TOWARD A MECHANISTIC UNDERSTANDING OF INFERENCE IN VISUAL CORTEX AND DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We describe a model of perceptual inference in primary visual cortex (V1) equivalent to a minimal diffusion model whose function can be readily understood from its parameters. We use a recurrent neural network to define the denoising network and train it with implicit differentiation. The recurrent dynamics are derived from generalizing classical sparse coding's inference dynamics by adding learnable pairwise interactions among latent units. After training on natural images, the model exhibits exceptionally good denoising performance, with the ability to restore image features such as extended contours amid extreme visual ambiguity, nearly matching the behavior of standard, black-box diffusion architectures. We then mechanistically analyze how the trained model resolves ambiguity. The structure of the learned pairwise interaction matrix links neurons of similar orientation tuning, mirroring the structure of horizontal connections in superficial layers of V1. Furthermore, analysis of the network's Jacobian reveals a mechanistic explanation for how the model assigns high probability over a continuous family of natural structural deformations, including those that allow the completion of contours. Intriguingly, a large fraction of latent variables learn to disconnect from visual input altogether yet remain connected to other latent units, essentially forming a hierarchical representation that appears to enforce global consistency among image features. Together, our results bridge two distinct domains: for neuroscience, they generate concrete, testable hypotheses regarding functional connectivity in V1 recurrent neural circuits during perceptual inference tasks; for machine learning, they elucidate the internal mechanisms learned by diffusion models that allow them to generate infinitely many novel images from a finite training set. Code is available at this anonymous repo.

1 INTRODUCTION

When viewing Figure 1A (left), human observers immediately and effortlessly perceive the contour embedded at the center, while the remaining Gabor elements, identical but randomly oriented, appear structureless (Field et al., 1993). Interestingly, a diffusion model trained on natural images perceives these gabor images similarly, as revealed by its output (Figure 1A, right). In particular, the principal eigenvectors of the model's Jacobian reveal the perturbations that induce high-probability variants of the input in image space (Figure 1B): in this case, spreading the discrete Gabor elements into a continuous contour. The question we address here is, why is this happening – both in visual cortex and diffusion models – and what specific neural computational mechanisms are needed to capture this in a model of perception?

While interrogation of a diffusion model reveals its probability gradient for specific images (Mohan et al., 2019), such analysis is only local and anecdotal, and we still lack an explanation for why this arises from the specific mechanisms within the neural network of the model. On the other hand, by studying the visual cortex of the brain, we can identify specific neuroanatomical structures that suggest how such contour completion occurs – namely, the network of horizontal connections in superficial layers that spread activation among neurons with spatially separate but similarly oriented (and in some cases co-linearly aligned) receptive fields (Hirsch & Gilbert, 1991; Fitzpatrick, 1996; Field et al., 1993; Garrigues & Olshausen, 2007; Lee & Mumford, 2003). It has long been theorized that this connectivity could underlie perceptual grouping phenomena such as contour integration (Hirsch & Gilbert, 1991; Field et al., 1993), whereby geometrically aligned features are perceived as belonging

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

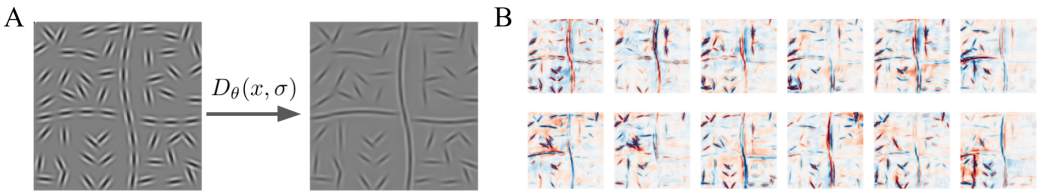


Figure 1: **Diffusion model perceives hidden contour. (A)** Left: visual stimulus where a hidden contour is embedded among randomly orientated gabor functions. Right: output of denoiser when fed the visual stimulus. **(B)** Visualization of the principal eigenvectors of the Jacobian $J(x)$ of the denoiser, evaluated at the hidden contour image and ranked by eigenvalue magnitude.

to a common object. This is also consistent with the evidence that neural dynamics perform sampling-based probabilistic inference (Orbán et al., 2016; Echeveste et al., 2020). Although numerous models have been proposed (Parent & Zucker, 1989; Grossberg et al., 1997; Li, 1998; George et al., 2025; Toosi & Miller, 2025; George et al., 2017), it has yet to be shown how this connectivity structure could emerge from a probabilistic, generative model trained to perform inference on natural images.

Here we approach this problem by building on the *sparse coding model* of V1, which has been proposed to account for its oriented, simple-cell receptive fields in terms of a linear generative model of natural images with a sparse, factorial prior over latent variables (Olshausen & Field, 1996; 1997; Field, 1987; 1994; Olshausen & Lewicki, 2014). Inspired by horizontal connections in visual cortex, we add to the sparse coding model an interaction matrix between latent variables with the goal of learning the co-occurrence statistics of features at different spatial locations, orientations and spatial scales (Simoncelli, 1999; Portilla & Simoncelli, 2000; Hyvärinen et al., 2009; Lee et al., 2003; Carlsson et al., 2008; Chen et al., 2018; 2022a). Previous theoretical and modeling work has shown that co-circularity of oriented elements constitutes an important aspect of natural scene statistics (Parent & Zucker, 1989; Geisler et al., 2001; Sigman et al., 2001). Here we reason that by training the model on natural images, it should be possible to incorporate these statistics into the prior, which should in turn improve the ability to resolve ambiguity in images.

In what follows, we first provide relevant background and then formally specify the proposed model. We then show that the learned interaction matrix weights are aligned with co-linear facilitation, and that the model's performance on denoising tasks is far superior to standard sparse coding and competitive with deep diffusion models. Local linear analysis of the model's behavior via its Jacobian reveals how activity spreads among latent variables in a context-dependent manner, leading to the completion of contours amidst ambiguity. Crucially, this recurrent spreading mechanism also explains how diffusion models assign high probability to realistic structural deformations, illuminating the internal mechanism that allows them to sample novel images.

2 Background and Preliminaries

Sparse coding model of primary visual cortex. The sparse coding model posits that neural activity represents latent variables, z, that encode an image, x, by descending an energy function $E(x, z) = \frac{1}{2\sigma^2}\|x - \Phi z\|_2^2 + \lambda||z||_1$. This energy function balances stimulus reconstruction error with neural sparsity. From a Bayesian perspective, the first and the second terms correspond to the likelihood and the prior, respectively. When trained on natural images to maximize the log-likelihood of the model, $p_\theta(x) \propto \int e^{-E(x,z)}dz$, the dictionary ($\Phi$) converges to a set of Gabor-like functions similar to V1 receptive fields (Olshausen & Field, 1997). The underlying assumption of statistical independence in the prior (corresponding to $||z||_1$) is a severe limitation, as natural image features are not statistically independent: oriented features exhibit strong dependencies across position, scale, and orientation (Hyvärinen & Hoyer, 2000; Portilla & Simoncelli, 2000; Schwartz & Simoncelli, 2001; Hyvärinen et al., 2001; Karklin & Lewicki, 2003; 2005). These dependencies are central to perceptual phenomena such as contour integration and completion, where spatially separated but geometrically aligned elements are perceived as grouped together. Several extensions have been proposed to model non-factorial structure in the latent variables. Hierarchical sparse coding models introduce additional latent variables that capture dependencies among lower-level features (Karklin & Lewicki, 2005;

Yu et al., 2011; Hosoya & Hyvärinen, 2015; 2016; Boutin et al., 2020). Other single-layer models impose structured priors, such as topographic or group sparsity, to encourage related features to co-activate (Hyvärinen & Hoyer, 2000; Paiton et al., 2020). These approaches capture some non-factorial structure, but often rely on rigid, hand-specified priors that cannot be learned. More flexible non-factorial models can learn dependencies between latent variables (Garrigues & Olshausen, 2007; 2010; Olshausen et al., 2007; Osindero et al., 2006), but their inference and learning procedures are typically more difficult and may require sampling over high-dimensional latent variables.

Training and sampling of denoising diffusion models. In recent years, Denoising Diffusion Models (DDMs) have emerged as an effective way to learn densities of natural images from finite datasets (Vincent et al., 2008; Sohl-Dickstein et al., 2015; Saremi & Hyvärinen, 2019; Ho et al., 2020; Song et al., 2021). At their core, these models parametrize the score function—the vector field pointing towards regions of higher data probability. Tweedie's formula implies that the true score can be approximated by a deep neural network trained with an MSE denoising objective (Robbins, 1956; Miyasawa, 1961; Hyvärinen, 2005; Efron, 2011; Raphan & Simoncelli, 2011; Vincent, 2011). While empirically powerful, the score function is typically parameterized by a black-box neural network, making it difficult to understand the parameterized image prior. In this work, we parametrize the denoiser with simple recurrent dynamics. The explicit nature of the model allows for direct, mechanistic analysis of its neural circuit, shedding light on the features learned by diffusion models.

Understanding feature learned in diffusion models in generalization regime With a sufficiently expressive backbone and enough data, diffusion models learn to model the underlying distribution from data (instead of memorizing the training data) (Kadkhodaie et al., 2023; Kamb & Ganguli, 2024; Niedoba et al., 2024; Liang et al., 2024). In this generalization regime, a natural question is *what features did the model extract from data during training?* One approach to answering this question is local linear analysis through eigen-decomposition of the Jacobian of the denoiser (Mohan et al., 2019). By Tweedie's formula, this Jacobian is analytically linked to the Hessian of the learned log-likelihood ($J(x) = I + \sigma^2 \nabla_x^2 \log p_\sigma(x)$), meaning its dominant eigenvectors point directly toward the high-probability states defined by the model. Empirically, these directions correspond to naturalistic structural deformations—specifically, oscillatory bases tied to image contours—which constitute the exact features that allow the model to generalize. While these analyses have revealed that the model learns these features, they fall short of offering a mechanistic explanation of how these features arise from the activities of individual neurons in the network. The simplicity of our model provides the opportunity to expand such Jacobian analyses to shed light on the mechanisms that enable generalization at a neural circuit level.

