Scale-mixture Langevin sampling in the subspaces of recurrent cortical circuit dynamics

Zimei Chen[1,2], **Yi Ren**[1,2], **Wen-Hao Zhang**[1,2*]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center.
[2]O'Donnell Brain Institute, UT Southwestern Medical Center.

Abstract

Recurrent cortical circuits are hypothesized to perform Bayesian posterior sampling through their stochastic dynamics. Yet identifying *which* sampling algorithm a circuit intrinsically implements, as opposed to mapping manually designing algorithms onto circuits, requires rigorous theoretical analysis of the biological circuit dynamics, and remains a fundamental open challenge because of the analytical intractability of nonlinear recurrent dynamics. Here, we tackle this challenge through comprehensive theoretical analysis of analytically tractable nonlinear recurrent circuit dynamics used in neuroscience research— the continuous attractor network (CAN). Surprisingly, we find that the circuit, without any non-Gaussian components in its connectivity or internal variability, intrinsically emerges heavy-tailed subspaces dynamics described as scale-mixture Langevin posterior sampling. The mechanism emerges from the circuit's own nonlinear dynamics: divisive normalization renders the overall population activation Gamma-distributed which multiplicatively modulates the time constant of the Langevin sampling in the stimulus feature subspace. Eventually, this scale-mixture structure produces a heavy-tailed (Student-t) sampling step size distribution, significantly accelerating the sampling speed. Moreover, the scale-mixture Langevin sampling approaches a Lévy-like regime in the weak input limit. Our results analytically identify an intrinsic circuit algorithm embedded in subspaces of nonlinear recurrent dynamics, and establish a minimal circuit mechanism for endogenous heavy-tail statistics without any explicit non-Gaussian components.

1 Introduction

A wide spectrum of perceptual and cognitive processes have been mathematically formulated as Bayesian inference [1], including visual perception [2], multisensory integration [3], sensorimotor control [4], and decision-making [5], establishing the "Bayesian brain" framework. A fundamental question is how neural circuits mechanistically implement this inference [6–8]. Among the proposed mechanisms, sampling-based Bayesian inference has attracted particular attention [7, 9–12]: the prominent variability of cortical responses is hypothesized to be a functional signature of stochastic posterior sampling [10–16]. Progress has been made in identifying specific sampling algorithms in recurrent circuits, including Gibbs sampling [12], Langevin sampling [17–19], Hamiltonian sampling [10, 11, 16, 20], and natural-gradient sampling [21, 22], etc.

Yet a conspicuous mismatch exists between these identified algorithms and empirical observations of circuit properties. Most theoretically identified circuit sampling algorithms assume Gaussian dynamics to gain analytical tractability (e.g., [9, 10, 12, 13, 16, 17]), whereas a growing body of

*Corresponding author. wenhao.zhang@utsouthwestern.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

evidence reveals that neural circuit properties are systematically *heavy-tailed*, including synaptic weights [23, 24], single neurons' firing rate [25–27], population response geometry [28], and hippocampal replay sequences during sharp-wave ripples [29, 30], etc. These suggest that heavy-tailed structure is not an incidental feature of neural circuits, but probably has functional significance.

A growing body of theoretical work has investigated how heavy-tailed neuronal responses are generated from circuit dynamics and their roles in Bayesian computation. Predominant strategies import non-Gaussian components into the circuit dynamics to facilitate heavy-tailed neuronal response generation, e.g., using heavy-tailed (log-normal) recurrent synaptic weights [31–34] or explicitly injecting Lévy-stable noise into the dynamics [35]. They found that heavy-tailed neuronal responses can explore the state space more efficiently due to the long-distance transition, and thus accelerate sampling-based Bayesian inference [34, 36]. Nevertheless, in such models, heavy-tailed circuit responses are more or less *inherited* from explicitly non-Gaussian connectivity and/or variability, and it is thus not surprising that these networks produce heavy-tailed responses. A more fundamental question remains unaddressed: can heavy-tailed structure emerge from a network without explicit non-Gaussian components, arising purely from the circuit's nonlinear recurrent dynamics? Furthermore, can such emergent heavy-tailed structure still benefit Bayesian sampling? Answering this requires rigorous analysis of the nonlinear stochastic circuit dynamics, while it remains generally challenging due to the analytical intractability.

The present study answers this question affirmatively through rigorous theoretical analysis of a *minimal*, analytically tractable recurrent circuit model–a continuous attractor network (CAN) widely used in neuroscience [37–40]. Surprisingly, our analysis reveals that the CAN *without* any heavy-tailed components collectively generates heavy-tailed structure in its subspaces satisfying the **scale-mixture Langevin posterior sampling**, an advanced sampling algorithm developed in ML with heavy-tailed step-size structure [41]. Notably, the heavy-tailed structure resides in the *sampling dynamics* with the step-size distribution being heavy-tailed (Student-t), while the *equilibrium distribution* of the sampling still converges to Gaussian posteriors. The circuit mechanism emerges in four steps:

- The bump-profile circuit dynamics has two dominant subspaces: the bump **position** (representing the stimulus feature) and the bump **height** (proportional to overall population activation) (Fig. 2B).
- Due to the nonlinear recurrent dynamics, the two subspaces are **coupled**: the bump height **multiplicatively** modulates the Langevin time constant of the position subspace (Fig. 2C).
- The position subspace executes Langevin sampling of the Gaussian stimulus posterior; the stationary bump-height distribution admits a **Gamma** approximation in the weak-activity regime, arising from divisive normalization (Fig. 2E), 3A).
- The approximately Gamma-distributed height acts as the scale "mixer" for the position step size, yielding a heavy-tailed Student-t step-size distribution, realizing scale-mixture Langevin sampling.

The above emergent structure, including Gamma scale mixer, Gaussian conditional, Student-t marginal, mathematically related to *Gaussian scale mixture* (GSM) model [42–46] widely used in heavy-tailed natural image statistics [47–50], suggesting a deep structural alignment between the circuit's intrinsic dynamics and the natural statistics. This emergent scale-mixture structure significantly accelerates sampling speed relative to a control with bump height fixed at its temporal mean (Fig. 2G), and even admits a Lévy-like regime in the weak input limit (Fig. 3B). Simulations confirm our theory, and reveal the robustness of the scale-mixture structure across a wide range of circuit parameters.

Significance and contribution. 1) A **minimal** circuit mechanism for the endogenous emergence of heavy-tailed sampling without any explicit non-Gaussian components in the circuit dynamics. **2)** The first mechanistic, analytical understanding of how the GSM structure emerges from nonlinear recurrent circuit dynamics. **3)** The first study analytically reveals the scale-mixture Langevin sampling in the circuit's subspace, deepening our understanding of advanced circuit algorithms.

2 The canonical recurrent circuit model

We consider a canonical recurrent circuit model based on continuous attractor networks [16, 37–39] or supralinear stabilization network [11, 51], which are widely used in neuroscience research and can be regarded as a cortical hypercolumn. Briefly, the circuit consists of excitatory (E) neurons and a pool of global inhibitory (I) neurons (Fig. 1). The E neurons tune to stimulus and are recurrently connected, while I neurons are not tuned for simplicity (details in Appendix B).

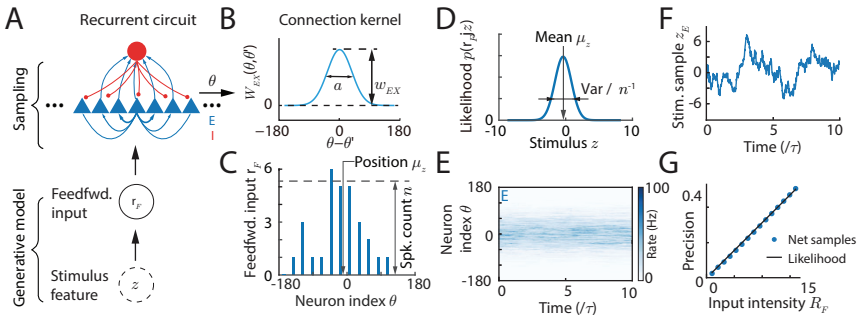


Figure 1: (A) The recurrent circuit of excitatory (E) neurons and inhibitory neurons and sampling-based Bayesian inference. (B) The recurrent and feedforward connection kernel between E neurons. (C) Feedforward (spiking) input with Gaussian tuning across the stimulus feature. (D) The stimulus feature likelihood given a feedforward input. (E) E population responses. (F) Stimulus feature samples from E neurons. (G) The circuit with fixed weights flexibly samples posteriors with different uncertainties in the equilibrium.

E neurons' dynamics. The E neurons in the circuit are selective for a 1D stimulus feature z, e.g., spatial location, and later we extend into the case of a 2D stimulus feature. Denote θ_j as the preferred stimulus feature of the j-th E neuron, and the preferred feature of all N_E E neurons, $\{\theta_j\}_{j=1}^{N_E}$ uniformly cover the entire stimulus feature space z, which is widely considered in neuroscience research. Mathematically, in the continuum limit ($\theta_j \to \theta$) corresponding to an infinite number of neurons, the dynamics of E neurons is,

$$\tau \dot{\mathbf{u}}_E(\theta,t) = -\mathbf{u}_E(\theta,t) + \rho \sum_{X=E,F} (\mathbf{W}_{EX} * \mathbf{r}_X)(\theta,t) + \sqrt{\tau \mathsf{F}[\mathbf{u}_E(\theta,t)]_+}\xi(\theta,t), \tag{1}$$

where $\mathbf{u}_E(\theta,t)$ and $\mathbf{r}_E(\theta,t)$ represent, respectively, the synaptic inputs and firing rates of neurons preferring $z=\theta$. $X=E,F$ denotes E neurons and feedforward sensory. τ is the time constant, $\rho = N/2\pi$ is the neuronal density covering the stimulus feature space, and $[x]_+ = \max(x,0)$. F is the Fano factor characterizing the internal Poisson variability mimicking spike generation.

Recurrent connection kernel. $\mathbf{W}_{YX}(\theta)$ is the recurrent connection kernel from neurons with type X to those with type Y, which are modeled as Gaussian functions in the model (Fig. 1B),

$$\mathbf{W}_{YX}(\theta) = w_{YX}\left(\sqrt{2\pi}a\right)^{-1} \exp(-\theta^2/2a^2), \tag{2}$$

where the scalar w_{YX} is the peak recurrent weight strength and is the only circuit parameter to be adjusted for realizing Bayesian sampling. and a the connection width across the stimulus feature space. The symbol $*$ denotes the convolution, i.e., $\mathbf{W}(\theta) * \mathbf{r}(\theta) = \int \mathbf{W}(\theta-\theta')\mathbf{r}(\theta')d\theta'$.

Divisive normalization from inhibitory neurons. I neurons are driven by E neurons and then provide global, unstructured inhibitory feedback. To facilitate the math analysis to obtain the circuit algorithm, we don't explicitly model I neurons' dynamics but absorb their effect into the divisive normalization that acts as E neurons' activation function, a canonical cortical operation [52–55].

$$\mathbf{r}_E(\theta,t) = [\mathbf{u}_E(\theta,t)]_+^2/(1+w_{EI}r_I), \;\; r_I = \rho \int [\mathbf{u}_E(\theta',t)]_+^2 d\theta', \tag{3}$$

where r_I (scalar) corresponds to the mean firing rate of the I neurons. w_{EI} (a positive scalar) characterizes the global inhibition strength. We assume that inhibitory activity relaxes rapidly relative to the collective bump-position and bump-height dynamics, allowing the normalizing signal to be approximated by its quasi-steady value. This assumption is motivated by the rapid signaling properties of parvalbumin-expressing (PV) interneurons [56] and their experimentally observed role in modulating cortical response gain [57, 58]. Finite inhibitory time constants and independent inhibitory variability are omitted from the present analysis.

Stochastic sensory feedforward input. The feedforward input $\mathbf{r}_F(\theta,t)$ (Eq. 1) is stochastically evoked by an external stimulus z, and is modeled as conditionally independent Poisson spikes with Gaussian tuning given a stimulus z (Fig. 1C), consistent with neural coding models (e.g., [6, 12]).

$$\mathbf{r}_F(\theta|z) \sim \text{Poisson}[\lambda_F(\theta|z)], \quad \lambda_F(\theta|z) = R_F \exp[-(\theta-z)^2/2a^2], \tag{4}$$

where $\lambda_F(\theta|z)$ is the mean firing rate. In simulating the rate-based circuit dynamics, $\mathbf{r}_F$ is approximated as a continuous Gaussian random variable with multiplicative noise.

2.1 The generative model and posteriors to be computed by the circuit

The stage from the external stimulus feature z to the feedforward input $\mathbf{r}_F$ is regarded as the generative process of the circuit's observation (Fig. 1A). To focus on the mechanistic understanding of the scale-mixture structure in the nonlinear recurrent dynamics, we consider a uniform prior over the stimulus feature z, as previous studies also suggested the recurrent circuits with translation-invariant connectivity also store an internal uniform prior [12]. Later we extend the circuit to the case of a non-uniform prior and show the scale-mixture structure still holds (Appendix G.1 and Fig. A2).

With the uniform prior, the stimulus posterior $\pi(z)$ is proportional to the likelihood $\mathcal{L}(z)$,

$$\pi(z) \equiv p(z|\mathbf{r}_F) \propto p(\mathbf{r}_F|z)\, p(z) \propto p(\mathbf{r}_F|z) \equiv \mathcal{L}(z). \tag{5}$$

From Eq. 4, the stimulus likelihood is derived as a Gaussian distribution (Appendix B.4),

$$\begin{aligned} &\mathcal{L}(z) \propto \textstyle\prod_\theta \mathrm{Poisson}[\lambda_F(\theta|z)] \propto \mathcal{N}\big(z|\mu_z, \Lambda^{-1}\big), \\ &\text{where } \ \mu_z = \textstyle\sum_j \mathbf{r}_F(\theta_j)\theta_j / \sum_j \mathbf{r}_F(\theta_j), \ \ \Lambda = a^{-2}\textstyle\sum_j \mathbf{r}_F(\theta_j) = \sqrt{2\pi}\rho a^{-1} R_F. \end{aligned} \tag{6}$$

Note that the Gaussian stimulus likelihood is not directly assigned, but resulted from the Gaussian profile of feedforward input tuning $\lambda_F(\theta|z)$ (Eq. 4) [6, 8]. Its mean μ_z and the precision Λ are regarded as $\mathbf{r}_F$'s location in the stimulus feature space and height (total input spike count) (Fig. 1C-D), and can be read out from $\mathbf{r}_F$ via a linear decoder called population vector [59, 60],

3 Naive Langevin sampling in the circuits

3.1 Theoretical analysis of the circuit dynamics to understand circuit algorithm

Mechanistic understanding of the circuit's algorithm requires rigorous theoretical analysis of the nonlinear recurrent dynamics, which is a major challenge in the field. The recurrent circuit model in the present study exists **analytical** solutions [39, 61], including attractor states, full eigenspectrum of the perturbation dynamics. This enables us to project the high-dimensional circuit dynamics onto the eigenmodes (subspaces), and then the subspace dynamics will be compared and linked to Bayesian sampling algorithms. Below, we briefly introduce the key steps and results (details in Appendix B).

Attractors. E neurons in canonical circuit dynamics have the following attractor states with a bump profile over the stimulus feature space (Fig. 1E; Appendix B),

$$\bar{\mathbf{u}}_E(\theta) = \bar{U}_E \exp[-(\theta - \bar{z}_E)^2/4a^2], \quad \bar{\mathbf{r}}_E(\theta) = \bar{R}_E \exp[-(\theta - \bar{z}_E)^2/2a^2]. \tag{7}$$

In contrast, I neurons don't have a spatial bump profile since they are not tuned to stimulus (Eq. 3).

