Gaussian Chaos Noise: Variational Noise Design for Reliable Deep Learning

Anonymous Authors[1]

Abstract

We develop a first-principles variational framework for noise injection that treats noise as a *design object*: instead of selecting an exogenous heuristic (e.g., i.i.d. dropout or hard masking), the noise structure is derived from minimal, task-driven desiderata encoded as variational optimality conditions. We refer to this design framework as the *Variational Kernel Design* (VKD), since it characterizes not only a noise distribution but also the correlation geometry (kernel) implied by the constraints.

Instantiating VKD under standard requirements for spatial perturbations yields *Gaussian Chaos Noise*: a mean-preserving positive multiplicative gate given by Wick-normalized exponentiation of a Laplacian/Green-correlated Gaussian field, with the Green kernel emerging as the canonical correlation structure from optimality. This answers a basic but under-addressed question: *what correlation structure should the "right" noise have* for a given representation under natural constraints? The resulting noise is scale-tolerant, geometrically stable, and well-suited to late semantic stages where long-range coherence makes hard discontinuities brittle.

Empirically, across major datasets and architectures, GCh consistently improves reliability metrics (NLL/ECE) without sacrificing classification accuracy, including in late-stage injection regimes where hard masking can degrade performance.

On clean ImageNet val (late-stage setting), GCh reduces ECE by 33% (0.030→0.020). On ImageNet-C, compared to the no-noise baseline, our method reduces ECE by 46% (0.105→0.056) and NLL by 3.3% (3.40→3.29) (Appendix Table 7).

We hope this work motivates a design-oriented view of noise and positions VKD as a practical blueprint for constructing noise mechanisms tailored to specific learning objectives.

1. Introduction

Noise injection is a ubiquitous tool in modern deep learning—used to improve generalization, calibration, and robustness—yet the structure of injected noise is still largely chosen exogenously by convention (e.g., Dropout (Srivastava et al., 2014), Stochastic Depth (Huang et al., 2016), block masking (Ghiasi et al., 2018)). This raises a simple but under-addressed question:

> *What noise structure is "right" for a given representation and objective under natural first-principles constraints?*

Core novelty: right correlation kernel is forced by desiderata. We do *not* treat the correlation kernel as a tunable modeling choice. Once deep-learning desiderata on spatial representations are stated as an operator-level constraint (locality + a Dirichlet-energy smoothness budget with gauge fixing), the correlation geometry is determined: the kernel must be the (Dirichlet) Green kernel (inverse Laplacian). MaxEnt enters only as a minimal extra-information principle within this constrained design class.

In this work we study noise as a *design variable* rather than a fixed heuristic, and develop a variational framework for its construction.

Noise appears throughout the pipeline. It acts as implicit regularization via additive perturbations (Bishop, 1995), as stochastic gating in hidden representations (Dropout (Srivastava et al., 2014), Stochastic Depth (Huang et al., 2016)), and as spatially structured occlusion in convolutional feature maps (Cutout (DeVries & Taylor, 2017), DropBlock (Ghiasi et al., 2018)). At the data level, popular augmentation pipelines can be viewed as structured noise processes (Mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), AutoAugment (Cubuk et al., 2019), RandAugment (Cubuk et al., 2020), AugMix (Hendrycks et al., 2020)). Meanwhile, modern deployment increasingly prioritizes *reliability* under distribution shift, including calibration (Guo et al., 2017) and robustness to common corruptions (Hendrycks & Dietterich, 2019), making the design of training-time uncertainty consequential.

Despite this ubiquity, injected noise is typically i.i.d. or hard-masked, implicitly assuming spatial independence and

ignoring the geometry of intermediate representations. This can become brittle in late semantic stages, where feature maps (or token grids) exhibit strong long-range coherence: discontinuous perturbations may destroy correlations and destabilize confidence, leading to depth-dependent behavior.

We therefore study noise injection as a *design problem*. We cast a noise mechanism as a triple $\mathcal{N} = (\mathcal{F}, K, \mathcal{T})$ (distribution family, correlation prior, injection operator), and derive the noise structure from learning desiderata via variational optimality. Instantiating this framework yields our *Gaussian Chaos Noise*: a smooth, spatially correlated, *mean-preserving positive* multiplicative gate given by Wick-normalized exponentiation of a Laplacian/Green-correlated Gaussian field, arising under a spatial smoothness budget and a maximum-entropy (equivalently, minimum-information) principle.

Note: We refer to our mechanism as Gaussian Chaos Noise, and denote it by **GCh** throughout (to avoid confusion with graph convolutional networks).

Empirically, GCh improves reliability metrics (NLL/ECE) while maintaining competitive classification accuracy, and remains effective in late stages where hard masking can degrade performance.

Contributions. Our contributions can be summarized as follows:

- **First-principles noise design.** We formulate noise injection as a variational design problem and derive its structure from minimal assumptions, rather than heuristic choices.
- **Canonical correlation is forced by desiderata.** We answer "what correlation structure is right?" by showing that once spatial desiderata are encoded as a local operator constraint (a Dirichlet-energy smoothness budget with gauge fixing), the correlation geometry is determined as the Dirichlet Green kernel $G = L_U^{-1}$. MaxEnt is used only as a minimal extra-information principle within this constrained design class.
- **Late-stage reliability and robustness.** We identify and empirically characterize a depth-dependent late-stage degradation of hard spatial masking, and demonstrate that GCh remains effective in late semantic stages, improving calibration and robustness without accuracy trade-offs.
- **Controlled ablations and mechanism.** Through causal controls, layer-wise ablations, and strength sweeps, we show that correlation, positivity, and injection depth are essential, and provide mechanistic evidence linking continuous correlated noise to preserved spectral and spatial structure.

1.1. Related Work

Noise injection and regularization in deep networks. Small additive noise is classically linked to Tikhonov-style regularization (Bishop, 1995). Dropout injects i.i.d. Bernoulli gating (Srivastava et al., 2014); stochastic depth drops residual branches (Huang et al., 2016) and is extended to Transformers via LayerDrop (Fan et al., 2019); ShakeDrop perturbs residual branches with randomized coefficients (Yamada et al., 2018). Spatial occlusion includes Cutout (DeVries & Taylor, 2017) and DropBlock (Ghiasi et al., 2018), while sample-level mixing includes Mixup and CutMix (Zhang et al., 2018; Yun et al., 2019). In ViTs, PatchDropout drops input patches and alters token topology (Liu et al., 2023). A recurring limitation is that noise is typically fixed a priori and often i.i.d. or hard-masked, which can mismatch late semantic representations where long-range coherence matters.

Calibration, reliability, and uncertainty-aware learning. Miscalibration is common and temperature scaling is a strong post-hoc baseline (Guo et al., 2017); nonparametric alternatives include BBQ (Naeini et al., 2015), and Dirichlet calibration extends beyond a single temperature (Kull et al., 2019). Dropout admits a Bayesian interpretation as approximate inference (Gal & Ghahramani, 2016), while deep ensembles remain a competitive uncertainty baseline (Lakshminarayanan et al., 2017). Under distribution shift, calibration can degrade substantially (Ovadia et al., 2019), and recent evidence shows calibration depends strongly on architecture/training recipe (Minderer et al., 2021). Label smoothing can improve calibration but is context-dependent (Müller et al., 2019). These results motivate improving NLL/ECE without sacrificing accuracy, especially under shift; our approach targets this directly by injecting uncertainty into representations as a positive, mean-preserving, correlated multiplicative gate rather than heuristic hard masking or purely i.i.d. perturbations.

Robustness under distribution shift and corruptions. For worst-case robustness, adversarial training (Madry et al., 2018) and TRADES (Zhang et al., 2019) formalize the robustness–accuracy trade-off. For average-case corruptions, ImageNet-C/P provide standardized benchmarks (Hendrycks & Dietterich, 2019), though robustness on synthetic corruptions may not transfer to natural shifts (Taori et al., 2020); broader OOD suites further emphasize heterogeneity across shift types (Hendrycks et al., 2021). Simple policies such as RandAugment and AugMix improve corruption robustness and uncertainty with low overhead (Cubuk et al., 2020; Hendrycks et al., 2020); properly tuned Gaussian/speckle noise can also be strong (Rusak et al., 2020). Noisy Student combines self-training with strong injected noise and improves ImageNet accuracy and robustness on

ImageNet-A/C/P (Xie et al., 2020).

