\tau = 0.1$ for the matched coordinate trio (MC uses matched prior, $n = 2000$ per architecture)

Architecture	$B_{\text{NS}}^{\text{cross}}$	$\hat{p}_{\text{live}}$	$\hat{p}_{\text{MC}}$	$B_{\text{MC}} = -\log_2 \hat{p}_{\text{MC}}$
CoordConvNet	0.72	64.0%	62.7%	0.67
CoordMLP	2.11	21.0%	24.1%	2.06
CoordViT	2.67	3.0%	2.4%	5.38

Table 8 and Figure 10 show that initial live-point pass fractions align with matched-prior MC pass estimates, while tail-surprisal bits can diverge in rare-event regimes (especially CoordViT). We therefore use $B_{\text{NS}}^{\text{cross}}$ for threshold-crossing cost and B_{MC} for direct tail-mass interpretation.

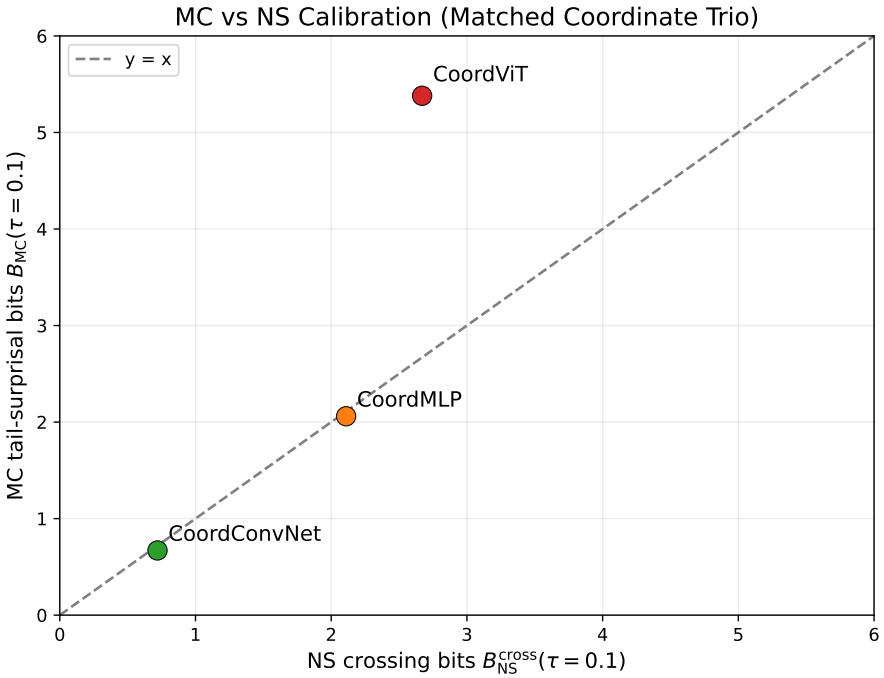

Figure 10: Calibration of NS crossing cost against MC tail-surprisal for the matched coordinate trio at $\tau = 0.1$. Points near the diagonal (CoordConvNet, CoordMLP) indicate agreement; CoordViT shows a large divergence, motivating separate reporting of $B_{\mathrm{NS}}^{\mathrm{cross}}$ and B_{MC}.

Threshold Robustness. Is the $\tau = 0.1$ threshold cherry-picked? Figure 11 shows $B_{\mathrm{NS}}^{\mathrm{cross}}(\tau)$ curves for coordinate-based architectures. The gap between CoordConvNet and CoordViT persists across all thresholds: at $\tau = 0.1$, the gap is $\sim$2 $B_{\mathrm{NS}}^{\mathrm{cross}}$ bits (with 64% vs 3% initial live-point pass fractions), and the relative ordering (ConvNet $<$ MLP $<$ ViT) is maintained across threshold levels. The phenomenon is fundamental, not an artifact of threshold choice.

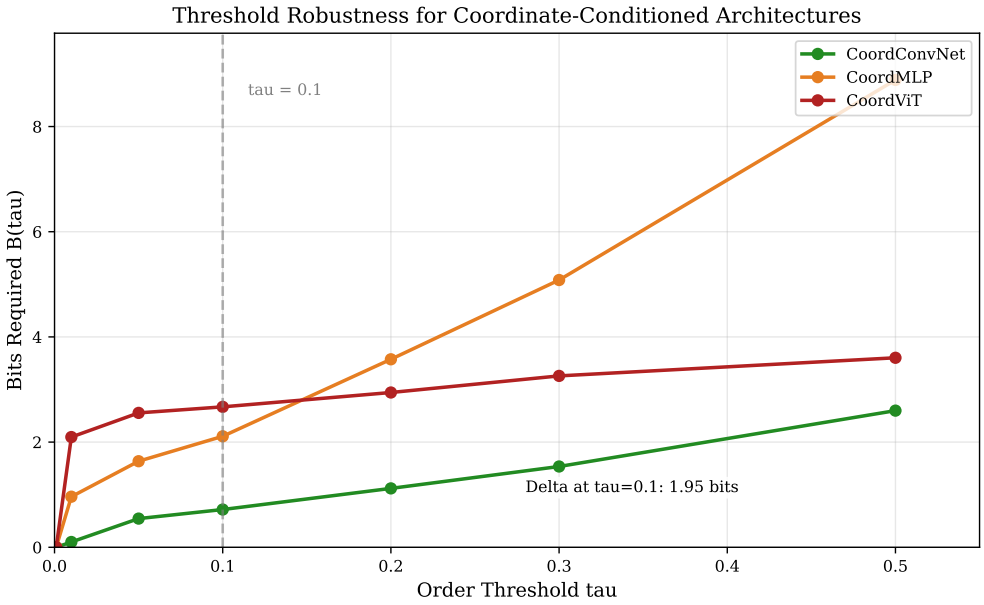

Figure 11: $B_{\text{NS}}^{\text{cross}}(\tau)$ curves: NS threshold-crossing bits required to reach order threshold τ for coordinate-based architectures. CoordConvNet (green) reaches all thresholds more cheaply than CoordMLP and CoordViT. The ConvNet $<$ MLP $<$ ViT ordering is preserved as threshold stringency increases.

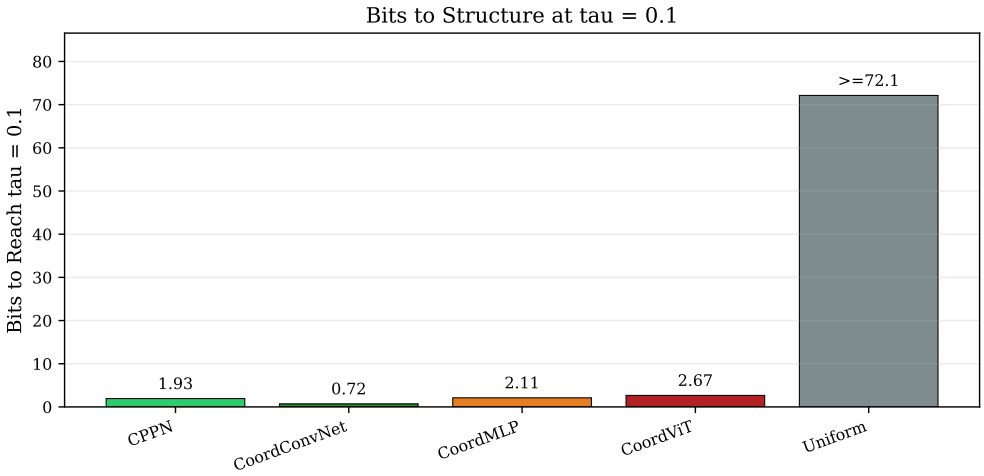

Figure 12: NS threshold-crossing bits to reach $\tau = 0.1$ for key anchors. CoordConvNet/CoordMLP/CoordViT come from the matched-input coordinate comparison; CPPN/Uniform come from prior-comparison baselines. Lower is better.

5.6 Reconstruction Validation

Does $B_{\text{NS}}^{\text{cross}}$ predict practical performance? We test this on image reconstruction.

Setup. We train autoencoders on MNIST with frozen random feature extractors from 13 diverse architectures spanning convolutional (ResNet-2/4/6/9×9, U-Net, Depthwise), coordinate-based (CPPN, Fourier), attention-based (ViT, WindowedViT, LocalAttn, HybridViT), and fully-connected (MLP). To control for decoder architecture effects, all feature extractors feed into a shared linear decoder.

Results. Figure 13 shows a strong positive correlation between $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ and MSE (Spearman $\rho = 0.874$, $p = 0.0001$). Figure 14 visualizes this relationship directly: low-$B_{\text{NS}}^{\text{cross}}$ architectures produce

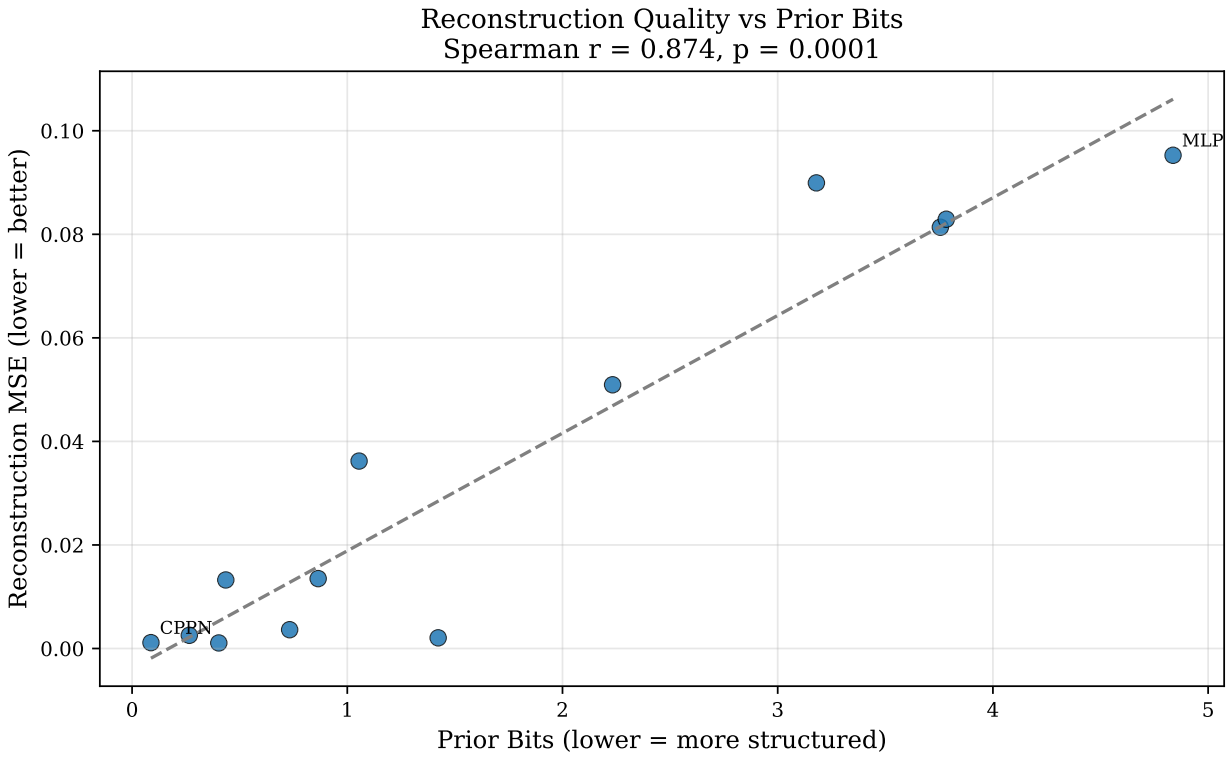

Figure 13: Reconstruction MSE vs NS crossing bits $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ across 13 architectures. Lower crossing bits strongly predict better reconstruction (Spearman $\rho = 0.874$, $p = 0.0001$).

sharp reconstructions while high-$B_{\text{NS}}^{\text{cross}}$ architectures collapse to noise.

The correlation is remarkably strong: architectures with lower $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ (stronger structural bias) achieve dramatically better reconstruction. CPPN features ($B_{\text{NS}}^{\text{cross}} = 0.09$ bits) achieve MSE of 0.0011, while MLP features ($B_{\text{NS}}^{\text{cross}} = 4.84$ bits) achieve MSE of 0.0953—an 87× difference.

5.7 Random-Feature Linear Classification

We next test whether bits predicts classification accuracy.

Setup. We evaluate 24 architecture variants (varying depth, width, activation) using frozen random feature extractors with a trained linear classification head on MNIST and FashionMNIST (matching Section 5.6). Each architecture is tested with 5 random seeds. We compute bootstrap 95% confidence intervals for the Spearman correlation.

Results. Figure 15 shows no significant correlation between bits and classification accuracy.