Sampling dynamics on the stimulus feature subspace. The first dominant eigenmode of the circuit dynamics corresponds to the change of bump position z_E (Appendix B, [39, 61]). z_E is regarded as the circuit's internal representation of stimulus feature. Projecting the E dynamics (Eq. 1) onto the position eigenvector (calculating the inner product of the circuit dynamics and the eigenvectors) yields the dynamics of z_E (details in Appendix B.3),

$$\text{Position eigenvec.}: \ \ \phi_1(\theta|z_E) \propto (\theta - z_E) e^{-(\theta - z_E)^2/4a^2} \tag{8a}$$

$$\text{Position dynamics}: \ \ \dot{z}_E \approx (\tau U_E)^{-1} U_{EF}(\mu_z - z_E) + \sigma_z (\tau U_E)^{-1/2} \xi_{z,t}, \tag{8b}$$

$$\text{where } \ U_{EF} = \rho w_{EF} R_F / \sqrt{2} \propto w_{EF}\Lambda$$

where U_E is the height of the bump attractor (Eq. 7) and determines the time constant of the z_E dynamics that will affect the average step size Δz_E over time (Fig. 2C). U_{EF} characterizes the magnitude (height) of the received feedforward input, and is proportional to the likelihood precision Λ (Eq. 6). This is a key property to link the position subspace dynamics to the Langevin sampling that will be shown below. Moreover, $\sigma_z^2 = 8a\mathsf{F}/(3\sqrt{3\pi})$ is a constant determining the strength of internal variability projected onto the position subspace, and doesn't change with network activities.

3.2 Naive Langevin sampling in the circuits

We explore how the position dynamics (Eq. 8b) can sample the posterior $\pi(z)$, i.e., its equilibrium distribution should converge to $\pi(z)$. By utilizing the proportionality between U_{EF} and Λ (Eq. 8b),

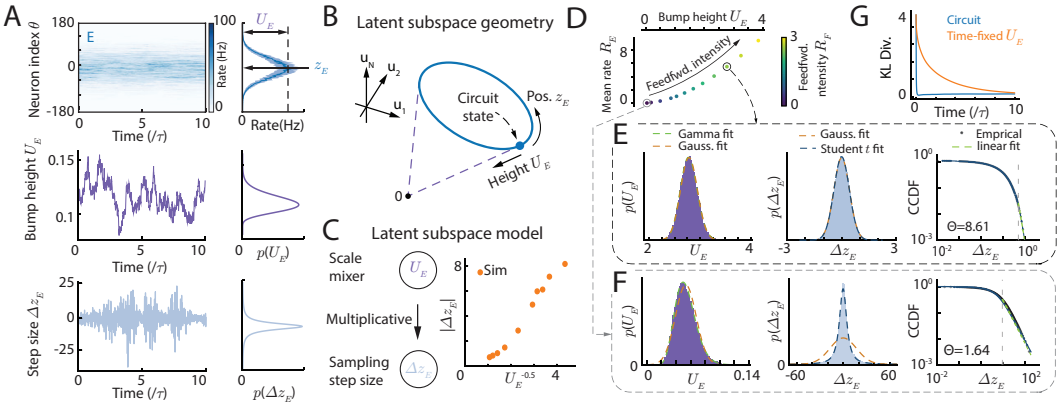


Figure 2: Scale-mixture Langevin sampling dynamics in circuit subspaces. (A) E population responses and the trajectories of bump height U_E and bump position step size Δz_E linearly read out from instant E responses. (B-C) The latent subspace geometry and model from theoretical analysis. The position and height subspaces are intertwined, where the height U_E multiplicatively modulates the position step size Δz_E. (D-F) The heavy-tailed sampling step size emerges at the limit of low circuit activation. (D): The population synaptic input height and population firing rate height with changing feedforward input intensity. (E-F): The distributions of bump height (left) and the sampling step size (middle), and the complementary cumulative distribution function of the step size (right). (G) The heavy-tailed step sizes accelerate sampling speed compared with the sampling with time-fixed U_E.

the position dynamics can be converted in conventional Langevin posterior dynamics [62, 63],

$$\dot{z}_E = (\tau U_E)^{-1} \underbrace{\left(a\, w_{EF} (2\sqrt{\pi})^{-1}\right)}_{\lambda_z} \underbrace{\Lambda(\mu_z - z_E)}_{\ln \pi(z_E)} + \sigma_z (\tau U_E)^{-1/2} \xi_t, \tag{9}$$

where the feedforward input $\mathbf{r}_F$ provides the likelihood score function (Eq. 9, drift term), and the internal Poisson variability (Eq. 1, last term) provides the variability for sampling (Eq. 9, diffusion term). To match the equilibrium distribution of z_E to the posterior $\pi(z)$, we calibrate the balance between positional drift and diffusion by setting the feedforward weight (Appendix Sec. B.5):

$$w^*_{EF} = \frac{\sigma_z^2}{\sqrt{2}\rho} \frac{\Lambda}{R_F} = \left(\frac{2}{\sqrt{3}}\right)^3 \mathsf{F}. \tag{10}$$

Once calibrated, the circuit can sample posteriors with different means and precisions without further changes to its synaptic weights (Fig. 1G). The internal Poisson variability enables the drift and diffusion terms in Eq. (9) share the same factor U_E, a necessary condition for Langevin sampling.

4 The scale-mixture Langevin sampling in the circuit

The above analyses of the Langevin sampling (Eq. 8b) implicitly treats the bump height U_E as a constant, free parameter. In practice, U_E is also a stochastic variable fluctuating over time due to the internal variability (Fig. 2A). Since the U_E governs the sampling time constant, its fluctuations will directly modulate the sampling step size and may generate complex sampling behaviors.

4.1 The bump height dynamics and its equilibrium distribution

We next analyze the dynamics of the bump height U_E and then find its fluctuation statistics. Previous studies identified the change of bump height U_E as the second dominant eigenmode of the circuit dynamics (Appendix B, [39, 61]), and its dynamics is derived by projecting the circuit dynamics onto the height eigenvector (details in Appendix B.3),

$$\text{Height eigenvec.}: \quad \phi_1(\theta|z_E) \propto e^{-(\theta - z_E)^2/4a^2} \tag{11a}$$

$$\text{Height dynamics}: \quad \dot{U}_E \approx \tau^{-1}\left(-U_E + U_{EE} + U_{EF}\right) + \sigma_U (\tau^{-1} U_E)^{1/2} \xi_{U,t}, \tag{11b}$$

where $U_{XY} = \rho w_{XY} R_Y/\sqrt{2}$. $\sigma_U^2 = \mathsf{F}/(\sqrt{3\pi}a)$ is a constant determining the strength of internal variability projected onto the height subspace, and doesn't change with network activities.

Dependency between position and height subspaces. Combining Eqs. (8b and 11b), it is worth noting that the U_E dynamics governs the step size of z_E dynamics (Fig. 2C), and hence the two dynamics are not independent, even if the eigenvectors of bump position and height (Eqs. 8a and 11a) are *locally* orthogonal. In the nonlinear attractor dynamics, both position and height eigenvectors depend on the instantaneous z_E (Eqs. 8a, 11a). The dependency over subspaces provides an important backbone supporting the scale-mixture structure in the sampling dynamics, as will be shown below.

The equilibrium bump height distribution. We use the Fokker-Planck approach to determine the equilibrium $p(U_E)$ that can be directly formulated based on the Langevin equation of U_E (Eq. 11b):

$$\frac{\partial}{\partial t}p(U_E,t) = -\frac{\partial}{\partial U_E}\left[\frac{1}{\tau}(-U_E + U_{EE} + U_{EF})\,p(U_E,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial U_E^2}\left[\frac{\sigma_U^2}{\tau}U_E\,p(U_E,t)\right].$$

Setting $\partial p/\partial t = 0$, the equilibrium distribution can be analytically solved (Appendix C.1),

$$p(U_E) \propto U_E^{\frac{\sqrt{2}\rho w_{EF}R_E}{\sigma_U^2}-1}(1+\rho w_{EI}U_E^2)^{\frac{w_{EE}}{\sqrt{2}w_{EI}\sigma_U^2}}\exp\Big(-\frac{2}{\sigma_U^2}U_E\Big). \tag{12}$$

In the circuit model, the global inhibition strength w_{EI} is typically a small positive value (e.g., $w_{EI} = 5\times10^{-4}$) to enable the circuit produces physiologically realistic firing rates. This leads to the $\rho w_{EI}U_E^2$ within the parenthesis way smaller than 1 in the main support of the distribution. Ignoring it allows us to obtain a convenient approximation of $p(U_E)$ that satisfies a Gamma distribution, i.e.,

$$p(U_E) \overset{\propto}{\approx} U_E^{\overbrace{\sqrt{2}\rho w_{EF}R_F/\sigma_U^2}^{\alpha_1}-1}\exp\Big(-\overbrace{\frac{2}{\sigma_U^2}}^{\beta_1}U_E\Big) \propto \mathrm{Gamma}\left(U_E|\alpha_1,\beta_1\right), \tag{13}$$

Moreover, to confirm $p(U_E)$ satisfies the Gamma distribution, we compare its goodness of fit to a Gaussian fit (Fig. A1B), which reveals that the Gamma distribution fits better than the Gaussian for most of the parameter regime especially the low-input regime. Lower input R_F leads to a smaller shape parameter α_1, resulting in a highly skewed distribution. Conversely, at higher input regimes, the increased α_1 causes the distribution to naturally converge toward a Gaussian-like bell curve.

4.2 Scale-mixture Langevin sampling in the position subspaces

Gaussian-conditional position step-size distribution. We next analyze how the height U_E fluctuations affect the position z_E dynamics. **First**, under the matching condition in Eq. (10), U_E fluctuations rescale positional drift and diffusion variance equally, preserving the posterior $\pi(z_E)$ within the leading-order reduction. This preservation does not depend on the Gamma approximation for the height distribution. **Second**, U_E fluctuations will significantly reshape the **sampling trajectories** of z_E, especially its increment or step size, $\Delta z_E(t) = z_E(t+\Delta t) - z_E(t)$. To see this effect, we discretize the position z_E dynamics using the Euler-Maruyama method with a time step Δt:

$$\Delta z_E \approx \underbrace{\left[\tau^{-1}\lambda_z\Lambda(\mu_z - z_E)\Delta t\right]}_{\text{Drift } f(z_E)}U_E^{-1} + \underbrace{(\sigma_z^2\tau^{-1}\Delta t)^{1/2}}_{\text{Diffusion } \sqrt{D}}U_E^{-1/2}\,\xi_{z,t}, \qquad \xi_{z,t}\sim\mathcal{N}(0,1)$$

where we define the drift term $f(z_E) = \tau^{-1}\lambda_z\Lambda(\mu_z - z_E)\Delta t$ and the diffusion variance parameter $D = \sigma_z^2\tau^{-1}\Delta t$. Conditional on the current height and position, the increment is Gaussian. Neglecting the drift term in the small Δt limit gives

$$p(\Delta z_E|U_E) = \mathcal{N}\left(f(z_E)U_E^{-1}, D\,U_E^{-1}\right) \approx \mathcal{N}\left(0, DU_E^{-1}\right)\ \text{(small } \Delta t \text{ limit)}. \tag{14}$$

It clearly shows that Δz_E depends on the bump height U_E in a **multiplicative** way (Fig. 2C), which is the key mechanism generating the scale-mixture structure in the sampling dynamics. In the limiting case when Δt is sufficiently small, the drift term $f(z_E)$ scales linearly with Δt, while the diffusion standard deviation $\sqrt{D}$ scales with $\sqrt{\Delta t}$. Thus, the drift term is of higher order than the diffusion term and becomes negligible. Eventually, the conditional distribution converges to a zero-mean Gaussian distribution with variance inversely scaling with U_E (Eq. 14). Fig. 2A confirms our theoretical derivation, showing that $p(\Delta z_E)$ can largely be approximated as a zero-mean Gaussian distribution with variance inversely scaling with U_E.

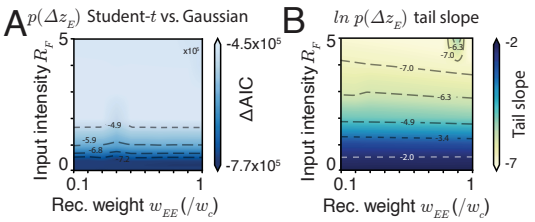


Figure 3: (A) The comparison of student-t vs Gaussian fit the sampling step size distribution $p(\Delta z_E)$. A negative ΔAIC means student-t provides a better fit. (B) Estimated log-log CCDF slopes across the same parameter grid. Under the Student-t approximation, $\Pr(|\Delta z_E| > r) \propto r^{-\nu_1}$ asymptotically, so the CCDF tail slope is $-\nu_1$ in Eq.(15).

Marginal position step-size Δz_E distribution. The marginal step-size distribution is obtained by integrating over the Gamma height distribution $p(U_E)$ (Eq. 13). This integration reveals a **Gaussian scale mixture** structure, where the bump height U_E serves as the scale "mixer" variable [42, 43, 45]:

$$p(\Delta z_E) = \int_0^\infty p(\Delta z_E \mid U_E) p(U_E) dU_E \propto \left(1 + \frac{\Delta z_E^2}{2\beta_1 D}\right)^{-\left(\alpha_1 + \frac{1}{2}\right)} \propto t_{\nu_1}(0, s_1^2) \tag{15}$$

Note that this calculation implicitly utilizes the independent fluctuations of U_E and z_E, i.e., $\xi_{z,t}$ and $\xi_{U,t}$ (Eqs. 8b and 11b) are independent. Although the height and position fluctuations, $\xi_{z,t}$ and $\xi_{U,t}$, arise from the same internal noise (Eq. 1, last term), they are independent because the position and height subspaces' eigenvectors are locally orthogonal (Eqs. 8a and 11a).

The resulting marginal $p(\Delta z_E)$ is a Student-t distribution defined by two parameters: degrees of freedom (DoF) $\nu_1 = 2\alpha_1$ and a scale parameter $s_1^2 = D\beta_1/\alpha_1$ (details in Appendix C.3).

- The DoF ν_1 uniquely controls the sharpness of the central peak and the heaviness of the tail. When ν_1 is small (near 1), it produces a much sharper central peak than a Gaussian, combined with very thick, heavy tails where the probability density decays polynomially rather than exponentially. This regime accommodates robust sampling and extreme outliers.
- The scale parameter s_1 controls the statistical dispersion, acting as a horizontal dilation or compression factor. Increasing s_1 stretches the distribution outwards from the center and simultaneously lowers the peak density to ensure the total area under the curve remains exactly 1.

To empirically validate the scale-mixture model, we analyze how feedforward input intensity R_F shapes the sampling geometry (Fig. 2D). Under weak input intensity R_F, the network operates in a superlinear regime where bump-height distribution $p(U_E)$ is more Gamma-like. This generates highly sparse, heavy-tailed marginal step-size distributions $p(\Delta z_E)$ that strongly favor a Student-t fit over a Gaussian baseline (Fig. 2F). Conversely, as R_F increases, the network exits this superlinear phase, and the step-size fluctuations become increasingly Gaussian-like (Fig. 2E).