2. Noise as a Design Object: A Unified Design Framework

High-level view. Rather than treating a noise mechanism as an exogenous choice (picked from a menu of standard perturbations), we view *noise* as a *design object*. The central question becomes: given a learning context and a set of desired properties, *what stochastic mechanism should be constructed* so that those properties hold by design?

Noise design triple. We parameterize a spatial noise mechanism by a triple

$$\boxed{\mathsf{N} \;=\; (\mathcal{F},\, K,\, \mathcal{T})}, \tag{1}$$

where each component captures a distinct axis of design:

(i) Distribution family $\mathcal{F}$ (what is sampled). $\mathcal{F}$ is a family of laws over random fields on a domain Ω (e.g. a grid V or its interior U):

$$\phi \sim F, \qquad F \in \mathcal{F}, \qquad \phi \in \mathbb{R}^{\Omega}.$$

This specifies admissible marginals/support (e.g. positivity), tail behavior, and global structure (Gaussian, chaos, mixtures, etc.).

(ii) Kernel K (how it correlates). K is a positive semidefinite kernel on $\Omega \times \Omega$ that encodes the intended correlation geometry (locality, smoothness, anisotropy, scale):

$K(x, y)$ encodes the prescribed dependence between $\phi(x)$ and $\phi(y)$.

In Gaussian designs, K coincides with the covariance; more generally it serves as the second-order descriptor that controls spatial coupling.

(iii) Injection operator $\mathcal{T}$ (where/how it acts). $\mathcal{T}$ injects a sampled field into the learning pipeline. Given a tensor F (e.g. a feature map at some layer), it produces a perturbed tensor

$$\widetilde{F} \;=\; \mathcal{T}(F; \phi), \tag{2}$$

covering multiplicative gating ($\widetilde{F} = F \odot \phi$), additive injection, and structured variants (channel-wise, block-wise, attention-compatible, multi-scale).

Here $\odot$ denotes the elementwise (Hadamard) product with spatial broadcasting. Concretely, if $F \in \mathbb{R}^{C\times H\times W}$ is a feature map and $\phi \in \mathbb{R}^{H\times W}$ is a spatial gate, then

$$(\widetilde{F})_{c,i,j} \;=\; F_{c,i,j}\, \phi_{i,j}, \tag{3}$$

for $c = 1, \ldots, C, \;\; i = 1, \ldots, H, \;\; j = 1, \ldots, W$. (Equivalently, ϕ is replicated across channels and multiplied pointwise.)

Design principle (from desiderata to mechanism). The triple $\mathsf{N} = (\mathcal{F}, K, \mathcal{T})$ is *derived*, not chosen. We start from learning-relevant desiderata (e.g. minimal semantic injection, unbiasedness, spatial coherence, locality, stability), translate them into mathematical constraints on admissible mechanisms, and then *solve for* (or *characterize*) the mechanisms consistent with those constraints. In this sense, the framework is a *blueprint*: it formalizes how empirical objectives and inductive biases are converted into a concrete stochastic design.

How this paper uses the framework. Within this blueprint, we provide a complete end-to-end design example: we specify a small set of learning-motivated constraints, show that they *force* a particular correlation geometry and distributional form, and then give an explicit implementation interface that turns the resulting mechanism into a drop-in noise module. The resulting construction yields a structured, positive, spatially correlated multiplicative noise (a Gaussian-chaos gate) on a finite grid.

3. Background and Notations

3.1. Context and notations

We consider a deep network and fix an injection site (layer) at which a feature map is perturbed. Let

$$h \in \mathbb{R}^{C\times H\times W}$$

denote the feature tensor at that site, where C is the number of channels and $H \times W$ is the spatial resolution. We index h by channel $c \in \{1, \ldots, C\}$ and spatial location $x = (i, j) \in V \;=\; \{1, \ldots, H\} \times \{1, \ldots, W\}$.

Spatial noise field. We focus on *spatial* perturbations: a random field acts on the $H \times W$ grid and is shared across channels. Concretely, we introduce a (typically positive) spatial field,

$$\nu : V \to (0, \infty),$$

for $V \;=\; \{1, \ldots, H\} \times \{1, \ldots, W\}$.

and apply it to all channels at the same spatial location. This matches the operating regime of spatial masking and gating regularizers.

Injection operators. The multiplicative injection operator induced by ν is defined by

$$\mathcal{T}_{\nu}(h)(c, x) \;=\; h(c, x)\, \nu(x), \tag{4}$$

i.e. elementwise multiplication with broadcasting across channels.

For numerical stability and controlled perturbation strength, we also consider a residual multiplicative gate:

$$\mathcal{T}_\nu^{\mathrm{res}}(h)(c,x) \;=\; h(c,x)\,\Big(1{+}\alpha\big(\nu(x){-}1\big)\Big), \qquad \alpha\in(0,1]. \tag{5}$$

Unless otherwise stated, we set $\alpha = 1$.

3.2. Discrete Gaussian free field on rectangular grid

Fix integers $H, W \geq 2$. Let

$$V \;=\; \{1,\dots,H\}\times\{1,\dots,W\}$$

be the vertex set (sites). Write $x = (i,j) \in V$. Equip V with the nearest-neighbor undirected edge set

$$E \;=\; \big\{\{x,y\}\subset V:\ \|x-y\|_1 = 1\big\}.$$

Optionally allow positive symmetric edge weights (conductances) $c_{xy} = c_{yx} > 0$ for $\{x,y\} \in E$; by unweighted case we mean $c_{xy} \equiv 1$.

Define the (vertex) boundary and interior sets

$$B \;=\; \{(i,j)\in V:\ i\in\{1,H\} \text{ or } j\in\{1,W\}\},$$

for

$$U \;=\; V\setminus B.$$

A field is a function $\phi : V \to \mathbb{R}$. We impose *Dirichlet boundary condition* $\phi|_B \equiv 0$ and identify ϕ with its restriction to U; hence also $\phi \in \mathbb{R}^U$.

Discrete Laplacian and Dirichlet energy. For any function $f : V \to \mathbb{R}$ define the weighted graph Laplacian $(Lf) : V \to \mathbb{R}$ by

$$(Lf)(x) \;=\; \sum_{y:\{x,y\}\in E} c_{xy}\,\big(f(x)-f(y)\big), \qquad x\in V.$$

Let L_U denote the *Dirichlet Laplacian* on U, i.e. the restriction of L to U acting on vectors $\phi \in \mathbb{R}^U$ with boundary values fixed at 0:

$$(L_U\phi)(x) \;=\; \sum_{y:\{x,y\}\in E} c_{xy}\,\big(\phi(x)-\widetilde{\phi}(y)\big), \qquad x\in U,$$

where $\widetilde{\phi}(y) = \phi(y)$ if $y \in U$ and $\widetilde{\phi}(y) = 0$ if $y \in B$.

The associated Dirichlet energy (quadratic form) is

$$\mathcal{E}(\phi) \;\coloneqq\; \frac{1}{2}\,\langle\phi, L_U\phi\rangle \;=\; \frac{1}{2}\sum_{\{x,y\}\in E} c_{xy}\,\big(\widetilde{\phi}(x)-\widetilde{\phi}(y)\big)^2, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product on $\mathbb{R}^U$, and $\widetilde{\phi}$ is the extension by 0 to B. Under Dirichlet boundary conditions, L_U is **symmetric positive definite**; hence $\mathcal{E}(\phi) > 0$ for $\phi \neq 0$.

Definition of the discrete GFF (Dirichlet). Fix a scale (inverse temperature) $\beta > 0$. The *(Dirichlet) discrete Gaussian free field* on U is the centered Gaussian vector

$$\phi \sim \mathcal{N}\big(0,\ (\beta L_U)^{-1}\big),$$

equivalently the probability density on $\mathbb{R}^U$ given by

$$\begin{aligned} p_\beta(\phi) \;&=\; \frac{1}{Z_\beta}\,\exp\Big(-\beta\,\mathcal{E}(\phi)\Big) \\ &=\; \Big(\frac{\det(\beta L_U)}{(2\pi)^{|U|}}\Big)^{\frac{1}{2}}\,\exp\Big(-\tfrac{1}{2}\,\phi^\top(\beta L_U)\phi\Big), \end{aligned} \tag{7}$$

where the normalizing constant is

$$Z_\beta = (2\pi)^{|U|/2}\det(\beta L_U)^{-1/2}.$$

Green function (correlation kernel). Define the (Dirichlet) Green matrix $G_U \in \mathbb{R}^{|U|\times|U|}$ by

$$G_U \;\coloneqq\; L_U^{-1}.$$

Then the covariance of the field is

$$\mathrm{Cov}\big(\phi(x),\phi(y)\big) \;=\; (\beta^{-1}G_U)(x,y), \qquad x,y\in U. \tag{8}$$

4. Design-to-math for our GCh mechanism.

We want to design a spatial noise mechanism on a finite $H \times W$ grid as a positive multiplicative gate $\xi : U \to (0,\infty)$ applied to intermediate feature maps. Our goal is to impose only learning-relevant desiderata—so that the resulting noise is *structured* yet *non-committal* beyond what is required.