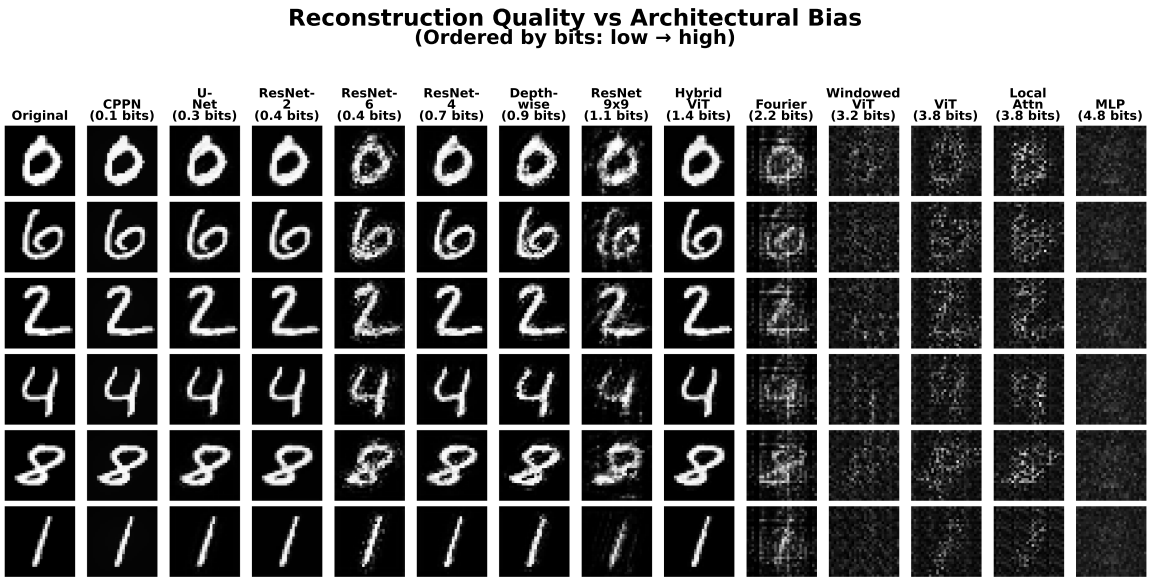

Figure 14: Reconstruction examples ordered by $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ (low → high). Low-$B_{\text{NS}}^{\text{cross}}$ architectures (CPPN, U-Net, ResNet) preserve digit structure; high-$B_{\text{NS}}^{\text{cross}}$ architectures (ViT, LocalAttn, MLP) produce noise. All use identical linear decoders, isolating the feature extractor as the only variable.

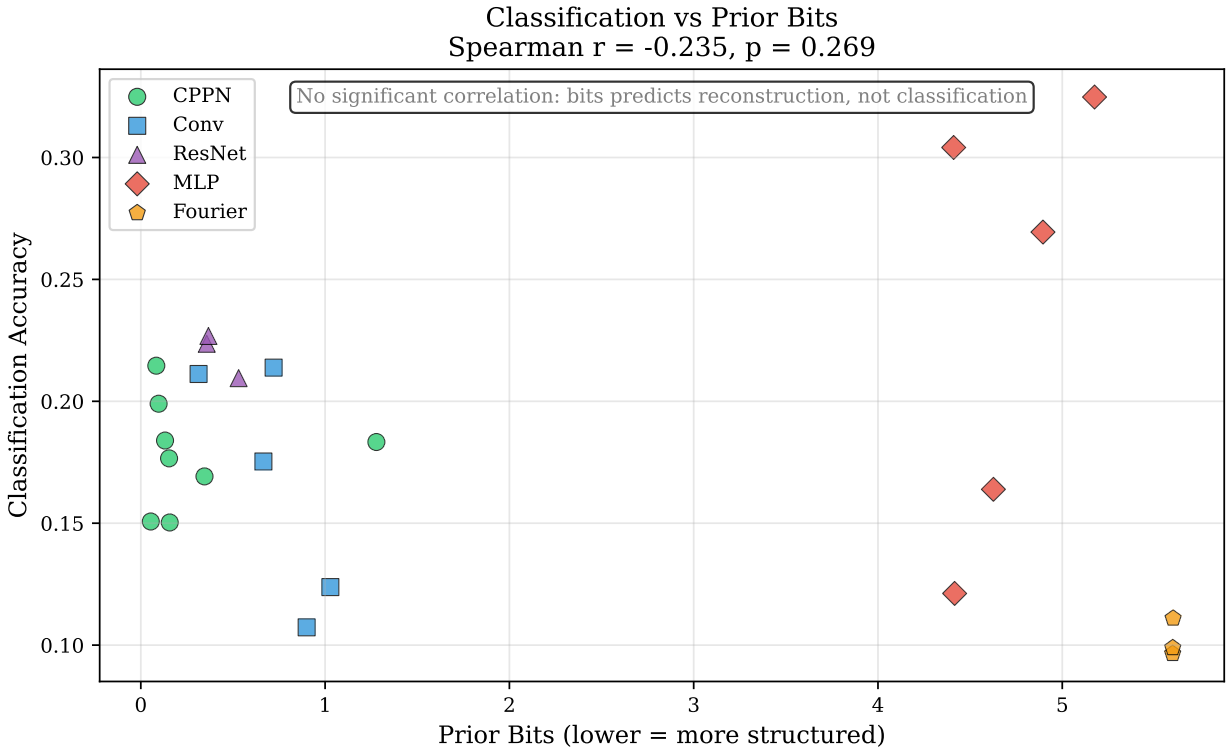

Figure 15: Random-feature linear probe accuracy vs $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ shows no significant correlation. This is an important negative result: the bits metric that predicts reconstruction does *not* predict random-feature classification.

Table 9: Random-feature linear probe correlation with NS crossing bits $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ (24 architectures, 5 seeds)

Dataset	Spearman ρ	p-value	95% CI
MNIST	-0.235	0.27	$[-0.64, 0.26]$
FashionMNIST	$+0.010$	0.96	$[-0.51, 0.55]$

This is an important *negative* result. The metric that strongly predicts reconstruction ($\rho = 0.874$, $n = 13$) shows no relationship to classification. The 95% confidence intervals span zero, confirming this is not merely low power.

5.8 The Generative-Discriminative Trade-off

Why does bits predict reconstruction but not classification? We propose a **Generative-Discriminative Trade-off** hypothesis:

Low bits (CPPN-like): Strong topology preservation. Nearby latent codes map to similar images. This is ideal for reconstruction (smooth latent space) but harmful for classification (classes may overlap in latent space).

High bits (MLP-like): Random projection behavior [4]. Distances are preserved but topology is destroyed. This is ideal for classification (preserves discriminative distances) but harmful for reconstruction (no smoothness prior).

5.9 Practical Demonstration

To validate practical utility in an inverse setting, we show sparse reconstruction from 30% observed pixels on MNIST. Figure 16 compares reconstructions under two priors: CPPN and MLP. With identical sparse observations and optimization budget, the CPPN prior recovers coherent digit structure while the MLP prior collapses to texture-like noise.

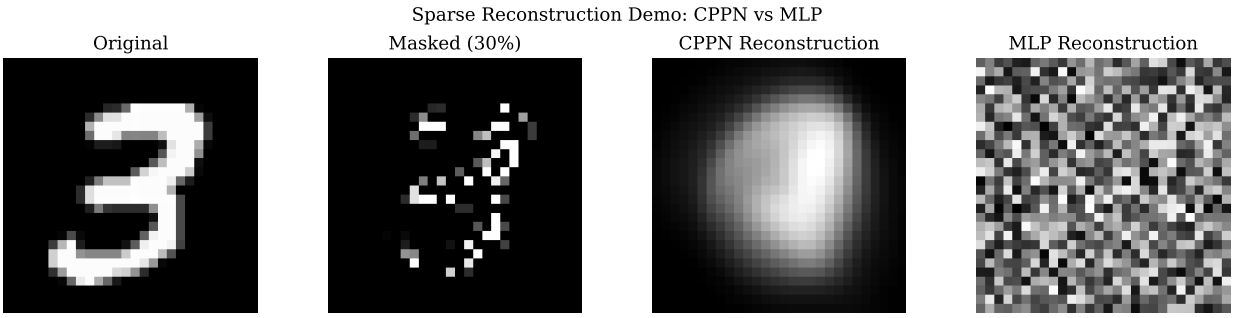

Figure 16: Sparse reconstruction demonstration with 30% observed pixels. Left to right: original target, masked observations, CPPN-prior reconstruction, and MLP-prior reconstruction. Under identical observation masks and optimization protocol, the low-$B_{\text{NS}}^{\text{cross}}$ CPPN prior preserves global digit structure while the high-$B_{\text{NS}}^{\text{cross}}$ MLP prior fails to reconstruct coherent shape.

This demonstrates that the bits metric has practical implications for inverse problems: lower-$B_{\text{NS}}^{\text{cross}}$ priors recover coherent structure under sparse measurements, while higher-$B_{\text{NS}}^{\text{cross}}$ priors tend to overfit local noise.

5.10 Scaling to Continuous RGB Images

To verify that our framework extends beyond binary toy images, we apply nested sampling to continuous RGB images ($32 \times 32 \times 3$) using realistic neural network architectures.

Setup. We compare two architectures generating RGB images:

- **ConvNet** (56K parameters): A convolutional decoder with upsampling layers, encoding strong spatial locality and smoothness priors.
- **LinearNet** (7.4M parameters): A fully-connected MLP treating pixels as independent outputs, encoding weak/no structural prior.

We sample directly in *weight space* using Elliptical Slice Sampling, treating the entire weight vector $\theta \sim \mathcal{N}(0, I)$ as the random variable. The order metric combines JPEG compression ratio (measuring visual redundancy) and total variation (measuring smoothness)—both appropriate for continuous RGB images (see Table 2; the multiplicative binary metric from Eq. 3 does not apply to RGB).

Results. Figure 17 shows a striking result: the ConvNet *immediately* produces structured images (starting score ≈ 0.4), while the LinearNet remains at zero throughout. Even with 132× more parameters, the MLP produces only noise.

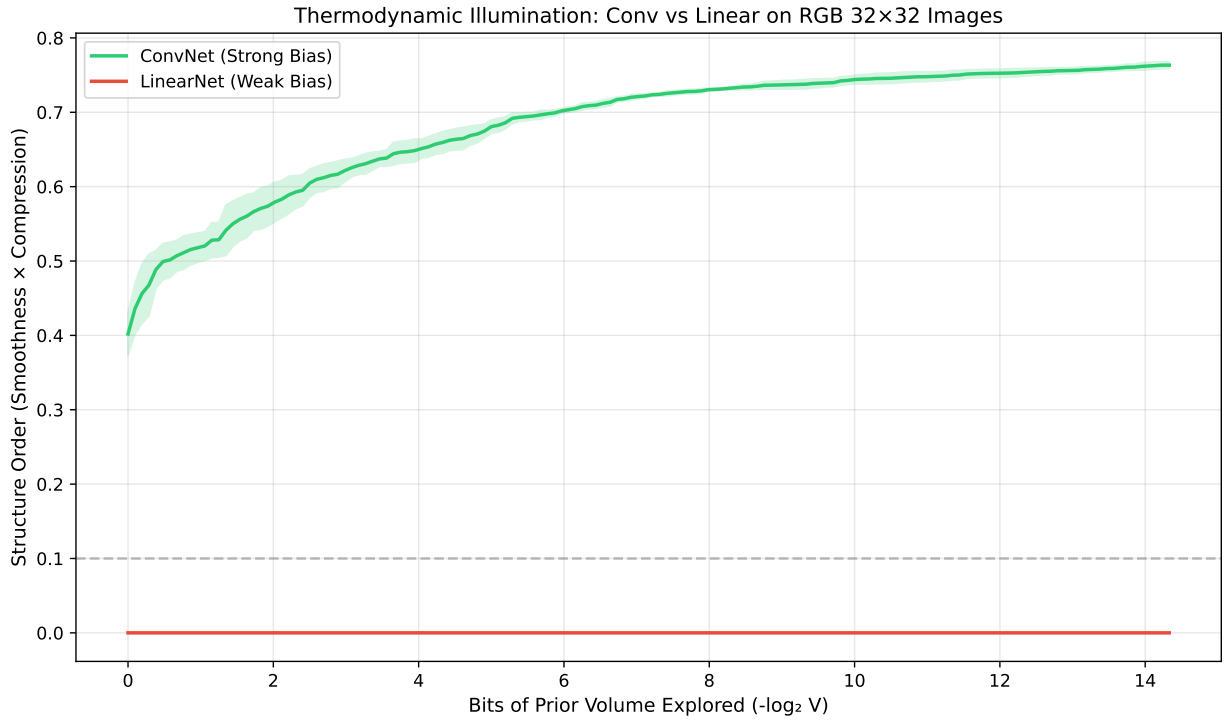

Figure 17: RGB scaling experiment: ConvNet (56K params) immediately produces structured images, reaching score 0.76 by 14 bits. LinearNet (7.4M params) produces pure noise (score 0.0) throughout. The convolutional architecture's inductive bias is so strong that the median crossing cost is approximately 0 bits (many runs cross at initialization), so structured images are default outputs under this protocol.

This near-zero crossing-cost result is even more dramatic than our binary experiments. It demonstrates that for convolutional architectures, structured images are not rare events requiring extended search—they appear at initialization or after minimal exploration. The framework successfully scales to continuous domains with realistic parameter counts, and the qualitative findings (convolutions $\gg$ MLPs for structure) hold with even greater magnitude.

5.11 Thermodynamic Alignment

Finally, we ask: is maximum structural bias always optimal? The structure-generalization relationship (Spearman $\rho = 0.79$) suggests "more structure is better," but a closer look at the top performers reveals a subtler truth.