Network simulations confirm the Student-t distribution of the position increments Δz_E across a wide range of parameters (Fig. 3A). This is demonstrated by comparing the difference in the Akaike Information Criterion (ΔAIC) between the Student-t and Gaussian models for the step-size segments. The ΔAIC is consistently and heavily negative across all evaluated feedforward input intensities R_F and recurrent weights w_{EE}, strongly penalizing the Gaussian assumption. This shows that the heavy-tailed Student-t distribution is a significantly better fit, particularly in weakly activated regimes where the disparity reaches its maximum magnitude.

4.3 Heavy-tailed dynamics accelerate sampling convergence

The heavy-tailed step-size Δz_E distribution enables the circuit sampling to execute occasional large jumps (Fig. 2F), facilitating rapid exploration of the stimulus subspace. As shown in Fig. 2G, the circuit sampling converges much faster to the posteriors compared to a Langevin sampler whose the bump height U_E is fixed at the temporal average. This suggests the bump height fluctuations can benefit the Bayesian sampling, rather than merely noise that degrades the sampling performance.

4.4 Lévy-like regime in the weak input limit

Systematic evaluation across the circuit parameter space reveals a distinct Lévy-like exploration phase (Fig. 3). This behavior is typically approximated using an α-stable distribution, where a stability parameter Θ between 0 and 2 defines the Lévy-like regime. To **robustly** quantify this

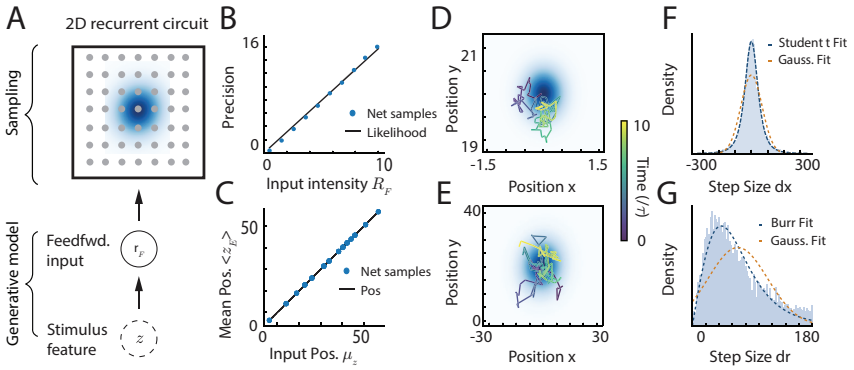


Figure 4: (A) The 2D recurrent circuit. (B) Precision of the network samples compared to the theoretical likelihood across different input intensities R_F. (C) Mean position of the network samples compared to the actual input position. (D-E) The decoded 2D trajectories of sampled positions over time. The network can sample different posteriors with varying precision, showing a more concentrated trajectory in with $R_F = 2.5$ in (D) and a wider spread with $R_F = 0.01$ in (E). Note the difference in their x-y coordinate range. (F) Probability density of step size on each dimension (dx), overlaid with Student-t and Gaussian fits. (G) The distribution of radial step size (dr), overlaid with custom Burr fit (Eq. 19) and Gaussian fit.

heavy-tailed behavior, we estimate the CCDF of the step sizes, $P(\Delta z > a) = \int_a^\infty p(\Delta z) d\Delta z$, which is a parameter-free method, unlike the estimation of empirical probability density function (PDF) that needs a bin size. Importantly, a power-law tail in the CCDF $\Delta z^{-\Theta}$ corresponds to a power-law probability density tail, $p(\Delta z) \propto \Delta z^{-(\Theta+1)}$ (Fig. 2F) [64]. Therefore we will estimate the probability density tail through estimating the tail of CCDF.

As shown in Fig. 3B, this Lévy-like behavior emerges in the weak input limit (low feedforward input R_F or weak recurrent weight w_{EE}). In this regime, the low baseline bump height allows multiplicative internal noise to induce massive transient fluctuations, propelling the network state across vast distances in the stimulus feature space.

5 Scale-mixture sampling in two-dimensional recurrent circuits

To demonstrate the generality of the identified scale-mixture sampling mechanism in recurrent circuits, we extend our analysis from the 1D recurrent circuit (the preferred stimulus θ is 1D, Eq. 1) to 2D recurrent circuits whose preferred stimulus $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in (-\pi, \pi)^2$ is 2D. Similarly, we consider $\boldsymbol{\theta}$ are uniformly distributed across the whole 2D stimulus feature space and we consider the periodic boundary of either θ_i to avoid boundary effect. The whole circuit dynamics is basically the same as the 1D case, as we only need to replace the 1D preferred stimulus θ and the external stimulus feature z by the 2D preferred stimulus $\boldsymbol{\theta}$ and $\mathbf{z} = (z_1, z_2)^\top$ respectively from Eqs. (1-4) (details in Appendix D). The precision and mean can be flexibly sampled by the network (Fig. 4B-C).

5.1 Dimensionality reduction for the joint bump position and height dynamics

Following the same theoretical approach in the 1D case, the dynamics of the 2D bump position $\mathbf{z}_E$ and height U_E (still 1D) can be similarly derived (details in Appendix B.1),

$$\text{Position}: \quad \dot{\mathbf{z}}_E \approx (\tau U_E)^{-1} U_{EF}(\boldsymbol{\mu}_z - \mathbf{z}_E) + \sigma_z (\tau U_E)^{-1/2} \boldsymbol{\xi}_{\mathbf{z},t} \tag{16a}$$

$$\text{Height}: \quad \dot{U}_E \approx \tau^{-1}\left(-U_E + U_{EE} + U_{EF}\right) + \sigma_U (\tau^{-1} U_E)^{1/2} \xi_{U,t} \tag{16b}$$

Several key parameters governing the amplitude dynamics remain one-dimensional (scalar) quantities, preserving the core structure of the 1D circuit: U_E represents the peak amplitude (height) of the 2D Gaussian bump. Regardless of the spatial dimensionality the bump spans, its maximum firing rate is a single scalar value. R_F is the maximum intensity (peak firing rate) of the external feedforward stimulus, which acts as a scalar scaling factor. U_{EF} represents the effective feedforward input drive. Because it is obtained by integrating the projection of the external input onto the bump's spatial profile over the 2D space, the resulting magnitude is inherently 1D.

Similarly, the bump height U_E multiplicatively modulates the time constant of the position dynamics, while leaving the equilibrium distribution of $\mathbf{z}_E$ invariant. Performing similar analyses as the 1D case, the equilibrium distribution of the bump height can be approximated as a Gamma distribution:

$$U_E \sim \text{Gamma}\left(\alpha_2, \beta_2\right), \quad \text{shape: } \alpha_2 = \frac{\rho w_{EF} R_F}{\sigma_U^2}, \quad \text{rate: } \beta_2 = \frac{2}{\sigma_U^2} \tag{17}$$

Empirical fits from network simulations confirm that U_E exhibits this expected long-tailed behavior rather than a symmetric Gaussian profile (Fig. A3A).

5.2 Gaussian scale mixture induces 2D heavy-tailed spatial exploration

The functional consequence of these bump height fluctuations is a direct modulation of the local 2D diffusivity. Over a short time interval Δt, conditioned on a fixed bump height U_E, the positional increment is an isotropic bivariate Gaussian (Appendix D). To determine the sampling step size statistics of the network, we integrate this conditional Gaussian over the Gamma distribution of U_E, resulting in an isotropic bivariate Student-t distribution (Fig. 4F):

$$\Delta \mathbf{z}_E \sim t_\nu \left(\mathbf{0},\ s_2^2 \mathbf{I}_2\right) \tag{18}$$

with degrees of freedom $\nu = 2\alpha_2$ and scale parameter $s_2^2 = D\beta_2/\alpha_2$.

A defining feature of the 2D sampling dynamics is the jump magnitude (radial displacement) $r = \|\Delta \mathbf{z}_E\|$ is a nonlinear function of the step size along each dimension. This transforms the bivariate Student-t distribution into another heavy-tailed distribution called the Burr distribution (Fig. 4G):

$$p_r(r) = \frac{2\alpha_2}{A} r \left(1 + \frac{r^2}{A}\right)^{-(\alpha_2+1)} \quad \text{where} \quad A = 2\beta_2 D. \tag{19}$$

These marginal distributions demonstrate that the scale-mixture mechanism generalizes robustly to higher dimensions (Fig. 4F-G). Crucially, the heavy-tailed radial jump length r ensures that the network is not constrained to slow, purely diffusive random walks. Instead, it allows the 2D network to routinely execute large, multi-dimensional exploratory leaps, facilitating the rapid traversal of complex, high-dimensional joint posterior landscapes.

6 Conclusion and Discussion

The present study rigorously reveals the nonlinear CAN dynamics without explicit non-Gaussian components can collectively emerge the scale-mixture Langevin dynamics to sample the stimulus posterior in the circuit's subspaces. The scale-mixture structure significantly speeds up the sampling convergence in the circuit, and it can enter the Lévy-like regime under the limit of low circuit activation. Our study provides a minimal circuit mechanism for heavy-tail statistics generation, suggesting a deep structural alignment between the circuit's intrinsic dynamics and the GSM used in describing natural statistics[2]. To the best of our knowledge, this is the first theoretical work to analytically identify the GSM structure in nonlinear recurrent circuit dynamics without non-Gaussian components.

Comparison with previous studies. First, previous studies have investigated heavy-tailed neural trajectories through a few different mechanisms: One is directly injecting Lévy-stable noise [35]; another is emerging from collective spiking-network dynamics, facilitated via heavy-tailed recurrent weights, whose operating points transit across the bifurcation boundary between the stable and traveling population activity back and forth [32–34, 36]. Our work complements these accounts by analytically identifying a scale-mixture mechanism in which bump-height fluctuations modulate positional drift and diffusion, linking circuit parameters to the resulting heavy-tailed increment statistics. Although a previous study also showed that the heavy-tailed structure can emerge in recurrent networks without non-Gaussian components [36], its circuit model requires extra neuronal adaptation that is absent in our model, and it didn't realize the posterior sampling that requires another constraint of the equilibrium distribution. **Second**, although our circuit model is similar to previous studies of circuit sampling [16, 19, 22], none of them explored the heavy-tailed sampling dynamics. **Third**, although heavy-tailed sampling algorithms were proposed in ML studies with top-down

[2] In our circuit, the GSM structure occurs on the sampling step size, while the posteriors sampled by the circuit are Gaussian.

design [41], these algorithms have been rarely embedded into biologically plausible recurrent circuit. Our work provides a minimal, biological circuit mechanism for emerging heavy-tailed sampling.

Limitations, generalizations, and future directions. **First**, on the circuit model level, the proposed circuit model doesn't include diverse types of inhibitory neurons in canonical cortical circuits, including PV, SOM, and VIP neurons. And the inhibitory neuron pool that provides unstructured, divisive inhibition in our circuit model is regarded as the PV neurons [53, 55]. We believe adding the SOM neurons that provide structured, subtractive inhibition into the circuit model could enrich the dynamical repertoire of the circuit dynamics and algorithms, with a likely potential of enlarging the Lévy-like regime in the circuit parameter space. This was suggested by a previous study [36] whose neural adaptation is similar to the inhibition from SOM neurons. Moreover, the VIP neurons exclusively inhibit SOM neurons and can act as a "knob" to switch the circuit algorithms. **Second**, on the types of posterior distributions, as a proof-of-concept example, we only demonstrate the 1D and 2D Gaussian posterior sampling in our circuit model, and emphasize how the circuit internally generates heavy-tailed sampling dynamics to sample Gaussian posteriors. Based on the neural coding theory, the circuit can be easily extended to utilize scale-mixture Langevin sampling to sample other types of posteriors, as long as we change the spatial profile of recurrent kernel (Eq. 2) and feedforward input tuning (Eq. 4). Here, the spatial profile effectively acts as the natural parameter of the sampling distribution [6, 12]. **Finally**, our circuit model theory produces testable experimental predictions. We show the scale-mixture sampling will have heavier tails as the circuit activation decreases (Fig. 2E). This is generally supported by spatial trajectories of hippocampal place cells: the replay sequences decoded during sharp-wave ripples, during which the mean firing rate is lower, exhibit heavier tails of step size than the ones of theta sequences when the mean firing rate is higher [29, 30, 65].

Acknowledgments and Disclosure of Funding

W.H.Z. is supported by the UT Southwestern Endowed Scholars program. The authors thank K.Y. Michael Wong and Ying Nian Wu on early formation of this project. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MPS-NITMB-00005320) to the NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB).

References

[1] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[2] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

[3] Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

[4] Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

[5] Jeffrey M Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K Churchland, Jamie Roitman, Michael N Shadlen, Peter E Latham, and Alexandre Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[6] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

[7] József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.

[8] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170, 2013.

[9] Patrik O Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. In *Advances in neural information processing systems*, pages 293–300, 2003.

[10] Laurence Aitchison and Máté Lengyel. The hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, 12(12), 2016.

[11] Rodrigo Echeveste, Laurence Aitchison, Guillaume Hennequin, and Máté Lengyel. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, 2020.

[12] Wen-Hao Zhang, Si Wu, Krešimir Josić, and Brent Doiron. Sampling-based bayesian inference in recurrent circuits of stochastic spiking neurons. *Nature Communications*, 14(1):7074, 2023.

[13] Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology*, 7(11):e1002211, 2011.

[14] 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.

[15] Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

[16] Eryn Sale and Wenhao Zhang. The bayesian sampling in a canonical recurrent circuit with a diversity of inhibitory interneurons. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[17] Cristina Savin and Sophie Denève. Spatio-temporal representations of uncertainty in spiking neural networks. *Advances in neural information processing systems*, 27, 2014.

[18] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D Miller. The dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.

[19] Wenhao Zhang, Tai Sing Lee, Brent Doiron, and Si Wu. Distributed sampling-based bayesian inference in coupled neural circuits. *bioRxiv*, 2020.

[20] Xingsi Dong, Zilong Ji, Tianhao Chu, Tiejun Huang, Wenhao Zhang, and Si Wu. Adaptation accelerating sampling-based bayesian inference in attractor neural networks. *Advances in Neural Information Processing Systems*, 35:21534–21547, 2022.

[21] Paul Masset, Jacob Zavatone-Veth, J Patrick Connor, Venkatesh Murthy, and Cengiz Pehlevan. Natural gradient enables fast sampling in spiking neural networks. *Advances in neural information processing systems*, 35:22018–22034, 2022.

[22] Zimei Chen, Yi Ren, Eryn Sale, Ying Nian Wu, and Wen-Hao Zhang. Natural gradient bayesian sampling automatically emerges in canonical cortical circuits. *bioRxiv*, pages 2025–07, 2025.

[23] Boris Barbour, Nicolas Brunel, Vincent Hakim, and Jean-Pierre Nadal. What can we learn from synaptic weight distributions? *TRENDS in Neurosciences*, 30(12):622–629, 2007.

[24] Christopher W Lynn, Caroline M Holmes, and Stephanie E Palmer. Heavy-tailed neuronal connectivity arises from hebbian self-organization. *Nature Physics*, 20(3):484–491, 2024.

[25] György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014.

[26] Luke Woloszyn and David L Sheinberg. Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. *Neuron*, 74(1):193–205, 2012.

[27] Kenji Mizuseki and György Buzsáki. Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex. *Cell reports*, 4(5):1010–1021, 2013.

[28] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

[29] Brad E Pfeiffer and David J Foster. PLACE CELLS. autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science*, 349(6244):180–183, July 2015.

[30] Daniel C McNamee, Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature neuroscience*, 24(6):851–862, 2021.

[31] Yoshiyuki Omura, Milena M Carvalho, Kaoru Inokuchi, and Tomoki Fukai. A lognormal recurrent network model for burst generation during hippocampal sharp waves. *The Journal of Neuroscience*, 35(43):14585–14601, 2015.