D1 **Least additional information (maximum entropy).**
Deep learning intuition. Regularization/augmentation noise should not inject unintended semantics; among all noises satisfying the required constraints, we choose the one that commits to the least extra information ("minimum information perturbation").
Mathematical interpretation. Among all admissible laws satisfying the constraints below, select the maximum-entropy distribution.

D2 **Positivity and multiplicative gating.**
Deep learning intuition. A gate should modulate feature amplitude without sign-flips or unnatural artifact patterns; positivity also makes the noise compatible with common multiplicative mechanisms (dropout-like masking, gain control, attention-style reweighting).
Mathematical interpretation. Model the noise as $\xi : U \to (0,\infty)$ and reparameterize via the log-field $\psi \coloneqq \log\xi \in \mathbb{R}^U$.

D3 **Unbiasedness (no systematic scale drift).**
Deep learning intuition. The noise should perturb features without introducing a persistent brightness/contrast (gain) shift that the network must waste capacity to cancel.
Mathematical interpretation. We enforce mean-one gating: $\mathbb{E}[\xi(x)] = 1$ for all sites, eliminating systematic scale drift. (Equivalently, this corresponds to a deterministic correction in the exponential gate; see Appendix D.)

D4 **Spatial coherence via a smoothness budget (structure without overfitting).**
Deep learning intuition. Pixelwise independent noise is overly high-frequency and destroys local structures; we want perturbations that are spatially coherent, respecting the grid locality inductive bias while controlling the overall "roughness budget".
Mathematical interpretation. Constrain the expected Dirichlet energy of the log-field:

$$\mathbb{E}\big[\tfrac{1}{2}\langle\psi, L_U\psi\rangle\big] = \varepsilon,$$

where L_U is the (local) Dirichlet Laplacian induced by the grid graph.

Remark. Other local quadratic energies (or boundary conditions) lead to the same MaxEnt-to-Gaussian principle, with the kernel given by the corresponding inverse operator; we focus on Dirichlet energy on grids as the canonical spatial choice and for sampling convenience (Appendix B).

D5 **Well-posedness / gauge fixing (remove unidentifiable global modes).**
Deep learning intuition. Global additive drifts (e.g. uniform gain/offset modes) are typically unidentifiable or absorbed by normalization layers; we fix this degree of freedom to make the noise mechanism stable and reproducible.
Mathematical interpretation. Adopt Dirichlet boundary conditions (or pinning / zero-mean constraint) so that $L_U \succ 0$ and the covariance operator is well-defined.

5. Main theorem: Design-forced Green-kernel log-field and the induced Gaussian-chaos gate

5.1. Kernel forcing via a minimal extra-information (MaxEnt) principle

We work on a finite grid domain U with Dirichlet gauge fixing, so the discrete Laplacian L_U is positive definite. We impose a smoothness/energy budget on the log-field ψ:

$$\mathbb{E}\left[\frac{1}{2}\langle\psi, L_U\psi\rangle\right] = \varepsilon, \mathbb{E}[\xi(x)] = 1 \text{ for all } x \in U, \quad (9)$$

with $\xi = \exp(\psi)$ with a deterministic correction.

Proposition 5.1 (Forced Green-kernel correlation geometry)**.** *Among all laws on $\psi \in \mathbb{R}^U$ satisfying the budget above, the maximum-entropy solution is Gaussian with precision βL_U for some $\beta > 0$ determined by ε. Equivalently,*

$$\psi \sim \mathcal{N}(0, (\beta L_U)^{-1}), \quad \mathrm{Cov}(\psi) = \beta^{-1} G_U,\ G_U := L_U^{-1}.$$

Proof sketch. Maximizing Shannon entropy under the quadratic moment constraint yields a Gibbs form $p(\psi) \propto \exp\{-\beta \cdot \frac{1}{2}\langle\psi, L_U\psi\rangle\}$ with a Lagrange multiplier β. Since the exponent is quadratic and $L_U \succ 0$, this Gibbs law is exactly a centered Gaussian with precision βL_U. Thus the correlation geometry is fixed as the inverse operator $G_U = L_U^{-1}$ (Dirichlet Green kernel), up to the scalar factor β^{-1} set by the budget ε. See Appendix B for the full variational derivation and the explicit $\beta(\varepsilon)$ relation.

5.2. From the forced GFF log-field to multiplicative Gaussian chaos

In Proposition 5.1 we have shown that under the minimal extra-information principle with the Dirichlet-energy budget, the *log-field* ψ is a discrete GFF with covariance $(\beta L_U)^{-1}$. (Here "discrete GFF" simply means a centered Gaussian field with covariance given by the Green operator $(\beta L_U)^{-1}$.)

Concretely, we work with a random field

$$\psi \in \mathbb{R}^U, \qquad \psi \sim \mathcal{N}(0, C), \quad C = (\beta L_U)^{-1}. \quad (10)$$

where $L_U \succ 0$ is the Dirichlet Laplacian on U, and $\beta > 0$ is the Lagrange multiplier (inverse temperature) induced by the energy budget.

Wick exponential (discrete Gaussian chaos). Fix $\gamma \in \mathbb{R}$. Define the (sitewise) Wick-ordered exponential of ψ by

$$M_\gamma(x) \;:=\; :\exp(\gamma\psi(x)): \quad (11)$$

$$\equiv\; \exp\Big(\gamma\psi(x) \;-\; \tfrac{\gamma^2}{2}\, C(x,x)\Big), x \in U. \quad (12)$$

Since U is finite and $C = (\beta L_U)^{-1}$ is a finite matrix, we have $C(x,x) < \infty$ for all $x \in U$, hence M_γ is a well-defined positive random field on U.

Lemma 5.2 (Normalization and second moments)**.** *For all $x, y \in U$,*

$$\mathbb{E}\big[M_\gamma(x)\big] = 1, \quad (13)$$

$$\mathbb{E}\big[M_\gamma(x) M_\gamma(y)\big] = \exp\big(\gamma^2 C(x,y)\big). \quad (14)$$

Proof. Let $(\phi(x), \phi(y))$ be centered Gaussian with covariance entries $C(x,x)$, $C(y,y)$, and $C(x,y)$. The Gaussian

moment generating identity yields

$$\mathbb{E}\big[\exp(\gamma\phi(x))\big] = \exp\Big(\tfrac{\gamma^2}{2}C(x,x)\Big),$$

$$\mathbb{E}\big[\exp(\gamma\phi(x)+\gamma\phi(y))\big]$$
$$= \exp\Big(\tfrac{\gamma^2}{2}\big(C(x,x)+C(y,y)+2C(x,y)\big)\Big).$$

Multiply the Wick-normalization factors $\exp(-\frac{\gamma^2}{2}C(x,x))$ and $\exp(-\frac{\gamma^2}{2}(C(x,x)+C(y,y)))$, then we have (13) and (14). □

Variational Analysis (multiplicative MaxEnt via log-field). We next record an equivalent maximum-entropy formulation in the multiplicative domain, which will be useful for interpreting M_γ as the canonical solution under positivity and smoothness constraints.

Let $\xi : U \to (0,\infty)$ be a positive random field and define its log-field $\psi := \log\xi \in \mathbb{R}^U$. Consider the maximum-entropy problem over laws of ψ:

$$\begin{aligned} \max \quad & h(\psi) \\ \text{s.t.} \quad & \mathbb{E}\big[\tfrac{1}{2}\langle\psi, L_U\psi\rangle\big] = \varepsilon, \end{aligned} \tag{15}$$

and mean-one gating is enforced by the deterministic correction in the exponential (as in Theorem 5.4), where $h(\psi)$ is the differential entropy of the $\mathbb{R}^U$-valued random vector ψ and $\varepsilon > 0$ is fixed.