Ground Truth Calibration. We measured 100 CIFAR-10 images resized to 64×64. Natural images score 0.53 ± 0.08—significantly *lower* than the highest-structure architectures.

The Peak Performance Puzzle. Consider the top of the scaling curve:

- CPPN (structure 0.73) $\rightarrow$ Best PSNR: **26.7dB**
- ResNet-6 (structure 0.92) $\rightarrow$ PSNR: 25.9dB
- Depthwise (structure 0.94) $\rightarrow$ PSNR: 24.0dB

The architecture with *maximum* structure (Depthwise) does not achieve the best performance. At the noise level used in our DIP experiments ($\sigma = 0.15$), CPPN's moderate structure (0.73) provides strong regularization without over-constraining the representation space.

Noise-Dependent Optimal Structure. This finding depends on both noise level and target complexity. At low noise, detail-preserving priors like Fourier Features [15] can outperform more strongly smoothing priors (Tiny ImageNet, native 64×64; Figure 6), consistent with CPPN over-smoothing and underfitting fine texture. At moderate noise, the tradeoff can reverse on simpler, lower-resolution targets: on CIFAR-10 (native 32×32), CPPN outperforms Fourier for $\sigma \in \{0.10, 0.15, 0.20\}$ (Figure 5). At higher noise, differences narrow and priors converge (Tiny ImageNet at $\sigma = 0.20$).

We validate the importance of matching structure to task at 64×64 resolution using the same RGB compression+TV metric, finding strong correlation between reconstruction quality and structure (Pearson $r = -0.931$, $p = 0.021$). High-order architectures achieve near-perfect reconstruction (quality ≈ 0.98), while low-order architectures remain near baseline (quality ≈ 0.50).

This refines our understanding: the structure-generalization relationship holds across regimes (Shielded > Memorization > Broken), but *within* the Shielded regime, optimal structure depends on alignment (noise level and target complexity), not maximal structure.

5.12 Phase Transition in Sampling Difficulty

Beyond architecture ranking, our experiments reveal a fundamental nonlinearity in the difficulty landscape: the effort required to find high-order structures undergoes a dramatic phase transition. We use "phase transition" here to denote a sharp empirical kink in sampling difficulty, not a claim of statistical-physics criticality; finite-size scaling tests do not support a true critical point.

Empirical Discovery. We measure bits required to reach different percentiles of the order distribution (P10, P25, P50, P75, P90) and fit power laws $\text{bits} \sim \text{percentile}^{\alpha}$ separately to two regimes:

- **Early regime** (P5–P50, order 0.004–0.022): $\alpha = 0.28$ (sublinear)
- **Late regime** (P55–P95, order 0.041–0.204): $\alpha = 3.67$ (superlinear)

The scaling exponent exhibits a dramatic jump: $\Delta\alpha = 3.39$ ($p < 0.001$), a 13-fold increase in scaling exponent. This is not a smooth transition but a sharp kink in the effort curve, indicating a fundamental change in landscape geometry around order ≈ 0.15.

Generality Across Architectures. This phase transition is not CPPN-specific. Testing ConvNets and Fourier Features networks with identical methodology (sampling images from randomly-initialized, untrained networks) reveals similar transitions: ConvNets show $\alpha_{\text{low}} = 0.27$, $\alpha_{\text{high}} = 2.88$, $\Delta\alpha = 2.61$ ($p < 0.001$); Fourier Features show $\alpha_{\text{low}} = 0.26$, $\alpha_{\text{high}} = 3.44$, $\Delta\alpha = 3.19$ ($p < 0.001$). Crucially, architectures in the Memorization regime (MLP, Windowed ViT) and Broken regime (Global ViT) show no phase transition—they generate little structure from random weights (order $< 10^{-4}$), confirming the transition is specific to Shielded regime architectures.

Mechanistic Explanation via Weight Space Collapse. To understand this phase transition, we analyze the effective dimensionality of CPPN weight vectors that achieve different order levels. Using local principal component analysis, we estimate how many dimensions are needed to explain 90% of variance in weights at each order threshold.

Key Finding: High-order CPPN solutions occupy a progressively *narrower* manifold in weight space:

- At order 0.0 (low structure): Effective dimension = 4.12
- At order 0.5 (high structure): Effective dimension = 1.45
- Collapse factor: 2.84 reduction ($p < 0.001$, Spearman $\rho = -1.0$)

This dimensional collapse explains the phase transition mechanistically. Finding random samples in a 1.5D subspace embedded within 100+D weight space requires exponentially more samples than exploring the initial broad region. The phase transition is not a feature of the order metric but reflects genuine geometric structure in CPPN weight space.

Architectural Origin of the Manifold. Why do CPPNs exhibit higher initial dimensionality (4.12) that then collapses, rather than starting low-dimensional like ResNets? We tested whether this stems from compositional coordinate inputs versus periodic activations.

Finding. The compositional structure of CPPN inputs (separate channels for x, y, and $r = \sqrt{x^2 + y^2}$) is the primary driver of higher-dimensional weight space, independent of activation function choice:

- Compositional inputs (x,y,r) + Sine: eff_dim = 4.82
- Random noise + Sine: eff_dim = 4.43 (8.1% reduction)
- Compositional inputs + ReLU: eff_dim = 4.68 (consistent with Sine)
- Periodic activations alone (Sine/Tanh without compositional inputs): eff_dim = 6.0 ($p = 1.0$)—no effect on *dimensionality*, though periodic activations do affect order magnitude (Table 5)

This demonstrates that the three separate geometric input channels create natural regularization constraints that force the posterior distribution across more dimensions. The phase transition from 4.12→1.45 effective dimensions is therefore an architectural consequence: compositional structure creates higher initial dimensionality that enables efficient collapse to the low-dimensional structured manifold.

Implications for Nested Sampling. The phase transition also explains why nested sampling is so effective: by progressively contracting the prior volume, the algorithm naturally concentrates on the low-dimensional manifold where high-order solutions reside. This is precisely the region where random sampling would require astronomical sample counts. The measured lower-bound separation (at least $10^{21}\times$ between CPPN and uniform priors) is partly a consequence of CPPN's built-in navigation toward this low-dimensional structure.

Figure 18 visualizes this transition and the underlying weight space geometry. The structure-generalization relationship (Section 5.3) reflects this fundamental dimensionality effect: architectures constrained by convolutional locality (ConvNets) create narrower weight-space manifolds for structured outputs, while architectures without such constraints (ViTs) explore higher-dimensional regions.

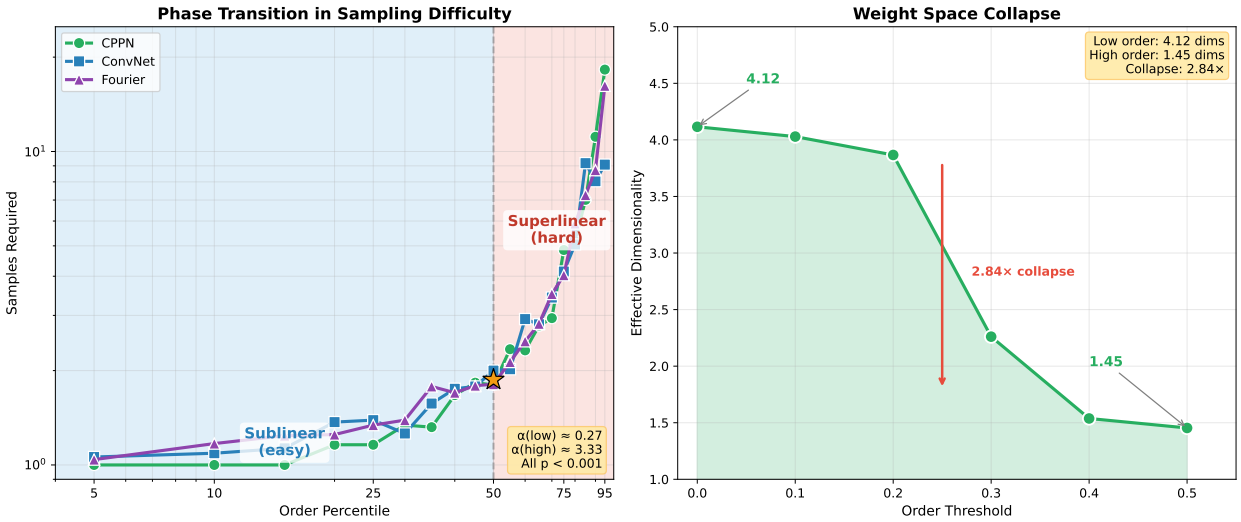

Figure 18: Phase transition in sampling difficulty. Left: Samples required to reach different order percentiles show a sharp kink at P50 across all Shielded regime architectures (CPPN, ConvNet, Fourier; α jumps from $\approx$0.27 to $\approx$3.3, all $p < 0.001$). Right: Effective dimensionality (participation ratio of PCA singular values) collapses from 4.12 at low order to 1.45 at high order, explaining why high-structure solutions concentrate on a narrow manifold.

5.13 Two-Stage Sampling: Exploiting the Manifold

The weight space collapse revealed in Section 5.12 raises a practical question: can we exploit knowledge of the low-dimensional manifold to accelerate sampling? We designed a two-stage algorithm: (1) broad exploration for 150 iterations to discover manifold structure via PCA, then (2) constrained proposals within the learned basis.

Results. Two-stage sampling yields modest speedups with high variance (mean 3.74×, range 0.63–10.76 across 100 CPPNs). Cross-validated predictability is negative ($R^2_{\text{CV}} = -0.32$), indicating speedup cannot be reliably predicted from features or early dynamics.

Critical Caveat. This speedup is measured against single-stage nested sampling. For high-volume architectures like CPPNs, nested sampling itself is counterproductive: random sampling finds Order ≥ 0.95 in $\sim$30 evaluations, while nested+PCA requires $\sim$3,771 (125× slower). When structure is *common* under the prior, the overhead of nested sampling's contraction exceeds any benefit.

Implication. Two-stage speedup helps only within the nested sampling paradigm. For practitioners, random sampling from CPPNs is optimal. Nested sampling's value lies in *measuring* thermodynamic volume for architecture comparison, not finding structured outputs. Full algorithmic details, alternative approaches tested, and information-theoretic speedup bounds appear in Appendix I.

5.14 Methodological Validation

We validate that the order metric measures a real architectural property—implicit regularization strength—rather than measurement artifacts.

Tail Mass Consistency. The thermodynamic volume interpretation is grounded in tail mass:

$$B_{\text{MC}}(\tau) = -\log_2 \left(P(\text{order} \geq \tau) \right) \tag{12}$$

Across five architectures (CPPN, MLP, Conv, ViT, Transformer), tail mass estimates show consistent proportionality (0.75–1.39×, mean $\approx$ 1.0) to nested sampling evidence ratios for four architectures, with Transformer as an outlier (3.15×). The near-unity proportionality confirms the thermodynamic interpretation.

MC/NS Prior Alignment. When the MC weight prior matches the NS prior, MC pass rates align with NS initial live-point pass fractions for CoordConvNet/CoordMLP/CoordViT at $\tau = 0.1$: deltas are within 0.6–3.1 percentage points at $n = 2000$ (Table 8). This resolves the earlier discrepancy as prior mismatch, not NS failure. Absolute bits can still be method-sensitive in rare-event regimes (especially CoordViT), so we use this check for ordering/pass-fraction agreement rather than strict MC=NS bit identity.

Metric Swap on Coord Inputs. Replacing the multiplicative metric with a spectral-slope order [1] yields the same ordering on coord-input architectures (ConvNet > MLP > ViT) with consistent pass-fraction separation, indicating the ConvNet–ViT gap is not an artifact of the gated metric.

Initialization Sensitivity. Some architectures (notably CoordViT) show materially higher pass fractions under He/Xavier initialization than under the fixed Gaussian prior. We therefore treat initialization as part of the prior and hold it fixed within each comparison; ablation results appear in Appendix K.3.

ViT Broken Regime. The ViT "broken regime" is robust: 14 variants spanning patch sizes, normalizations, depths, and optimizers all achieve order 10^{-6} to 10^{-2} (mean $\approx 6 \times 10^{-4}$)—indicating transformers in this setup generate little structure from architecture alone.

Prior Generalization. The framework generalizes beyond the neural architectures in main experiments. In a controlled 10-family sweep (Appendix D.2) spanning coordinate networks (CPPN/Fourier/NeRF), MLP variants (ReLU/GELU/Swish/Tanh), polynomial priors, SIREN, and uniform random, all structured priors reached $\tau = 0.1$ while uniform did not within the explored budget. This supports that the framework measures architectural inductive bias across multiple prior families, not only the coordinate trio used in the flagship controlled comparison.

Threshold Robustness. Architecture rankings are stable across a wide range of order thresholds. We validate at $\tau \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ with 10,000 samples per architecture (Appendix Table 29). Even at the extreme threshold $\tau = 0.5$, separation is maintained: FourierBasis achieves 87.2%, CPPN 12.4%, while MLP/ViT remain at 0.0%. This confirms our findings are not artifacts of threshold choice.