[32] Guozhang Chen and Pulin Gong. Computing by modulating spontaneous cortical activity patterns as a mechanism of active visual processing. *Nature communications*, 10(1):4915, 2019.

[33] Yifan Gu, Yang Qi, and Pulin Gong. Rich-club connectivity, diverse population coupling, and dynamical activity patterns emerging from local cortical circuits. *PLoS computational biology*, 15(4):e1006902, 2019.

[34] Yang Qi and Pulin Gong. Fractional neural sampling as a theory of spatiotemporal probabilistic computations in neural circuits. *Nature communications*, 13(1):1–19, 2022.

[35] Umut Simsekli. Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov chain Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3200–3209. PMLR, 2017.

[36] Xingsi Dong, Tianhao Chu, Tiejun Huang, Zilong Ji, and Si Wu. Noisy adaptation generates lévy flights in attractor neural networks. *Advances in neural information processing systems*, 34:16791–16804, 2021.

[37] R Ben-Yishai, R Lev Bar-Or, and H Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[38] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *The Journal of Neuroscience*, 16(6):2112–2126, 1996.

[39] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016.

[40] Mikail Khona and Ila R. Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, December 2022.

[41] Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. Bayesian sampling using stochastic gradient thermostats. *Advances in neural information processing systems*, 27, 2014.

[42] Martin J Wainwright and Eero P Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems*, volume 12, pages 855–861, 2000.

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

[44] Odelia Schwartz, Terrence J Sejnowski, and Peter Dayan. Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural computation*, 18(11):2680–2718, 2006.

[45] Ruben Coen-Cagli, Peter Dayan, and Odelia Schwartz. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS computational biology*, 8(3):e1002405, 2012.

[46] Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11):1648–1655, 2015.

[47] Daniel L Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

[48] 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, 1996.

[49] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[50] Yan Karklin and Michael S Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–86, 2009.

[51] Daniel B Rubin, Stephen D Van Hooser, and Kenneth D Miller. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015.

[52] Sophie Deneve, Peter E Latham, and Alexandre Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2(8):740–745, 1999.

[53] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[54] Wen-Hao Zhang, Aihua Chen, Malte J Rasch, and Si Wu. Decentralized multisensory information integration in neural systems. *Journal of Neuroscience*, 36(2):532–547, 2016.

[55] Cristopher M Niell. Cell types, circuits, and receptive fields in the mouse visual cortex. *Annual review of neuroscience*, 38:413–431, 2015.

[56] Hua Hu, Jian Gan, and Peter Jonas. Fast-spiking, parvalbumin⁺ gabaergic interneurons: From cellular design to microcircuit function. *Science*, 345(6196):1255263, 2014.

[57] Nathan R. Wilson, Caroline A. Runyan, Forea Wang, and Mriganka Sur. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature*, 488:343 – 348, 2012.

[58] Bassam V. Atallah, William Bruns, Matteo Carandini, and Massimo Scanziani. Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. *Neuron*, 73(1):159–170, 2012.

[59] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[60] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press, 2001.

[61] C. C Alan Fung, K. Y. Michael Wong, and Si Wu. A moving bump in a continuous manifold: A comprehensive study of the tracking dynamics of continuous attractor neural networks. *Neural Computation*, 22(3):752–792, 2010.

[62] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[63] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient mcmc. *Advances in neural information processing systems*, 28, 2015.

[64] Aaron Clauset, Cosma Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 11 2009.

[65] Mengni Wang, David J Foster, and Brad E Pfeiffer. Alternating sequences of future and past behavior encoded within hippocampal theta oscillations. *Science*, 370(6513):247–250, 2020.

Appendix

A Appendix Figures **16**

B A single canonical circuit and 1D natural gradient sampling: theory **17**

B.1 Continuous attractor network dynamics 17

B.2 Network's attractor states 17

B.3 Dimensionality reduction by projecting on dominant modes 18

B.4 The probabilistic generative model embedded in the circuit model 19

B.4.1 The stimulus likelihood 19

B.4.2 Uniform stimulus prior in the circuit 19

B.5 Conditions for realizing Langevin sampling in the circuit 20

C How to derive the distributions of bump height and position increments **20**

C.1 Fokker-Planck equation for the bump-height dynamics 20

C.2 Approximation of the bump-height distribution 21

C.3 Gaussian scaled mixture model 22

D Dimensionality reduction in a 2D CANN **23**

D.1 2D continuous attractor network dynamics 23

D.2 Network's attractor state 24

D.3 Dimensionality reduction by projecting on dominant modes 24

D.4 The probabilistic generative model embedded in the circuit model 25

D.4.1 The 2D stimulus likelihood with noise 25

D.4.2 Uniform stimulus prior in the 2D circuit 26

D.5 Conditions for realizing 2D Langevin sampling in the circuit 26

E The distribution of step size in 2d case **28**

E.1 2D Fokker–Planck equation for the bump-height dynamics 28

E.2 Gamma approximation of the 2D bump-height distribution 28

E.3 2D Gaussian scale mixture model for bump-position increments 28

E.4 Distribution of the displacement magnitude 29

F Circuit simulation parameters and details **30**

F.1 Critical weight 30

F.1.1 1D case 30

F.1.2 2D case 30

F.2 Simulation and Statistical Fitting Procedures 31

F.2.1 Network Simulation Protocol 31

F.2.2 Statistical Distribution Fitting 31

F.2.3 Heavy Tail Quantification 32

G Generalization **33**

G.1 Encoding a Non-Uniform Prior 33

A Appendix Figures

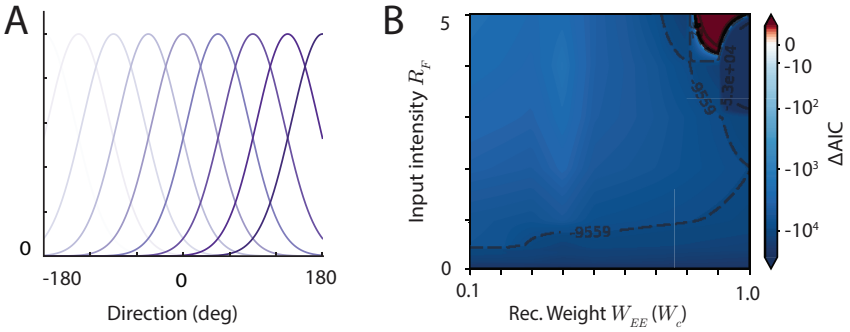


Figure A1: (A) Tuning curves of the neural population across the network space, mapping the preferred directions from $-180°$ to $180°$. (B) Heatmap of $\Delta\text{AIC} = \text{AIC(Gamma)} - \text{AIC(Gaussian)}$ evaluating the goodness of fit for the bump height distribution across varying input intensities R_F and recurrent weights W_{EE}.

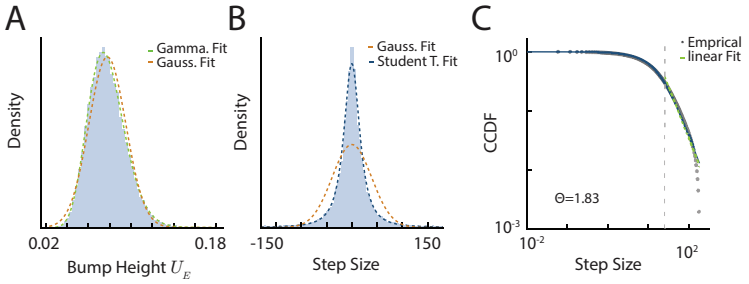


Figure A2: Sampling dynamics with a spatially shifted prior input. The prior input is provided at the same intensity as the likelihood input, but its position is shifted $5°$ away from the likelihood. This configuration still successfully induces long-tail sampling behavior (see details in Sec. G.1). (A) Probability density of the Bump Height (U_E), overlaid with Gamma and Gaussian fits. (B) Probability density of the step size, overlaid with Gaussian and Student-t fits. The Student-t fit better captures the heavy-tailed nature of the distribution. (C) Complementary Cumulative Distribution Function (CCDF) of the step size plotted on a log-log scale. The empirical data is overlaid with a linear fit in the tail region, revealing a tail exponent of $\Theta = 1.83$.

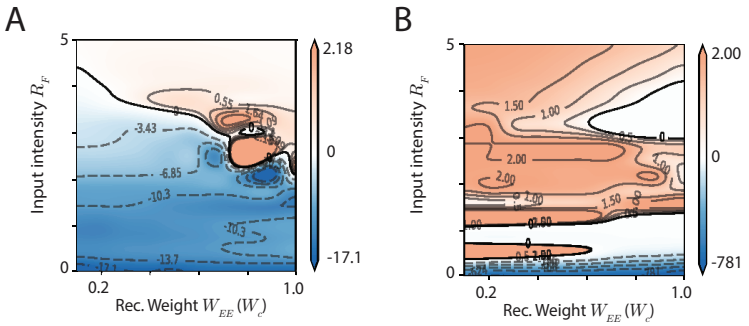


Figure A3: Goodness of fit analysis in the 2D network model. (A) Heatmap of $\Delta\text{AIC} = \text{AIC(Gamma)} - \text{AIC(Gaussian)}$ evaluating the goodness of fit for the bump height (U) distribution across varying input intensities R_F and recurrent weights W_{EE}. (B) Heatmap of $\Delta\text{AIC} = \text{AIC(Student-}t\text{)} - \text{AIC(Gaussian)}$ evaluating the goodness of fit for the 1D step size (dx) distribution across varying input intensities R_F and recurrent weights W_{EE}.

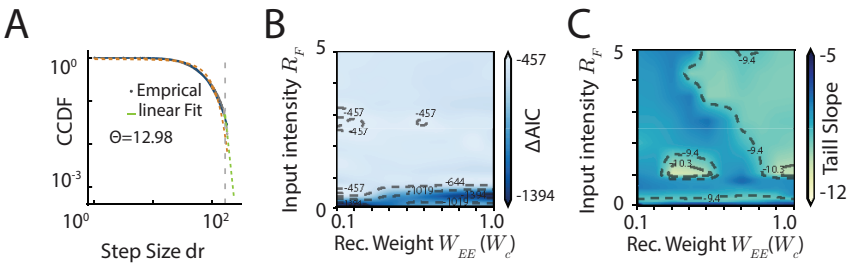


Figure A4: (A) Log-log plot of the Complementary Cumulative Distribution Function (CCDF) for the radial step size (dr), showing an empirical linear fit with a tasil exponent of $\Theta = 12.98$. (B) Heatmap of ΔAIC comparing the goodness of fit for step size distributions across different input intensities R_F and recurrent weights W_{EE}. (C) Tail slope of the step size distribution across different input intensities R_F and recurrent weights W_{EE}.

B A single canonical circuit and 1D natural gradient sampling: theory

In this section, we detail the mathematical foundation and theoretical properties of our simplified recurrent circuit model, which comprises excitatory (E) and inhibitory (I) neurons driven by continuous attractor network (CAN) dynamics.

B.1 Continuous attractor network dynamics

For clarity, we restate the governing equations of the network(Eq. 1). The temporal evolution of the excitatory population is described by:

$$\tau \frac{\partial \mathbf{u}_E(\theta,t)}{\partial t} = -\mathbf{u}_E(\theta,t) + \rho \sum_{X=E,F} (\mathbf{W}_{EX} * \mathbf{r}_X)(\theta,t) + \sqrt{\tau \mathsf{F}[\mathbf{u}_E(\theta,t)]_+}\xi(\theta,t), \tag{B1}$$

and the divisive normalization provided by I neurons (Eq. 3),

$$\mathbf{r}_E(\theta) = \frac{[\mathbf{u}_E(\theta)]_+^2}{1 + \rho w_{EI} \int_{-\pi}^{\pi} [\mathbf{u}_E(\theta)]_+^2 d\theta}, \tag{B2}$$

and the recurrent connection kernel $\mathbf{W}_{EX}$ (Eq. 2)

$$\mathbf{W}_{YX}(\theta) = w_{YX}\left(\sqrt{2\pi}a_{XY}\right)^{-1} \exp(-\theta^2/2a_{XY}^2). \tag{B3}$$

B.2 Network's attractor states

We verify the proposed Gaussian ansatz of the attractor states of E neurons (Eq. 7),

$$\bar{\mathbf{u}}_E(\theta) = \bar{U}_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right]. \tag{B4}$$

First, we substitute it into the divisive normalization (Eq. B2), yielding the following expression for the firing rate of E neurons,

$$\bar{\mathbf{r}}_E(\theta) = \frac{[\bar{\mathbf{u}}_E^2(\theta)]_+^2}{1 + \rho w_{EI} \int [\bar{\mathbf{u}}_E(\theta)]_+^2 d\theta} = \underbrace{\frac{\bar{U}_E^2}{1 + \rho w_{EI} \bar{U}_E^2 \sqrt{2\pi} a_E}}_{\bar{R}_E} \exp\left[-\frac{(\theta - z_E)^2}{2a_E^2}\right]. \tag{B5}$$

Then we use the above E firing rate (Eq. B5) to calculate the recurrent input from the neuronal population of type Y to the one with type X in the circuit model,

$$\begin{aligned} \mathbf{u}_{XY}(\theta) &= \rho \mathbf{W}_{XY} * \mathbf{r}_Y(\theta) \\ &= \frac{\rho w_{XY} R_Y}{\sqrt{2\pi} a_{XY}} \int \exp\left[-\frac{(\theta' - \theta)^2}{2a_{XY}^2} - \frac{(\theta' - z_Y)^2}{2a_Y^2}\right] d\theta' \\ &= \rho w_{XY} R_Y \frac{a_Y}{\sqrt{a_{XY}^2 + a_Y^2}} \exp\left[-\frac{(\theta - z_Y)^2}{2(a_{XY}^2 + a_Y^2)}\right]. \end{aligned} \tag{B6}$$

Specifically, based on Eq. (B6), the recurrent E population input is

$$\mathbf{u}_{EE}(\theta) = \rho \mathbf{W}_{EE} * \mathbf{r}_E(\theta) = \underbrace{\frac{\rho}{\sqrt{2}} w_{EE} R_E}_{U_{EE}} \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right], \tag{B7}$$

and the feedforward population input is

$$\mathbf{u}_{EF}(\theta) = \rho \mathbf{W}_{EF} * r_F(\theta) = \underbrace{\frac{\rho}{\sqrt{2}} w_{EF} R_F}_{U_{EF}} \exp\left[-\frac{(\theta - \mu_z)^2}{4a_E^2}\right]. \tag{B8}$$

. It can be checked the proposed Gaussian ansatz (Eq. B4) is indeed the sum of the recurrent input (Eq. B7) and the feedforward input (Eq. B8), i.e.,

$$\bar{\mathbf{u}}_E(\theta) = \mathbf{u}_{EE}(\theta) + \mathbf{u}_{EF}(\theta)$$

$$\bar{U}_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] = U_{EE} \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] + U_{EF} \exp\left[-\frac{(\theta - \mu_z)^2}{4a_E^2}\right]$$

and implies

$$\mathbf{r}_E(\theta, t) = U_{EE} + U_{EF}, \quad z_E = \mu_z.$$

This completes the recurrent loop of the dynamics, and verify the validity of the Gaussian ansatz (Eq. 7).