Iterative inference hypothesis. Modern neural networks are based on deep feedforward architectures, which are extremely difficult to understand (LeCun et al., 2015; He et al., 2016; Zeiler & Fergus, 2014; Zeiler et al., 2010). One theory is that each layer does not learn a distinct transformation but rather a step of recursive refinement. This is supported by evidence such as the diminishing norm and early exiting residual stream (Jastrzębski et al., 2017; Moon et al., 2024; Gudovskiy et al., 2026). This suggests a much simpler architecture that could replace the feedforward architecture: instead of having 100 layers, we could have 10 layers repeated 10 times. Such recurrent architectures can match the performance of deep feedforward networks (Bai & Melas-Kyriazi, 2024; Wang et al., 2025; Zhu et al., 2025). While prior work emphasizes capability and efficiency, we see this as an opportunity for mechanistic understanding: it reduces the number of distinct layers to inspect, while the recurrent connections make the architecture more similar to biological neural circuits.

3 Non-factorial sparse coding as a minimal diffusion model

3.1 Training non-factorial sparse coding model via denoising score matching

We construct a minimal diffusion model by defining its denoiser through fixed points of a simple recurrent dynamical system:

$$D_\theta(x, \sigma) = \Phi z^*(x, \sigma; \theta), \qquad \text{where} \quad z^* = T_\theta(z^*; x, \sigma). \tag{1}$$

Let y denote the clean image and $x = y + \sigma\epsilon$ denote the Gaussian-corrupted noisy image with noise level σ. The fixed point $z^*(x, \sigma; \theta)$ is an implicit function of the noisy input x, the noise level σ, and the model parameters θ. We then optimize the implicit function's model parameters θ such that the

linear decoding of the fixed point $\Phi z^*(x, \sigma; \theta)$ predicts the clean image, i.e.

$$\mathcal{L}_{\text{DSM}}(\theta) = \mathbb{E}_{\sigma,y,\epsilon}\left[w(\sigma)\left\|\Phi z^*(x;\sigma,\theta) - y\right\|_2^2\right], \quad \text{where} \quad x = y + \sigma\epsilon \tag{2}$$

By Tweedie's formula, the trained denoiser yields the score estimate $s_\theta(x, \sigma) := [\Phi z^*(x, \sigma; \theta) - x]/\sigma^2 \approx \nabla_x \log p_\sigma(x)$, which guides the reverse diffusion process to generate samples. The recurrent updates are derived by minimizing the energy function of a non-factorial sparse coding model:

$$E_{\theta,\sigma}(x, z) = \frac{1}{2\gamma_\psi(\sigma)}\|x - \Phi z\|_2^2 + \|\lambda \circ z\|_1 + \frac{1}{2} z^\top M z. \tag{3}$$

where the first two terms correspond to the energy function of the classical sparse coding model. The last term models higher-order interactions between latent units, resolving a severe limitation of classical sparse coding: the assumption that natural image features are statistically independent. We define the recurrent update as the iterative shrinkage-thresholding algorithm (ISTA) (Beck & Teboulle, 2009; Rozell et al., 2008) that corresponds to this energy function:

$$z_{t+1} = T_\theta(z_t; x, \sigma) = \text{ReLU}\left(z_t + \eta\left[\underbrace{\Phi^\top x}_{\text{Feedforward Drive}} - \underbrace{\left(\Phi^\top\Phi + \gamma_\psi(\sigma)M\right) z_t}_{\text{Recurrent Drive}} - \underbrace{\gamma_\psi(\sigma)\lambda}_{\text{Sparsity Bias}}\right]\right) \tag{4}$$

where $\eta > 0$ is the step size and $\gamma_\psi(\sigma)$ controls the strength of the prior as a function of noise level. These dynamics can be separated into three terms: A feedforward input drive that projects the noisy input x via the decoding dictionary Φ. The recurrent drive with two recurrent weight matrices: a Gram matrix $\Phi^\top\Phi$ that encourages reconstruction, and a noise-weighted $\gamma(\sigma)M$ that favors or suppresses co-active latent variables according to the non-factorial prior. Lastly, ReLU nonlinearity and bias term $\gamma(\sigma)\lambda$ that encourages sparsity. Note that in order for the ISTA dynamics to constantly lower the energy function in Equation 3, we need to constrain M to be symmetric and positive semi-definite. However, we did not enforce these constraints, to further improve the expressivity of the model.

3.2 Differentiation through the dynamics with implicit function theorem

Both training the denoising objective and analyzing the model via the Jacobian $J(x)$ require differentiating through the convergent state $z^*(x; \sigma, \theta)$. Rather than unrolling the entire iterative trajectory of Equation 4 via Backpropagation Through Time (BPTT), which incurs prohibitive memory and computational costs, we exploit the fact that z^* converges to a stable fixed point (verified in Figure S1). By defining the fixed-point residual $F(z^*; x) = z^* - T(z^*; x) = 0$, the Implicit Function Theorem (IFT) allows us to compute the exact analytical gradients in a single, closed-form step:

$$\frac{\partial z^*}{\partial \theta} = -\left[\nabla_z F\right]^{-1} \nabla_\theta F, \qquad \frac{\partial z^*}{\partial x} = -\left[\nabla_z F\right]^{-1} \nabla_x F \tag{5}$$

The first gradient $\frac{\partial z^*}{\partial \theta}$ enables efficient parameter updates during training via the chain rule. The second gradient $\frac{\partial z^*}{\partial x}$ is the link for our subsequent analysis: because our explicit model formulates the denoised output as a linear synthesis of the latent state ($f_\theta(x) = \Phi z^*$), applying the chain rule directly yields the exact analytical denoising Jacobian $J(x) = \Phi \frac{\partial z^*}{\partial x}$ (see A.2).

The approach of training recurrent neural networks that converge to fixed points using the IFT was originally described by Pineda (1987) and Almeida (1990). Recent work has applied this method to deep learning architectures (Bai et al., 2019; El Ghaoui et al., 2021) to achieve impressive results. We use recent improvements to this algorithm to efficiently estimate the gradient (Geng et al., 2021; Fung et al., 2022) without computing the inverse of $\nabla_z F$ (details in A.4).

3.3 Mechanistically Decomposing the Jacobian

Due to the simplicity of the model, we can further open the black box by decomposing the denoising Jacobian $J(x)$. By defining the effective recurrent weight matrix as $W = I - \eta(\Phi^\top\Phi + \gamma(\sigma)M)$, we can expand the implicit gradient in Equation (5) to decompose the denoising Jacobian in the following form:

$$J(x) = \frac{dx_{\text{clean}}}{dx_{\text{noise}}} = \eta\Phi\underbrace{\left[I - \Sigma W\right]^{-1}\Sigma}_{J_z}\Phi^\top = \eta\Phi J_z \Phi^\top \tag{6}$$

where $\Sigma = \text{diag}(1_{z^*>0})$ acts as a binary gating matrix representing the currently active sparse code. This formulation explicitly reveals a structural change of basis: the pixel-space Jacobian $J(x)$ is the Jacobian in latent-space J_z projected through the dictionary Φ. We can further expand the matrix inverse into an infinite Neumann series ($[I - A]^{-1} = \sum_{k=0}^{\infty} A^k$), as $J_z = \eta \sum_{k=0}^{\infty} (\Sigma W)^k \Sigma$.

To understand J_z, we examine a single column $J_{z,i} = J_z e_i$, which represents the network's impulse response to perturbing the i-th latent variable. Because the binary diagonal gating matrix Σ only passes active neurons ($\Sigma e_i = e_i$ if $z_i^* > 0$, and 0 otherwise), any perturbation to an inactive neuron is immediately silenced. For an active neuron, we can use the Neumann series expansion to decompose its impulse response in the following highly interpretable manner:

$$J_{z,i} = \underbrace{e_i}_{\text{Sparse gating}} + \underbrace{\Sigma W e_i}_{\text{1st-order spread}} + \underbrace{\Sigma W \Sigma W e_i}_{\text{2nd-order spread}} + \ldots \quad (\text{for } i \text{ s.t. } z_i^* > 0, \text{ else } 0) \tag{7}$$

Together with the change-of-basis in Equation 6, this formulation provides a transparent, mechanistic explanation for how the network processes incoming perturbations. First, a pixel-space perturbation is projected to and from the latent space by the dictionary Φ. The Σ matrix then acts as a mask that silences perturbations on inactive neurons, ensuring only contextually relevant signals survive. Next, during **context-aware spread** ($\Sigma W \Sigma$), these surviving signals propagate via lateral connections W. The outer Σ strictly clamps this activity spreading, ensuring cross-talk between neurons only leaks into currently activated neurons. Furthermore, the infinite tail of the sum $(\Sigma W)^k$ represents higher-order, continuous spreading of neural activity within the context.

4 Results

4.1 Emergence of contour completion and long-range horizontal connections

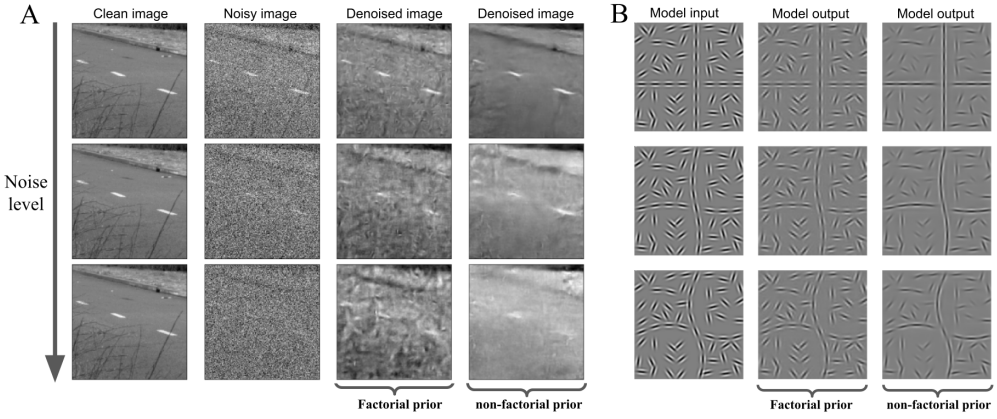


Figure 2: **Emergence of contour completion and denoising performance.** **(A)** Denoising comparison across increasing noise levels (rows) showing clean, noisy, factorial, and non-factorial sparse coding (SC) reconstructions. **(B)** Model's response to a contour-completion illusion, showing the illusion image and the factorial and non-factorial SC reconstructions.