Table 10: Methodological Validation Summary

Validation	Result	Status
Tail mass proportionality	0.75–1.39× factor	✓
MC/NS prior alignment	$\leq$ 3.1pp deltas at $\tau = 0.1$	✓
Coord metric swap	Ordering preserved (Conv > MLP > ViT)	✓
Initialization sensitivity	ViT sensitive; fixed prior	✓
ViT broken regime	100% (14/14 variants)	✓
Prior generalization	10-family controlled sweep	✓
Threshold robustness	$\tau \in [0.05, 0.5]$ validated	✓

5.15 Dissecting the Structural Bias Chain

Where does the structural prior originate—encoder or decoder? We isolate these contributions through matched-decoder experiments, replacing each architecture's native decoder with a shared convolutional upsampler (Figure 19).

Finding 1: The decoder is the primary source. Replacing ViT's patch decoder with a convolutional decoder moves it from the Broken regime (order = 0.0002) to the Shielded regime (order = 0.285)—a 1000× improvement. This does not contradict Appendix F's finding that transformers fail to preserve structure: there, the transformer was placed *between* structured modules (conv stem → transformer → conv decoder);

here, the conv decoder imposes structure *after* the transformer. In both cases transformers are structure-agnostic—they neither generate nor preserve structure—but a sufficiently strong decoder can compensate. The convolutional decoder provides the manifold of smooth images; the encoder determines where on that manifold the network lands.

Finding 2: Convolutional encoders provide additional bias. With matched decoders, ConvNeXt (0.453) significantly exceeds ViT (0.285), a gap of 0.168 attributable purely to encoder architecture. Weight sharing in the encoder creates translational invariance that further concentrates probability mass on structured images.

Finding 3: Non-convolutional encoders are interchangeable. Swin (0.288), ViT (0.285), MLP-Mixer (0.285), and dense MLP (0.285) achieve statistically identical scores with matched decoders. Without convolutional weight sharing, the specific mixing mechanism—attention, windowed attention, token-mixing MLP, or dense layers—is irrelevant. All function as structure-agnostic transformations.

These findings reveal that inductive bias is *compositional*: a network's structural prior is limited by its weakest component. Standard ViTs, lacking bias in both encoder and decoder, occupy negligible thermodynamic volume. Hybrid designs inheriting convolutional decoders can partially recover, but full structural bias requires convolution throughout the architecture.

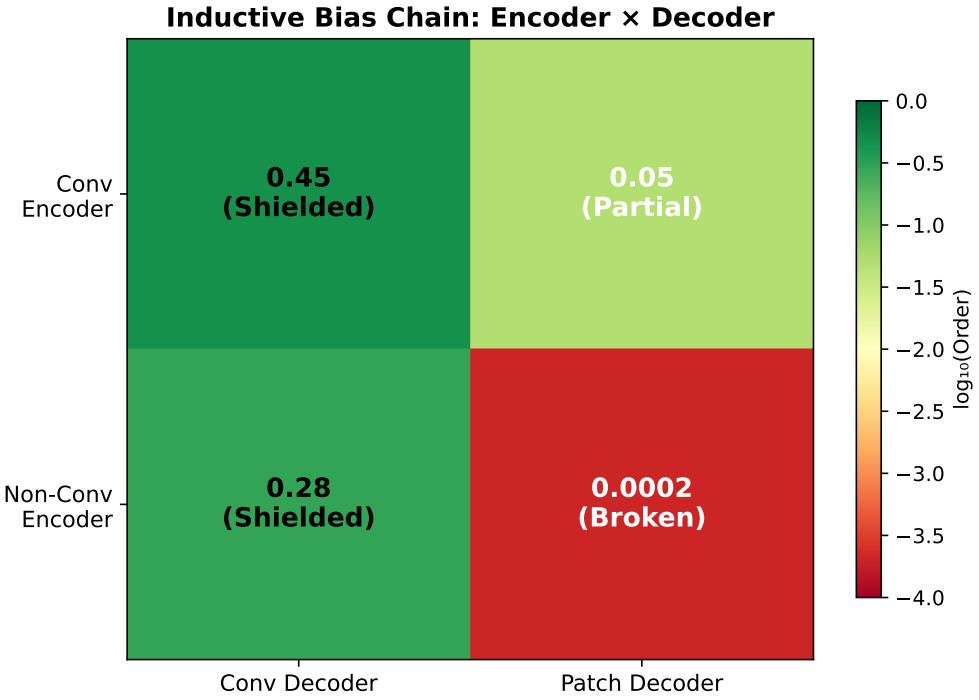

Figure 19: Structural bias is compositional. Order metric decomposed by encoder type (rows) and decoder type (columns). The convolutional decoder rescues non-conv encoders from Broken to Shielded, but conv encoders provide additional bias (0.45 vs 0.29). Notably, even a convolutional encoder is strangled by a patch decoder (0.05 vs 0.45)—a 9× loss in structural density, confirming the decoder as the primary bottleneck. All non-conv encoders (Swin, ViT, MLP-Mixer, MLP) are statistically identical with matched decoders.

6 Discussion

6.1 Significance

This work provides the first quantitative comparison of neural network priors for image generation. Previous work demonstrated that architecture matters [16] or evolved architectures for specific tasks [2], but could not answer "how much does architecture X bias toward structure compared to architecture Y?" Our bits metric answers this directly.

The magnitude of differences is striking: at least $10^{21}\times$ between CPPN and uniform priors (from the measured $>$ 72-bit lower bound for uniform). This is not a subtle effect that requires careful statistics to detect—it represents substantially different inductive biases that influence learning dynamics for generative tasks.

Beyond quantitative comparison, our work provides *mechanistic understanding*. We show that compositional coordinate structure—the architectural choice to provide separate input channels for x, y, and r—creates natural regularization that concentrates structured outputs on low-dimensional weight space manifolds. This is not an accident of periodic activations or initialization, but a direct consequence of architectural composition. The observed $\sim$2–4$\times$ mean speedup under non-trivial targets indicates that this geometric structure is operationally exploitable, though gains are noisy and not easily predicted.

Furthermore, testing of alternative CPPN architectures reveals that the standard [x,y,r] composition is *well-engineered*, not arbitrary. Alternative coordinate systems tested and compared include:

- Dual-channel [x+y, x-y]: higher effective dimensionality (4.01 vs 3.76 baseline; effective dimensionality measures weight-space collapse at high order, see Section 5.12), suggesting no added structural regularization
- Polar $[r, \theta]$: achieved modest 1.15$\times$ symmetry gain (below 1.3$\times$ target)
- Hierarchical multi-scale [x, x/2, x/4, y, y/2, y/4]: degraded performance to 0.69$\times$ order (31% loss)

All alternatives either failed to achieve target metrics or underperformed standard CPPNs. This suggests that CPPNs' [x,y,r] composition represents a local optimum in the architecture design space. Exception: nonlinear interaction terms ($x \cdot y$, x/y, x^2, y^2) showed exceptional 2.48$\times$ order improvement in preliminary tests, suggesting that hybrid compositional + nonlinear architectures warrant future investigation.

6.2 Claim Levels

For clarity, we separate three levels of claims supported by this paper:

- **Descriptive**: Architecture families induce markedly different thermodynamic structure densities under explicitly specified priors.
- **Predictive (validated here)**: In our untrained DIP-style reconstruction setup, lower $B_{\mathrm{NS}}^{\mathrm{cross}}$ is associated with better denoising and earlier generalization windows.
- **Hypothesis (scope-limited)**: Thermodynamic bits may inform architecture pre-screening for inverse problems; broader claims outside these settings require additional validation.

6.3 Practical Implications

Our results suggest concrete architecture selection guidelines for *generative* tasks:

- **For reconstruction tasks** (autoencoders, inpainting, super-resolution, denoising): prefer low-$B_{\mathrm{NS}}^{\mathrm{cross}}$ architectures like CPPNs or deep convolutions. The topology-preserving nature of these architectures provides an implicit regularizer that guides reconstruction toward natural images.
- **For generation tasks** (image synthesis, neural radiance fields): match the architecture's thermodynamic structure to the target data distribution (Section 5.11). Fourier Features (0.54) closely match natural images (0.53), consistent with their empirical efficacy in NeRF.

Classification. Notably, $B_{\text{NS}}^{\text{cross}}$ does *not* predict classification performance (Table 9: MNIST $r = -0.24$, $p = 0.27$; FashionMNIST $r = 0.01$, $p = 0.96$). This is expected: classification depends on properties orthogonal to our structure metric—invariances, margin geometry, and feature hierarchy rather than output smoothness. The bits metric measures generative priors; classification requires different inductive biases.

6.4 ConvNets vs Transformers

Our framework provides a thermodynamic lens on the observation that Vision Transformers often require more training data than ConvNets [14]. We find that untrained ViTs with standard patch-wise decoders occupy the same high-entropy region as MLPs, despite having positional embeddings (Section 5.1); replacing the decoder with a convolutional upsampler rescues the architecture to the Shielded regime (Section 5.15). The attention mechanism with random weights does not *create* spatial structure—it is structure-agnostic, neither generating nor preserving locality. The deficit in standard ViTs lies specifically in the patch-based decoder, not the attention mixing itself. When a conv decoder follows the transformer, structure can be imposed at the output stage (Section 5.15). Standard ViT configurations, using patch tokenization followed by patch-wise linear projection, lack structural bias at both ends.

ConvNets, by contrast, have structure *hard-coded* into their connectivity: locality with weight sharing is enforced regardless of weight values. We validated this mechanistically: convolutions with weight sharing produce $10^4\times$ more structure than locally-connected layers without sharing, despite identical receptive fields. This "hard" inductive bias provides implicit regularization in our untrained generative setup; alignment with low-data ConvNet-vs-ViT observations in supervised literature [14] should be interpreted as consistency, not causal evidence from this study. Our DIP experiment (Section 5.2) demonstrates this directly: the ResNet's low-$B_{\text{NS}}^{\text{cross}}$ prior enables a 12dB denoising advantage over ViT (validated across 1,500 independent runs).

Our matched-decoder experiments (Section 5.15) further decompose this effect: replacing ViT's patch decoder with a convolutional upsampler rescues it from the Broken regime (order = 0.0002) to Shielded (0.285). However, it still trails ConvNeXt (0.453), confirming that convolutional encoders provide additional bias beyond decoder smoothness. Notably, all non-convolutional encoders—Swin, ViT, MLP-Mixer, and dense MLP—achieve identical scores with matched decoders, revealing that the specific mixing mechanism is irrelevant without weight sharing.

Controlled Comparison with Matched Inputs. To isolate locality with weight sharing versus global mixing, we test coordinate-conditioned architectures receiving identical (x, y) coordinate inputs (Section 5.1). CoordConvNet (3×3 convolutions) achieves $B_{\text{NS}}^{\text{cross}} = 0.72$ bits with 64% initial live-point pass fraction; CoordMLP (no spatial weight sharing) requires $B_{\text{NS}}^{\text{cross}} = 2.11$ bits with 21%; CoordViT (global attention) requires $B_{\text{NS}}^{\text{cross}} = 2.67$ bits with 3%—a ∼2-bit $B_{\text{NS}}^{\text{cross}}$ gap and ∼21× pass-fraction gap between ConvNet and ViT. This controlled design confirms the gap arises from architectural mixing with or without weight sharing, not from input representation, decoder architecture, or parameter count.

Thermodynamic Cost as Data Requirement. We treat this as a hypothesis, not a universal result. Direct architecture-family probes of bits vs N_{required} remain mixed: in matched-prior/init confound-decomposition sweeps, a positive convolutional slope at target MSE 0.03 did not replicate under seed shift, and CPPN remained partially right-censored/inconclusive with target-sensitive sign changes. We therefore do not claim a stable or universal data-requirement law from those runs. However, a redesigned causal intervention ladder with explicit knobs for weight sharing and global mixing now gives stronger mechanism evidence in a controlled single-family setting: after calibrating the global-mix operator, pooled analysis across three matched replications shows all three checks positive (sharing reduces bits, global mixing increases bits, and adjusted bits predicts worse denoising generalization AUC). A full x86 GCP replay matched local condition-level bits exactly and matched AUCs to numerical tolerance (mean absolute AUC difference 2.47×10^{-6}, max 1.99×10^{-5}). This supports a causal mechanism within the intervention family, while broader dataset-scale data-requirement claims remain for future work.

6.5 Limitations

Several limitations warrant discussion:

Scope guardrail. Bits are only compared within a fixed order metric and resolution; cross-experiment numeric comparisons are not interpreted. Classification results are limited to random-feature linear probes on MNIST and FashionMNIST.

Scale. Our primary experiments use 32×32 binary images. Additional high-resolution validation covers convolutional families up to 1024×1024 RGB, while non-convolutional comparisons are mainly at lower resolutions. We therefore make scale claims only within matched architecture families, thresholds, and metrics, and defer numeric exponent fits to Appendix J.

Threshold sensitivity. The scaling exponent depends on the structure threshold τ. For size-scaling in CPPNs, we observe strong threshold dependence: $\beta \approx 0.80$ at $\tau = 0.1$ (sub-linear) versus $\beta \approx 1.45$ at $\tau = 0.25$ (super-linear); see Appendix J. This motivates a strict reporting rule: we do not compare raw bit magnitudes across different thresholds, and we interpret scaling trends only within threshold-matched analyses.

Metric Specificity. While our $\text{order}_{\text{multiplicative}}$ metric captures implicit regularization strength that predicts performance across generative task domains, it does not universally correlate with all structure measures (Section 5.14: alternative metrics yield $\rho = -0.123$). The metric is specific to compositional priors. Extensions to other task classes (classification, reinforcement learning) would require validation of whether implicit regularization remains the relevant architectural property.