B.3 Dimensionality reduction by projecting on dominant modes

We substitute Eqs. (B4-B8) into the Eq. (B1),

$$\begin{aligned}
&\tau \dot{U}_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] + \frac{\tau U_E}{2a_E} \dot{z}_E \left(\frac{\theta - z_E}{a_E}\right) \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] \\
&= -U_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] + \frac{\rho}{\sqrt{2}} w_{EE} R_E \exp\left[-\frac{(\theta - z_E)^2}{4a_E^2}\right] \\
&\quad + \frac{\rho}{\sqrt{2}} w_{EF} R_F \exp\left[-\frac{(\theta - \mu)^2}{4a_E^2}\right] + \sqrt{\tau \mathsf{F} U_E} \exp\left[-\frac{(\theta - z_E)^2}{8a_E^2}\right] \xi(\theta, t)
\end{aligned} \tag{B9}$$

Previous studies analytically calculated the first two dominant eigenvectors (modes) [39, 61], corresponding to the change of the position and height of the Gaussian ansatz respectively,

$$\text{Position}: \ \phi_1(\theta | z_E) \propto \nabla_z \bar{\mathbf{u}}_E(\theta) \propto (\theta - z_E) \exp[-(\theta - z_E)^2 / 4a^2], \tag{B10a}$$

$$\text{Height}: \quad \phi_2(\theta | z_E) \propto \quad \bar{\mathbf{u}}_E(\theta) \propto \exp[-(\theta - z_E)^2 / 4a^2]. \tag{B10b}$$

Projecting the dynamics Eq. (B9) into these 2 motion modes (Eq. B10), which means calculate the inner product $\int f(\theta)\phi(\theta|z_E)d\theta$ with $f(\theta)$ a term in Eq. (B9),

$$\tau U_E \dot{z}_E = \frac{\rho}{\sqrt{2}} w_{EF} R_F (\mu - z_E) \exp\left[-\frac{(\mu - z_E)^2}{8a_E^2}\right] + \sigma_z \sqrt{\tau U_E} \xi$$

$$\tau \dot{U}_E = -U_E + \frac{\rho}{\sqrt{2}} w_{EE} R_E + \frac{\rho}{\sqrt{2}} w_{EF} R_F \exp\left[-\frac{(\mu - z_E)^2}{8a_E^2}\right] + \sigma_U \sqrt{\tau U_E} \xi$$

where

$$\sigma_z^2 = \frac{8a\mathsf{F}}{3\sqrt{3\pi}}, \quad \sigma_U^2 = \frac{\mathsf{F}}{\sqrt{3\pi}a}. \tag{B11}$$

When the bump position z_E is near the input position, i.e., $\mu - z_E \ll a_E$, which is usually the case in the circuit model, the exponential term $\exp[-(\mu - z_E)^2/8a_E^2]$ is close to one and can be safely ignored,

$$\dot{z}_E = (\tau U_E)^{-1} \frac{\rho}{\sqrt{2}} w_{EF} R_F (\mu_z - z_E) + \sigma_z (\tau U_E)^{-1/2} \xi_t \tag{B12a}$$

$$\dot{U}_E = \tau^{-1} \left[-U_E + \frac{\rho}{\sqrt{2}} (w_{EE} R_E + w_{EF} R_F)\right] + \sigma_U (\tau^{-1} U_E)^{1/2} \xi_t \tag{B12b}$$

Furthermore, by using the notation

$$U_{EF} = \rho w_{EF} R_F/\sqrt{2}, \quad U_{EE} = \rho w_{EE} R_E/\sqrt{2},$$

we arrive at Eqs. (8b and 11b) in the main text.

B.4 The probabilistic generative model embedded in the circuit model

B.4.1 The stimulus likelihood

We study how feedforward input defines the latent stimulus likelihood, i.e., $\mathcal{L}(z) \propto p(\mathbf{r}_F|z)$. From the Eq. (4), the feedforward input $\mathbf{r}_F$ is modeled as a set of independent Poisson spike trains, where each neuron's firing rate is Gaussian-tuned to the stimulus [6]:

$$\mathbf{r}_F(\theta|z) \sim \text{Poisson}[\lambda_F(\theta|z)], \quad \lambda_F(\theta|z) = R_F \exp[-(\theta - z)^2/2a^2], \tag{B13}$$

where $\lambda_F(\theta|z)$ is the mean firing rate of the neuron with stimulus preference θ. $\mathbf{r}_F$ denotes the peak input rate, and a specifies the tuning width. Explicitly writing the Poisson distribution of feedforward input spikes (we discretize the continuous θ into equally spaced θ_j),

$$p(\mathbf{r}|z) = \prod_{j=1}^{N_E} \text{Poisson}\,(\mathbf{r}_j|\lambda_j \Delta t) = \prod_{j=1}^{N_E} \frac{(\lambda_j \Delta t)^{\mathbf{r}_j}}{\mathbf{r}_j!} \exp(-\lambda_j \Delta t). \tag{B14}$$

Taking the logarithm,

$$\begin{aligned} \ln p(\mathbf{r}|z) &= \textstyle\sum_j \left[\mathbf{r}_j \ln(\lambda_j \Delta t) - \ln(\mathbf{r}_j!) - \lambda_j\right], \\ &= \textstyle\sum_j \mathbf{r}_j \ln(\lambda_j \Delta t) + \text{const.} \end{aligned} \tag{B15}$$

The const. in the above equation is under the assumption that the sum of population firing rate $\sum_j \lambda_j$ is a constant irrelevant to latent stimulus z , which is true in a homogeneous population with a large number of neurons. Substituting the expression of the Gaussian tuning,

$$\ln p(\mathbf{r}|z) = -\sum_j \mathbf{r}_j \frac{(\theta - z)^2}{2a^2} + \text{const} = -\frac{1}{2}\Lambda(z - \mu_z)^2 + \text{const}, \tag{B16}$$

where

$$\mu_z = \frac{\sum_j \mathbf{r}(\theta_j)\theta_j}{\sum_j \mathbf{r}(\theta_j)}, \;\; \Lambda = a^{-2} \sum_j \mathbf{r}(\theta_j) \approx \sqrt{2\pi}\rho a^{-1} R_F. \tag{B17}$$

This implies the latent stimulus likelihood for the latent stimulus feature z given an observed feedforward input $\mathbf{r}_F$ is derived as a Gaussian distribution,

$$\mathcal{L}(z) = \mathcal{N}(z|\mu_z, \Lambda^{-1}),$$

which is the Eq. (6) in the main text. Notably, the Gaussian distribution comes from the profile of the Gaussian tuning (Eq. B13) [6].

B.4.2 Uniform stimulus prior in the circuit

To convert the E bump position dynamics (Eq. B12a) into the Langevin sampling dynamics, it immediately suggests that the circuit stores a uniform (uninformative) stimulus prior, i.e., $p(z)$ is uniform. This is because the gradient of the log-likelihood ($\nabla\mathcal{L}(z)$, Eq. 6) has the same form as the drift term in the E position dynamics

$$\begin{aligned} &\text{Likelihood gradient:} && \nabla \ln \mathcal{L}(z) = \Lambda(\mu_z - z) \\ &\text{E bump position drift term:} && U_{EF}(\mu_z - z_E) \end{aligned}$$

suggesting the gradient of the prior is zero, i.e., $\nabla \ln p(z) = 0$. This uniform prior arises from the circuit's homogeneous neurons (uniformly distributed in feature space) and its translation-invariant connection profile. Consequently, for the circuit to store a non-uniform prior, it must break this inherent symmetry in its neural organization and connectivity.

B.5 Conditions for realizing Langevin sampling in the circuit

The circuit sampling of the likelihood means the equilibrium distribution of the bump position (Eq. B12a) should match with the likelihood (Eq. 6). We copy the circuit bump position dynamics and the likelihood Langevin sampling dynamics in below for comparison,

$$\text{Circuit:} \quad \dot{z}_E = (\tau U_E)^{-1} \underbrace{\frac{\rho w_{EF} R_F}{\sqrt{2}\Lambda}}_{\lambda_z} \Lambda(\mu_z - z_E) + \sigma_z (\tau U_E)^{-1/2} \xi_t,$$

$$\text{Langevin:} \quad \dot{z}_t = \tau_L^{-1} \Lambda(\mu_z - z) + (2\tau_L^{-1})^{1/2} \xi_t.$$

The σ_z is a constant that doesn't change with neuronal activities. Therefore, the likelihood Langevin sampling in the circuit can be realized by setting the feedforward weight w_{EF} appropriately to make the ratio of the drift and diffusion coefficients the same as the Langevin sampling dynamics. The optimal feedforward weight can be found as (by using Eq. B17)

$$\frac{\sigma_z^2}{\lambda_z} = 2 \quad \Leftrightarrow \quad w_{EF}^* = \frac{\sigma_z^2}{\sqrt{2}\rho} \frac{\Lambda}{R_F} = \left(\frac{2}{\sqrt{3}}\right)^3 \mathsf{F} \tag{B18}$$

Furthermore, the time constant of the z_E dynamics is

$$\tau_z = \lambda_z^{-1} \tau U_E = \frac{2\sqrt{\pi}}{a w_{EF}} \tau U_E, \tag{B19}$$

which is proportional to the E bump height U_E. Finally, the equation of bump position (Eq. B12a) can be converted into the same form as a standard Langevin sampling,

$$\dot{z}_E = \tau_z^{-1} \Lambda(\mu - z_E) + (2\tau_z^{-1})^{1/2} \xi_t$$

C How to derive the distributions of bump height and position increments

C.1 Fokker-Planck equation for the bump-height dynamics

To derive the stationary distribution of the bump height, we start from the reduced stochastic dynamics of U_E,

$$dU_E = A(U_E)\, dt + B(U_E)\, dW_t, \tag{C20}$$

with drift and diffusion terms

$$A(U_E) = \tau^{-1} \left[-U_E + \frac{\rho}{\sqrt{2}} \left(\frac{w_{EE} U_E^2}{1 + \rho w_{EI} U_E^2} + w_{EF} R_F \right) \right], \tag{C21}$$

$$B(U_E) = \sigma_U \left(\tau^{-1} U_E\right)^{1/2}. \tag{C22}$$

For a one-dimensional Langevin equation of the form $dX = a(X)\, dt + b(X)\, dW_t$, the corresponding Fokker–Planck equation for the probability density $p(x,t)$ is

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\big[a(x)p(x,t)\big] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big[b^2(x)p(x,t)\big]. \tag{C23}$$

Applying Eq. (C23) to Eq. (C20) gives

$$\frac{\partial p(U_E,t)}{\partial t} = -\frac{\partial}{\partial U_E}\big[A(U_E)p(U_E,t)\big] + \frac{1}{2}\frac{\partial^2}{\partial U_E^2}\big[B^2(U_E)p(U_E,t)\big]. \tag{C24}$$

At steady state, $\partial_t p = 0$, so the stationary density $p(U_E)$ satisfies

$$0 = -\frac{d}{dU_E}\big[A(U_E)p(U_E)\big] + \frac{1}{2}\frac{d^2}{dU_E^2}\big[B^2(U_E)p(U_E)\big]. \tag{C25}$$

Assuming zero probability flux at the boundaries, we integrate once to obtain

$$A(U_E)p(U_E) - \frac{1}{2}\frac{d}{dU_E}\big[B^2(U_E)p(U_E)\big] = 0. \tag{C26}$$

Rearranging gives

$$\frac{d}{dU_E}\big[B^2(U_E)p(U_E)\big] = 2A(U_E)p(U_E), \tag{C27}$$

and therefore

$$\frac{d}{dU_E}\ln\big[B^2(U_E)p(U_E)\big] = \frac{2A(U_E)}{B^2(U_E)}. \tag{C28}$$

Integrating both sides yields the standard stationary solution

$$p(U_E) = \frac{C}{B^2(U_E)}\exp\left[2\int^{U_E}\frac{A(u)}{B^2(u)}\,du\right], \tag{C29}$$

where C is a normalization constant.

For the present dynamics,

$$B^2(U_E) = \sigma_U^2\tau^{-1}U_E, \tag{C30}$$

so that

$$\frac{A(U_E)}{B^2(U_E)} = \frac{\tau^{-1}\Big[-U_E + \frac{\rho}{\sqrt{2}}\Big(\frac{w_{EE}U_E^2}{1+\rho w_{EI}U_E^2} + w_{EF}R_F\Big)\Big]}{\sigma_U^2\tau^{-1}U_E} \tag{C31}$$

$$= \frac{1}{\sigma_U^2}\left[-1 + \frac{\rho w_{EE}}{\sqrt{2}}\frac{U_E}{1+\rho w_{EI}U_E^2} + \frac{\rho w_{EF}R_F}{\sqrt{2}U_E}\right]. \tag{C32}$$

The integral in Eq. (C29) can then be evaluated term by term:

$$2\int\frac{A(U_E)}{B^2(U_E)}\,dU_E = \frac{2}{\sigma_U^2}\int\left[-1 + \frac{\rho w_{EE}}{\sqrt{2}}\frac{U_E}{1+\rho w_{EI}U_E^2} + \frac{\rho w_{EF}R_F}{\sqrt{2}U_E}\right]dU_E. \tag{C33}$$

Using

$$\int\frac{U_E}{1+\rho w_{EI}U_E^2}\,dU_E = \frac{1}{2\rho w_{EI}}\ln\big(1+\rho w_{EI}U_E^2\big), \qquad \int\frac{1}{U_E}\,dU_E = \ln U_E, \tag{C34}$$

we obtain

$$2\int\frac{A(U_E)}{B^2(U_E)}\,dU_E = \frac{2}{\sigma_U^2}\left[-U_E + \frac{\rho w_{EE}}{2\sqrt{2}\rho w_{EI}}\ln\big(1+\rho w_{EI}U_E^2\big) + \frac{\rho w_{EF}R_F}{\sqrt{2}}\ln U_E\right]. \tag{C35}$$

Substituting this result back into Eq. (C29) and absorbing all constants into the normalization factor gives

$$\boxed{p(U_E) \propto U_E^{\frac{2\rho w_{EF}R_F}{\sqrt{2}\sigma_U^2}-1}\big(1+\rho w_{EI}U_E^2\big)^{\frac{\rho w_{EE}}{\sqrt{2}\rho w_{EI}\sigma_U^2}}\exp\left(-\frac{2U_E}{\sigma_U^2}\right)}. \tag{C36}$$

It is convenient to rewrite this stationary law as

$$p(U_E) \propto U_E^{\alpha_1-1}\big(1+\rho w_{EI}U_E^2\big)^{\tilde{c}}\,e^{-\beta_1 U_E}, \tag{C37}$$

with

$$\alpha_1 = \frac{2\rho w_{EF}R_F}{\sqrt{2}\sigma_U^2}, \qquad \tilde{c} = \frac{\rho w_{EE}}{\sqrt{2}\rho w_{EI}\sigma_U^2}, \qquad \beta_1 = \frac{2}{\sigma_U^2}. \tag{C38}$$

C.2 Approximation of the bump-height distribution

This distribution can be viewed as a Gamma-like kernel multiplied by an additional correction term that arises from recurrent excitation under divisive normalization.