When trained on images of natural scenes (Van Hateren & van der Schaaf, 1998), the model learns a dictionary Φ of localized, oriented Gabor-like basis functions, mirroring V1 simple cell receptive fields (Figure 3A). The inference dynamics (Equation 4) decompose the lateral interactions into two mechanisms: the Gram matrix ($\Phi^\top \Phi$), and the learned non-factorial structural prior (M). As visualized in Figure 3B, M learns *collinear facilitation*. It develops strong excitatory connections between spatially offset neurons whose preferred orientations broadly co-align along a shared axis, alongside strong self-inhibition to enforce sparsity. This emergent collinear connectivity parallels the 'association field' theory proposed by Field et al. (1993). Computationally, M acts as the mechanism to bind local edge segments into extended, continuous contours. By contrast, the Gram matrix exclusively mediates more localized interactions between neurons with overlapping receptive fields (quantified in Figure S4). To understand the practical gains of learning a non-factorial prior relative

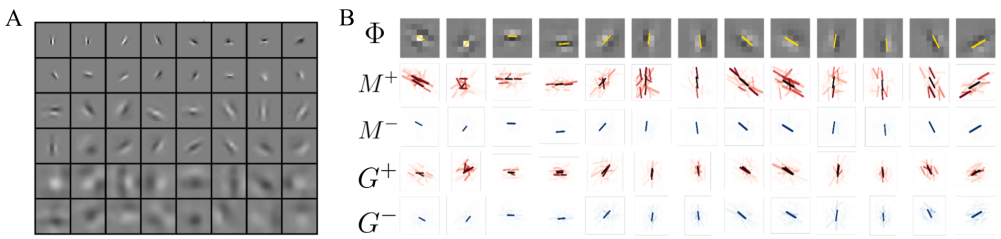


Figure 3: **Emergence of V1-like receptive fields and connectivity. (A)** Visualization of learned dictionary elements in Φ (for each scale, only 16 are shown). **(B)** Visualization of learned latent interactions for a subset of dictionary elements. Each column is centered on one reference basis function Φ_i shown in the top row. The yellow/black bar indicates its position and preferred orientation. Rows below show needle plots of its interactions with other basis functions Φ_j: excitatory and inhibitory components of the learned pairwise matrix M_{ij}, followed by excitatory and inhibitory components of the Gram matrix $G_{ij} = \Phi_i^\top \Phi_j$. Each colored bar is placed at the spatial location and orientation of Φ_j, and color saturation indicates interaction magnitude.

to a factorial model, we trained another model which has fixed $M_{ij} = 0 \; \forall ij$, but is otherwise identical. As shown in Figure 2.A, we find that the non-factorial model achieves significantly better denoising than its factorial counterpart (quantified in Figure S5). Due to the factorial prior, the model hallucinates dense, spurious edges in noisy, unstructured regions. In contrast, the non-factorial prior leverages M to explicitly enforce the continuity of true structural edges while actively suppressing uncoordinated background noise.

4.2 Mechanistic decomposition of implicit prior in diffusion model

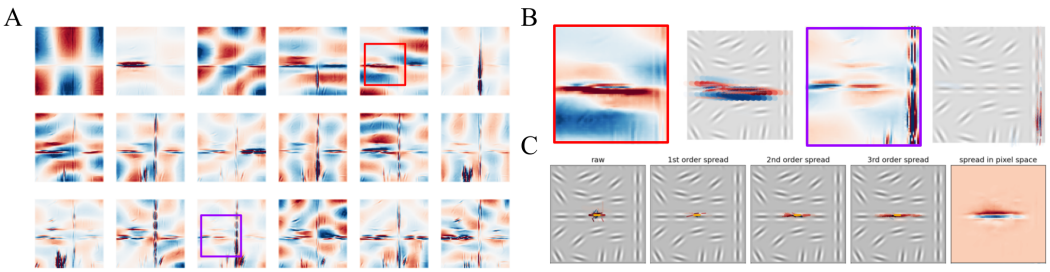


Figure 4: **Mechanistic decomposition of the denoising Jacobian.** (A) Dominant eigenvectors of the pixel-space Jacobian $J(x)$ for the non-factorial sparse coding model ($\sigma = 0.17$), x is the illusion image in the first row of Fig. 2.B. (B) Paired pixel-space and latent-space eigenvectors from the highlighted regions. Latent eigenvectors are visualized using the same needle plots as in Fig.3.B. (C) Panel 1: the connections between a neuron and nearby neurons. Panels 2-4: visualizing a single column of the latent Jacobian (Equation 7) at different orders of approximation (red:excitatory; blue:inhibitory). Panel 5: single column of corresponding pixel Jacobian (Additional visualizations in Figures S6 and S7.)

Beyond the connectivity pattern, we identify the exact neural circuit that allows the model to process visual ambiguity. First, eigen-analysis of the denoising Jacobian reveals that the model prefers a small number of directions that span the subspace of naturalistic deformations, similar to those of U-Nets. Many of these deformations appear to stitch two aligned but separated edges in the visual illusion. However, exactly how the neural network constructs these highly structured eigenvectors remains a black box. Thanks to the simplicity of our model, we can mechanistically explain how exactly these bases are constructed. To uncover this mechanism, Equation 6 allows us to analyze eigenvectors

directly within the latent space via a change of basis.[1] Visualizing the latent Jacobian J_z reveals that the dominant eigenvectors are concentrated on neurons with preferred orientations and locations that strictly align with the contour. To further understand J_z, we visualize the decomposition of the i-th column of J_z, which is equivalent to perturbing the i-th neuron at the fixed point of the dynamics. This can be further decomposed into how the recurrent dynamics propagate the perturbation (Equation 7), as iterative recurrent spreading of activity, explicitly gated by the currently active neurons (Σ).

As shown in Figure 4C, if an active neuron spreads its signal based on raw connectivity (We_i), the activation exhibits a collinear bias but still suffers from diffuse off-axis leakage. However, by applying this learned collinear spread and sparse gating (ΣWe_i) in an alternating order, the signal is confined to the set of active neurons. Since every column of the latent Jacobian looks like a spread of activity along the active sparse neurons, their macroscopic integration inherently produces eigenvectors that are strictly tied to this geometry. Crucially, this spatially bounded recurrence naturally gives rise to the emergent oscillatory bases observed in Figure 4A. By interpolating these bases, the model can locally deform a contour, such as shifting, bending or stitching broken parts of a contour. Since these eigenvectors span the high-probability directions of the learned prior, any linear interpolation maps to another highly likely state, allowing the model to assign high probability to an infinite, continuous family of structurally valid deformations. Since the training data is finite, some of these deformations lead to valid configurations that never appear in the training set, which is the fundamental mechanism for generalization.

4.3 Beyond contour completion: Generalization and Emergent Hierarchy

An excellent example of such deformation can be seen in the face dataset (CelebA): the eigenvectors of the denoising Jacobian yield a set of perturbation directions that seamlessly transition the input into a different, yet structurally coherent, face (Figure 5A,B). What is even more intriguing is that when trained on highly structured datasets like the face dataset, both standard U-Nets and our minimal model exhibit semantic-level generalization. While the mechanisms detailed in Section 4.2 explain how the recurrent neural circuit constructs the local oscillatory bases responsible for these naturalistic deformations, they do not inherently explain global coherence. Instead of independent, disconnected local deformations, the learned eigenvectors manifest as globally coordinated deformations—such as synchronously shifting both eyes—ensuring all modifications remain structurally consistent with the high-level concept of a "face" (Figure 5A). This raises the question: how does a single-layer circuit coordinate these local deformations across the entire image to enforce global constraints?

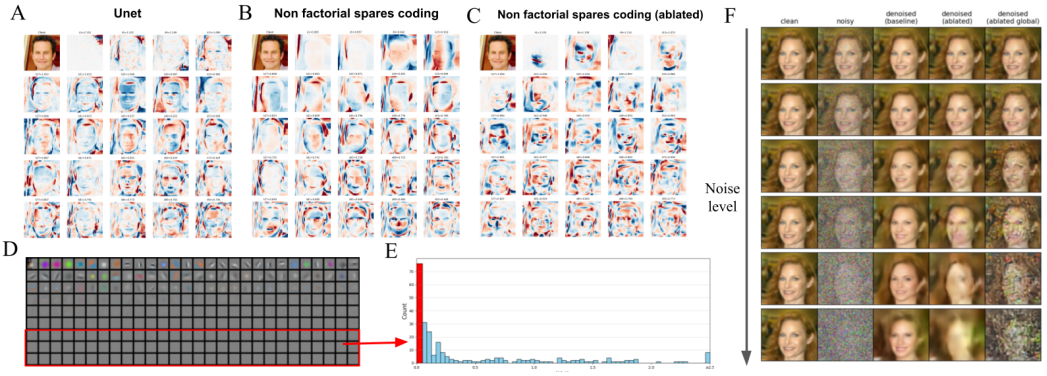


Figure 5: **Semantic generalization and the emergence of detached neurons.** (A-B) Dominant eigenvectors of the pixel-space Jacobian $J(x)$ for a standard U-Net (A) and our non-factorial sparse coding model (B), showing coordinated semantic deformations. (C) Dominant eigenvectors with hidden neurons ablated. (D) Learned dictionary elements (Φ) sorted by norm magnitude. (E) Dictionary norm histogram revealing a structural bifurcation: $\sim$30% of the population learns near-zero weights, forming functionally decoupled "detached neurons." (F) Qualitative ablation results across increasing noise levels.

[1] Projecting pixel eigenvectors via $\Phi^\top$ bypasses the prohibitive cost of computing the full latent Jacobian. This proxy mathematically assumes Φ forms a tight frame ($\Phi^\top \Phi \propto I$) and serves as an effective approximation for visualization.