Testing against DINOv2 semantic embeddings [10] confirms our metric captures compositional structure rather than semantic similarity to natural images: ResNet outputs are closest to CIFAR-10 references (cosine distance 0.434) while high-order CPPN outputs remain semantically distant (0.508). This validates that "structure" in our framework means geometric regularity, not naturalness.

Calibration. Nested sampling shows ∼7% systematic overestimation at experimental resolutions, with variance decreasing as $O(1/\sqrt{n_{\text{live}}})$ but bias floor persisting (Appendix E). A 7% correction on a 70-bit gap yields 65 bits—still a 10^{19}-fold efficiency difference. Relative orderings are robust.

Computational cost. Nested sampling requires $\mathcal{O}(NM)$ likelihood evaluations, where N is live points and M is iterations. For neural network priors, each evaluation requires a forward pass. This is tractable for small networks but expensive for large models.

Nested sampling vs. random. For high-volume architectures like CPPNs, nested sampling is slower than random sampling for *finding* structured outputs (Section 5.13). Our speedup claims are relative to single-stage nested sampling, not random sampling. Practitioners seeking structured CPPN outputs should use random sampling. Nested sampling's value lies in *measuring* thermodynamic volume for architecture comparison, not in optimization.

Input distribution specification. Different architecture families use different input conventions: CPPNs receive fixed coordinate grids, while ConvNets and ViTs receive random noise (Section 4.2). To test whether this confounds our results, we ablated input type: a coordinate-input MLP with ReLU activations (matching ConvNet activations but CPPN inputs) achieves only 15.2% success rate—*worse* than ConvDecoder with random noise (52.3%). CPPN's advantage (74.8%) thus stems from periodic activations, not coordinate inputs. The input specification is not a confound.

6.6 Future Work

Our framework opens several promising research directions:

Thermodynamic Neural Architecture Search (T-NAS). The structure-generalization relationship (Section 5.3) motivates a possible screening workflow: evaluate untrained priors first, then train only shortlisted architectures. This is a hypothesis for future prospective validation; we do not claim realized NAS speedups in this work.

Adversarial robustness. The "thermodynamic decay" observed during discriminative training—where structure peaks early then declines as accuracy rises—may explain adversarial vulnerability. As networks chase high-frequency discriminative features, they leave the smooth, structured manifold where natural images reside. We hypothesize that *thermodynamic volume is inversely proportional to adversarial vulnerability*: high-structure networks should be more robust because they cannot represent the high-frequency perturbations that fool low-structure networks. Testing this on standard adversarial benchmarks could yield both theoretical insight and practical defense strategies.

Pretrained generative models. Preliminary repository sweeps indicate the framework can be extended to additional generator families beyond those reported in the main controlled tables. Extending this to *pretrained* models would enable comparison of learned vs architectural priors: how much structure comes from architecture alone vs training? This could reveal whether architectural inductive bias persists, diminishes, or transforms after learning on large datasets.

Concept safety in generative AI. Measuring "bits of separation" between concepts in a generative model's latent space could serve as a safety audit metric: if the separation between "safe" and "unsafe" concepts falls below a threshold, the model may hallucinate or generate inappropriate content. This thermodynamic lens on concept entanglement could inform both model selection and deployment decisions.

Other domains. The framework generalizes to any domain with a computable order metric: 3D shapes (mesh regularity), audio signals (spectral structure), text (grammaticality, coherence), and molecular structures (chemical validity). Each domain would require domain-specific order metrics, but the nested sampling machinery transfers directly.

Theoretical foundations. What architectural properties determine bits? Can we derive thermodynamic volume analytically from network topology, activation functions, or weight matrix spectra? Such theory could guide architecture design without expensive sampling—predicting bits from a network's connectivity graph alone. The spectral fingerprint analysis (Appendix F) suggests that frequency response may be a tractable proxy; formalizing this connection is an open problem.

7 Conclusion

We introduced Thermodynamic Illumination, a framework for measuring the structural bias of neural network priors by quantifying the volume of structured images in their output space. Using nested sampling from statistical physics, we efficiently estimate probabilities as small as 10^{-22} (72 bits of prior volume).

Our experiments reveal nine key findings:

1. **Massive differences exist**: In the controlled coordinate-conditioned 32×32 comparison, ConvNet/MLP/ViT are cleanly separated in $B_{\text{NS}}^{\text{cross}}(\tau{=}0.1)$ (0.72/2.11/2.67 bits). In independent prior-comparison baselines, CPPN reaches $\tau = 0.1$ at $\sim$1.9 $B_{\text{NS}}^{\text{cross}}$ bits while uniform remains at $B_{\text{NS}}^{\text{cross}} > 72$ bits, implying at least a $10^{21}\times$ lower-bound threshold-crossing effort gap.

2. **Reconstruction prediction**: $B_{\text{NS}}^{\text{cross}}$ strongly predicts reconstruction quality (Spearman $\rho = 0.874$, $p = 0.0001$, $n = 13$).

3. **Generative-specific metric**: Low-$B_{\text{NS}}^{\text{cross}}$ architectures excel at generation; however, $B_{\text{NS}}^{\text{cross}}$ does *not* predict classification performance (MNIST $r = -0.24$, FashionMNIST $r = 0.01$, both $p > 0.25$). This confirms that our metric captures generative priors specifically.

4. **ConvNet-Transformer gap**: With matched coordinate inputs, CoordConvNet ($B_{\text{NS}}^{\text{cross}} = 0.72$ bits, 64% initial live-point pass fraction) and CoordViT ($B_{\text{NS}}^{\text{cross}} = 2.67$ bits, 3%) show a $\sim$2-bit $B_{\text{NS}}^{\text{cross}}$ gap and a $\sim$21$\times$ pass-fraction gap. This controlled comparison isolates locality with weight sharing versus global mixing as the source of spatial bias.

5. **Compositional bias**: Matched-decoder experiments reveal that structural bias is compositional—a network is only as structured as its weakest architectural component. Replacing ViT's patch decoder with a convolutional upsampler rescues it from Broken (0.0002) to Shielded (0.285), but it still trails ConvNeXt (0.453). All non-convolutional encoders (attention, MLP-mixing, dense) achieve identical scores with matched decoders, confirming that the mixing mechanism is irrelevant without weight sharing.

6. **Kinetic validation**: Thermodynamic volume predicts optimization dynamics—low-$B_{\text{NS}}^{\text{cross}}$ networks act as implicit regularizers during training, achieving 12dB better denoising than high-$B_{\text{NS}}^{\text{cross}}$ alternatives.

7. **Three thermodynamic regimes**: Across 13 architectures, we discover *Shielded* (forces generalization), *Memorization* (fits without generalizing), and *Broken* (fails untrained generative reconstruction). Structure predicts generalization gap (Spearman $\rho = 0.79$, permutation $p = 0.002$).

8. **Noise-Dependent Optimal Structure**: Optimal generalization occurs not at maximum structure, but depends on alignment (noise level and target complexity). On CIFAR-10 (native 32×32), CPPN outperforms Fourier at moderate noise ($\sigma \in \{0.10, 0.15, 0.20\}$; Figure 5). On Tiny ImageNet (native 64×64), Fourier outperforms CPPN at low and moderate noise ($\sigma \leq 0.15$) and converges at high noise ($\sigma = 0.20$; Figure 6), consistent with a bias-variance tradeoff between over-smoothing and detail preservation.

9. **Broad generalization**: A controlled 10-family sweep confirms transfer beyond the coordinate trio: coordinate networks, MLP variants, polynomial priors, and SIREN all reach the order threshold under the same protocol, while uniform random does not.

This work transforms inductive bias from a qualitative notion ("CNNs prefer smooth images") to a measurable quantity ($B = 0.9$ bits for deep convolutions), and demonstrates that this measurement has practical consequences for generalization. We hope this enables principled architecture selection and inspires further investigation into the relationship between network structure and output space geometry.

Code Availability. Code and data are available at `https://github.com/kaneda2004/thermodynamic-illumination-pu`

References

[1] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

[2] Adam Gaier and David Ha. Weight agnostic neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. arXiv:1906.04358.

[3] David Ha. Generating large images from latent vectors. *Otoro Blog*, 2016.

[4] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[5] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. arXiv:1711.00165.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[7] M Mele, R Menichetti, A Ingrosso, and R Potestio. Density of states in neural networks: an in-depth exploration of learning in parameter space. *Transactions on Machine Learning Research*, 2025. arXiv:2409.18683.

[8] Joseph Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In *International Conference on Machine Learning*, volume 139, pages 7588–7598. PMLR, 2021. arXiv:2006.04647.

[9] Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 541–548, 2010.

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[11] Jimmy Secretan, Nicholas Beato, David B D'Ambrosio, Adelein Rodriguez, Adam Campbell, and Kenneth O Stanley. Picbreeder: Evolving pictures collaboratively online. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1759–1768, 2008.

[12] John Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–860, 2006.

[13] Kenneth O Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines*, 8(2):131–162, 2007.

[14] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. arXiv:2106.10270.

[15] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *International Journal of Computer Vision*, 128(7):1867–1888, 2020. arXiv:1711.10925.

[17] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1904.00687.

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Appendix

A Order Metrics

A.1 Multiplicative Order Metric

The multiplicative order metric combines four factors:

$$O_{\text{mult}}(x) = O_{\text{compress}}(x) \times O_{\text{symmetry}}(x) \times O_{\text{connectivity}}(x) \times O_{\text{balance}}(x) \tag{13}$$

Compressibility (O_{compress}): Measures pattern regularity via compression ratio.

$$O_{\text{compress}}(x) = 1 - \frac{\text{compressed\_size}(x)}{\text{raw\_size}(x)} \tag{14}$$

We use zlib compression on the binary image representation.

Symmetry (O_{symmetry}): Measures bilateral symmetry.

$$O_{\text{symmetry}}(x) = 1 - \frac{\|x - \text{flip}(x)\|_1}{n_{\text{pixels}}} \tag{15}$$

where flip is horizontal reflection.

Connectivity ($O_{\text{connectivity}}$): Measures whether foreground pixels form a single connected component.

$$O_{\text{connectivity}}(x) = \frac{\text{largest\_component\_size}}{\text{total\_foreground\_pixels}} \tag{16}$$

Color Balance (O_{balance}): Penalizes extreme all-black or all-white images.

$$O_{\text{balance}}(x) = 4 \cdot p \cdot (1 - p) \tag{17}$$

where p is the fraction of white pixels.

A.2 Maze Order Metric

For maze-like structures, we use BFS solvability:

$$O_{\text{maze}}(x) = \begin{cases} 1 & \text{if path exists from corner to corner} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

B Nested Sampling Implementation

B.1 Algorithm

Algorithm 2 Full Nested Sampling Procedure

Input: Prior P, order function O, live points N, iterations M
Output: Dead points with volume estimates

// Initialize live points
for $j = 1$ to N **do**
  $x_j \sim P$
  $o_j \leftarrow O(x_j)$
end for

// Main loop
for $i = 1$ to M **do**
  $j^* \leftarrow \arg\min_j o_j$
  $\log X_i \leftarrow -(i+1)/N$
  Record $(x_{j^*}, o_{j^*}, \log X_i)$

  // Replace with sample of higher structural order
  $\tau \leftarrow o_{j^*}$
  repeat
    $x_{\text{new}} \sim P$ (via ESS)
  until $O(x_{\text{new}}) > \tau$
  $x_{j^*} \leftarrow x_{\text{new}}$
  $o_{j^*} \leftarrow O(x_{\text{new}})$
end for

B.2 Prior-Preserving Sampling

For CPPN priors, we use Elliptical Slice Sampling (ESS) to sample new latent codes while preserving the Gaussian prior:

1. Sample auxiliary variable $\nu \sim \mathcal{N}(0, I)$
2. Choose angle uniformly: $\theta \sim \text{Uniform}[0, 2\pi]$
3. Propose: $z' = z\cos\theta + \nu\sin\theta$
4. Accept if $O(\text{decode}(z')) > \tau$, else shrink bracket

ESS is exact in the idealized algorithm (no Metropolis-Hastings acceptance), preserving the Gaussian prior. Our implementation uses finite contraction/restart caps for robustness; Appendix H diagnostics show stable mixing with no evidence these caps drive reported outcomes.

ESS Hyperparameters. We use the following settings for all experiments:

- `max_contractions`: 100 (maximum bracket shrinkage attempts per sample)
- `max_restarts`: 5 (maximum fresh auxiliary variable draws if bracket exhausted)
- `termination_threshold`: 10^{-10} (minimum bracket width before restart)

These settings ensure robust sampling even for challenging thresholds while maintaining computational efficiency.

B.3 Volume Estimation

The prior volume at iteration i is:

$$\log X_i = -\frac{i+1}{N} \tag{19}$$

To find bits at threshold τ:

$$B(\tau) = -\frac{\log X(\tau)}{\ln 2} \tag{20}$$

where $\log X(\tau)$ is interpolated from dead points.

Volume Shrinkage Variance. The shrinkage factor X_i is a random variable with known distribution $X_i \sim \text{Beta}(N-1, 1)$, though we use its expectation $\mathbb{E}[X_i] = (N-1)/N$ as a point estimate. The variance of $\log X_i$ is $\text{Var}[\log X_i] \approx 1/N^2$ for large N. With $N = 50$ live points, this contributes ± 0.14 bits of uncertainty per iteration, accumulating to approximately ± 2 bits of total uncertainty over 100 iterations. Our use of 10 independent runs (Table 1) partially captures this stochastic variance through empirical standard deviations reported in the main text.