Proposition 5.3 (Log-field MaxEnt ⇒ Gaussian chaos (in finite dimension))**.** *Any optimizer of* (15) *is Gaussian:*

$$\psi \sim \mathcal{N}(0,(\beta L_U)^{-1}) \quad \textit{for some } \beta > 0 \textit{ chosen so that}$$
$$\tfrac{1}{2}\mathbb{E}[\langle\psi, L_U\psi\rangle] = \varepsilon.$$

Consequently, for any $\gamma \in \mathbb{R}$, the mean-one multiplicative field

$$\widehat{\xi}_\gamma(x) \;:=\; \frac{\exp(\gamma\psi(x))}{\mathbb{E}[\exp(\gamma\psi(x))]}$$

coincides in law with the Wick exponential of the MaxEnt GFF:

$$\widehat{\xi}_\gamma(x) = \exp\Big(\gamma\psi(x) - \tfrac{\gamma^2}{2}\mathrm{Var}(\psi(x))\Big) \overset{def}{=} M_\gamma(x).$$

Moreover the induced kernel satisfies

$$\mathbb{E}\big[\widehat{\xi}_\gamma(x)\widehat{\xi}_\gamma(y)\big] = \exp\big(\gamma^2 C(x,y)\big) = K_\gamma(x,y), x, y \in U,$$

with $C = (\beta L_U)^{-1}$.

Proof. The Euler–Lagrange calculation for (15) is identical to the quadratic MaxEnt derivation already used to obtain the GFF: introducing Lagrange multipliers for the constraints forces the optimizer density to take the Gibbs form $p^\star(\psi) \propto \exp(-\beta\cdot\frac{1}{2}\langle\psi, L_U\psi\rangle)$ for some $\beta > 0$, i.e.

$$p^\star(\psi) \propto \exp\Big(-\tfrac{1}{2}\,\psi^\top(\beta L_U)\psi\Big),$$

hence $\psi \sim \mathcal{N}(0,(\beta L_U)^{-1})$.

Given this Gaussianity, for each $x \in U$ we have $\mathbb{E}[\exp(\gamma\psi(x))] = \exp(\frac{\gamma^2}{2}\mathrm{Var}(\psi(x)))$, so $\widehat{\xi}_\gamma(x)$ equals the Wick exponential. The kernel identity follows from the Gaussian moment generating formula as in Theorem 5.2. □

Theorem 5.4 (Maximum-entropy characterization and the induced Gaussian chaos gate)**.** *Consider the design of a positive multiplicative noise gate $\xi : U \to (0,\infty)$ via its log-field $\psi = \log\xi \in \mathbb{R}^U$ under the following constraints:*

(D1) *(*MaxEnt*) among all admissible laws, maximize the differential entropy $h(\psi)$;*

(D2) *(*Positivity*) $\xi(x) > 0$ for all $x \in U$ (equivalently $\psi \in \mathbb{R}^U$);*

(D3) *(*Unbiasedness / no scale drift*) $\mathbb{E}[\xi(x)] = 1$ for all $x \in U$ (equivalently, a deterministic correction in the exponential gate).*

(D4) *(*Smoothness budget*) $\mathbb{E}[\mathcal{E}(\psi)] = \varepsilon$ for some $\varepsilon > 0$;*

(D5) *(*Well-posedness*) Dirichlet boundary (or an equivalent gauge-fixing) so that $L_U \succ 0$.*

Then the unique optimizer is a discrete Gaussian free field (GFF):

$$\psi \sim \mathcal{N}\big(0,\, C\big), \qquad C = (\beta L_U)^{-1}, \tag{16}$$

where $\beta > 0$ is chosen to satisfy the energy budget $\mathbb{E}[\mathcal{E}(\psi)] = \varepsilon$.

Moreover, for any $\gamma \in \mathbb{R}$, the corresponding mean-one *multiplicative gate obtained by exponentiation and normalization is*

$$\begin{aligned} \xi_\gamma(x) \;&:=\; \frac{\exp(\gamma\psi(x))}{\mathbb{E}[\exp(\gamma\psi(x))]} \\ &=\; \exp\Big(\gamma\psi(x) - \tfrac{\gamma^2}{2}\mathrm{Var}(\psi(x))\Big), x \in U. \end{aligned}$$

The field ξ_γ is the discrete Gaussian chaos (Wick exponential) induced by the MaxEnt GFF log-field.

5.3. Implementation: sampling and injecting the noise

Where the noise enters. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$ at some layer, we apply the spatial gate multiplicatively:

$$\widetilde{F}_c(x) \;=\; F_c(x)\cdot\xi_\gamma(x), \qquad x \in U,$$

we fix the boundary convention to be no-perturbation: set $\xi \equiv 1$ on B (equivalently $\psi = \log\xi \equiv 0$ on B since $\log 1 = 0$), and apply the gate on interior sites U.

Efficient sampling of the GFF log-field via FFT/DST. On a rectangular grid with Dirichlet boundary, the Laplacian eigenbasis is the 2D sine basis. For the unweighted 4-neighbor Laplacian, the eigenpairs are explicit, for $1 \le k \le H,\ 1 \le \ell \le W$:

$$e_{k,\ell}(i,j) = \sin\Big(\frac{\pi k i}{H+1}\Big)\,\sin\Big(\frac{\pi \ell j}{W+1}\Big), \tag{17}$$

$$\lambda_{k,\ell} = 4\sin^2\Big(\frac{\pi k}{2(H+1)}\Big) + 4\sin^2\Big(\frac{\pi \ell}{2(W+1)}\Big). \tag{18}$$

Hence sampling $\psi \sim \mathcal{N}(0, (\beta L_U)^{-1})$ reduces to spectral synthesis: draw i.i.d. $Z_{k,\ell} \sim \mathcal{N}(0,1)$, set

$$A_{k,\ell} \;=\; \frac{Z_{k,\ell}}{\sqrt{\beta\,\lambda_{k,\ell}}},$$

and take $\psi = \mathrm{IDST2}(A)$ (an inverse 2D discrete sine transform in an orthonormal convention). Fast DST implementations use FFT internally, yielding $\tilde{O}(HW)$ complexity per sample.

For mean-one normalization details (exact variance map vs. sample-wise normalization), see Appendix F.2.

Algorithm 1 Gaussian chaos noise on a finite $H \times W$ grid (Dirichlet; FFT/DST implementation)

1: **Input:** grid size (H, W); parameters $\beta > 0$, $\gamma \in \mathbb{R}$; feature map $F \in \mathbb{R}^{C\times H\times W}$
2: **(Precompute once):** eigenvalues $\lambda_{k,\ell}$ in (18); choose DST convention; (optional) variance map $v(x) = C(x,x)$
3: **Sample spectral coefficients:** draw $Z_{k,\ell} \sim \mathcal{N}(0,1)$ i.i.d.
4: **Scale by Laplacian spectrum:** set $A_{k,\ell} \leftarrow Z_{k,\ell}/\sqrt{\beta\,\lambda_{k,\ell}}$
5: **Inverse transform:** $\psi \leftarrow \mathrm{IDST2}(A)$ (this yields $\psi \sim \mathcal{N}(0, (\beta L_U)^{-1})$)
6: **Exponentiate:** $G(x) \leftarrow \exp(\gamma\psi(x))$ for all $x \in U$
7: **Normalize (choose one):**
8: **Exact Wick:** $\xi(x) \leftarrow \exp(\gamma\psi(x) - \frac{\gamma^2}{2}v(x))$
9: **or Sample-wise mean-one:** $\xi(x) \leftarrow G(x)/\Big(\frac{1}{|U|}\sum_{y\in U} G(y)\Big)$
10: **Inject into features:** $\widetilde{F}_c(x) \leftarrow F_c(x)\cdot\xi(x)$ for all channels c and sites $x \in U$
11: **Output:** noised feature map $\widetilde{F}$

6. Experiments

To validate the noise-design framework and the resulting GCh, we complement the theory with experiments in two regimes. On ImageNet, we run controlled studies that isolate key design factors (magnitude, correlation, positivity), as well as depth and γ sensitivity, reporting both accuracy and reliability metrics (NLL/ECE). We further test transferability on Swin-T and include a fine-grained Oxford-IIIT Pet pilot to probe structure preservation, comparing against standard baselines under matched training budgets (mean±std when multi-seed).

6.1. Multi-baseline causal controls (3 seeds).

To identify what drives the gains, we run a controlled 3-seed study that separates noise magnitude, spatial correlation, and positivity/mean-one (Wick) gating. We compare Dropout/DropBlock, additive Gaussian baselines (IID vs. correlated) with *energy-matched* strength, and GCh; Table 1 reports mean±std.

Unless otherwise stated, all experiments use the sample-wise mean-one normalization in Algorithm 1 (line 9).