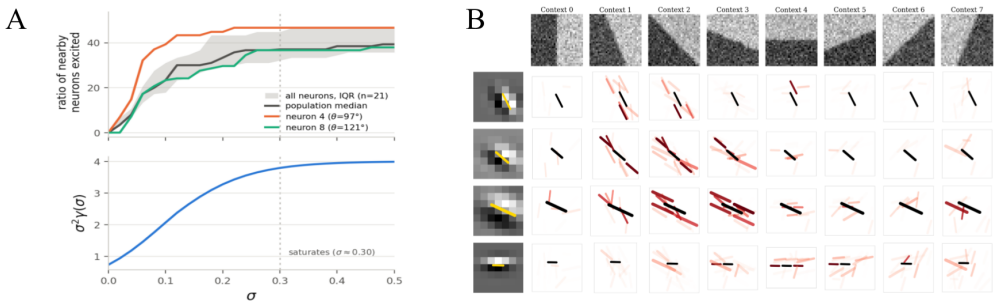


Figure 6: **State-dependent excitation and dynamic geometric routing.** (A) Top: Ratio of the 30 neighbors that the reference neuron excites rather than suppresses, as a function of input noise σ. Bottom: the learned scaling factor $\gamma(\sigma)$ that modulates the network's recurrent drive ($\Phi^\top\Phi+\gamma(\sigma)M$). (B) Context-dependent excitatory spread for individual target neurons across varying visual stimuli. The top row displays eight noisy line contexts overlapping the target receptive fields. Subsequent rows visualize the target neurons' (black needle) excitatory spread to neighboring neurons (red needle), gated by Σ. When context is misaligned with a neuron's preferred orientation, the collinear spread is suppressed.

To understand how a single-layer sparse coding model captures these rules, we examined the magnitude of the learned dictionary elements (Φ) and found that over 30% of the neural population learns near-zero norm weights (Figure 5B-C). We term these *detached neurons* because they are functionally decoupled from the ambient pixel space. Reconstructing the image exclusively from their activity yields no visible features (Figure S2.A).

However, these neurons are computationally useful. Ablating their recurrent interactions causes a significant drop in denoising performance, particularly at extreme noise levels (SNR $\ll 0$) where the network must hallucinate missing structures using global priors (Figure 5F and Figure S2.B). Furthermore, without these detached neurons, the global eigenvectors remain tied to the contours of the input image, but they lose the coordinated, long-range structure required to maintain the semantic integrity of a face (Figure 5A, right). This suggests a functional organization in which the network uses detached neurons to support hierarchical computation in an otherwise single-layer model. By actively decoupling from sensory input, these detached neurons operate as an emergent latent "second layer," utilizing lateral horizontal connections (M) to mediate long-range semantic dependencies and constrain low-level geometric features into globally consistent concepts.

4.4 Hypotheses for surround excitation in V1

Beyond elucidating the mechanisms of generalization in diffusion models, our framework offers testable hypotheses for experimental neuroscience. Our model predicts how activity propagates in neural circuits in the presence of visual stimuli. These predictions are derived from the latent Jacobian J_z, which describes the spread of neural activity following a single-neuron perturbation after the network has converged to a fixed point. This setup is conceptually similar to two-photon optogenetic experiments in layer 2/3 of mouse V1. Chettih & Harvey (2019) showed that perturbing a neuron during high-contrast stimulation mostly suppresses nearby neurons when the stimulus aligns with the target neuron's preferred orientation. Our model is consistent with this finding, but suggests a more detailed hypothesis. We predict that perturbing a neuron suppresses neighboring neurons when the stimulus is clean and high-contrast, but gradually excites them as the stimulus becomes noisier. This transition is captured mechanistically by the recurrent drive $\left(\Phi^\top\Phi+\gamma(\sigma)M\right)z_t$ and the learned modulation function (Figure 6A.bottom) in the network dynamics (Equation 4). As the visual ambiguity increases, the network transitions from relying on the likelihood term $\Phi^\top\Phi$ to relying on M (Figure 6A). This noise-dependent shift is supported by physiological evidence in V1, where low-contrast stimuli flip the influence of the surround from suppressive to facilitatory (Angelucci et al., 2017).

In addition, our model predicts that this excitatory facilitation is context-dependent and not uniform. In Figure 6B, we perturb different neurons under different stimuli and visualize only the excitatory patterns. When the target neuron's preferred orientation aligns with the stimulus context, perturbation

results in an extended, collinear spread of neural activity. This contrasts with the raw anatomical connectivity pattern in Figure 3B. Conversely, if the stimulus does not align with the target neuron's preferred orientation, such collinear spread will be suppressed. This can be explained by the decomposition of latent Jacobian in Equation 7. When the target neuron's preferred orientation aligns with the stimulus context, the sparse gate Σ restricts the spread of neural activation to a collinear path. Misaligned contexts partially shut these primary gates to allow only weak spread, while orthogonal contexts completely extinguish spread. This context-dependent non-uniform excitation pattern is consistent with a recent experiment showing that optogenetically activating ensembles of neurons with similar orientation tuning excites nearby similarly tuned neurons in the absence of visual stimuli (Oldenburg et al., 2024).

4.5 Generative Parity with Deep Learning Architectures

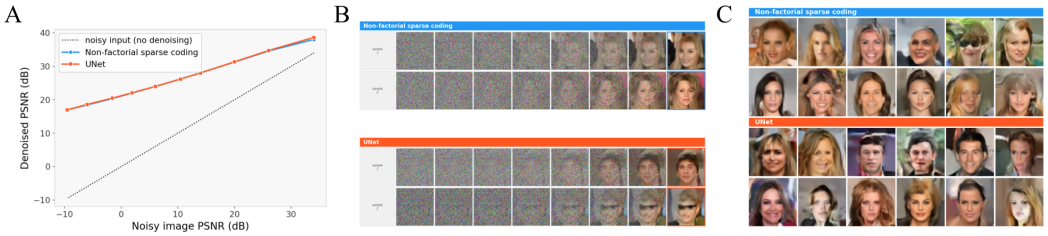


Figure 7: **Generation Performance.** (A) Denoising PSNR curves of our non-factorial SC versus a parameter-matched U-Net. (B) Reverse diffusion sampling trajectories. (C) Final generated samples.

As shown in the previous section, the non-factorial sparse coding model is consistently similar to diffusion models with a feedforward deep neural network backbone (Ronneberger et al., 2015), i.e., similar contour completion capabilities (Figures 1, 4) and similar semantic-level generalization on the face dataset (Figure 5A). To confirm this functional equivalence extends to raw generative capability, we perform a final empirical sanity check against a parameter-matched U-Net (see detailed architecture in Appendix A.10). Despite its architectural simplicity, our model achieves similar performance (with sampling algorithm described in Appendix A.6). Both models yield visually indistinguishable samples and comparable denoising PSNRs across noise levels (Figure 7A-C).

5 Limitations and Discussion

We recognize that the model we trained in this paper is small-scale and that the neural circuit we discovered comes from a diffusion model with a recurrent backbone. It is possible that models with different architectures could learn completely different neural circuits to solve the same problem. However, we also recognize striking functional parallels among the visual cortex, deep neural networks (U-Nets), and a non-factorial sparse coding model. This convergence points toward a universal principle: these networks all capture intrinsic statistics of natural images. The deformation neural circuit we discovered could also occur within black-box neural networks, offering a new direction for mechanistic interpretability (Wang et al., 2022; Lindsey et al.).

To fully understand this convergence, we must better characterize the structure of the natural image itself. We hypothesize that natural images act as sparse signals on a continuous manifold (Chen et al., 2018; 2022b; Roweis & Saul, 2000; Tenenbaum et al., 2000; He et al., 2026). The manifold represents continuous, elementary structures (like edges), while the sparse signal specifies their composition. Approximating the distribution of natural images thus requires first parameterizing this manifold, and then estimating the density of these compositional functions. Sparse coding is interpretable precisely because it approximates this manifold in the simplest possible way: by dropping discrete "landmarks" onto it and interpolating the signal between them. It appears that visual cortex uses a similar strategy, likely driven by the metabolic necessity of sparsity (Atick & Redlich, 1992; Field, 1994; Karklin & Simoncelli, 2011). Thus, the models developed in the computational neuroscience and deep learning communities do not dwell in isolated domains, but can be viewed as solving the same geometric problem. Such insights are essential for building simpler, more efficient AI models and deepening our understanding of the brain.

AI USE STATEMENT

We used generative AI tools for writing assistance and for coding assistance. For writing, we used them to polish prose and to check grammar; all scientific content, claims, and arguments originated with the authors, and every AI-edited passage was reviewed and, where necessary, rewritten by the authors. We did not use generative AI tools to generate research ideas, to retrieve or summarize related work, or to produce any of the results, figures, or analyses reported here. All AI-assisted code was read, tested, and verified by the authors against the mathematical derivations given in the main text and appendices. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have provided a complete code release in this anonymous repo, and have verified that every figure in this paper is reproducible. We also described the engineering details of training and constructing the minimal diffusion model in Appendices A.3, A.6, and A.8, and list all hyperparameters here.

REFERENCES

Luis B Almeida. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Artificial neural networks: concept learning*, pp. 102–111. 1990.

Alessandra Angelucci, Maryam Bijanzadeh, Lauri Nurminen, Frederick Federer, Sam Merlin, and Paul C Bressloff. Circuits and mechanisms for surround modulation in visual cortex. *Annu. Rev. Neurosci.*, 40(1):425–451, July 2017.

Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in neural information processing systems*, 32, 2019.

Xingjian Bai and Luke Melas-Kyriazi. Fixed point diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9430–9440, 2024.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Victor Boutin, Angelo Franciosini, Franck Ruffier, and Laurent Perrinet. Effect of top-down connections in hierarchical sparse coding. *Neural Computation*, 32(11):2279–2309, 2020.

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1):1–12, 2008.

Rakesh Chalasani, Jose C Principe, and Naveen Ramakrishnan. A fast proximal method for convolutional sparse coding. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–5. IEEE, 2013.