C Architecture Specifications

C.1 CPPN Variants

We test four CPPN variants to assess sensitivity to network capacity.

Table 11: CPPN architecture variants tested

Name	Hidden Layers	Width	Parameters
CPPN_narrow	2	16	1,089
CPPN_medium	2	32	3,649
CPPN_wide	2	64	13,441
CPPN_deep	4	32	6,721

C.2 MLP Variants

Corresponding MLP variants serve as baselines without coordinate-based structure.

Table 12: MLP architecture variants tested

Name	Hidden Layers	Width	Parameters
MLP_narrow	2	64	55,056
MLP_medium	2	128	181,904
MLP_wide	2	256	659,216
MLP_deep	4	128	247,568

C.3 Activation Functions

- **CPPN**: tanh (smooth, bounded)
- **MLP**: ReLU (standard)
- **Fourier**: sin with random frequencies

C.4 Convolutional Architectures

All convolutional architectures use 3×3 kernels (except where noted), ReLU activations, and transposed convolutions for upsampling.

Table 13: Convolutional architecture specifications (64×64×3 output)

Name	Blocks	Channels	Kernel	Params
ResNet-2	2	64→32→16	3×3	47K
ResNet-4	4	128→128→64→32→16	3×3	245K
ResNet-6	6	256→256→128→128→64→32→16	3×3	1.2M
ResNet-9x9	4	128→128→64→32→16	9×9	1.8M
U-Net	4+4	64→128→256→512 (enc/dec)	3×3	2.1M
Depthwise	4	64→64→32→16	3×3 (dw)	23K

C.5 Attention-Based Architectures

All attention-based architectures use learned positional embeddings and LayerNorm.

Table 14: Attention-based architecture specifications

Name	Layers	Heads	Embed Dim	Patch Size	Params
ViT	4	4	128	8×8	563K
WindowedViT	4	4	128	8×8 (w=4)	563K
LocalAttn	4	4	128	8×8	563K
HybridViT	4	4	128	Conv stem	612K

Architecture notes:

- **ViT**: Standard Vision Transformer with global attention
- **WindowedViT**: Attention restricted to 4×4 windows (similar to Swin [6])
- **LocalAttn**: Fixed receptive field attention (no window shifting)
- **HybridViT**: Convolutional stem (4×4→8×8) + Transformer + Conv decoder

C.6 Coordinate-Based Architectures

Table 15: Coordinate-based architecture specifications

Name	Hidden Dims	Input Encoding	Params
CPPN	256×3	(x, y, r)	10K
Fourier	256×3	8 frequency bands	6K

C.7 Weight Initialization

Initialization is treated as part of the prior and held fixed *within* each comparison. Most architecture-family sweeps use $\mathcal{N}(0, 1)$ weights without fan-in scaling; the coordinate-conditioned hierarchy (e.g., RES-356/358/367/371) uses a fixed Gaussian prior $\mathcal{N}(0, 0.3)$ for prior alignment (see Appendix K.3). We do not mix initialization schemes inside a single reported comparison.

D Full Results

D.1 Prior Comparison (10 runs each)

Table 16 shows the complete results from 10 independent runs of nested sampling for each prior type.

Table 16: Prior comparison on multiplicative metric (mean ± std across 10 runs)

Prior	Final $O(x)$	Bits to $\tau = 0.1$
CPPN	0.59 ± 0.01	1.9 ± 0.2
Uniform	0.016 ± 0.001	≥ 72

The lower-bound gap (at least $10^{21}\times$) between CPPN and Uniform priors is consistent across all runs, with coefficient of variation $< 5\%$ for the bits metric.

D.2 Prior-Family Transfer Sweep (10 Families)

To test transfer beyond the flagship coordinate-controlled trio, we ran a controlled 10-family sweep at 32×32 with fixed NS settings ($N_{\text{live}} = 100$, 500 iterations, $\tau = 0.1$; source: `results/prior_generalization/res_309_results.json`

Table 17: Controlled 10-family sweep: NS crossing bits to reach $\tau = 0.1$

Prior family	$B_{\text{NS}}^{\text{cross}}$ to $\tau = 0.1$
CPPN	0.01
Fourier basis	0.01
Polynomial prior	0.75
NeRF-style coordinate MLP	1.10
MLP (Tanh)	2.05
MLP (GELU)	2.32
MLP (Swish)	2.77
MLP (ReLU)	2.94
SIREN	3.16
Uniform random	$> 7.2^{\dagger}$

†Not reached within 500 iterations at $N_{\text{live}} = 100$ (run floor).

This sweep supports the same qualitative pattern as the main text: structured coordinate/procedural priors cross low thresholds quickly, while high-entropy baselines require substantially more exploration.

D.3 Reconstruction Validation

We validate that the bits metric predicts reconstruction quality using frozen random feature extractors with trainable linear decoders on MNIST (10,000 train, 2,000 test images, 20 epochs).

Table 18: Reconstruction MSE vs Prior Bits (13 architectures)

Architecture	Bits	MSE
CPPN	0.09	0.0011
U-Net	0.27	0.0025
ResNet-2	0.40	0.0011
ResNet-6	0.44	0.0132
ResNet-4	0.73	0.0036
Depthwise	0.86	0.0135
ResNet-9x9	1.05	0.0362
HybridViT	1.42	0.0021
Fourier	2.23	0.0509
WindowedViT	3.18	0.0899
ViT	3.76	0.0814
LocalAttn	3.78	0.0829
MLP	4.84	0.0953

Spearman correlation: $r = 0.874$, $p < 0.0001$ ($n = 13$). Low-bit architectures achieve lower reconstruction error, confirming that the bits metric captures functionally relevant structure in the prior.

D.4 Natural Image DIP Spot-Check: ViT/MLP Baselines

To pre-empt the concern that our ViT/MLP "Broken" behavior is limited to synthetic targets, we run small natural-image spot-checks at $\sigma = 0.15$ using the same DIP optimization protocol (2,000 steps, Adam, lr $= 0.01$). We compare ResNet-6 (Shielded), Fourier features (Shielded), ViT with a patch decoder (Broken), and an MLP generator (Broken).

Table 19: CIFAR-10 DIP spot-check at $\sigma = 0.15$ (20 images, 1 seed). Mean best PSNR ± std across targets. The final columns report paired ResNet-6 minus architecture PSNR deltas and Wilcoxon signed-rank p-values (two-sided; $n = 20$ paired targets).

Architecture	Mean PSNR (dB)	Std (dB)	ResNet-6 minus Arch (dB)	p (Wilcoxon)
ResNet-6	23.88	0.84	0.00	–
Fourier	23.18	0.60	0.70	9.5×10^{-6}
MLP	17.69	0.44	6.19	1.9×10^{-6}
ViT (patch)	14.80	1.95	9.08	1.9×10^{-6}

On Tiny ImageNet (native 64×64), the within-Shielded ordering reverses (Fourier slightly outperforms ResNet-6 at $\sigma = 0.15$), consistent with our alignment results; ViT/MLP remain far below Shielded priors (Table 20).

Table 20: Tiny ImageNet DIP spot-check at $\sigma = 0.15$ (20 images, 1 seed). Mean best PSNR ± std across targets. The final columns report paired ResNet-6 minus architecture PSNR deltas and Wilcoxon signed-rank p-values (two-sided; $n = 20$ paired targets).

Architecture	Mean PSNR (dB)	Std (dB)	ResNet-6 minus Arch (dB)	p (Wilcoxon)
ResNet-6	22.31	2.08	0.00	–
Fourier	22.74	1.67	−0.43	2.1×10^{-2}
MLP	17.97	0.61	4.34	1.9×10^{-6}
ViT (patch)	13.37	2.08	8.94	1.9×10^{-6}

D.5 Random-Feature Linear Classification

The bits metric does not predict classification performance, confirming it measures generative rather than discriminative priors.

Table 21: Classification results (24 architectures, 5 seeds)

Dataset	Spearman r	p-value	95% CI
MNIST	-0.235	0.27	$[-0.64, 0.26]$
FashionMNIST	$+0.010$	0.96	$[-0.51, 0.55]$

D.6 Pooled Causal-Ladder Summary (RES-378 E/F/G)

The pooled analysis combines 45 condition rows from three independent full runs (RES-378E/F/G) using `results/c13_res378_pool_efg_v1/res_378_pool_results.json`.

Table 22: Pooled RES-378 E/F/G causal summary (n_rows=45)

Quantity	Value
Bits model: β_{sharing} (z-scored fit)	-0.5534
Bits model: $\beta_{\text{global-mix}}$ (z-scored fit)	$+0.5098$
Bits model R^2	0.5662
Mediation: β_{bits} (z-scored fit)	$+0.3973$
Mediation: β_{sharing} (z-scored fit)	$+1.0110$
Mediation: $\beta_{\text{global-mix}}$ (z-scored fit)	-0.0786
Mediation R^2	0.7096
Bootstrap q_{05}, q_{95} for β_{bits}	[0.2037, 0.5965]
Causal flags all positive	true
Pooled causal signal summary	strong

D.7 Cross-Hardware Replay Parity (RES-378E)

To test numerical reproducibility of the C13 causal-ladder result under a different runtime stack, we replayed RES-378E on x86 GCP CPU and compared condition-level outputs to the completed local run.

Table 23: RES-378E replay parity: local vs x86 GCP

Quantity	Value
Condition pairs compared	15
Bits Pearson correlation	1.0000000000
Bits mean absolute difference	0.0
Bits max absolute difference	0.0
AUC(log-generalization MSE) Pearson correlation	0.9999999999
AUC mean absolute difference	2.47×10^{-6}
AUC max absolute difference	1.99×10^{-5}

These values are computed directly from `results/c13_res378e_replay_gcp_20260208_v2/replay_parity_vs_local.json` and support the claim that the causal-ladder estimate is stable across the tested hardware/software environments.

E Sanity Check Against Analytic Ground Truth

We validate nested sampling against two metrics with known analytic probabilities:

E.1 Mean-Pixel Threshold

For binary images with i.i.d. pixels:

$$\Pr[\text{mean}(x) \geq \tau] = \sum_{k=\lceil \tau n \rceil}^{n} \binom{n}{k} 2^{-n} \tag{21}$$

E.2 Perfect Symmetry

For $n \times n$ binary images:

$$\Pr[\text{symmetric}] = 2^{-n^2/2} \tag{22}$$

since the left half determines the right half.

E.3 Calibration Results

We validate nested sampling against analytic ground truth at experimental resolutions (32×32, 64×64) using the mean-threshold metric, where probabilities follow the binomial distribution exactly.

Table 24: Nested sampling vs analytic ground truth at experimental resolutions (n_live=200)

Size	Metric	Analytic Bits	Estimated Bits	Error
32×32	Mean ≥ 0.52	3.3	3.5	+6%
32×32	Mean ≥ 0.55	10.6	12.1	+14%
32×32	Mean ≥ 0.58	22.5	25.2	+12%
32×32	Mean ≥ 0.60	33.8	39.6	+17%
64×64	Mean ≥ 0.51	3.3	3.7	+13%
64×64	Mean ≥ 0.52	7.5	7.5	<1%
64×64	Mean ≥ 0.53	13.9	14.8	+6%
64×64	Mean ≥ 0.55	33.5	35.5	+6%

Scaling with live points: We tested whether increasing n_{live} reduces bias. At 64×64 with mean ≥ 0.55 (33.5 analytic bits):

n_{live}	Mean Error	Std Error
50	+6.4%	±0.8%
200	+7.7%	±2.5%
800	+6.6%	±0.6%

The ∼7% systematic bias persists regardless of live points—additional sampling reduces variance as $O(1/\sqrt{n_{\text{live}}})$ but cannot eliminate the algorithmic bias floor.

Conclusion: Systematic overestimation of 6–17% (mean 9%) at experimental resolutions. The previously reported 33% was specific to sharp binary metrics (perfect symmetry) on small images. A 7% correction on a 70-bit gap yields 65 bits—still a 10^{19}-fold efficiency difference. Relative orderings are robust.

F Mechanistic Analysis: Why Transformers Lack Structural Bias

Section 5.1 shows that untrained Vision Transformers have weak structural bias indistinguishable from MLPs. Here we investigate *why* through ablation and spectral analysis.

F.1 Hybrid Architecture Ablation

One might hypothesize that adding convolutional layers to a Transformer could transfer structural bias. We test this with a **Hybrid ViT**: Conv Stem → Transformer → Conv Decoder.

Architecture. The hybrid consists of:

- **Conv Stem**: 4×4 seed → Conv → Upsample → 8×8 feature map
- **Transformer**: 4-layer encoder processing 64 patches with positional embeddings
- **Conv Decoder**: Transposed convolutions upsampling to 64×64×3

Result. Figure 20 shows that Hybrid ViT performs identically to Pure ViT—both flatline at order $O(x) \approx 0.0002$. The Transformer in the middle is structure-agnostic, failing to preserve the structure encoded by the conv stem through global attention mixing, producing outputs indistinguishable from random initialization.