To make this correction explicit, we rewrite

$$\big(1+\rho w_{EI}U_E^2\big)^{\tilde{c}} = \exp\big[\tilde{c}\ln\big(1+\rho w_{EI}U_E^2\big)\big]. \tag{C39}$$

When $\rho w_{EI}U_E^2 \ll 1$, we use the Taylor expansion

$$\ln(1+x) = x - \frac{x^2}{2} + O(x^3), \tag{C40}$$

with $x = \rho w_{EI} U_E^2$, to obtain

$$\begin{aligned}\tilde{c}\ln\big(1+\rho w_{EI}U_E^2\big) &= \tilde{c}\left[\rho w_{EI}U_E^2 - \frac{(\rho w_{EI})^2 U_E^4}{2} + O\big((\rho w_{EI})^3 U_E^6\big)\right] \\ &= \frac{\rho w_{EE}}{\sqrt{2}\sigma_U^2}U_E^2 - \frac{\rho w_{EE}\rho w_{EI}}{2\sqrt{2}\sigma_U^2}U_E^4 + O\big((\rho w_{EI})^2 U_E^6\big)\,. \end{aligned} \tag{C41}$$

Therefore,

$$\big(1+\rho w_{EI}U_E^2\big)^{\tilde{c}} \approx \exp\left[\frac{\rho w_{EE}}{\sqrt{2}\sigma_U^2}U_E^2 + O(\rho w_{EI}U_E^4)\right]. \tag{C42}$$

Substituting this into Eq. (C37) gives the local approximation

$$p(U_E) \propto U_E^{\alpha_1 - 1}\exp\left[-\beta_1 U_E + \frac{\rho w_{EE}}{\sqrt{2}\sigma_U^2}U_E^2 + O(\rho w_{EI}U_E^4)\right]. \tag{C43}$$

Equation (C43) shows that the exact stationary density is a Gamma-like distribution with a quadratic exponential correction. In the regime relevant for the leading-order analysis, we further assume that this correction is weak over the dominant support of the distribution, so that

$$\frac{\rho w_{EE}}{\sqrt{2}\sigma_U^2}U_E^2 \ll 1, \qquad \rho w_{EI}U_E^2 \ll 1. \tag{C44}$$

Then

$$\exp\left[\frac{\rho w_{EE}}{\sqrt{2}\sigma_U^2}U_E^2 + O(\rho w_{EI}U_E^4)\right] = 1 + O(U_E^2), \tag{C45}$$

and to leading order we may neglect this factor entirely. The stationary law then reduces to

$$p(U_E) \approx \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}U_E^{\alpha_1-1}e^{-\beta_1 U_E}, \qquad U_E > 0, \tag{C46}$$

with

$$\alpha_1 = \frac{2\rho w_{EF}R_F}{\sqrt{2}\sigma_U^2}, \qquad \beta_1 = \frac{2}{\sigma_U^2}. \tag{C47}$$

Thus, the Gamma approximation should be understood as the leading-order truncation of the more general stationary density in Eq. (C36). The retained exponential correction $\exp(\tilde{c}_{corr}U_E^2)$, with $\tilde{c}_{corr} = \rho w_{EE}/(\sqrt{2}\sigma_U^2)$, may be viewed as the first higher-order deviation from the Gamma form. Keeping this correction gives a refined approximation, whereas dropping it yields the simple Gamma law used in the main text and in the Gaussian scale mixture argument for the position increments.

C.3 Gaussian scaled mixture model

We begin with the reduced neural dynamics for the bump position:

$$\dot{z}_E = (\tau U_E)^{-1}\frac{\rho}{\sqrt{2}}w_{EF}R_F(\mu_z - z_E) + \sigma_z(\tau U_E)^{-1/2}\xi_{z,t}, \tag{C48}$$

For a small time interval Δt, the Euler–Maruyama approximation gives

$$\Delta z_E \approx \underbrace{\tau^{-1}\lambda_z\Lambda(\mu_z - z_E)\Delta t}_{\text{Drift } f(z_E)}U_E^{-1} + \underbrace{(\sigma_z^2\tau^{-1}\Delta t)^{1/2}}_{\text{Diffusion } \sqrt{D}}U_E^{-1/2}\,\xi_{z,t}, \qquad \xi_{z,t}\sim\mathcal{N}(0,1)$$

Hence, conditioned on U_E, the displacement follows a Gaussian distribution:

$$\Delta z_E \mid U_E \sim \mathcal{N}\big(f(z_E)U_E^{-1},\, DU_E^{-1}\big)\,, \tag{C49}$$

where $D = \sigma_z^2\tau^{-1}\Delta t$. In the short-time or diffusion-dominated regime (as $\Delta t \to 0$), the drift contribution $f(z_E) \propto \Delta t$ is subleading compared to the diffusion term $\sqrt{D} \propto \sqrt{\Delta t}$. Thus, we approximate:

$$\Delta z_E \mid U_E \approx \mathcal{N}\big(0,\, DU_E^{-1}\big)\,. \tag{C50}$$

As the bump height is approximated as Gamma distributed:

$$p(U_E) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} U_E^{\alpha_1 - 1} e^{-\beta_1 U_E}, \qquad U_E > 0, \tag{C51}$$

the marginal displacement density is obtained by integrating out U_E:

$$p(\Delta z_E) = \int_0^\infty \frac{1}{\sqrt{2\pi(D/U_E)}} \exp\left[-\frac{\Delta z_E^2}{2(D/U_E)}\right] \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} U_E^{\alpha_1 - 1} e^{-\beta_1 U_E}\, dU_E. \tag{C52}$$

Using $(2\pi D U_E^{-1})^{-1/2} = U_E^{1/2}/\sqrt{2\pi D}$, we obtain

$$p(\Delta z_E) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)\sqrt{2\pi D}} \int_0^\infty U_E^{\alpha_1 - \frac{1}{2}} \exp\left[-U_E\left(\beta_1 + \frac{\Delta z_E^2}{2D}\right)\right] dU_E. \tag{C53}$$

Rewriting the power as $U_E^{(\alpha_1 + \frac{1}{2}) - 1}$ shows that this is a standard Gamma integral:

$$p(\Delta z_E) \propto \int_0^\infty U_E^{\left(\alpha_1 + \frac{1}{2}\right) - 1} \exp\left[-U_E\left(\beta_1 + \frac{\Delta z_E^2}{2D}\right)\right] dU_E, \tag{C54}$$

so that

$$p(\Delta z_E) \propto \left(\beta_1 + \frac{\Delta z_E^2}{2D}\right)^{-(\alpha_1 + \frac{1}{2})}. \tag{C55}$$

Factoring out β_1 gives

$$p(\Delta z_E) \propto \left(1 + \frac{\Delta z_E^2}{2\beta_1 D}\right)^{-\frac{2\alpha_1 + 1}{2}}. \tag{C56}$$

Comparing this with the standard Student-t distribution form:

$$p(x) \propto \left(1 + \frac{x^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}}, \tag{C57}$$

we can identify the degrees of freedom:

$$\boxed{\nu_1 = 2\alpha_1} \tag{C58}$$

and the scale parameter:

$$\nu_1 s_1^2 = 2\beta_1 D \qquad \Rightarrow \qquad \boxed{s_1 = \sqrt{\frac{\beta_1}{\alpha_1} D}}. \tag{C59}$$

Therefore, the marginal step size follows a Student-t distribution:

$$\Delta z_E \sim \text{Student-}t(\nu_1, s_1), \qquad \nu_1 = 2\alpha_1, \qquad s_1 = \sqrt{\frac{\beta_1}{\alpha_1} D}. \tag{C60}$$

This rigorous derivation shows that the heavy tail arises directly from a Gaussian scale mixture: conditioned on a fixed bump height, the displacement is locally Gaussian, but integrating over the stochastic network fluctuations in U_E produces a heavy-tailed Student-t marginal law.

D Dimensionality reduction in a 2D CANN

D.1 2D continuous attractor network dynamics

E: $$\tau \dot{\mathbf{u}}_E = -\mathbf{u}_E(\theta, t) + \rho \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} d\theta' \mathbf{W}_{YX}(\theta, \theta') r(\theta', t) + \sqrt{\tau \mathsf{F}[\mathbf{u}_E(\theta, t)]_+}\, \xi(\theta, t) \tag{D61a}$$

DivNom. $$r(\theta, t) = \mathbf{u}_E(\theta, t)^2 / (1 + \rho w_{EI} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\theta' \mathbf{u}_E(\theta', t)^2) \tag{D61b}$$

Rec. weight: $$\mathbf{W}_{YX} = \frac{W_{YX}}{2\pi a^2} \exp\left[-\frac{(\mathbf{x} - \mathbf{x}')^2}{2a_{YX}^2}\right] \tag{D61c}$$

Feedfwd. $$r_j \sim \text{Poisson}(\lambda_j(\mathbf{z})\Delta t), \qquad \lambda_j(\mathbf{z}) = R_F \exp\left[-\frac{\|\boldsymbol{\theta}_j - \mathbf{z}\|^2}{2a^2}\right], \tag{D61d}$$

D.2 Network's attractor state

To derive a low-dimensional description of the 2D network, we assume that the excitatory activity remains close to a Gaussian bump whose amplitude and center evolve slowly in time. We approximate the bump by

$$\mathbf{u}_E(\boldsymbol{\theta}, t) = U_E(t) \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_E(t)\|^2}{4a^2}\right], \tag{D62}$$

where $\mathbf{z}_E(t) = (z_{E,1}(t), z_{E,2}(t))$ denotes the bump center. Under divisive normalization, the corresponding firing-rate profile is

$$\mathbf{r}_E(\boldsymbol{\theta}, t) = R_E(t) \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_E(t)\|^2}{2a^2}\right], \qquad R_E(t) = \frac{U_E(t)^2}{1 + 2\pi a^2 \rho w_{EI} U_E(t)^2}. \tag{D63}$$

For a Gaussian kernel

$$\mathbf{W}_{YX}(\boldsymbol{\theta} - \boldsymbol{\theta}') = \frac{w_{YX}}{2\pi a_{YX}^2} \exp\left[-\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2a_{YX}^2}\right], \tag{D64}$$

and a Gaussian firing profile centered at $\mathbf{z}_X$ with width a_X, direct convolution yields

$$\rho(\mathbf{W}_{YX} * \mathbf{r}_X)(\boldsymbol{\theta}) = \rho w_{YX} R_X \frac{a_X^2}{a_{YX}^2 + a_X^2} \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_X\|^2}{2(a_{YX}^2 + a_X^2)}\right]. \tag{D65}$$

In the matched-width case $a_{XY} = a_X = a$, this reduces to

$$\rho(\mathbf{W}_{YX} * \mathbf{r}_X)(\boldsymbol{\theta}) = \frac{\rho}{2} w_{YX} R_X \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_X\|^2}{4a^2}\right]. \tag{D66}$$

D.3 Dimensionality reduction by projecting on dominant modes

The dominant slow modes are the height mode

$$\phi_0(\boldsymbol{\theta}|\mathbf{z}_E) \propto \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_E\|^2}{4a^2}\right], \tag{D67}$$

and the two translational modes

$$\phi_i(\boldsymbol{\theta}|\mathbf{z}_E) \propto (\theta_i - z_{E,i}) \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_E\|^2}{4a^2}\right], \qquad i = 1, 2. \tag{D68}$$

Thus the slow manifold is three-dimensional, parametrized by $(U_E, z_{E,1}, z_{E,2})$.

Substituting Eq. (D62) into the network dynamics and projecting onto the height mode and the two translational modes yields the reduced stochastic dynamics

$$\tau \dot{U_E} = \left[-U_E + U_{EE} + U_{EF} \exp\left(-\frac{\|\boldsymbol{\mu}_z - \mathbf{z}_E\|^2}{8a^2}\right)\right] + \sigma_U \sqrt{\tau U_E}\, d\xi, \tag{D69}$$

$$\tau U_E \dot{\mathbf{z}_E} = U_{EF} \exp\left(-\frac{\|\boldsymbol{\mu}_z - \mathbf{z}_E\|^2}{8a^2}\right) (\boldsymbol{\mu}_z - \mathbf{z}_E) + \sigma_z \sqrt{\tau U_E}\, d\xi, \tag{D70}$$

where W_U is a scalar Wiener process and $\mathbf{W}_z = (W_{z,1}, W_{z,2})^\top$ is a two-dimensional Wiener process with independent components. For the Gaussian bump ansatz, the projected noise strengths are

$$\sigma_U^2 = \frac{\mathsf{F}}{3\pi a^2}, \qquad \sigma_z^2 = \frac{8\mathsf{F}}{9\pi}. \tag{D71}$$

Here

$$U_{EE} = \frac{\rho}{2} w_{EE} R_E, \qquad U_{EF} = \frac{\rho}{2} w_{EF} R_F. \tag{D72}$$

When $\|\boldsymbol{\mu}_z - \mathbf{z}_E\| \ll a$, the equations simplify to

$$\tau \dot{U_E} = (-U_E + U_{EE} + U_{EF})\, dt + \sigma_U \sqrt{\tau U_E}\, \xi, \tag{D73}$$

$$\tau U_E \dot{\mathbf{z}}_E = U_{EF}(\boldsymbol{\mu}_z - \mathbf{z}_E)\, dt + \sigma_z \sqrt{\tau U_E}\, d\xi. \tag{D74}$$

D.4 The probabilistic generative model embedded in the circuit model

D.4.1 The 2D stimulus likelihood with noise

We extend the feedforward population-code likelihood to a two-dimensional latent stimulus $\mathbf{z} = (z_1, z_2)^\top$. Let each feedforward neuron have preferred stimulus $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2})^\top$, and let its spike count over a small time bin Δt be conditionally independent with Gaussian tuning,

$$r_j \sim \text{Poisson}(\lambda_j(\mathbf{z})\Delta t)\,, \qquad \lambda_j(\mathbf{z}) = R_F \exp\left[-\frac{\|\boldsymbol{\theta}_j - \mathbf{z}\|^2}{2a^2}\right], \tag{D75}$$

where $\|\boldsymbol{\theta}_j - \mathbf{z}\|^2 = (\theta_{j1} - z_1)^2 + (\theta_{j2} - z_2)^2$, R_F is the peak input rate, and a is the shared tuning width.

The likelihood of the population response $\mathbf{r} = \{r_j\}_{j=1}^{N_E}$ is

$$p(\mathbf{r} \mid \mathbf{z}) = \prod_{j=1}^{N_E} \frac{(\lambda_j(\mathbf{z})\Delta t)^{r_j}}{r_j!} \exp[-\lambda_j(\mathbf{z})\Delta t]\,. \tag{D76}$$

Taking the logarithm gives

$$\log p(\mathbf{r} \mid \mathbf{z}) = \sum_j r_j \log(\lambda_j(\mathbf{z})\Delta t) - \sum_j \log(r_j!) - \Delta t \sum_j \lambda_j(\mathbf{z}). \tag{D77}$$

For a homogeneous dense population, $\sum_j \lambda_j(\mathbf{z})$ is approximately independent of $\mathbf{z}$, so the last two terms can be absorbed into the constant. Substituting the Gaussian tuning curve then yields

$$\log p(\mathbf{r} \mid \mathbf{z}) = -\frac{1}{2a^2} \sum_j r_j \|\boldsymbol{\theta}_j - \mathbf{z}\|^2 + \text{const}. \tag{D78}$$

Expanding the quadratic form in $\mathbf{z}$,

$$\log p(\mathbf{r} \mid \mathbf{z}) = -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_z)^\top \Lambda (\mathbf{z} - \boldsymbol{\mu}_z) + \text{const}, \tag{D79}$$

where

$$\boldsymbol{\mu}_z = \frac{\sum_j r_j \boldsymbol{\theta}_j}{\sum_j r_j}, \tag{D80}$$

and

$$\Lambda = \frac{\sum_j r_j}{a^2} \mathbf{I}_2. \tag{D81}$$

Therefore, conditioned on the observed noisy feedforward response, the latent stimulus likelihood is

$$\mathcal{L}(\mathbf{z}) = \mathcal{N}\big(\mathbf{z} \mid \boldsymbol{\mu}_z, \Lambda^{-1}\big)\,. \tag{D82}$$

To make the noise explicit, let the true latent stimulus be $\mathbf{z}_t$, and define the total feedforward count

$$N_F = \sum_j r_j. \tag{D83}$$

In the continuum limit for a homogeneous population of density ρ,

$$\bar{N}_F = \mathbb{E}[N_F] \approx \rho \int_{\mathbb{R}^2} R_F \exp\left[-\frac{\|\boldsymbol{\theta} - \mathbf{z}_t\|^2}{2a^2}\right] d\boldsymbol{\theta} = 2\pi\rho a^2 R_F. \tag{D84}$$

Allowing a Fano factor F for the feedforward spike-count variability,

$$\text{Var}(N_F) \approx \mathsf{F}\, \bar{N}_F. \tag{D85}$$

Hence the precision fluctuates around its mean as

$$\Lambda = \bar{\Lambda}\, \mathbf{I}_2 + \delta\Lambda\, \mathbf{I}_2, \qquad \bar{\Lambda} = \frac{\bar{N}_F}{a^2} = 2\pi\rho R_F, \tag{D86}$$

with

$$\mathrm{Var}(\delta\Lambda) = \frac{\mathrm{Var}(N_F)}{a^4} \approx \frac{2\pi\rho \mathsf{F} R_F}{a^2}. \tag{D87}$$

Similarly, a first-order expansion of the population vector gives

$$\boldsymbol{\mu}_z = \mathbf{z}_t + \boldsymbol{\eta}_\mu, \qquad \boldsymbol{\eta}_\mu \sim \mathcal{N}\left(\mathbf{0},\, \frac{\mathsf{F}}{2\pi\rho R_F}\mathbf{I}_2\right), \tag{D88}$$

so that the decoded stimulus mean fluctuates isotropically around the true stimulus.