Yubei Chen, Dylan M. Paiton, and Bruno A. Olshausen. The sparse manifold transform. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10534–10545, 2018.

Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Bag of image patch embedding behind the success of self-supervised learning. *arXiv preprint arXiv:2206.08954*, 2022a.

Yubei Chen, Zeyu Yun, Yi Ma, Bruno Olshausen, and Yann LeCun. Minimalistic unsupervised representation learning with the sparse manifold transform. In *The Eleventh International Conference on Learning Representations*, 2022b.

Selmaan N. Chettih and Christopher D. Harvey. Single-neuron perturbations reveal feature-specific competition in v1. *Nature*, 567(7748):334–340, Mar 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0997-6. URL https://doi.org/10.1038/s41586-019-0997-6.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature neuroscience*, 23(9):1138–1149, 2020.

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

Laurent El Ghaoui, Fangda Gu, Bertrand Travacca, Armin Askari, and Alicia Tsai. Implicit deep learning. *SIAM Journal on Mathematics of Data Science*, 3(3):930–958, 2021.

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.

David J Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 1994.

David J Field, Anthony Hayes, and Robert F Hess. Contour integration by the human visual system: evidence for a local "association field". *Vision research*, 33(2):173–193, 1993.

David Fitzpatrick. The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cerebral cortex*, 6(3):329–341, 1996.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. Jfb: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6648–6656, 2022.

Pierre Garrigues and Bruno Olshausen. Learning horizontal connections in a sparse coding model of natural images. *Advances in neural information processing systems*, 20, 2007.

Pierre Garrigues and Bruno Olshausen. Group sparse coding with a laplacian scale mixture prior. *Advances in neural information processing systems*, 23, 2010.

Wilson S Geisler, Jeffrey S Perry, BJ Super, and DP Gallogly. Edge co-occurrence in natural images predicts contour grouping performance. *Vision research*, 41(6):711–724, 2001.

Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. *Advances in neural information processing systems*, 34:24247–24260, 2021.

Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368):eaag2612, 2017.

Dileep George, Miguel Lázaro-Gredilla, Wolfgang Lehrach, Antoine Dedieu, Guangyao Zhou, and Joseph Marino. A detailed theory of thalamic and cortical microcircuits for predictive visual inference. *Science Advances*, 11(6):eadr6698, 2025.

Stephen Grossberg, Ennio Mingolla, and William D Ross. Visual brain and visual perception: how does the cortex do perceptual grouping? *Trends in neurosciences*, 20(3):106–111, 1997.

Denis Gudovskiy, Wenzhao Zheng, Tomoyuki Okuno, Yohei Nakata, and Kurt Keutzer. Odet (odel): Shortcutting the time and the length in diffusion and flow models for faster sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6111–6120, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ye He, Yitong Qiu, and Molei Tao. Diffusion model's generalization can be characterized by inductive biases toward a data-dependent ridge manifold. *arXiv preprint arXiv:2602.06021*, 2026.

Judith A Hirsch and Charles D Gilbert. Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, 11(6):1800–1809, 1991.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Haruo Hosoya and Aapo Hyvärinen. A hierarchical statistical model of natural images explains tuning properties in v2. *Journal of Neuroscience*, 35(29):10412–10428, 2015.

Haruo Hosoya and Aapo Hyvärinen. Learning visual spatial pooling by strong pca dimension reduction. *Neural computation*, 28(7):1249–1264, 2016.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.*, volume 39. Springer Science & Business Media, 2009.

Stanisław Jastrzębski, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. Residual connections encourage iterative inference. *arXiv preprint arXiv:1710.04773*, 2017.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *arXiv preprint arXiv:2310.02557*, 2023.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Yan Karklin and Michael S Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14(3):483–499, 2003.

Yan Karklin and Michael S Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural computation*, 17(2):397–423, 2005.

Yan Karklin and Eero Simoncelli. Efficient coding of natural images with a population of noisy linear-nonlinear neurons. *Advances in neural information processing systems*, 24, 2011.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Ann B Lee, Kim S Pedersen, and David Mumford. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(1):83–103, 2003.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Zhaoping Li. A neural model of contour integration in the primary visual cortex. *Neural computation*, 10(4):903–940, 1998.

Qiyao Liang, Ziming Liu, and Ila Fiete. Do diffusion models learn semantically meaningful and efficient representations? *arXiv preprint arXiv:2402.03305*, 2024.

J Lindsey, W Gurnee, E Ameisen, B Chen, A Pearce, and NL Turner. & batson, j.(2025). on the biology of a large language model. *Transformer Circuits Thread*.

K Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 38:181–188, 1961.

Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019.

Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. *arXiv preprint arXiv:2408.05927*, 2024.

Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024.

Ian Antón Oldenburg, William D Hendricks, Gregory Handy, Kiarash Shamardani, Hayley A Bounds, Brent Doiron, and Hillel Adesnik. The logic of recurrent circuits in the primary visual cortex. *Nature neuroscience*, 27(1):137–147, 2024.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

Bruno A Olshausen and Michael S Lewicki. What natural scene statistics can tell us about cortical representation. *The New Visual Neurosciences,(London)*, pp. 1247–1262, 2014.

Bruno A Olshausen, Charles Cadieu, Jack Culpepper, and David K Warland. Bilinear models of natural images. In *Human Vision and Electronic Imaging XII*, volume 6492, pp. 67–76. SPIE, 2007.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

Simon Osindero, Max Welling, and Geoffrey E Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2006.

Dylan M Paiton, Steven Shepard, Kwan Ho Ryan Chan, and Bruno A Olshausen. Subspace locally competitive algorithms. In *Proceedings of the 2020 Annual Neuro-Inspired Computational Elements Workshop*, pp. 1–8, 2020.

P. Parent and S.W. Zucker. Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):823–839, 1989. doi: 10.1109/34.31445.

Fernando Pineda. Generalization of back propagation to recurrent and higher order neural networks. In *Neural information processing systems*, 1987.

Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

M Raphan and E P Simoncelli. Least squares estimation without priors or supervision. *Neural Computation*, 23(2):374–420, 2 2011.

H Robbins. An empirical bayes approach to statistics. In *Proc Third Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pp. 157–163. University of CA Press, 1956.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL `http://arxiv.org/abs/1505.04597`.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. *Neural computation*, 20(10):2526–2563, 2008.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

Odelia Schwartz and Eero P Simoncelli. Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8):819–825, 2001.

Mariano Sigman, Guillermo A Cecchi, Charles D Gilbert, and Marcelo O Magnasco. On a common circle: natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences*, 98(4): 1935–1940, 2001.

Eero P Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Wavelet Applications in Signal and Image Processing VII*, volume 3813, pp. 188–195. SPIE, 1999.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

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

Tahereh Toosi and Kenneth D Miller. Generative inference unifies feedback processing for learning and perception in natural and artificial vision. *bioRxiv*, pp. 2025–10, 2025.

J Hans Van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Kai Yu, Yuanqing Lin, and John Lafferty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR 2011*, pp. 1713–1720. IEEE, 2011.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pp. 2528–2535. IEEE, 2010.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

A Appendix

A.1 Noise Distribution and Loss Weighting

Having established the weighted reconstruction objective (Equation 2), we must define the noise sampling distribution $p(\sigma)$ and the loss weighting schedule $w(\sigma)$. Following the preconditioning framework of Elucidating Diffusion Models (EDM) (Karras et al., 2022), we sample the noise scale σ from a log-normal distribution during training:

$$\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2) \tag{8}$$

where P_{mean} and P_{std} are hyperparameters governing the median and width of the noise distribution, respectively. This sampling strategy ensures the network focuses its capacity on the intermediate noise levels that are most critical for learning the data manifold, explicitly avoiding the extremes of pure noise or perfectly clean data. To stabilize training and ensure that the expected gradient magnitudes remain uniform across all noise levels, we apply the EDM weighting schedule:

$$w(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma \cdot \sigma_{\text{data}})^2} \tag{9}$$

where σ_{data} is a hyperparameter representing the standard deviation of the underlying data distribution. By heavily penalizing errors at low noise levels (where $\sigma \to 0$) and down-weighting errors at high noise levels, this schedule dynamically balances the learning signal, allowing the optimal sparse coding dictionary Φ and latent state z^* to be efficiently jointly optimized across all noise scales.

A.2 Implicit gradient

Let $z^*(x; \sigma, \theta) := \lim_{t\to\infty} z_t(x; \sigma, \theta)$ denote the convergent state of the recurrent inference in Equation (4), and let θ denote all model parameters, i.e., $\theta := \{\Phi, M, \lambda, \gamma\}$. Both training the model under the denoising objective (Equation 2) and calculating the Jacobian of the score function $H(x)$ for analysis require differentiating through the convergent state $z^*(x; \sigma, \theta)$. Typically, computing gradients through a recurrent circuit requires Backpropagation Through Time (BPTT), a process that unrolls the entire iterative trajectory and incurs prohibitive computational and memory costs. However, because z^* converges to a stable fixed point of the inference dynamics, we can completely bypass BPTT. The Implicit Function Theorem allows us to compute the exact analytical gradients in closed form:

$$F(z; x, \sigma, \theta) := z - T_{\theta,\sigma}(z; x), \qquad F(z^*; x, \sigma, \theta) = 0. \tag{10}$$

Assuming z^* is an isolated stable fixed point and that F is locally differentiable with a non-singular Jacobian $\nabla_z F(z^*; x, \sigma, \theta)$, the implicit function theorem yields

$$\frac{\partial z^*}{\partial \theta} = -\left[\nabla_z F(z^*; x, \sigma, \theta)\right]^{-1} \nabla_\theta F(z^*; x, \sigma, \theta), \tag{11}$$

$$\frac{\partial z^*}{\partial x} = -\left[\nabla_z F(z^*; x, \sigma, \theta)\right]^{-1} \nabla_x F(z^*; x, \sigma, \theta). \tag{12}$$

We use the first implicit gradient $\frac{\partial z^*}{\partial \theta}$ to compute the exact gradient of model parameters for training, by chain rule: $\frac{d\mathcal{L}_{\text{DSM}}}{d\theta} = \frac{\partial \mathcal{L}_{\text{DSM}}}{\partial \theta} + \frac{\partial \mathcal{L}_{\text{DSM}}}{\partial z^*}\frac{\partial z^*}{\partial \theta}$. We utilize the second implicit gradient, $\frac{\partial z^*}{\partial x}$, to compute the exact analytical Jacobian of the denoiser. Because the denoised output is a linear function of the convergent state of inference, $f_\theta(x) = \Phi z^*(x; \sigma, \theta)$, applying the chain rule directly yields the denoising Jacobian: $J(x) = \Phi \frac{\partial z^*}{\partial x}$.