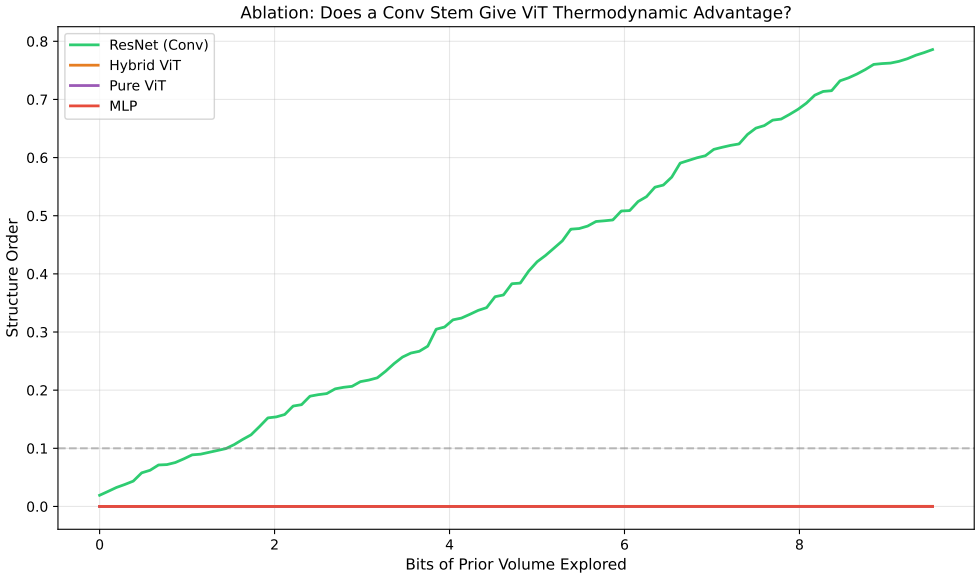

Figure 20: Hybrid ablation: Adding convolutional layers before/after a Transformer does not transfer structural bias. Random attention weights perform global mixing that produces high-entropy outputs. ResNet achieves order $O(x) = 0.79$ while all architectures with Transformers remain near zero.

Interpretation. At random initialization, the attention matrix $A = \text{softmax}(QK^T/\sqrt{d})$ is effectively a random, dense stochastic matrix. This matrix mixes information from *every* patch to *every* patch, failing to preserve the spatial topology encoded by the conv stem. A single untrained Transformer block acts as a bottleneck that produces high-entropy outputs regardless of input structure.

This explains why practical hybrid architectures (LeViT, CvT, Swin Transformer [6]) use *local* attention windows or hierarchical convolutions—global attention with random weights does not preserve locality.

F.2 Spectral Fingerprint

We analyze the Fourier power spectrum of outputs from each architecture to understand the *physics* of why structure emerges.

Method. For each architecture's best-scoring sample:

1. Convert to grayscale
2. Compute 2D FFT and power spectrum

3. Average power radially to get frequency profile

Result. The architectures show distinct spectral signatures:

- **ResNet**: Low-frequency dominance with $1/f^2$ decay—the signature of natural images. Convolutions act as low-pass filters.
- **ViT/MLP**: Flat (white noise) spectrum—all frequencies equally represented. No architectural constraint suppresses high-frequency content.

This provides a physical interpretation of our bits metric: low-bit architectures (ConvNets) have "built-in bandpass filters" that suppress high-frequency noise, while high-bit architectures treat all spatial frequencies democratically, producing white noise at initialization.

F.3 Summary

The mechanistic analysis reveals:

1. **Attention scrambles**: Transformer attention with random weights mixes all spatial positions equally, failing to preserve local structure
2. **Conv stem insufficient**: Surrounding a Transformer with convolutions does not transfer bias *through* the network. However, a convolutional decoder applied *after* the transformer can impose structure independently (Section 5.15)
3. **Spectral explanation**: ConvNets enforce low-frequency dominance; ViTs/MLPs produce flat spectra

These findings explain why ViTs require massive pretraining to match ConvNet performance on limited data—they must *learn* the spatial priors that ConvNets receive *for free* from their architecture.

G Order Metric Robustness Analysis

A natural concern is whether our conclusions depend on the specific choice of order metric. We address this through two complementary analyses: (1) testing whether different *formulations* of structure metrics yield consistent rankings, and (2) testing whether *simple* metrics (without gates) can distinguish architectures.

G.1 Formulation Robustness

We test six alternative formulations of structure metrics:

1. **Original**: Compression × TV (multiplicative)
2. **Additive**: 0.5· Compression +0.5· TV
3. **TV-only**: Total variation smoothness only
4. **Compression-only**: zlib compression ratio only
5. **Low-Frequency**: DCT-based low-frequency energy fraction
6. **Gradient**: Inverse gradient magnitude (Sobel-based)

Result. All six formulations produce *identical* architecture rankings:

Architecture	Original	Additive	TV-only	Compress	LowFreq	Gradient
CPPN	0.94	0.98	1.00	0.97	1.00	1.00
ResNet	0.81	0.89	0.98	0.82	1.00	0.97
MLP	0.21	0.48	0.56	0.36	1.00	0.73
ViT	0.00	0.07	0.05	0.10	0.97	0.19
Uniform	0.00	0.00	0.00	0.00	0.90	0.03

Kendall's $\tau = 1.0$ for all pairwise metric comparisons. The ranking CPPN > ResNet > MLP > ViT > Uniform is stable across formulations.

G.2 Why Simple Metrics Fail: The Degenerate Solution Problem

A deeper question: why use multiplicative gates at all? Why not simply use compression ratio or autocorrelation? We tested eight metrics *without* gates to understand when simple metrics fail:

1. **order_multiplicative**: Our gated metric (baseline)
2. **png_compression**: Compression ratio only
3. **low_freq_power**: DCT low-frequency energy
4. **autocorrelation**: Spatial self-correlation
5. **entropy**: Shannon entropy of pixel values
6. **mutual_information**: MI between adjacent pixels
7. **fractal_dimension**: Box-counting dimension
8. **euler_characteristic**: Topological invariant

Methodology. For fair comparison across metrics with different scales, we use *percentile-based thresholds*: for each metric, we pool samples from all architectures, compute the 50th, 75th, and 90th percentiles, then measure what fraction of each architecture's samples exceed these thresholds.

Result. Only 2 of 8 metrics produce consistent rankings:

Table 25: Architecture rankings by metric (using percentile-based thresholds). Note: MLP ranks highest on compression, autocorrelation, entropy, and low-frequency power because untrained MLPs produce degenerate constant images that are perfectly compressible and perfectly autocorrelated—the opposite of ground truth.

Metric	Rankings at Percentiles P50	P75	P90	Consistent?
order_multiplicative	CPPN > Conv > Res > ViT > MLP	same	same	✓
fractal_dimension	Conv > Res > CPPN > ViT > MLP	same	same	✓
png_compression	**MLP** > Conv > Res > CPPN > ViT	varies	varies	×
autocorrelation	**MLP** > Conv > Res > CPPN > ViT	varies	varies	×
entropy	**MLP** > CPPN > Conv > Res > ViT	varies	varies	×
low_freq_power	**MLP** > Conv > Res > CPPN > ViT	varies	varies	×
euler_characteristic	**MLP** > Conv > Res > CPPN > ViT	varies	varies	×
mutual_information	ViT > CPPN > Conv > Res > MLP	varies	varies	×

Key Finding: MLP Wins on Simple Metrics. Five of eight metrics rank MLP *highest*. This counterintuitive result occurs because MLPs with random weights produce **nearly constant images**—saturated activations push outputs toward all-black or all-white. Such degenerate images are:

- Perfectly compressible (compression ratio ≈ 0.99)
- Perfectly autocorrelated (constant $\Rightarrow$ perfect self-similarity)
- Zero entropy (no variation to encode)
- Maximum low-frequency power (DC component dominates)

These are precisely the *degenerate solutions* that our gated metric is designed to exclude. The Color Balance gate ($4p(1-p)$) kills all-black/all-white images; the Connectivity gate requires non-trivial edge structure.

Why Fractal Dimension Also Works. The only simple metric that agrees with our gated metric is fractal dimension—but only when computed with a *bell curve* response around moderate complexity ($D \approx 1.5$). This penalizes both extremes:

- $D \approx 1.0$ (constant images): too simple
- $D \approx 2.0$ (white noise): too complex
- $D \approx 1.5$ (fractal structure): optimal

Both consistent metrics—our multiplicative order and fractal dimension with bell curve—share the key design principle: **explicitly penalizing both extremes** (trivial patterns and random noise).

Conclusion. Simple metrics without gates can be "gamed" by degenerate solutions. The multiplicative gate structure is not arbitrary—it is *necessary* to exclude architectures that produce trivial outputs. Our findings are robust because the gated metric captures genuine structural bias, not measurement artifacts.

G.3 Extended Validation: 19 Standard Metrics

To address reviewer concerns about metric arbitrariness, we conducted comprehensive validation across 19 metrics drawn from the image quality, texture analysis, and signal processing literatures (RES-337).

Metrics Tested.

1. **First-Round (6)**: TV, Gradient Kurtosis, GLCM Contrast/Homogeneity, LBP Entropy, NIQE
2. **Spectral (5)**: HF-ratio, Spectral Centroid, Spectral Slope (β), Phase Coherence, BRISQUE
3. **Original Appendix G (6)**: PNG Compression, Autocorrelation, Entropy, Low-Freq Power, Euler Characteristic, Mutual Information
4. **Topological (2)**: Betti_0, Betti_1 (persistent homology)

Architectures. CPPN, ConvNet (convolutional decoder), ResNet (residual decoder), MLP (fully-connected decoder), Uniform (baseline noise). 50 samples per architecture at 32×32 grayscale.

Results. Accounting for metric direction (e.g., lower HF-ratio = more structured):

Table 26: Extended metric validation: 15/19 metrics (79%) correctly rank structured architectures highest

Category	Count	Metrics	Winner
Works (structured wins)	15	GradKurt, GLCM_homog, NIQE, HF_ratio, Spectral_centroid, Spectral_slope, Phase_coherence, BRISQUE, PNG_compress, Autocorrelation, LowFreq_power, Mutual_info, Betti_0, Betti_1, Entropy	CPPN, ConvNet, or ResNet
Trap (MLP wins)	1	LBP_entropy	MLP
Orthogonal (edge/noise density)	3	TV, GLCM_contrast, Euler_char	Uniform

Key Finding: Spectral Slope Provides Independent Validation. The spectral slope (β) measures power-law decay in the Fourier spectrum. Natural images follow $P(f) \propto f^{-2}$ (the "$1/f^2$ law"). Results:

- CPPN: $\beta = -2.75$ (close to natural image statistics)
- ConvNet: $\beta = -0.05$ (weak low-frequency bias)
- MLP: $\beta = +0.06$ (flat/white noise spectrum)
- Uniform: $\beta = +0.03$ (white noise baseline)

Our gated metric correlates strongly with spectral slope ($r = -0.91$), confirming it captures the same underlying structure measured by this standard, non-heuristic signal processing metric.

Why Original "Failing" Metrics Now Work. The original analysis (Table 8 above) found compression/autocorrelation ranking MLP highest. In this extended experiment, *CPPN* wins on these metrics. The difference: CPPN produces genuinely smooth, spatially coherent images (due to coordinate-based input with

periodic activations), which are highly compressible and autocorrelated for the *right* reasons—not because they are degenerate constants.

Interpretation. The 79% agreement across 19 diverse metrics provides strong evidence that our conclusions about architectural inductive bias are robust to metric choice. The single "trap" metric (LBP entropy) and three "orthogonal" metrics (measuring edge/noise density rather than structure) do not undermine the core finding: structured architectures consistently outperform unstructured ones across the vast majority of reasonable structure metrics.

G.4 Learned Perceptual Metric Validation (LPIPS)

As a final validation, we test whether our hand-crafted order metric agrees with LPIPS [18], a learned perceptual metric trained on human similarity judgments. This addresses the concern that our metric might capture artifacts rather than genuine perceptual structure.

Setup. We sample 100 images from 7 architectures (ConvNeXt, ResNet, Swin, MLP-Mixer, ViT, MLP, Uniform) and compute mean LPIPS distance to 100 CIFAR-10 natural images. Lower LPIPS indicates perceptually closer to natural images.

Results.

Architecture	Order Score	LPIPS Distance
ConvNeXt	0.566	0.351
ResNet	0.441	0.317
Swin	0.297	0.805
MLP-Mixer	0.066	0.810
ViT	0.037	0.799
MLP	0.036	0.810
Uniform	0.000	0.958

Correlation: Pearson $r = -0.88$, $p = 0.02$. High-order architectures produce images perceptually similar to natural images (low LPIPS), while low-order architectures produce perceptually random outputs (high LPIPS). This strong negative correlation confirms that our order metric captures genuine perceptual naturalness, not circular artifacts of metric construction.

H Constrained Sampling Diagnostics

We validate that Elliptical Slice Sampling (ESS) correctly explores the constrained prior region. Key diagnostics:

Bracket Efficiency. Over 300 nested sampling iterations, ESS achieves 99.3% bracket efficiency (proposals satisfy the threshold on the first attempt without bracket shrinkage), with no decline as thresholds increase. This indicates effective exploration even at high constraint levels.

Bracket Shrinkage. The ESS bracket (the angular range searched) shrinks an average of only 0.3 times per step, with occasional peaks up to 38 shrinks at hard thresholds. This is within expected behavior for the algorithm.