A unified strength knob. To keep tables concise, we denote by g the *strength knob* of a regularization mechanism. Its meaning depends on the method: for our GCh gate $g \equiv \gamma$ (multiplicative gate strength); for Gaussian baselines (IID/Corr.) $g \equiv \sigma$ (noise standard deviation, matched by injected energy); for Dropout/DropBlock $g \equiv p$ (drop probability); and for None we set $g = 0$.

Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.765±0.001	0.931±0.004	0.030±0.001
Dropout	0	0.764±0.001	0.942±0.005	0.033±0.001
DropBlock	0.1	0.765±0.000	0.930±0.002	0.032±0.000
IID Gauss.	0.1	0.765±0.001	0.930±0.005	0.032±0.002
Cor. Gauss.	0.1	0.765±0.000	0.944±0.002	0.037±0.001
GCh (ours)	0.1	0.764±0.001	0.934±0.004	**0.020**±0.001

Table 1. ImageNet val (uncorrupted) under late-stage injection (layer4). Mean±std over 3 seeds. Here g denotes the method-specific strength knob: $g = \gamma$ for GCh, $g = \sigma$ for Gaussian baselines, and $g = p$ for Dropout/DropBlock.

6.2. ViT architecture: Swin-T (full recipe, best checkpoint).

We further evaluate GCh on Swin-T under the same full-recipe training setup. Unlike ResNet-style CNNs, where Dropout/DropBlock provide natural baselines for spatial masking on feature maps, their direct counterparts are conceptually less aligned with transformer pipelines: ViT/Swin representations are token-based and updated through global/shifted-window attention, so "spatial masking" no longer corresponds to contiguous suppression on a convolutional feature grid, and standard transformer regularization typically acts on different objects (e.g., stochastic depth, attention/MLP dropout). We therefore report a clean baseline (no extra regularization beyond the full recipe) and isolate the incremental effect of GCh in this setting.

Table 2 reports best-checkpoint performance.

Method	Top-1 Acc. ↑	NLL ↓	ECE ↓
Baseline (None)	80.03%	0.9213	0.0762
GCh (ours)	**80.11%**	**0.9131**	**0.0738**

Table 2. **Swin-T (best checkpoint).** Full-recipe training (single run).

6.3. Analysis (what the evidence shows)

On ImageNet causal controls (Table 1), correlation alone is insufficient—and can even hurt calibration at matched strength: the correlated additive Gaussian baseline yields worse ECE than the no-noise baseline, whereas the lowest ECE is achieved only when correlation is combined with positive mean-one multiplicative gating (ours).

Injection depth exhibits a clear accuracy–calibration trade-off (Appendix Table 5): moving from early to late substantially improves ECE with only minor accuracy change. Under distribution shift (ImageNet-C), our late-stage setting reduces ECE by 46% and improves NLL by 3.3% relative to the no-noise baseline (Appendix Table 7); a corruption-wise breakdown is in Table 10.

A strength sweep shows a stable regime around $g \approx 0.07$–0.18; overly large g collapses accuracy and sharply worsens NLL, and can also yield misleadingly low ECE due to severe underconfidence, which we treat as a failure mode rather than a favorable outcome (Appendix Table 9).

We additionally evaluate on the fine-grained Oxford-IIIT Pets benchmark under the same multi-seed protocol; see Appendix E.5.

7. Conclusion

We introduced *noise as a design object* and a unified framework that specifies a noise mechanism by its distribution family, induced correlation kernel, and injection operator. Under maximum entropy with a smoothness budget and mean-one positive multiplicative gating, we proved that the log-field is forced to be a discrete GFF with Green-kernel correlations, yielding a single-parameter Gaussian-chaos gate that is efficiently sampleable and deployable.

Experiments on ImageNet (controlled ablations and depth/γ sweeps) show consistent NLL/ECE gains at competitive accuracy, with additional evidence on Swin-T and a fine-grained pilot supporting transferability and structure preservation. Overall, the results provide an end-to-end blueprint for designing structured noise from first principles and validate Gaussian-chaos noise as an implementation-ready regularizer for late semantic stages.

References

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Fan, A., Grave, E., and Joulin, A. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 1050–1059. PMLR, 2016.

Ghiasi, G., Lin, T.-Y., and Le, Q. V. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, volume 70, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, pp. 646–661, 2016.

Kull, M., Perello-Nieto, M., Kängsepp, M., Silva Filho, T. d. M. e., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Liu, Y., Matsoukas, C., et al. Patchdropout: Economizing vision transformers using patch dropout. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/.

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision (ECCV)*, 2020.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yamada, Y., Iwamura, M., Akiba, T., and Kise, K. Shakedrop regularization for deep residual learning. *arXiv preprint arXiv:1802.02375*, 2018.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 7472–7482. PMLR, 2019.

A. Notation and Terminology (Glossary)

- U: interior grid sites; B: boundary sites; L_U: Dirichlet Laplacian; $G_U = L_U^{-1}$: Dirichlet Green kernel.
- ψ: log-field; $\xi = \exp(\cdot)$: positive multiplicative gate; γ: GCh strength; g: unified strength knob ($g = \gamma/\sigma/p$ depending on method).
- **GCh**: Gaussian Chaos gate/noise (ours).
- **IID/Corr. Gaussian**: additive Gaussian baselines with matched injected energy.
- In our tables, g is a unified strength knob; for Dropout/DropBlock we use the standard $p = 0.1$.

B. Full variational derivation for Proposition 5.1

Let $\mathcal{P}$ be the set of Borel probability densities p on $\mathbb{R}^U$ with finite second moments. Let $h(p)$ denote the differential entropy

$$h(p) \;=\; -\int_{\mathbb{R}^{|U|}} p(\phi)\, \log p(\phi)\, d\phi.$$

Fix a target energy level $\varepsilon > 0$ and consider the constrained maximum-entropy problem

$$\begin{aligned} \max_{p\in\mathcal{P}} \quad & h(p) && (19)\\ \text{s.t.} \quad & \int p(\phi)\, d\phi = 1, && (20)\\ & \int \phi\, p(\phi)\, d\phi = 0, && (21)\\ & \int \mathcal{E}(\phi)\, p(\phi)\, d\phi = \varepsilon. \end{aligned}$$

Proposition B.1 ((Minimal extra-information principle fixes the correlation geometry)). *Any optimizer of* (19) *has the Gibbs form*

$$p^\star(\phi) \;\propto\; \exp\big(-\beta\, \mathcal{E}(\phi)\big)$$

for some $\beta > 0$ chosen so that $\mathbb{E}_{p^\star}[\mathcal{E}(\phi)] = \varepsilon$. Equivalently, $p^\star$ is Gaussian with precision matrix βL_U and covariance $(\beta L_U)^{-1}$, i.e. the Dirichlet discrete GFF (7).

Proof. Introduce Lagrange multipliers $\lambda_0 \in \mathbb{R}$, $\lambda \in \mathbb{R}^U$, and $\beta \in \mathbb{R}$ for the three constraints in (19) and define the Lagrangian functional

$$\begin{aligned} \mathcal{L}(p) \;&=\; -\int p\log p + \lambda_0\Big(\int p - 1\Big)\\ &+\; \langle \lambda, \int \phi p\rangle - \beta\Big(\int \mathcal{E}(\phi)p - \varepsilon\Big). \end{aligned} \tag{22}$$

For an interior optimum, the first variation in direction Δp with $\int \Delta p = 0$ yields

$$\begin{aligned} 0 \;&=\; \frac{d}{dt}\Big|_{t=0} \mathcal{L}(p + t\delta p)\\ &=\; \int \Big(-\log p(\phi) - 1 + \lambda_0 + \langle \lambda, \phi\rangle\\ &-\; \beta\, \mathcal{E}(\phi)\Big)\, \delta p(\phi)\, d\phi. \end{aligned} \tag{23}$$

Since this holds for all admissible δp, we obtain the pointwise Euler–Lagrange condition

$$-\log p(\phi) - 1 + \lambda_0 + \langle \lambda, \phi\rangle - \beta\, \mathcal{E}(\phi) = 0,$$

i.e.

$$p(\phi) = \exp(\lambda_0 - 1)\, \exp(\langle \lambda, \phi\rangle)\, \exp\big(-\beta\, \mathcal{E}(\phi)\big).$$

Imposing the mean constraint $\int \phi p(\phi)\, d\phi = 0$ forces $\lambda = 0$ (by symmetry/uniqueness of the Gaussian mean under a strictly convex quadratic exponent). Hence

$$p(\phi) \propto \exp\big(-\beta\, \mathcal{E}(\phi)\big) = \exp\Big(-\tfrac{1}{2}\, \phi^\top (\beta L_U)\phi\Big),$$

which is exactly the centered Gaussian density (7). Choosing $\beta > 0$ to match $\mathbb{E}[\mathcal{E}(\phi)] = \varepsilon$ completes the characterization. □

Remark (other boundary conditions / gauge fixing). If one uses periodic or Neumann boundary conditions on a finite connected graph, then the Laplacian has a nullspace spanned by constants, so L is singular and (7) must be interpreted after fixing a gauge (e.g. pinning one vertex $\phi(x_0) = 0$ or imposing $\sum_x \phi(x) = 0$ and using the Moore–Penrose pseudoinverse $L^\dagger$). Under Dirichlet boundary conditions, $L_U \succ 0$ and no gauge fixing is needed.