A.3 Other training details

To stabilize the highly nonlinear, recurrent dynamics during training and to ensure smooth convergence of the generative prior, we maintain an exponential moving average (EMA) of the model weights. Rather than evaluating the network using the highly fluctuating active weights during inference, we use the EMA parameters. The decay rate β is adjusted dynamically based on the global batch size (N_{batch}) and a predefined half-life (k_{img}) to ensure consistent smoothing independent of hardware scaling:

$$\beta = 0.5^{\frac{N_{\text{batch}}}{\max(k_{\text{img}} \times 1000, 1)}}$$

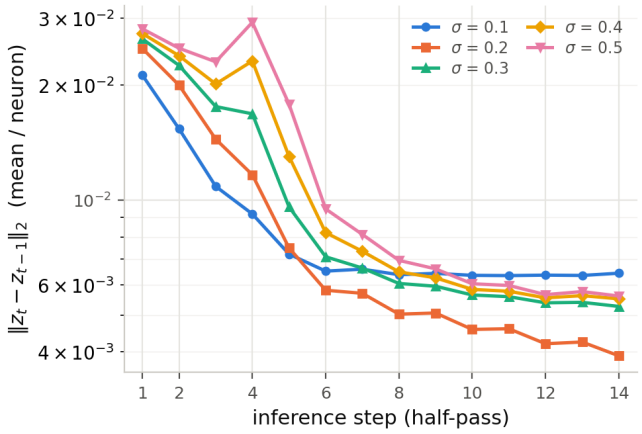


Figure S1: **The inference dynamics converge to a fixed point.** residual $\|z_t - z_{t-1}\|_2$ of the inference dynamic against inference step, for five input noise levels, averaged over 32 held-out for model trained on (Van Hateren & van der Schaaf, 1998) dataset. The residual decreases and plateaus at every noise level.

This practice is standard in state-of-the-art diffusion models, prevents mode collapse, and significantly improves the perceptual quality of the completed contours.

A.4 Approximating Implicit Gradient

Given the implicit gradient for updating parameter as follows:

$$\frac{\partial z^*}{\partial \theta} = -\left[\nabla_z F(z^*; x, \sigma, \theta)\right]^{-1} \nabla_\theta F(z^*; x, \sigma, \theta), \tag{13}$$

Computing the matrix inverse of the fix point update's dynamics $\left[F(z^*; x, \sigma, \theta)\right]^{-1}$ is expensive, and computationally intractable if z^* is extremely high dimension like convolution sparse code for images. In practice, we approximate $\left[F(z^*; x, \sigma, \theta)\right]^{-1}$. Recall we define

$$F(z; x, \sigma, \theta) := z - T_{\theta,\sigma}(z; x), \qquad F(z^*; x, \sigma, \theta) = 0.$$

So the Jacobian of F is $[I - J]^{-1}$, where J is the Jacobian of the single step ISTA update $T_{\theta,\sigma}(z; x)$. Assuming the ISTA update is stable (the spectral radius of J is strictly less than 1), then we can approximate $[I - J]^{-1}$ using truncated neumann-series expansion: $\sum_{k=0}^{m} J^k$. While this Neumann-series Phantom Gradient (NPG) can be explicitly computed via iterative vector-Jacobian products, it is mathematically equivalent to simply unrolling the forward dynamics. Therefore, following the method derived in (Geng et al., 2021; Fung et al., 2022), we implement an Unrolling-based Phantom Gradient (UPG): after rapidly finding the detached fixed point z^* without gradient tracking, we unroll the dynamics from z^* for m additional iterations with gradients enabled to compute $z_m = T^m(z^*; x, \sigma, \theta)$. During backpropagation, evaluating $\frac{\partial z_m}{\partial \theta}$ via the standard chain rule naturally accumulates the exact Neumann series expansion. We generally set m between 1 and 3 steps during training. This allows us to efficiently approximate the implicit gradient by leveraging standard autograd engines, bypassing the severe memory constraints of full backpropagation through time.

The Neumann expansion converges only if the fixed-point iteration is stable. We do not enforce this constraint explicitly during training; instead we verify it empirically. Figure S1 reports the residual of the inference dynamics, averaged over 32 held out Van Hateren images. At every noise level, the residual falls by roughly 5 times over first six steps and then flattens, indicating the convergence of the dynamics empirically.

A.5 Detailed intuition on the mechanistic decomposition of the denoising Jacobian

We provide a more detailed interpretation on the decomposition of the denoising Jacobian in Equation 6 and Equation7:

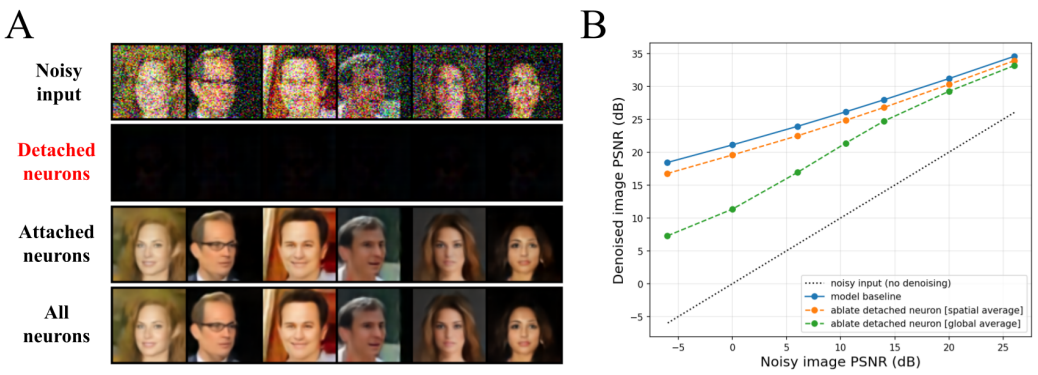


Figure S2: **Detached neurons contribute in recurrent dynamics but not linear decoding** (A) Sub-population reconstructions: detached neurons yield no visible features, whereas attached neurons capture local geometry. (B) Denoising PSNR drops catastrophically at high noise when detached neurons are ablated (replaced by spatial or global means).

$$J(x) = \frac{dx_{\text{clean}}}{dx_{\text{noise}}} = \eta\Phi \underbrace{\left[I - \Sigma W\right]^{-1}\Sigma}_{J_z} \Phi^\top = \eta\Phi J_z \Phi^\top$$

$$J_{z,i} = \underbrace{e_i}_{\text{Sparse gating}} + \underbrace{\Sigma W e_i}_{\text{1st-order spread}} + \underbrace{\Sigma W \Sigma W e_i}_{\text{2nd-order spread}} + \ldots \quad (\text{for } i \text{ s.t. } z_i^* > 0, \text{ else } 0)$$

1. **Projection and Inhibition ($\Sigma\Phi^\top$):** The perturbed input is projected onto the dictionary. The Σ mask instantly silences inactive elements, ensuring that only perturbations relevant to the current context survive, analogous to coring or hard thresholding in signal processing.
2. **Context-Aware Spread (ΣW):** Surviving signals spread via lateral connections W. The outer Σ clamps this expansion, isolating local geometric cross-talk by ensuring perturbations only spread to activated neurons.
3. **Higher-Order Context-Aware Spread $(\Sigma W)^k$:** The infinite tail of the sum temporally unrolls the dynamics. Just as raising an adjacency matrix to the k-th power computes k-hop context-aware spread of activity defined by lateral connection. Higher-order term $(\Sigma W)^k$ allows the neural activity not only spread but ricochets globally yet strictly confined among mutually active neurons.
4. **Synthesis (Φ):** The context-modulated signals are projected back into ambient pixel space.

A.6 Sampling algorithm

Let $\nabla_x \log p_t(x_t)$ denote the score function of the perturbed data distribution, following the deterministic formulation of Karras et al. (2022), we can map generation process to a Probability Flow Ordinary Differential Equation (ODE). By parameterizing the time variable directly by the noise standard deviation σ, the deterministic ODE governing the data evolution is given by:

$$\mathrm{d}x = -\sigma \nabla_x \log p_{\theta,\sigma}(x)\, \mathrm{d}\sigma \tag{14}$$

In practice, the score is parameterized via a denoiser model $D_\theta(x, \sigma)$, which relates to the score via Tweedie's formula, $D_\theta(x, \sigma) = x + \sigma^2 \nabla_x \log p_{\theta,\sigma}(x)$. Substituting this relationship into Equation 14 yields the ODE in terms of the denoiser:

$$\mathrm{d}x = \frac{x - D_\theta(x, \sigma)}{\sigma}\, \mathrm{d}\sigma \tag{15}$$

In our framework, evaluating the denoiser $D_\theta(x, \sigma)$ is parametrize as the fix point of the recurrent dynamics to approximate the fixed point, i.e., $D_\theta(x, \sigma) = \Phi z^*(x, \sigma)$.

To numerically integrate Equation 15, we employ a second-order Heun (improved Euler) scheme. The continuous noise level σ is discretized into a monotonically descending sequence $\{\sigma_0, \sigma_1, \ldots, \sigma_N\}$, starting from pure noise $\sigma_0 = \sigma_{\max}$ and terminating at a minimum noise level $\sigma_N = \sigma_{\min}$. At each integration step i, the solver performs a predictor-corrector update. First, an Euler step predicts the intermediate state at σ_{i+1}. Then, a corrector step evaluates the denoiser at this predicted state. The final update is computed using the trapezoidal rule by averaging the derivatives. The full sampling procedure is summarized in Algorithm 1.