Thus, in the minimal 2D extension, the noisy feedforward population encodes an isotropic Gaussian likelihood whose center and precision both fluctuate from trial to trial:

$$\mathcal{L}(\mathbf{z}) = \mathcal{N}\big(\mathbf{z} \mid \mathbf{z}_t + \boldsymbol{\eta}_\mu,\, (\bar{\Lambda} + \delta\Lambda)^{-1}\mathbf{I}_2\big)\,. \tag{D89}$$

D.4.2 Uniform stimulus prior in the 2D circuit

For the minimal 2D extension, the feedforward population encodes an isotropic Gaussian likelihood over the latent stimulus $\mathbf{z} \in \mathbb{R}^2$,

$$\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_z, \Lambda^{-1}), \qquad \Lambda = \lambda\mathbf{I}_2, \tag{D90}$$

where $\boldsymbol{\mu}_z$ is the decoded 2D population vector and, in the continuum limit,

$$\lambda \approx 2\pi\rho R_F. \tag{D91}$$

Therefore,

$$\nabla_{\mathbf{z}} \ln \mathcal{L}(\mathbf{z}) = \Lambda(\boldsymbol{\mu}_z - \mathbf{z}). \tag{D92}$$

Comparing this likelihood gradient with the drift term in the reduced E-bump position dynamics shows that the circuit again stores a uniform (uninformative) prior over the 2D stimulus space, i.e., $p(\mathbf{z})$ is constant. This is because the circuit is homogeneous and translation invariant on the 2D manifold, so there is no preferred location in stimulus space. Thus,

$$\begin{aligned} &\text{Likelihood gradient:}\ \ \nabla_{\mathbf{z}} \ln \mathcal{L}(\mathbf{z}) = \Lambda(\boldsymbol{\mu}_z - \mathbf{z}), \\ &\text{E bump drift term:}\quad U_{EF}(\boldsymbol{\mu}_z - \mathbf{z}_E), \end{aligned}$$

which implies

$$\nabla_{\mathbf{z}} \ln p(\mathbf{z}) = \mathbf{0}. \tag{D93}$$

Hence, in the minimal 2D circuit, the posterior is determined entirely by the likelihood. As in the 1D case, a non-uniform prior would require breaking the translational symmetry of the network through non-homogeneous neural density, non-uniform external bias, or structured recurrent connectivity.

D.5 Conditions for realizing 2D Langevin sampling in the circuit

To realize likelihood sampling in the minimal 2D circuit, the equilibrium distribution of the bump position should match the likelihood in Eq. (D90). Near the stimulus, the linearized reduced bump-position dynamics can be written as

$$\dot{\mathbf{z}}_E = (\tau U_E)^{-1} U_{EF}(\boldsymbol{\mu}_z - \mathbf{z}_E) + \sigma_z(\tau U_E)^{-1/2}\boldsymbol{\xi}_t, \tag{D94}$$

where $\boldsymbol{\xi}_t$ is a two-dimensional white-noise process with independent components,

$$\langle \xi_i(t)\xi_j(t')\rangle = \delta_{ij}\delta(t - t'), \tag{D95}$$

and

$$U_{EF} = \frac{\rho w_{EF} R_F}{2}. \tag{D96}$$

Using the 2D likelihood gradient,

$$\nabla_{\mathbf{z}} \ln \mathcal{L}(\mathbf{z}) = \Lambda(\boldsymbol{\mu}_z - \mathbf{z}), \tag{D97}$$

Eq. (D94) can be rewritten as

$$\dot{\mathbf{z}}_E = (\tau U_E)^{-1} \underbrace{\frac{U_{EF}}{\Lambda}}_{\lambda_z} \Lambda(\boldsymbol{\mu}_z - \mathbf{z}_E) + \sigma_z (\tau U_E)^{-1/2} \boldsymbol{\xi}_t, \tag{D98}$$

where, in the isotropic minimal extension,

$$\lambda_z = \frac{U_{EF}}{\Lambda} = \frac{\rho w_{EF} R_F}{2\Lambda}. \tag{D99}$$

For comparison, the standard overdamped Langevin dynamics that samples the posterior $p(\mathbf{z} \mid \mathbf{r}_F) \propto \mathcal{L}(\mathbf{z}) p(\mathbf{z})$ is

$$\dot{\mathbf{z}}_t = \tau_L^{-1} \nabla_{\mathbf{z}} \ln p(\mathbf{z} \mid \mathbf{r}_F) + (2\tau_L^{-1})^{1/2} \boldsymbol{\xi}_t. \tag{D100}$$

Because the prior is uniform, this reduces to

$$\dot{\mathbf{z}}_t = \tau_L^{-1} \Lambda(\boldsymbol{\mu}_z - \mathbf{z}_t) + (2\tau_L^{-1})^{1/2} \boldsymbol{\xi}_t. \tag{D101}$$

Therefore, the 2D circuit realizes Langevin sampling when its drift and diffusion coefficients satisfy the same ratio as in Eq. (D101). Matching the diffusion terms gives

$$\tau_L^{-1} = \frac{\sigma_z^2}{2\tau U_E}. \tag{D102}$$

Matching the drift terms then requires

$$\frac{U_{EF}}{\tau U_E} = \tau_L^{-1} \Lambda = \frac{\sigma_z^2}{2\tau U_E} \Lambda. \tag{D103}$$

Equivalently,

$$\frac{\sigma_z^2}{\lambda_z} = 2, \tag{D104}$$

which yields the optimal feedforward weight

$$w_{EF} = \frac{\sigma_z^2 \Lambda}{\rho R_F}. \tag{D105}$$

Using the isotropic 2D likelihood precision

$$\Lambda \approx 2\pi\rho R_F, \tag{D106}$$

we obtain

$$w_{EF} = 2\pi\sigma_z^2. \tag{D107}$$

If we further substitute the projected 2D bump-position noise strength

$$\sigma_z^2 = \frac{8\mathsf{F}}{9\pi}, \tag{D108}$$

then the optimal feedforward weight becomes

$$w_{EF} = \frac{16}{9}\mathsf{F}. \tag{D109}$$

The corresponding sampling time constant is

$$\tau_z = \lambda_z^{-1} \tau U_E = \frac{\Lambda}{U_{EF}} \tau U_E = \frac{4\pi}{w_{EF}} \tau U_E, \tag{D110}$$

which is again proportional to the E bump height U_E. Thus the bump-position dynamics can be written in the standard 2D Langevin form

$$\dot{\mathbf{z}}_E = \tau_z^{-1} \Lambda(\boldsymbol{\mu}_z - \mathbf{z}_E) + (2\tau_z^{-1})^{1/2} \boldsymbol{\xi}_t. \tag{D111}$$

E The distribution of step size in 2d case

E.1 2D Fokker–Planck equation for the bump-height dynamics

For the reduced 2D CANN, the bump height satisfies

$$dU_E = A_{2D}(U_E)\,dt + B_{2D}(U_E)\,dW_t, \tag{E112}$$

with

$$A_{2D}(U_E) = \tau^{-1}\left[-U_E + \frac{\rho}{2}\left(\frac{w_{EE}U_E^2}{1+2\pi a^2\rho w_{EI}U_E^2} + w_{EF}R_F\right)\right], \tag{E113}$$

$$B_{2D}(U_E) = \sigma_U(\tau^{-1}U_E)^{1/2}. \tag{E114}$$

The corresponding Fokker–Planck equation is

$$\frac{\partial p(U_E,t)}{\partial t} = -\frac{\partial}{\partial U_E}\big[A_{2D}(U_E)p(U_E,t)\big] + \frac{1}{2}\frac{\partial^2}{\partial U_E^2}\big[B_{2D}^2(U_E)p(U_E,t)\big]. \tag{E115}$$

At stationarity, imposing zero probability flux gives

$$p(U_E) = \frac{C}{B_{2D}^2(U_E)}\exp\left[2\int^{U_E}\frac{A_{2D}(u)}{B_{2D}^2(u)}\,du\right]. \tag{E116}$$

Since

$$B_{2D}^2(U_E) = \sigma_U^2\tau^{-1}U_E, \tag{E117}$$

we obtain

$$\boxed{p(U_E) \propto U_E^{\alpha_2-1}\left(1+\eta_{2D}U_E^2\right)^{\tilde{c}_2}\exp\!\left(-\frac{2U_E}{\sigma_U^2}\right),} \tag{E118}$$

where

$$\eta_{2D} = 2\pi a^2\rho w_{EI}, \qquad \alpha_2 = \frac{\rho w_{EF}R_F}{\sigma_U^2}, \qquad \tilde{c}_2 = \frac{\rho w_{EE}}{2\eta_{2D}\sigma_U^2}. \qquad \beta_2 = \frac{2}{\sigma_U^2}. \tag{E119}$$

E.2 Gamma approximation of the 2D bump-height distribution

When $\eta_{2D}U_E^2 \ll 1$, we write

$$\left(1+\eta_{2D}U_E^2\right)^{\tilde{c}_2} = \exp\big[\tilde{c}_2\ln(1+\eta_{2D}U_E^2)\big] \approx \exp\left[\frac{\rho w_{EE}}{2\sigma_U^2}U_E^2 + O(\eta_{2D}U_E^4)\right]. \tag{E120}$$

If this quadratic correction is weak over the dominant support, then to leading order

$$p(U_E) \approx \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}U_E^{\alpha_2-1}e^{-\beta_2 U_E}, \qquad \beta_2 = \frac{2}{\sigma_U^2}. \tag{E121}$$

E.3 2D Gaussian scale mixture model for bump-position increments

The reduced bump-position dynamics in 2D is:

$$d\mathbf{z}_E = \frac{1}{\tau U_E}\mathbf{F}(\mathbf{z}_E)\,dt + \sigma_z(\tau U_E)^{-1/2}d\mathbf{W}_t, \qquad \mathbf{F}(\mathbf{z}_E) = U_{EF}(\boldsymbol{\mu}_z - \mathbf{z}_E). \tag{E122}$$

Over a short interval Δt, the increment conditioned on U_E is:

$$\Delta\mathbf{z}_E \mid U_E \sim \mathcal{N}_2\big(\mathbf{f}(\mathbf{z}_E)U_E^{-1},\, DU_E^{-1}\mathbf{I}_2\big), \qquad \mathbf{f}(\mathbf{z}_E) = \frac{\mathbf{F}(\mathbf{z}_E)\Delta t}{\tau}. \tag{E123}$$

In the diffusion-dominated regime (small Δt), we neglect the subleading drift term $\mathbf{f}(\mathbf{z}_E)$ and use:

$$\Delta\mathbf{z}_E \mid U_E \approx \mathcal{N}_2\big(\mathbf{0},\, DU_E^{-1}\mathbf{I}_2\big). \tag{E124}$$

Mixing over the Gamma law of U_E yields:

$$p(\Delta\mathbf{z}_E) = \int_0^\infty \frac{1}{2\pi(D/U_E)} \exp\left[-\frac{\|\Delta\mathbf{z}_E\|^2}{2(D/U_E)}\right] \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} U_E^{\alpha_2-1} e^{-\beta_2 U_E}\, dU_E \tag{E125}$$

$$= \frac{\alpha_2}{2\pi\beta_2 D}\left(1 + \frac{\|\Delta\mathbf{z}_E\|^2}{2\beta_2 D}\right)^{-(\alpha_2+1)}. \tag{E126}$$

This is an isotropic bivariate Student-t distribution:

$$\boxed{\Delta\mathbf{z}_E \sim t_{\nu_2}\big(\mathbf{0},\ s_2^2\mathbf{I}_2\big), \qquad \nu_2 = 2\alpha_2, \qquad s_2^2 = \frac{\beta_2}{\alpha_2} D.} \tag{E127}$$

Each coordinate marginal therefore obeys:

$$\Delta z_{E,i} \sim \text{Student-}t(\nu_2, s_2), \qquad i = 1, 2. \tag{E128}$$

E.4 Distribution of the displacement magnitude

Define the jump length (radial step size):

$$r = \|\Delta\mathbf{z}_E\| = \sqrt{\sum_{i=1}^{2} (\Delta z_{E,i})^2}. \tag{E129}$$

Under the short-time, diffusion-dominated approximation,

$$\Delta\mathbf{z}_E \mid U_E \approx \mathcal{N}_2\big(\mathbf{0},\ DU_E^{-1}\mathbf{I}_2\big), \tag{E130}$$

Equivalently,

$$\Delta z_{E,i} = \sqrt{DU_E^{-1}}\, Z_i, \qquad Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{E131}$$

Hence

$$r = \sqrt{DU_E^{-1}}\sqrt{Z_1^2 + Z_2^2}. \tag{E132}$$

Since $Z_1^2 + Z_2^2 \sim \chi_2^2$, it follows that

$$r \mid U_E \sim \sqrt{DU_E^{-1}}\, \chi_2. \tag{E133}$$

Equivalently, $r \mid U_E$ follows a Rayleigh distribution with density:

$$p(r \mid U_E) = \frac{U_E}{D}\, r\, \exp\left(-\frac{U_E r^2}{2D}\right), \qquad r \geq 0. \tag{E134}$$

Using the Gamma approximation for the bump-height distribution,

$$p(U_E) = \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} U_E^{\alpha_2-1} e^{-\beta_2 U_E}, \qquad U_E > 0, \tag{E135}$$

the marginal density becomes

$$p_r(r) = \int_0^\infty p(r \mid U_E) p(U_E)\, dU_E \tag{E136}$$

$$= \frac{\beta_2^{\alpha_2} r}{\Gamma(\alpha_2) D} \int_0^\infty U_E^{\alpha_2} \exp\left[-U_E\left(\beta_2 + \frac{r^2}{2D}\right)\right] dU_E. \tag{E137}$$

Applying the Gamma integral yields:

$$p_r(r) = \frac{\Gamma(\alpha_2+1)}{\Gamma(\alpha_2)} \frac{\beta_2^{\alpha_2} r}{D} \left(\beta_2 + \frac{r^2}{2D}\right)^{-(\alpha_2+1)}. \tag{E138}$$

Using $\Gamma(\alpha_2 + 1) = \alpha_2\Gamma(\alpha_2)$ and defining

$$A_2 = 2\beta_2 D, \tag{E139}$$

we obtain

$$\boxed{p_r(r) = \frac{2\alpha_2}{A_2}\, r \left(1 + \frac{r^2}{A_2}\right)^{-(\alpha_2+1)}, \qquad r \geq 0.} \tag{E140}$$

F Circuit simulation parameters and details

F.1 Critical weight

To scale the connection strengths in our network model, we use a critical recurrent connection strength as a reference point. This critical strength is defined as the smallest value that allows the network to maintain persistent activity even when there is no feedforward input.