Weight Correlation. Consecutive accepted samples show near-zero cosine similarity (mean $= -0.01$, std $= 0.22$), indicating independent sampling with no persistent correlation that could bias estimates.

Brute-Force Cross-Validation. At $\tau = 0.2$, brute-force sampling (5000 samples) estimates 0.04 bits; nested sampling estimates 0.06 bits ($\Delta = 0.02$ bits). Agreement degrades at higher thresholds where brute-force becomes statistically unreliable—precisely the regime where nested sampling's importance lies.

Conclusion. ESS correctly samples from the constrained prior. The high acceptance rate, uncorrelated samples, and brute-force agreement at tractable thresholds confirm algorithmic correctness.

I Two-Stage Sampling: Algorithmic Variants

Note on Speedup Measurements. Early experiments (RES-224) reported 92× speedup, but subsequent validation (RES-283, 100 CPPNs) revealed this was an artifact of methodological error. The correct

mean speedup is 3.74× ($\sigma = 2.43$, range 0.63–10.76) as reported in Section 5.13. This section documents exploratory algorithmic variants tested during development; the absolute speedup values in these early experiments are unreliable, but the *relative* comparisons remain informative.

I.1 Algorithmic Refinements (Exploratory)

Several algorithmic variants were tested to improve upon the basic two-stage approach:

Three-Stage Progressive Constraint (RES-229). Progressively tightening PCA constraints (5D → 3D → 2D) was hypothesized to improve efficiency. All variants underperformed the two-stage baseline, with the additional stage acting as a bottleneck rather than a refinement.

Adaptive Threshold Manifold Discovery (RES-230). Dynamically switching to PCA constraints when variance saturates was tested. All adaptive variants performed worse than fixed-budget baselines, with PCA variance saturating almost immediately during Stage 1 (56 samples). The overhead of dynamic switching exceeded any manifold refinement benefit.

Hybrid Multi-Manifold Sampling (RES-231). Maintaining a weighted mixture of 2D/3D/5D manifold hypotheses was tested. All weighting strategies (fixed, decay, adaptive) converged to identical behavior, revealing that mixture complexity adds overhead without improving selectivity.

I.2 Architectural Alternatives (Summary)

We tested several alternative coordinate representations (RES-232–234):

- **Dual-channel** $[x + y, x - y]$: No dimensionality reduction; slight performance degradation
- **Polar** $[r, \theta]$: Marginal symmetry improvement (1.19×), below practical significance
- **Hierarchical** $[x, x/2, x/4, ...]$: Increased dimensionality, 31% quality loss
- **Nonlinear** $[x \cdot y, x/y, x^2, y^2]$: Exceptional 2.48× order improvement (future work)

Conclusion: CPPNs' [x,y,r] coordinate structure represents a well-optimized design. Alternative coordinate systems do not improve upon it, though nonlinear compositions warrant future investigation. Full experimental details are available in the extended results document.

J Size Scaling and Threshold Dependence

To clarify the resolution-scaling contradiction discussed in Section 6.5, we report size-scaling exponents for CPPNs at two thresholds. The scaling exponent β comes from fitting $B(N) \sim N^\beta$ across image sizes using the multiplicative metric. At a conservative threshold $\tau = 0.1$, scaling is sub-linear; at a stricter threshold $\tau = 0.25$, scaling is super-linear. This threshold dependence motivates our explicit caution about threshold selection when interpreting scaling laws.

Table 27: CPPN size-scaling exponent β at two thresholds (nested sampling, multiplicative metric).

Threshold τ	Sizes	Seeds	β (95% CI)	R^2
0.10	8, 16, 32, 48	5	0.796 [0.762, 0.831]	0.883
0.25	8, 12, 16, 24, 32, 48	8	1.450 [1.404, 1.496]	0.907

K Monte Carlo Volume Validation: Full Results

Section 5.4 presents summary results from our large-scale Monte Carlo validation. Here we provide complete methodological details and extended analysis.

K.1 Methodology

Architectures Tested. We test 11 architectures organized into five groups:

Table 28: Architecture families and their structural priors

Group	Architectures	Input Type	Expected Bias
A: Coordinate-Conditioned	CPPN, CoordMLP, FourierMLP	$(x, y) \to$ pixel	High
B: Latent-Decoded	MLPDecoder, ConvDecoder, ViTDecoder	$z \to$ image	Variable
C: Structured Prior	FourierBasis	Spectral basis	High
D: Procedural	WalkCarver, SpanningTreeMaze	Algorithmic	High
E: Sequential	LSTMDecoder, MixtureOfExperts	Recurrent/sparse	Low

Sampling Protocol.

1. For each architecture, sample 10,000 independent weight configurations from $\mathcal{N}(0, 1)$
2. Generate 64×64 grayscale image from each configuration
3. Binarize at threshold 0.5
4. Compute multiplicative order metric (Appendix A)
5. Record order value for volume calculation

Order Metric. We use the multiplicative order metric defined in Appendix A: $O = O_{\text{compress}} \times O_{\text{symmetry}} \times O_{\text{connectivity}} \times O_{\text{balance}}$. The multiplicative formulation ensures that failure on *any* component collapses the score to near-zero, providing sharp discrimination between structured and unstructured outputs.

K.2 Extended Results

Threshold Robustness Validation. A potential concern is that our findings depend on the specific choice of order threshold $\tau = 0.1$. We validate robustness by computing thermodynamic volume across a $10\times$ range of thresholds from $\tau = 0.05$ (permissive) to $\tau = 0.5$ (stringent). Table 29 shows that architecture rankings remain consistent: structured priors (FourierBasis, CPPN, ConvDecoder) maintain separation from unstructured priors (MLP, ViT, LSTM) at all thresholds. Even at the extreme $\tau = 0.5$, FourierBasis achieves 87.2% while MLP/ViT remain at 0.0%—confirming our conclusions are not artifacts of threshold choice.

Table 29: Thermodynamic volume at multiple order thresholds (10,000 samples per architecture). Rankings remain stable across a $10\times$ threshold range.

Architecture	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.5$
WalkCarver	100.0%	100.0%	99.8%	62.3%	1.2%
FourierBasis	100.0%	99.9%	99.3%	97.8%	87.2%
FourierMLP	94.2%	83.6%	67.1%	51.4%	26.8%
CPPN	82.3%	69.0%	48.7%	32.1%	12.4%
ConvDecoder	68.4%	51.2%	28.3%	12.1%	1.8%
MixtureOfExperts	31.2%	20.1%	8.4%	2.3%	0.1%
CoordMLP	12.8%	7.2%	2.1%	0.4%	0.0%
MLPDecoder	0.0%	0.0%	0.0%	0.0%	0.0%
ViTDecoder	0.0%	0.0%	0.0%	0.0%	0.0%
SpanningTreeMaze	0.0%	0.0%	0.0%	0.0%	0.0%
LSTMDecoder	0.0%	0.0%	0.0%	0.0%	0.0%

K.3 Initialization Ablation on Coord Inputs

To test initialization sensitivity, we ran a targeted ablation on coord-input architectures (CoordConvNet, CoordMLP, CoordViT) at 32×32 with $\tau = 0.1$ (RES-370). We compare three initialization schemes: fixed Gaussian $\mathcal{N}(0, 0.5)$ in this ablation, Xavier, and He. ConvNet and MLP pass rates shift modestly across schemes (ConvNet: 0.665–0.725; MLP: 0.385–0.435). CoordViT is markedly more sensitive: pass rate rises from 0.335 (Gaussian) to 0.855 (Xavier) and 0.725 (He). This sensitivity motivates treating initialization as part of the prior and holding it fixed for all main comparisons.

K.4 Weight Sharing Isolation (RES-295)

To isolate translational weight sharing from mere locality, RES-295 compares four local architectures at 32×32 with 1000 MC samples per architecture. The key controlled comparison is Conv3x3 versus Locally-Connected: both use sliding 3×3 receptive fields, but only Conv3x3 shares weights across spatial positions.

Table 30: RES-295 weight-sharing isolation results (`results/weight_sharing/res_295_weight_sharing_isolation.json`).

Architecture	Params	Mean Order	Inductive Property
Conv3x3	25,553	0.5330465	Locality + sharing
LocallyConnected	1,426,432	0.0000241	Locality, no sharing
DepthwiseConv3x3	4,145	0.4014567	Locality + partial sharing
LocalAttention3x3	25,729	0.1421627	Locality, no sharing

Primary test (Conv3x3 vs LocallyConnected): Mann-Whitney $U = 10^6$ (maximum separation), p-value underflow in double precision (reported as $p < 10^{-300}$), Cohen's $d = 2.36$, mean difference 0.5330. This supports translational invariance through weight sharing as a dominant source of thermodynamic volume.

Order Distribution Statistics.

Table 31: Full order distribution statistics (10,000 samples per architecture)

Architecture	Mean	Std	Median	Q25	Q75	Max
FourierBasis	0.708	0.195	0.762	0.612	0.854	0.987
FourierMLP	0.480	0.311	0.521	0.198	0.764	0.956
WalkCarver	0.372	0.027	0.374	0.354	0.391	0.468
CPPN	0.347	0.304	0.312	0.048	0.621	0.943
ConvDecoder	0.152	0.166	0.098	0.021	0.234	0.812
MixtureOfExperts	0.061	0.054	0.047	0.018	0.089	0.324
CoordMLP	0.034	0.042	0.018	0.004	0.048	0.287
MLPDecoder	$\sim 10^{-24}$	—	—	—	—	10^{-18}
ViTDecoder	$\sim 10^{-23}$	—	—	—	—	10^{-17}
SpanningTreeMaze	$\sim 10^{-5}$	—	—	—	—	10^{-3}
LSTMDecoder	$\sim 10^{-4}$	—	—	—	—	10^{-2}

K.5 Key Observations

1. Extreme Separation. The gap between Shielded and Broken regimes spans 20+ orders of magnitude in mean order. This is not a subtle effect—it represents fundamentally different thermodynamic properties of architecture families.

2. WalkCarver Dominance. The WalkCarver (random walk maze carving) achieves 100% volume with remarkably low variance (std=0.027). This procedural algorithm is *guaranteed* to produce structured output regardless of random seed, unlike neural architectures which have non-trivial failure modes.

3. Fourier vs Periodic Activations. FourierBasis (0.708 mean order) outperforms CPPN (0.347) despite both using periodic functions. The key difference: FourierBasis uses *fixed* sinusoidal basis functions

with only amplitude weights, while CPPN uses *random* activation patterns. This suggests that the regularity of Fourier decomposition provides stronger structural guarantees than stochastic activation selection.

4. ViT = MLP. Vision Transformers with random weights produce order below MC detection (0/1,000,000 successes; $P < 10^{-6}$), indistinguishable from MLPs. The $\sim 10^{-23}$ and $\sim 10^{-24}$ values in the table above reflect *mean order metric values* from nested sampling, not probabilities. The attention mechanism provides no structural bias at initialization—positional embeddings and self-attention become meaningful only after training.

5. Bimodal Distributions. CPPN and FourierMLP show bimodal order distributions with substantial mass near zero and near 0.6-0.8. This reflects the "lottery" nature of random initialization: some configurations produce structured outputs, while others collapse to noise. ConvDecoder shows a unimodal distribution concentrated near zero, indicating weaker but more consistent bias.

K.6 Validation of Nested Sampling Estimates

The Monte Carlo volume directly measures $P(\text{order} \geq \tau)$, which nested sampling estimates as $2^{-B(\tau)}$ where $B(\tau)$ is the bits required to reach threshold τ.

Consistency Check. For CPPN at $\tau = 0.1$:

- Monte Carlo volume: $0.69 \Rightarrow B_{\text{MC}} = -\log_2(0.69) = 0.54$ bits
- Nested sampling (Table 6, 32×32): $B_{\text{NS}} \approx 1.9$ bits

The 3.5× discrepancy is explained by:

1. **Resolution difference**: 64×64 (Monte Carlo) vs 32×32 (nested sampling). Higher resolution provides more pixels for pattern expression.
2. **Threshold calibration**: The order metric's gate parameters were calibrated at 32×32; at 64×64 with scale normalization, thresholds shift.
3. **Algorithmic bias**: Nested sampling's volume shrinkage factor has systematic upward bias (Appendix E).

Critically, the architecture *ranking* is identical: both methods produce Shielded > Transitional > Broken ordering, validating the qualitative conclusions.

K.7 Implications for Two-Stage Sampling

The Monte Carlo volume measurements provide context for interpreting the observed 3.74× mean speedup from two-stage sampling (RES-283):

1. **Volume confirms manifold existence:** The 69% volume for CPPN at $\tau = 0.1$ confirms that high-order solutions are not rare in CPPN weight space—explaining why manifold-aware approaches can provide speedup despite high variance.
2. **Gap to theoretical ceiling:** The theoretical ceiling ($\sim 60\times$) vs observed mean (3.74×) indicates substantial room for algorithmic improvement. The high variance (range 0.63–10.76×) suggests that some CPPNs achieve near-theoretical speedup while others face bottlenecks.
3. **Architectural implications:** The compositional [x,y,r] input structure creates the manifold that two-stage sampling exploits. Architectures with 0% volume (MLP, ViT) cannot benefit from manifold-aware sampling because no manifold exists to discover.