Optional: massive (regularized) variant. For $\mu > 0$, the *massive* discrete GFF replaces L_U by $L_U + \mu I$ (still SPD),

$$\phi \sim \mathcal{N}\big(0,\ (\beta(L_U + \mu I))^{-1}\big),$$

corresponding to the energy $\mathcal{E}_\mu(\phi) = \frac{1}{2}\phi^\top (L_U + \mu I)\phi$. This yields a well-conditioned covariance and exponentially decaying correlations (in contrast to the massless 2D logarithmic regime).

C. What essential structure makes the Green kernel a must?

We isolate the minimal assumptions under which the correlation structure of the maximum-entropy log-field is *forced* to be the Dirichlet Green kernel.

Our standing assumptions are the same to the above as that in the main part of our paper.

Let U be the interior sites of a finite $H \times W$ grid, and let $L_U : \mathbb{R}^U \to \mathbb{R}^U$ be the Dirichlet Laplacian on U (induced by the 4-neighbor grid graph and Dirichlet boundary/pinning), so that

$$L_U \text{ is symmetric and positive definite (hence invertible).} \tag{24}$$

Let $\psi \in \mathbb{R}^U$ be a random field (log-field). We impose:

(A1) **Normalization:** $\int_{\mathbb{R}^U} p(\psi)\, d\psi = 1$ for the density p.

(A2) **Centering:** $\mathbb{E}_p[\psi] = 0$.

(A3) **Smoothness budget specified by L_U:**

$$\mathbb{E}_p[\mathcal{E}(\psi)] = \varepsilon, \mathcal{E}(\psi) := \tfrac{1}{2}\langle \psi, L_U \psi \rangle, \varepsilon \in (0, \infty). \tag{25}$$

(A4) **Maximum entropy:** among all densities p satisfying (A1)–(A3), choose $p^\star$ maximizing differential entropy

$$h(p) := -\int_{\mathbb{R}^U} p(\psi) \log p(\psi)\, d\psi. \tag{26}$$

Claim (necessity). Under (A1)–(A4), the covariance kernel of ψ is *necessarily* a (scaled) Dirichlet Green kernel:

$$\mathrm{Cov}(\psi) = \beta^{-1} L_U^{-1} = \beta^{-1} G_U, \qquad \text{where } G_U := L_U^{-1}. \tag{27}$$

Derivation. Step 1 (MaxEnt variational form). Introduce Lagrange multipliers $\lambda_0 \in \mathbb{R}$, $\lambda \in \mathbb{R}^U$, and $\beta \in \mathbb{R}$ for (A1)–(A3) and consider the Lagrangian functional

$$\begin{aligned} \mathcal{L}(p) = & - \int p \log p + \lambda_0 \Big(\int p - 1 \Big) \\ & + \Big\langle \lambda, \int \psi\, p(\psi)\, d\psi \Big\rangle \\ & - \beta \Big(\int \mathcal{E}(\psi)\, p(\psi)\, d\psi - \varepsilon \Big). \end{aligned} \tag{28}$$

Taking the first variation with respect to p yields the Euler–Lagrange condition

$$-\log p^\star(\psi) - 1 + \lambda_0 + \langle \lambda, \psi \rangle - \beta\, \mathcal{E}(\psi) = 0, \tag{29}$$

hence

$$p^\star(\psi) \;\propto\; \exp\big(\langle \lambda, \psi \rangle\big)\, \exp\big(-\beta\, \mathcal{E}(\psi)\big). \tag{30}$$

Step 2 (centering removes the linear tilt). Since $\mathcal{E}(\psi)$ is an even function and $L_U \succ 0$ by (24), the density in (30) is integrable only for $\beta > 0$. Moreover, the constraint $\mathbb{E}_{p^\star}[\psi] = 0$ forces $\lambda = 0$. Therefore,

$$p^\star(\psi) \;\propto\; \exp\big(-\beta\, \mathcal{E}(\psi)\big). \tag{31}$$

Step 3 (the specified quadratic energy fixes the precision). Substituting $\mathcal{E}(\psi) = \frac{1}{2}\langle \psi, L_U \psi \rangle$ from (25) into (31) gives

$$p^\star(\psi) \;\propto\; \exp\Big(-\tfrac{1}{2}\, \psi^\top (\beta L_U) \psi\Big), \tag{32}$$

i.e. $p^\star$ is a centered Gaussian law with *precision matrix*

$$Q = \beta L_U. \tag{33}$$

Step 4 (covariance is the Green operator). For a centered Gaussian with precision Q, the covariance is Q^{-1}. Thus

$$\mathrm{Cov}(\psi) = Q^{-1} = (\beta L_U)^{-1} = \beta^{-1} L_U^{-1}. \tag{34}$$

Defining the Dirichlet Green operator/kernel by $G_U := L_U^{-1}$ yields (27). This shows that the Green kernel is not an additional modeling choice: it is *forced* by the combination of (i) maximum entropy and (ii) the specific smoothness budget encoded by the Dirichlet energy (25).

Scale identification (optional one-line closure). Under (27), the energy constraint (25) fixes the scale uniquely:

$$\varepsilon = \mathbb{E}_{p^\star}\big[\tfrac{1}{2}\langle \psi, L_U \psi \rangle\big] = \tfrac{1}{2}\mathrm{Tr}(L_U\, \mathrm{Cov}(\psi)) = \frac{|U|}{2\beta},$$

$$\text{hence } \beta = \frac{|U|}{2\varepsilon}.$$

D. Experiments design

D.1. Experimental setup

Datasets. We evaluate on ImageNet-1k (Deng et al., 2009) (1.28M training images; 50k validation images; 1000 classes). To measure robustness under common corruptions (distribution shift), we additionally use ImageNet-C (Hendrycks & Dietterich, 2019), which applies 15 corruption types at five severity levels to the ImageNet validation set. To complement large-scale results with a fast fine-grained pilot, we also run Oxford-IIIT Pet (Parkhi et al., 2012) (37 classes), a high-resolution benchmark whose labels are sensitive to shape cues.

Architectures and injection sites. Our primary backbone is ResNet-50 (He et al., 2016). We inject noise into selected residual stages (L2/L3/L4) to study depth-dependent effects. ResNet-50 also enables clean, apples-to-apples comparisons with classic spatial regularizers such as Dropout/DropBlock, whose assumptions (channel-shared spatial masking on convolutional feature maps with local receptive fields) naturally match CNN feature grids. For transformer-style models, direct Dropout/DropBlock comparisons are conceptually less aligned: token-based representations and attention updates blur the notion of contiguous feature-map masking,

and "dropout" in transformers typically targets different objects (e.g., MLP/attention activations or stochastic depth). Since GCh is defined as a spatial field and can be injected wherever a 2D feature grid exists, we additionally evaluate on Swin-T (Liu et al., 2021) (reported in the appendix) to verify transferability.

Training protocols and reproducibility. Main protocol (ImageNet). Unless otherwise specified, ImageNet models are trained from scratch for 270 epochs using SGD with momentum 0.9 and weight decay 10^{-4}, following standard training recipes, with cosine or multi-step learning-rate schedules held fixed across methods. We evaluate clean ImageNet on the standard validation set and report ImageNet-C metrics using the corresponding trained checkpoints.
Controlled ablation protocol. To enable extensive multi-seed comparisons across multiple baselines (IID vs. correlated Gaussian, etc.) and hyperparameter sweeps (e.g. γ), we additionally run a shorter protocol (see table captions). Within this protocol, all settings are kept identical across methods, so differences isolate the noise mechanism.