F.1.1 1D case

In the absence of feedforward input ($R_F = 0$), the stationary state of the 1D circuit's excitatory bump height is determined solely by the recurrent input:

$$U_E = U_{EE} = \frac{\rho}{\sqrt{2}} w_{EE} R_E. \tag{F1}$$

Furthermore, the firing rate of the E population, R_E, is related to its input U_E by the divisive normalization activation function. Since the spatial integral of the squared 1D Gaussian bump is $\int_{-\pi}^{\pi} u_E(\theta)^2 d\theta \approx \sqrt{2\pi} a U_E^2$, the firing rate is:

$$R_E = \frac{U_E^2}{1 + \sqrt{2\pi} a \rho w_{EI} U_E^2}. \tag{F2}$$

Substituting Eq. (F2) into Eq. (F1) allows us to write an equation solely in terms of U_E:

$$U_E = \frac{\rho}{\sqrt{2}} w_{EE} \frac{U_E^2}{1 + \sqrt{2\pi} a \rho w_{EI} U_E^2}.$$

Assuming $U_E \neq 0$ (for persistent activity), we can divide by U_E and rearrange the equation into a quadratic form for U_E:

$$\sqrt{2\pi} a \rho w_{EI} U_E^2 - \frac{\rho}{\sqrt{2}} w_{EE} U_E + 1 = 0.$$

This quadratic equation for U_E has real solutions if and only if its discriminant is non-negative ($\Delta \geq 0$):

$$\frac{\rho^2}{2} w_{EE}^2 - 4\sqrt{2\pi} a \rho w_{EI} \geq 0.$$

The critical weight, w_c, is the smallest value of w_{EE} that permits non-zero persistent activity, which occurs when the discriminant is exactly zero:

$$w_c^2 = \frac{8\sqrt{2\pi} a w_{EI}}{\rho} \quad \Rightarrow \quad w_c = \sqrt{\frac{8\sqrt{2\pi} a w_{EI}}{\rho}}. \tag{F3}$$

F.1.2 2D case

In the 2D network extension, the stationary state of the bump height without feedforward input is similarly driven by the 2D recurrent input projection:

$$U_E = U_{EE} = \frac{\rho}{2} w_{EE} R_E. \tag{F4}$$

For the 2D divisive normalization, the spatial integral of the squared 2D Gaussian bump over the feature space is $\iint u_E(\boldsymbol{\theta})^2 d\boldsymbol{\theta} \approx 2\pi a^2 U_E^2$. Thus, the peak firing rate becomes:

$$R_E = \frac{U_E^2}{1 + 2\pi a^2 \rho w_{EI} U_E^2}. \tag{F5}$$

Substituting Eq. (F5) into Eq. (F4) yields:

$$U_E = \frac{\rho}{2} w_{EE} \frac{U_E^2}{1 + 2\pi a^2 \rho w_{EI} U_E^2}.$$

Assuming $U_E \neq 0$, we divide by U_E and rearrange to form the 2D quadratic equation:

$$2\pi a^2 \rho w_{EI} U_E^2 - \frac{\rho}{2} w_{EE} U_E + 1 = 0.$$

For persistent 2D activity to exist, the discriminant of this quadratic must be non-negative:

$$\frac{\rho^2}{4} w_{EE}^2 - 8\pi a^2 \rho w_{EI} \geq 0.$$

Setting the discriminant to zero gives the 2D critical weight w_c:

$$w_c^2 = \frac{32\pi a^2 w_{EI}}{\rho} \quad \Rightarrow \quad w_c = \sqrt{\frac{32\pi a^2 w_{EI}}{\rho}}. \tag{F6}$$

F.2 Simulation and Statistical Fitting Procedures

F.2.1 Network Simulation Protocol

To systematically investigate the network dynamics across different dynamical regimes, simulations were conducted over a comprehensive grid of parameters, specifically varying the feedforward input intensity (R_f), the recurrent excitatory synaptic weight (w_{ee}), and the global normalization constant (k).

For each parameter combination, the network dynamics were numerically integrated over a fixed simulation duration. To ensure the analysis accurately reflected the steady-state equilibrium properties of the network, an initial burn-in period (t_{steady}) was discarded from the time-series data to eliminate transient initialization artifacts. From this raw steady-state time series, we extracted the macroscopic variables governing the network's continuous attractor dynamics. The bump height (U) was calculated as the spatial integral of the excitatory synaptic input across the network, normalized by the theoretical bump width and neuron density. The center of mass (bump position, z for 1D, or $\mathbf{r} = (x, y)$ for 2D) was decoded at each time step, from which the raw coordinate step sizes between successive time points were extracted for statistical analysis.

F.2.2 Statistical Distribution Fitting

To characterize the stochastic nature of the network's continuous attractor, we evaluated the empirical distributions of the macroscopic variables against several theoretical probability density functions. Parameters for all theoretical distributions were estimated using Maximum Likelihood Estimation (MLE). Model selection was performed by computing the Akaike Information Criterion (AIC) for each fit, where a lower (or negative relative) AIC indicates a superior model fit.

Bump Height (U): The steady-state distribution of the integrated bump height U exhibits strictly positive properties. We compared fits between a symmetric Gaussian distribution and a Gamma distribution. The superiority of the Gamma distribution was determined via $\Delta\text{AIC} = \text{AIC}_{\text{Gamma}} - \text{AIC}_{\text{Gaussian}}$.

Coordinate Step Increments (Δz, Δx, Δy): For 1D spatial dynamics, the coordinate step size Δz was extracted. For 2D spatial dynamics, the step vector was decomposed into its orthogonal components, Δx and Δy. To capture the anomalous diffusion and heavy-tailed jumping behaviors observed in specific parameter regimes, the empirical distributions of these signed increments were fitted to a Student's t-distribution (characterized by degrees of freedom ν, location μ, and scale σ) and compared against a standard Gaussian baseline.

2D Step Magnitude (Δr): In the 2D network, the absolute step magnitude $\Delta r = \sqrt{\Delta x^2 + \Delta y^2}$ is strictly non-negative and requires a specialized radial distribution to accurately capture heavy-tailed radial jumps. We modeled this using a custom Burr distribution, defined by the probability density function:

$$P(r) = \frac{2\kappa}{A} r \left(1 + \frac{r^2}{A}\right)^{-(\kappa+1)}, \quad \text{for } r > 0 \tag{F7}$$

where κ dictates the shape (tail heaviness) and A governs the scale. The parameters were derived via numerical optimization of the negative log-likelihood. The AIC of this custom burr distribution was then compared against the AIC of a baseline Gaussian radial fit.

F.2.3 Heavy Tail Quantification

To systematically quantify the presence of heavy tails in the step-size distributions independently of global MLE fits, we analyzed the CCDF of the absolute step magnitudes.

The empirical CCDF, $P(X > x)$, was computed and evaluated in log-log space. Rather than relying on a fixed heuristic threshold, the onset of the heavy tail was identified algorithmically. We iteratively performed linear regression on the log-log empirical CCDF across a sweeping range of candidate upper-percentile thresholds. The optimal tail threshold was selected as the point that maximized the Pearson correlation coefficient (R^2), indicating the most linear tail regime in log-log space. The heavy-tail index Θ was subsequently extracted from the negative slope of this optimal linear fit.

A.4 Simulation Parameters

The core network simulations were carried out using the parameters listed in Table 1. The parameter sets for 1D and 2D spatial environments are largely identical, with specific scalar adjustments made to the global normalization constant, total simulation time, and the analytical feedforward scale to maintain consistent bump dynamics across dimensions.

Table 1: Core Network and Simulation Parameters

Parameter	**Variable**	**Value**
Integration & Timing		
Time Constant	τ	1
Integration time step	dt	0.1 τ
Total simulation time (1D)	T	500 τ
Total simulation time (2D)	T	500 τ
Burn-in period	t_{steady}	50 τ
Network Architecture		
Excitatory time constant	τ_E	1.0
Number of neurons	N	180
Spatial domain		$[-180°, 180°)$
Excitatory connection width	a_E	$40°$
Dynamics & Synaptic Weights		
Fano factor (noise intensity)	F	0.5
Baseline recurrent weight	w_{ee}	$0.4 w_c$
Normalization constant (1D)	k	5×10^{-4}
Normalization constant (2D)	k	5×10^{-5}
Feedforward scale (1D)	w^{EF}	0.716
Feedforward scale (2D)	w^{EF}	0.605

G Generalization

G.1 Encoding a Non-Uniform Prior

To successfully encode a non-uniform marginal prior, the network must break the functional symmetry of its excitatory (E) neurons. Unlike standard models where tuning curves are homogeneous and translation-invariant across the stimulus space, storing a complex prior requires introducing heterogeneity. This can involve varying the tuning widths, amplitudes, and spatial distributions of the E neurons across the stimulus subspace.

Biologically, this required tuning heterogeneity can emerge through two primary circuit mechanisms:

- **Internal Network Structure:** Embedding the prior directly into the local circuit by superimposing a heterogeneous weight structure over the standard translation-invariant recurrent weight matrix.
- **External Top-Down Input:** Supplying a sustained prior signal from higher cortical regions.

Focusing on the external input mechanism, we can demonstrate that an applied prior signal successfully preserves the critical relationship between the bump amplitude and the Fisher Information (FI). For analytical simplicity, we model this external prior as a time-invariant Gaussian profile added to the excitatory network dynamics:

$$\rho\textstyle\sum_X(\mathbf{W}_{Ep} * \mathbf{r}_p)(\theta,t), \quad \mathbf{r}_p(\theta,t) = R_p \exp[-(\theta - z_p)^2/2a^2]$$

Incorporating this into the primary circuit dynamics yields:

$$\tau\dot{\mathbf{u}}_E(\theta,t) = -\mathbf{u}_E(\theta,t)+\rho\textstyle\sum_X(\mathbf{W}_{EX}*\mathbf{r}_X)(\theta,t)+\rho\sum_X(\mathbf{W}_{Ep}*\mathbf{r}_p)(\theta,t)+\sqrt{\tau\mathsf{F}[\mathbf{u}_E(\theta,t)]_+}\xi(\theta,t)$$

Because the external prior input functions mathematically in the same manner as the sensory feedforward input, the governing equations for the bump's spatial position (z_E) and amplitude (U_E) can be updated directly. The resulting stochastic differential equations are:

$$\text{Position}: \ \dot{z}_E \approx (\tau U_E)^{-1}U_{EF}(\mu_z - z_E) + (\tau U_E)^{-1}U_{EP}(\mu_p - z_E) + \sigma_z(\tau U_E)^{-1/2}\xi_t \quad \text{(G1a)}$$

$$\text{Height}: \quad \dot{U}_E \approx \tau^{-1}[-U_E + U_{EE} + U_{EF} + U_{EP}] + \sigma_U(\tau^{-1}U_E)^{1/2}\xi_t, \quad \text{(G1b)}$$

Here, U_{EP} and μ_p represent the precision and the mean of the non-uniform marginal prior, respectively. The drift term in the position dynamics now integrates both the sensory feedforward input and the external prior. As a result, when the bump position z reaches equilibrium, the network correctly samples from a posterior distribution that combines both sources of information:

$$p(z) \propto \exp\left[-\frac{\sqrt{2\pi}}{2}\rho R_f a^{-1}(\mu_z - z_E)^2 - \frac{\sqrt{2\pi}}{2}\rho R_p a^{-1}(\mu_p - z_E)^2\right]$$

To validate this external prior configuration, we simulated the network dynamics with a spatially shifted prior. The prior input and the likelihood input were assigned equal intensities ($R_p = R_f = 0.01$). However, the prior's mean ($\mu_p = 5$) was deliberately shifted 5° away from the likelihood's mean ($\mu_z = 0$).

As shown in the analysis of the sampling dynamics(Fig.A2), this offset configuration successfully preserves the network's long-tail sampling behavior. The probability density of the bump amplitude (U_E) exhibits right-skewed fluctuations that are best captured by a Gamma distribution, rather than a standard Gaussian fit. And the distribution of the sampling step sizes significantly deviates from simple diffusion. It tightly aligns with a Student-t distribution, confirming the presence of heavy-tailed jump dynamics. Plotting the CCDF of the step sizes on a log-log scale reveals a distinct linear relationship in the tail region. The empirical data yields a tail exponent of $\Theta = 1.83$, mathematically demonstrating that the network utilizes long-range, Lévy-flight-like jumps to sample the modified posterior space efficiently.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We perform mathematically rigorous theoretical analysis on the nonlinear circuit dynamics (Sections 3-5) and verify our theoretical results via numerical simulations (Figures 2 and 3) to directly support our core claim: the intrinsic emergence of scale-mixture Langevin sampling without explicit non-Gaussian components.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We explicitly discuss the limitations of our study in the "Conclusion and Discussion" section under the "Limitations, generalizations, and future directions" paragraph.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the full set of assumptions (e.g., uniform priors, translation invariance, and weak divisive normalization) and complete mathematical derivations for all theoretical claims, including the Fokker-Planck solutions and Gaussian scale mixture derivations, in the comprehensive Appendices (Appendices A through D).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main text clearly outlines the mathematical formulation and assumptions of the neural circuit model. Furthermore, Appendix F provides comprehensive details on the simulation protocols and the statistical fitting procedures used to quantify the heavy-tailed distributions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The custom code used for all network simulations and statistical fitting is included in the supplementary materials. Furthermore, comprehensive instructions, parameter values, and evaluation protocols required to faithfully reproduce these results are detailed in Appendix F ("Circuit simulation parameters and details").

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Our study focuses on the theoretical analysis and numerical simulation of stochastic differential equations governing continuous attractor networks, rather than training machine learning models on datasets. Therefore, concepts like data splits or optimizers do not apply.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: While this work relies on numerical simulations of nonlinear circuit dynamics rather than physical experiments, we provide comprehensive statistical evaluations to substantiate our claims. This includes Maximum Likelihood Estimation (MLE) for distribution fitting, ΔAIC heatmaps to demonstrate the statistical superiority of the Student-t and Gamma models over Gaussian baselines, and algorithmic tail exponent extraction (Θ) using CCDFs (as discussed in the text for Figures 2 and 4). Complete details of these statistical procedures are provided in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: They are provided in the Supplementary information Sec.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research are conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work is a foundational, purely theoretical study focused on mathematically modeling the biological dynamics of recurrent continuous attractor networks (CANs) in the cortex. Because the research aims to explain intrinsic neural mechanisms rather than introducing a deployed artificial intelligence system or applied machine learning tool, a discussion of direct societal impacts falls outside the scope of this paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper focuses purely on theoretical models of biological neural circuits and does not involve the creation or release of high-risk AI models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers for all adapted models and code used, allowing readers to locate the original versions.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper focuses purely on theoretical mathematical modeling and numerical simulation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper is a theoretical neuroscience study based exclusively on mathematical modeling and numerical simulations. It does not involve human subjects, crowdsourcing, or behavioral experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This is the theoretical work so we don't involve these risks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The theoretical development and analyses do not involve any LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.