Algorithm 1 Annealed Heun Sampling

Require: Denoiser D_θ, recurrent iterations K, noise schedule $\{\sigma_0, \sigma_1, \ldots, \sigma_N\}$
Sample $x_0 \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I})$
for $i = 0$ **to** $N - 1$ **do**
  $\hat{x}_0 \leftarrow D_\theta(x_i, \sigma_i)$ {Evaluated as $\Phi z^*(x_i, \sigma_i)$}
  $d_i \leftarrow \frac{x_i - \hat{x}_0}{\sigma_i}$
  $x_{\text{euler}} \leftarrow x_i + (\sigma_{i+1} - \sigma_i) d_i$
  % Corrector Step
  $\hat{x}_1 \leftarrow D_\theta(x_{\text{euler}}, \sigma_{i+1})$ {Evaluated as $\Phi z^*(x_{\text{euler}}, \sigma_{i+1})$}
  $d'_{i+1} \leftarrow \frac{x_{\text{euler}} - \hat{x}_1}{\sigma_{i+1}}$
  % Heun Update
  $x_{i+1} \leftarrow x_i + (\sigma_{i+1} - \sigma_i) \left(\frac{1}{2} d_i + \frac{1}{2} d'_{i+1}\right)$
end for
$x_{\text{final}} \leftarrow D_\theta(x_N, \sigma_N)$ {Final projection at $\sigma_{\min}$}
return x_{final}

A.7 Negative Log likelihood Estimation

For our model, which directly predicts the clean data point x_0 via a parameterized denoising network $D_\theta(x, \sigma)$, the probability flow ODE with respect to the continuous noise scale σ takes the following form:

$$\frac{\mathrm{d}x}{\mathrm{d}\sigma} = \frac{x - D_\theta(x, \sigma)}{\sigma} =: \tilde{f}_\theta(x, \sigma)$$

Using the instantaneous change of variables formula, we can compute the marginal log-likelihood of the data at the minimum noise level σ_{min} by integrating the divergence of the empirical drift function up to the maximum noise level σ_{max}:

$$\log p_{\sigma_{\text{min}}}(x(\sigma_{\text{min}})) = \log p_{\sigma_{\text{max}}}(x(\sigma_{\text{max}})) + \int_{\sigma_{\text{min}}}^{\sigma_{\text{max}}} \nabla_x \cdot \tilde{f}_\theta(x(\sigma), \sigma) \mathrm{d}\sigma$$

where the random variable $x(\sigma)$ as a function of σ is obtained by solving the probability flow ODE. In high-dimensional spaces, exactly computing the divergence $\nabla_x \cdot \tilde{f}_\theta(x, \sigma)$ is computationally prohibitive. Following Grathwohl et al. (2018), we approximate it using the unbiased Skilling-Hutchinson trace estimator:

$$\nabla_x \cdot \tilde{f}_\theta(x, \sigma) = \mathbb{E}_{p(\epsilon)}[\epsilon^\top \nabla_x \tilde{f}_\theta(x, \sigma) \epsilon]$$

where $\nabla_x \tilde{f}_\theta$ denotes the Jacobian of $\tilde{f}_\theta(\cdot, \sigma)$, and the random projection vector $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The vector-Jacobian product $\epsilon^\top \nabla_x \tilde{f}_\theta(x, \sigma)$ can be efficiently evaluated in a single backward pass using reverse-mode automatic differentiation. To accelerate the evaluation of the trace integral, we solve the probability flow ODE using a first-order Euler method over a geometric (log-linear) schedule of $N = 20$ discrete steps from σ_{min} to σ_{max}. We assume the prior distribution at the boundary condition σ_{max} is a standard isotropic Gaussian, yielding $p_{\sigma_{\text{max}}}(x) = \mathcal{N}(0, \sigma_{\text{max}}^2 \mathbf{I})$. Finally, to report the likelihood results, we convert the computed continuous log-likelihoods to discrete Bits Per Dimension (BPD). We divide the log-likelihood by $D \ln(2)$ (where D is the data dimensionality) and apply an 8.0 bit dequantization offset to account for the uniform binning of standard 8-bit [0, 1] scaled image data.

A.8 Practical implemention of model's dynamics: Multi-Scale Convolutional energy function, and Asymmetric update.

While Equation 4 provides a theoretically exact formulation for a single-scale, flatten vector space, adapting this inference dynamics to high-dimensional natural images requires a multi-scale, convolutional architecture. In our practical implementation, the dictionary Φ and the prior interaction matrix M are distributed across multiple spatial scales (akin to a Gaussian pyramid) to capture both fine local edges and coarse global contours. Furthermore, we employ an asymmetric, alternating update schedule to accelerate the propagation of information across the network.

Multi-Scale Likelihood via Laplacian Pyramid Instead of a single dense dictionary, the network utilizes a set of scale-specific convolutional dictionaries (Chalasani et al., 2013) $\Phi^{(i)}$ for scales $i = 0, \ldots, L-1$ (where $i = 0$ is the finest spatial resolution). To map these latent variables to the image space, we define P as the multiscale Laplacian pyramid synthesis operator. The multiscale reconstruction is formulated as $P(\Phi z)$, allowing us to rewrite the joint energy function's likelihood term as:

$$E_{\text{likelihood}} = \frac{1}{\gamma_\psi(\sigma)} \|x - P\Phi z\|_2^2$$

To optimize this, the gradient with respect to z requires applying the chain rule, yielding $-\Phi^\top P^\top (x - P\Phi z)$. Here, $P^\top$ acts as the Laplacian pyramid analysis operator, decomposing the image-space residual into scale-specific frequency bands. The feedforward drive for a specific scale i is then computed by passing this banded residual through the corresponding transposed dictionary:

$$\text{Feedforward Drive}^{(i)} = \Phi^{(i)\top} \left[P^\top (x - P\Phi z)\right]^{(i)}$$

To explicitly model the non-factorial co-occurrence statistics across different spatial scale, the global prior matrix M is defined as a block-tridiagonal operator. This restricts interactions strictly to intra-scale horizontal connections and adjacent inter-scale vertical connections. We can then decompose the learned recurrent connection contribution to the dynamics (Wz) in the following way. For any scale i,

1. Lateral Scale Connections ($W_{\text{lat}}^{(i)}$): Captures intra-scale horizontal co-occurrences (e.g., collinear edges).
2. Up-Scale Connections ($W_{\text{up}}^{(i)}$): Routes feedforward contextual signals from the finer, higher-resolution scale $i-1$.
3. Up-Down Connections ($W_{\text{down}}^{(i)}$): Routes feedback signals from the coarser, lower-resolution scale $i+1$.

which results in the following ISTA inference update step for multiscale version of non-factorial sparse coding:

$$z_{t+1}^{(i)} = \text{ReLU}\left(z_t^{(i)} + \eta \left[\underbrace{\Phi^{(i)\top}\left[P^\top r_t\right]^{(i)}}_{\text{Feedforward Drive}} - \gamma_\psi(\sigma) \underbrace{\left(W_{\text{lat}}^{(i)} z_t^{(i)} + W_{\text{up}}^{(i)} z_t^{(i-1)} + W_{\text{down}}^{(i)} z_t^{(i+1)} \right)}_{\text{Hierarchical Recurrent Prior}} - \underbrace{\gamma_\psi(\sigma)\lambda^{(i)}}_{\text{Sparsity Bias}} \right] \right) \tag{16}$$

where $r_t = x - P\Phi z_t$ is the current global image-space reconstruction residual. In addition, all the matrix multiplication in this model is implemented using convolution, and transpose of the matrix is implemented using transposed convolution, as illustrated Figure S3.

Asymmetric (Sweeping) Updates In a standard, synchronous ISTA implementation (Jacobi-style updates), information strictly propagates one scale per iteration, creating a latency of L steps for the finest scale to influence the coarsest scale. To overcome this and simulate rapid biological inference, our implementation utilizes an asymmetric sweeping update schedule (Gauss-Seidel optimization). A

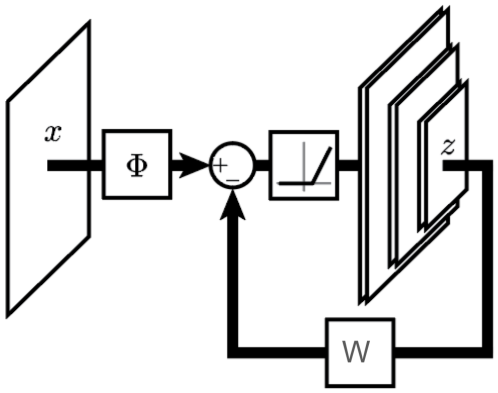


Figure S3: Schematic of the multiscale, convolutional sparse coding's inference dynamics.

single inference iteration consists of a continuous two-pass sweep: Bottom-Up Sweep: The network updates scales sequentially from finest to coarsest ($i = 0 \rightarrow L - 1$). During this pass, the calculation for $z^{(i)}$ utilizes the already updated values from $z_{t+1}^{(i-1)}$, allowing fine-grained local evidence to immediately percolate up the hierarchy. Top-Down Sweep: The network reverses direction, updating from coarsest to finest ($i = L - 1 \rightarrow 0$). Here, the calculation for $z^{(i)}$ utilizes the already updated coarse context from $z_{t+1}^{(i+1)}$, rapidly broadcasting global shape constraints down to the local feature detectors. Empirically, this update rule help both model to get to fix point faster and parameter convergence faster.