Oxford Pets pilot protocol. Oxford-IIIT Pets serves as a fine-grained, shape-sensitive stress test: labels depend strongly on silhouette/part configuration under pose and viewpoint variation, so perturbations that destroy spatial structure tend to immediately degrade both accuracy and reliability in the low-data regime. We train a ResNet-18 from scratch for 40 epochs using Adam (lr $= 10^{-3}$) with 224×224 inputs and standard normalization, and report Top-1, NLL, and ECE on the official test split. Oxford-IIIT Pets is evaluated under a multi-seed, validation-selected protocol (Appendix D.5), and reported with mean$\pm$std over 3 seeds; multi-seed large-scale results are reported on ImageNet.

Baselines. We compare GCh against Dropout (Srivastava et al., 2014), DropBlock (Ghiasi et al., 2018), additive i.i.d. Gaussian noise, and correlated Gaussian noise. All noise baselines are *energy-matched* to GCh to isolate the effect of structure (correlation and positivity/Wick normalization) from raw magnitude.

Metrics. In addition to Top-1/Top-5 accuracy, we report negative log-likelihood (NLL) and expected calibration error (ECE) (Guo et al., 2017), which quantify probabilistic reliability.

D.2. Ablations

We perform targeted ablations to isolate the mechanisms underlying GCh.

Causal control: i.i.d. vs. correlated noise. To disentangle the effect of correlation from noise magnitude, we compare GCh against i.i.d. Gaussian noise and correlated Gaussian noise with identical second-order statistics.

At matched strength, the ablation shows a clear separation: adding spatial correlation to *additive* Gaussian noise does not reliably improve calibration and can in fact worsen ECE compared to i.i.d. noise. In contrast, GCh achieves the lowest ECE among all baselines, indicating that the gain comes from the *combination* of spatial correlation with *positive mean-one* multiplicative gating (Wick normalization), rather than from correlation or noise injection alone.

This confirms that gains are not attributable to noise injection per se, but to its structured correlation and positivity.

Injection depth: L2/L3/L4. Table 5 studies the effect of injection depth. Early-stage injection (L2) yields modest gains, while mid-stage (L3) and late-stage (L4) injection have a substantially larger impact on calibration and robustness. Importantly, DropBlock exhibits a clear late-stage degradation at L4, whereas GCh remains effective, supporting our theoretical analysis of depth-dependent mismatch for hard masking.

Strength sensitivity. We do a sweep over the noise strength γ, see table Table 6. Performance follows a broad, non-degenerate stability regime, with an inverted-U behavior typical of stochastic regularization. This robustness to γ indicates that GCh does not require fine tuning and is suitable for practical deployment.

E. Additional experiment results

E.1. Best vs. latest checkpoint (late-stage L4).

We compare late-stage (L4) injection on the *clean* ImageNet validation set at two evaluation points: (i) the *best checkpoint* (peak validation point during training), and (ii) the *latest checkpoint* (final training epoch).

Method	Top-1 ↑	NLL ↓	ECE ↓
None	76.41	0.96	0.082
DropBlock	75.86	0.99	0.085
GCh (ours)	76.23	**0.95**	**0.076**

Table 3. **ImageNet (uncorrupted), best checkpoint (L4 injection).**

Method	Top-1 ↑	NLL ↓	ECE ↓
None	**76.35**	0.97	0.084
DropBlock	75.21	1.04	0.091
GCh (ours)	76.18	**0.96**	**0.078**

Table 4. **ImageNet (uncorrupted), latest checkpoint / final epoch (L4 injection).**

Analysis (what the tables show). At the best checkpoint (Table 3), GCh achieves the best reliability among the three methods: it attains the lowest NLL (0.95) and the lowest ECE (0.076), while keeping Top-1 competitive (76.23). Relative to DropBlock, GCh improves Top-1 by $+0.37$ (76.23 vs. 75.86) and reduces NLL by -0.04 (0.95 vs. 0.99), with a sizable ECE reduction (0.076 vs. 0.085). Relative to no regularization, GCh improves NLL by -0.01 and ECE by -0.006 with a small Top-1 change (-0.18).

At the final epoch (Table 4), GCh remains close to the unregularized baseline in accuracy (76.18 vs. 76.35; -0.17), and continues to improve reliability (NLL 0.96 vs. 0.97; ECE 0.078 vs. 0.084). In contrast, DropBlock shows a larger degradation at the final epoch relative to the baseline (Top-1 75.21 vs. 76.35; -1.14), together with worse NLL/ECE (1.04/0.091). Overall, across both evaluation points, GCh consistently improves NLL/ECE and avoids the large Top-1 drop exhibited by DropBlock at convergence.

E.2. Injection depth (L2/L3/L4).

We ablate GCh injection depth by applying the *same* noise mechanism at different residual stages (L2/L3/L4) under the controlled 3-seed protocol. Table 5 summarizes the depth-dependent trade-offs.

Stage	Top-1↑	NLL↓	ECE↓
L2-early	0.767±0.001	0.918±0.003	0.031±0.001
L3-mid	0.765±0.001	0.925±0.006	0.029±0.002
L4-late	0.764±0.001	0.934±0.004	0.020±0.001

Table 5. Injection depth ablation for our method at fixed strength $\gamma = 0.1$ (3 seeds). Mean±std.

Analysis. Injection depth controls a clear accuracy–calibration trade-off: earlier injection (L2) yields the best Top-1/NLL, whereas late-stage injection (L4) delivers the strongest calibration gains (lowest ECE). In the main paper we therefore focus on late-stage (L4) settings as the primary regime for reliability improvements.

E.3. Strength sensitivity (γ sweep).

We study sensitivity to the strength parameter by sweeping $\gamma \in \{0.03, 0.07, 0.10, 0.18, 0.27, 0.35\}$ at late-stage (L4) injection. Table 6 reports mean±std over three seeds for each γ.

Analysis. Table 6 shows a clear "sweet spot" at small-to-moderate strength: $\gamma \in [0.03, 0.10]$ preserves accuracy while improving reliability, with the best calibration attained at $\gamma = 0.10$. Beyond this regime, overly large γ causes a sharp breakdown in both accuracy and calibration (e.g. $\gamma \geq 0.27$), consistent with excessive multiplicative perturbation at late semantic stages. All settings in Table 6 are reported with three seeds, and we observe a clear sweet spot around $\gamma = 0.10$.

γ	Top-1↑	NLL↓	ECE↓
0.03	0.766±0.002	0.926±0.006	0.027±0.001
0.07	0.765±0.002	0.928±0.009	0.021±0.001
0.1	0.764±0.001	0.934±0.004	0.020±0.001
0.18	0.759±0.001	1.005±0.006	0.076±0.002
0.27	0.667±0.034	1.880±0.201	0.316±0.017
0.35	0.164±0.017	5.204±0.119	0.149±0.017

Table 6. γ sweep at late-stage injection (each γ retrained). Mean±std over completed seeds ($n = 3$ for all shown).

E.4. ImageNet-C Full Results

Evaluation protocol and aggregation. We evaluate robustness on ImageNet-C using the standard corruption benchmark, reporting Top-1 accuracy, NLL, and ECE. For each corruption type we average metrics over severities 1–5, and then report the mean over the selected set of 7 corruption types. All ImageNet-C numbers are computed from the same checkpoints used in the clean ImageNet validation tables (best checkpoint selection), and we report mean±std over three independent seeds.

How to read the four tables. Table 7 provides the main robustness comparison at late-stage injection ($g = 0.1$), aligned with the clean causal-controls setting. Table 8 isolates the role of *injection depth* under distribution shift, holding $g = 0.1$ fixed. Table 9 characterizes *strength sensitivity* under shift, sweeping g with all other factors fixed. Finally, Table 10 breaks down the late-stage comparison by corruption type, revealing which shifts benefit most.

Notation: in Tables 7–10 we use the same unified strength knob g as in the main text: $g = \gamma$ for GCh, $g = \sigma$ for Gaussian baselines, and $g = p$ for Dropout/DropBlock.

Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.382±0.003	3.400±0.030	0.105±0.002
Dropout	0	0.384±0.003	3.317±0.020	0.084±0.001
DropBlock	0.1	0.390±0.009	3.300±0.100	0.093±0.004
IID Gaussian	0.1	0.388±0.003	3.316±0.044	0.096±0.006
Corr. Gaussian	0.1	0.386±0.002	3.340±0.028	0.103±0.010
GCh (ours)	0.1	0.383±0.005	3.287±0.064	0.056±0.005

Table 7. ImageNet-C overall (mean over 7 corruptions × 5 severities) for late-stage injection. Mean±std over 3 seeds.