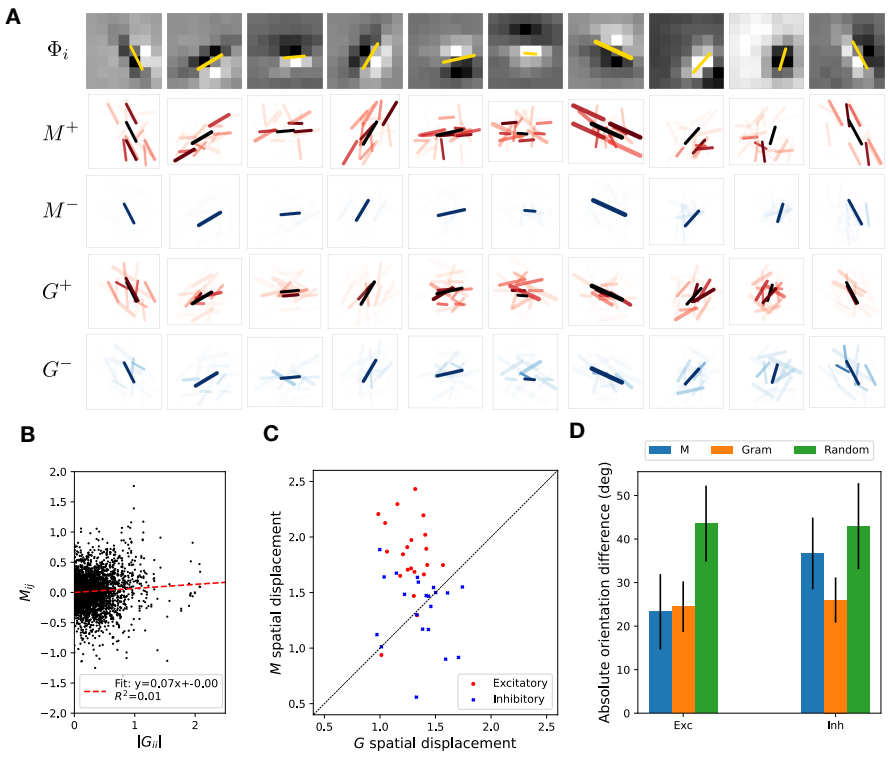


Figure S4: **(A)** Additional learned interaction maps using the same needle-plotting convention as Figure 2B. Columns correspond to reference basis functions Φ_i and colored bars show interactions with other basis functions Φ_j. (B) Joint scatter plot comparing learned pairwise interactions M_{ij} with Gram interactions $G_{ij} = \Phi_i^\top \Phi_j$. The red dashed line shows the linear fit. Its positive slope indicates that learned interactions are biased toward facilitation **(C)** Spatial displacement of the top-10 largest-magnitude interactions for each neuron, comparing learned interactions M with Gram interactions G separately for excitatory and inhibitory components. Learned excitatory interactions span larger spatial displacements than Gram interactions, indicating longer-range facilitation. A less significant difference **(D)** Average absolute orientation difference between neurons and its top-10 connected units for M, G, and randomly assigned connections. Error bars indicate one standard deviation. Learned excitatory interactions connect units with similar preferred orientations, whereas learned inhibitory interactions show weaker orientation similarity.

A.9 More visualization eigenvectors of Jacobian and decomposition

See Figure S7.

A.10 U-Net Architecture Details

For our experiments, we employ a U-Net architecture based on the guided-diffusion implementation (Dhariwal & Nichol, 2021). Specifically, we use the following configuration for processing 64×64 resolution images: Base Architecture: The model takes 3-channel inputs and outputs 3 channels, starting with a base feature width of 32 channels. Depth and Channel Scaling: The network consists of 4 resolution levels with channel multipliers of $(1, 2, 3, 4)$. This results in feature map widths of 32, 64, 96, and 128 at each respective stage of the encoder and decoder. Residual Blocks: Each resolution level contains 1 residual block. Downsampling and upsampling operations are also parameterized as residual blocks. We apply dropout with a probability of 0.1 within the residual blocks, and we do not use scale-shift normalization. Attention: To maintain a fully convolutional, minimal architectural footprint, all spatial self-attention blocks are disabled. Timestep Conditioning: The noise scale (or

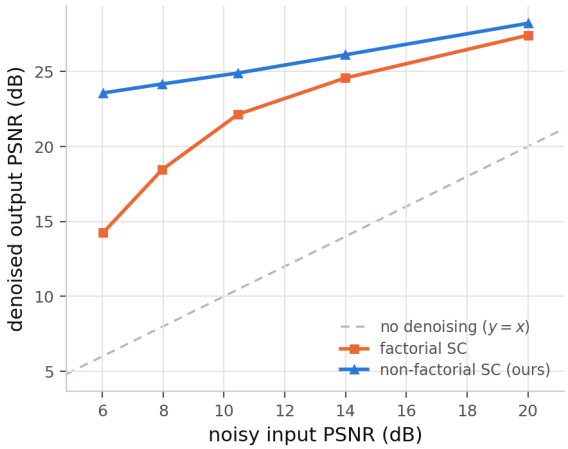


Figure S5: **Denoising performance.** Mean PSNR over 100 van Hateren test images at five noise levels. The non-factorial model outperforms the factorial sparse coding baseline at every noise level, and the gap widens as noise increases.

timestep) is projected into a 64-dimensional embedding space (using a factor of 2 relative to the base width) and is injected into every residual block via standard addition.

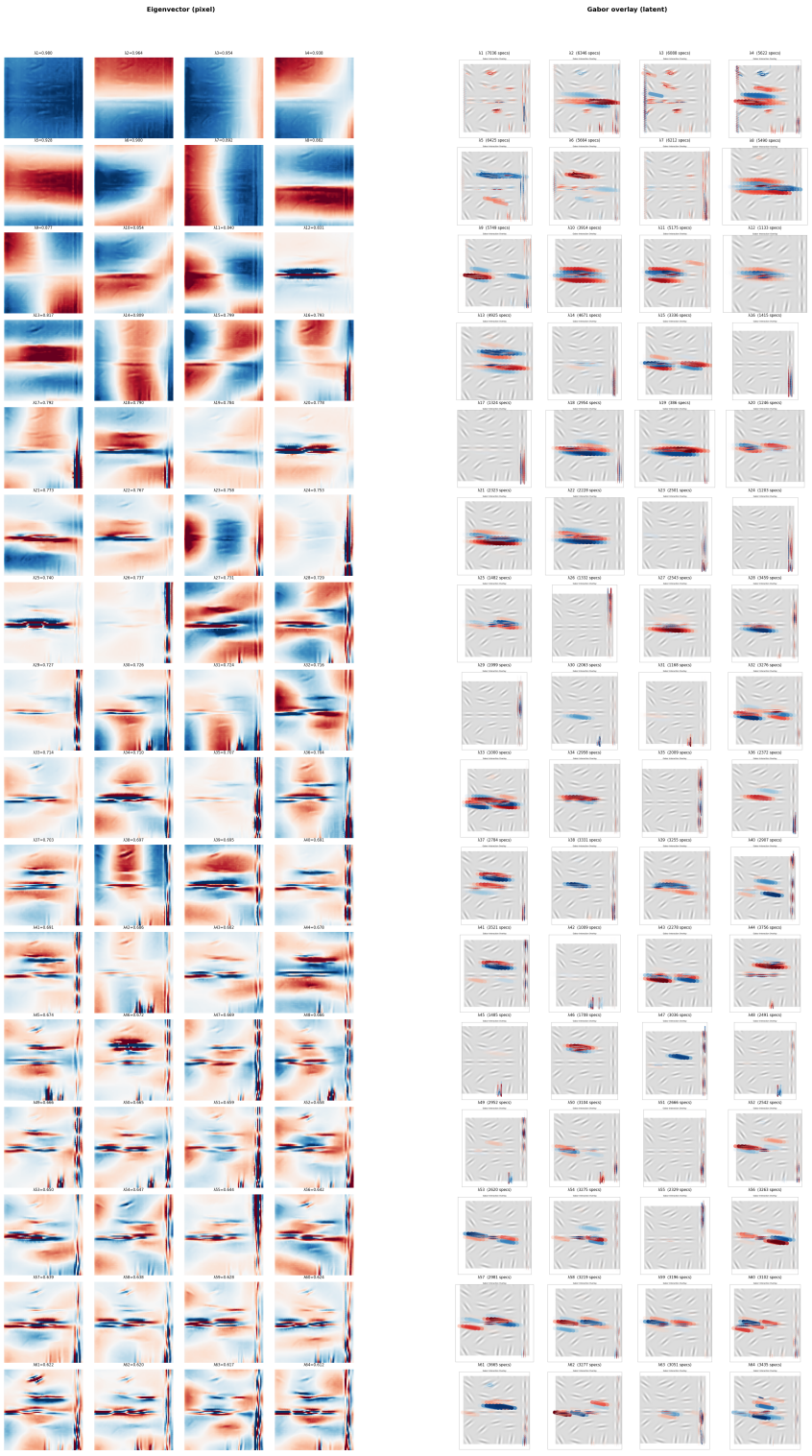


Figure S6: More pair of eigenvectors of pixel vs latent Jacobian of non-factorial sparse coding model visualized in Figure 4.A.

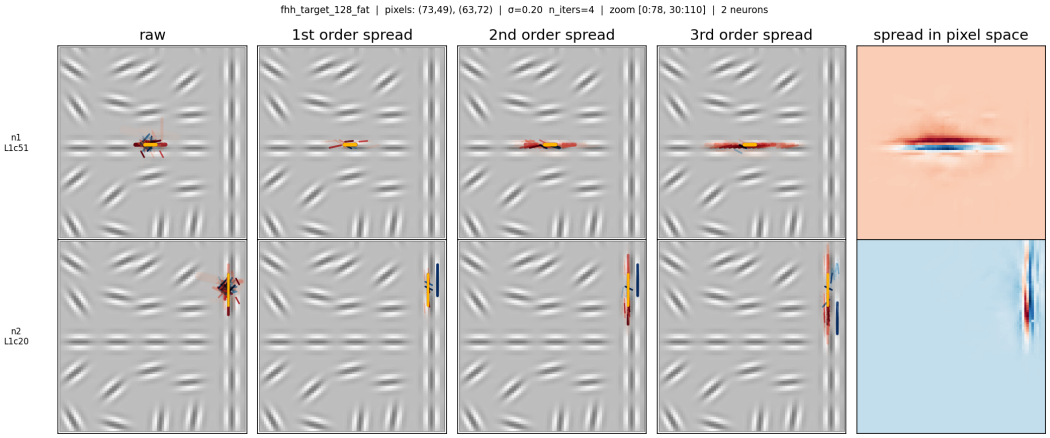


Figure S7: More latent Jacobian impulse decomposition visualized in Figure 4.D.