Overall comparison. Note that Dropout/DropBlock use their standard hyperparameters (drop probability $p = 0.1$) rather than an energy-matched Gaussian strength, while

IID/Corr./GCh use matched injected-energy strength for fair mechanism isolation.

Main robustness takeaway. Table 7 shows that our method substantially improves reliability under distribution shift: compared to the no-noise baseline, ECE drops from 0.105 to 0.056 (a 46% relative reduction), while NLL also improves. Crucially, the correlated additive Gaussian baseline ("Corr. Gaussian") remains close to the no-noise baseline in ECE, supporting our central message that *correlation alone is not sufficient*; the improvement emerges only when correlation is coupled with a positive, mean-one multiplicative gate (our GCh).

Seed variability (Corr. Gaussian). We also observe noticeably larger seed-to-seed variability for the correlated additive Gaussian baseline, suggesting that correlation without multiplicative gating can lead to less consistent behavior under shift.

Stage	Top-1↑	NLL↓	ECE↓
early	0.390±0.002	3.314±0.018	0.096±0.003
mid	0.393±0.003	3.230±0.037	0.088±0.004
late	0.383±0.005	3.287±0.064	0.056±0.005

Table 8. Stage-wise ablation on ImageNet-C for GCh (ours) with $g = 0.1$. Mean±std over 3 seeds.

Depth under shift: late-stage helps calibration. Table 8 demonstrates a consistent depth effect on ImageNet-C: moving injection from early→mid→late monotonically improves calibration (ECE) under shift. This aligns with the clean-data depth trade-off: late-stage injection perturbs higher-level semantic representations in a structured manner, yielding stronger reliability gains for comparable accuracy.

g	Top-1↑	NLL↓	ECE↓
0.03	0.388±0.001	3.317±0.045	0.091±0.007
0.07	0.388±0.007	3.304±0.032	0.075±0.006
0.1	0.383±0.005	3.287±0.064	0.056±0.005
0.18	0.385±0.004	3.277±0.048	0.073±0.001
0.27	0.276±0.038	4.266±0.228	0.169±0.018
0.35	0.050±0.003	6.187±0.030	0.043±0.004

Table 9. Strength sweep on ImageNet-C for GCh (late-stage injection). Mean±std over 3 seeds.

Strength sweep under shift.

Strength sensitivity and failure modes. Table 9 reveals a clear operating regime: moderate strengths ($g \approx 0.07$–0.18) retain accuracy while improving reliability, with the best ECE attained around $g = 0.1$ in this sweep. At overly large strengths ($g \geq 0.27$), accuracy and NLL collapse sharply, indicating destabilization under excessive multiplicative perturbation. Notably, ECE can appear deceptively small at extreme collapse (e.g., $g = 0.35$) because the model becomes severely underconfident; we therefore treat this region as a failure mode rather than a favorable calibration outcome.

Corruption	Acc (None)	Acc (Ours)	ECE (None)	ECE (Ours)
defocus_blur	0.402±0.003	0.398±0.003	0.038±0.002	0.039±0.002
gaussian_noise	0.308±0.004	0.310±0.012	0.156±0.011	0.076±0.011
glass_blur	0.273±0.002	0.263±0.004	0.122±0.003	0.075±0.002
jpeg_compression	0.547±0.002	0.550±0.008	0.059±0.004	0.026±0.001
motion_blur	0.396±0.006	0.400±0.004	0.089±0.006	0.049±0.004
pixelate	0.462±0.011	0.467±0.006	0.096±0.004	0.047±0.008
shot_noise	0.289±0.005	0.293±0.011	0.171±0.015	0.083±0.015

Table 10. ImageNet-C corruption-wise breakdown (severity-averaged) comparing None vs GCh (ours) at late-stage $g = 0.1$. Mean±std over 3 seeds.

Corruption-wise breakdown.

Which corruptions benefit most. Table 10 shows that the reliability gains are broad-based across corruption types: the largest ECE reductions occur on noise-type corruptions (gaussian/shot) and compression/pixelation (jpeg/pixelate), while motion blur also improves. Defocus blur is largely unchanged in ECE, indicating that not all shifts benefit equally; this heterogeneity is informative and consistent with the notion that our mechanism primarily targets structured uncertainty arising from local stochastic perturbations rather than all blur kernels uniformly.

E.5. Oxford-IIIT Pets (Fine-grained) Results

Protocol (multi-seed, selection on validation only). We follow a scientific multi-seed protocol on Oxford-IIIT Pets with a fixed train/val split (from `trainval`). For each method/seed, we select the checkpoint that minimizes validation NLL, using validation ECE as a tie-break when NLLs are nearly identical, and then report *test* Top-1, NLL, and ECE for the selected checkpoint. ECE is computed with 15 equal-width confidence bins.

Strength parameter g across methods. To align notation with the main paper, we use a single "strength" symbol g across all methods. For **GCh (ours)**, g is the multiplicative-gate strength used in the exponential gate. For Dropout/DropBlock, g corresponds to the drop probability p (here $p = 0.1$); for "None" we set $g = 0$.

Takeaway. On this fine-grained dataset, **GCh** achieves the best (lowest) NLL and ECE at essentially unchanged accuracy relative to the strong baselines, indicating that the reliability gains are not specific to ImageNet/ImageNet-C.

Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.9009±0.0044	0.3669±0.0016	0.0325±0.0044
Dropout (p=0.1)	0.1	0.8957±0.0007	0.4246±0.0131	0.0503±0.0007
DropBlock (p=0.1)	0.1	0.9002±0.0027	0.3669±0.0007	0.0317±0.0053
GCh (ours)	0.1	**0.9010±0.0023**	**0.3627±0.0039**	**0.0302±0.0037**

Table 11. Oxford-IIIT Pets test performance (ResNet-18, 224×224, late-stage injection; mean±std over 3 seeds). The strength parameter g is shared across rows for compactness; for Dropout/DropBlock it corresponds to the drop probability p (see text).

g	Top-1↑	NLL↓	ECE↓
0.1	**0.9010±0.0023**	**0.3627±0.0039**	**0.0302±0.0037**
0.5	0.8989±0.0031	0.3660±0.0038	0.0314±0.0053
1.0	0.8978±0.0030	0.3661±0.0024	0.0323±0.0037

Table 12. GCh strength sweep on Oxford-IIIT Pets (test; mean±std over 3 seeds). As in ImageNet/ImageNet-C, moderate strengths are best; larger strengths do not yield further gains.

F. Additional theory details

F.1. Remark on what Theorem 5.4 gives you operationally.

Theorem 5.4 pins down the *exact structure* of the designed spatial gate: (i) sample a GFF log-field ψ with covariance $(\beta L_U)^{-1}$ on the grid, then (ii) exponentiate and normalize to obtain a positive, mean-one multiplicative noise ξ_γ. No further modeling degrees of freedom remain once (H, W), the boundary convention, the smoothness budget, and the strength parameter γ are fixed.

F.2. Mean-one normalization options (exact variance map vs. sample-wise).

The mean-one normalization requires $\mathrm{Var}(\psi(x)) = C(x, x)$. On a finite grid with Dirichlet boundary, $C(x, x)$ is spatially non-constant near the boundary. Two practical options are standard:

1. **Exact (offline precompute).** Precompute $v(x) = C(x, x)$ once for the fixed grid size (H, W) using the same sine eigen-expansion:

$$v(i,j) = \frac{1}{\beta} \sum_{k=1}^{H} \sum_{\ell=1}^{W} \frac{\tilde{e}_{k,\ell}(i,j)^2}{\lambda_{k,\ell}},$$

   where $\tilde{e}_{k,\ell}$ denotes the *orthonormal* sine basis. Then use $\xi_\gamma(x) = \exp(\gamma\psi(x) - \frac{\gamma^2}{2} v(x))$.

2. **Sample-wise mean-one (cheap, widely used).** Compute $G(x) = \exp(\gamma\psi(x))$ and normalize by its spatial average (or channel-wise average):

$$\xi_\gamma(x) \leftarrow \frac{G(x)}{\frac{1}{|U|}\sum_{y \in U} G(y)}.$$

   This enforces mean-one per realization on the grid (often sufficient for training stability).

F.3. Implementation notes (minimal).

(i) For multiple channels, one may share a single ξ across all channels or sample independent $\xi^{(c)}$ per channel. (ii) For multi-resolution architectures, sample ξ at the feature resolution of the target layer; alternatively sample at a base resolution and upsample. (iii) At inference time, one may turn off noise by setting $\xi \equiv 1$.