Variational Kernel Design for Internal Noise: Gaussian Chaos Noise, Representation Compatibility, and Reliable Deep Learning

Ziran Liu[1,2,3,†]

[1]Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS), Shanghai 200433, China
[2]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai 200433, China
[3]Institute for Intelligent Computing@SJTU, Shanghai 200433, China
[†]Correspondence to zliu@simis.cn

Abstract

Internal noise in deep networks is usually inherited from heuristics such as dropout, hard masking, or additive perturbation. We ask two questions: what correlation geometry should internal noise have, and is the implemented perturbation compatible with the representations it acts on? We answer these questions through Variational Kernel Design (VKD), a framework in which a noise mechanism is specified by a law family, a correlation kernel, and an injection operator, and is derived from learning desiderata. In a solved spatial subfamily, a quadratic maximum-entropy principle over latent log-fields yields a Gaussian optimizer with precision given by the Dirichlet Laplacian, so the induced geometry is the Dirichlet Green kernel. Wick normalization then gives a canonical positive mean-one gate, Gaussian Chaos Noise (GCh). For the sample-wise gate used in practice, we prove exact Gaussian control of pairwise log-ratio deformation, margin-sensitive ranking stability, and an exact expected intrinsic roughness budget; hard binary masks instead induce singular or coherence-amplified distortions on positive coherent representations. On ImageNet and ImageNet-C, GCh consistently improves calibration and under shift also improves NLL at competitive accuracy.

Keywords: Noise Design, Deep Learning Reliability, Calibration, Distribution Shift, Gaussian Multiplicative Chaos

1. Introduction

Noise injection is one of the most widely used yet least principled components of deep learning. It appears as additive perturbation, stochastic gating, masking, augmentation, corruption-aware training, and uncertainty regularization, and it is routinely used to improve generalization, calibration, and robustness. Yet one central design choice is usually left heuristic: *what structure should the noise have*? In most pipelines, that choice is inherited from familiar templates such as i.i.d. dropout (Srivastava et al., 2014), stochastic depth (Huang et al., 2016), or hard spatial masking (Ghiasi et al., 2018), rather than derived from the geometry of the representation or the objective of the learner.

This raises a more structural question:

> *If internal noise is to be used as part of representation learning, what aspects of that noise should be derived from first principles rather than fixed by convention?*

Our answer is to treat internal noise as a *design object*. We call the resulting program *Variational Kernel Design* (VKD). In VKD, a noise mechanism is specified by a triple

$$N = (\mathcal{F}, K, \mathcal{T}),$$

consisting of a law family, a correlation kernel, and an injection operator. A realization map then turns a sampled latent field into an implemented perturbation. The mechanism is therefore not just "a distribution"; it is a compositional system that separates *what is sampled*, *what geometry it must respect*, and *where and how it is deployed*.

This viewpoint reveals that there are really two linked questions. The first is a *design question*: once locality, smoothness, and mean-preserving positivity are encoded as operator-level constraints, what perturbation geometry is canonically induced? The second is a *compatibility question*: once such a mechanism is implemented in a deep network, what does it actually do to the geometry of positive semantic representations, and how does that differ from hard masking? The paper is built around this two-layer split. The first layer derives the mechanism. The second studies the induced action of the implemented perturbation on a target representation regime.

The design layer leads to a solved quadratic VKD program. In the spatial setting of this paper, a maximum-entropy log-field under a Dirichlet-energy budget is Gaussian with precision βL_U, hence covariance $\beta^{-1} L_U^{-1}$. In other words, within the chosen local quadratic design class, the Dirichlet Green kernel is not an additional modeling choice; it is the inverse operator forced by the constraints. Exponentiating that field with exact Wick normalization yields a positive mean-one multiplicative gate, which we call GCh.

The compatibility layer is where the practical distinction emerges. For the sample-wise gate actually used in our experiments, we prove exact Gaussian control of pairwise log-ratio deformations, explicit margin-sensitive ranking stability, and an exact expected intrinsic roughness budget. For hard binary masks, we prove a qualitatively different behavior: incompatibility with finite log-ratio geometry, a margin-blind ranking law for inverted dropout, and a coherence-sensitive distortion term whose relative size diverges as the underlying representation becomes increasingly smooth. This is the rigorous form of the informal claim that smooth positive multiplicative perturbations are better matched to coherent late-stage semantics than hard deletion.

A key theme throughout the paper is that these two layers belong together. The contribution is not just that a Gaussian field can be derived from a quadratic maximum-entropy problem—that isolated fact is classical. The contribution is that the variational solution is used as an operator-level design map for training-time noise, *and then* analyzed as an implemented mechanism acting on coherent positive evidence maps. In short: first derive the geometry; then ask whether the realized perturbation is compatible with the representation regime of interest.

Scope of the theoretical claims. The paper does *not* claim that deeper is always better, or that hard masking is universally inferior on every architecture, layer, or objective. The claims are conditional and operational: when a layer carries positive region-level or token-level evidence and becomes increasingly coherent in the late-semantic sense, the mathematically relevant quantities are relative log-ratios, ranking stability, and aggregate geometric roughness. In that regime, we show that the implemented GCh gate yields finite, margin-aware Gaussian deformations, whereas hard binary masks yield singular or coherence-amplified distortions.

This perspective yields both a principled mechanism and a practical prediction. If later layers encode increasingly decisive relative evidence between regions or tokens while also becoming more spatially coherent, then a margin-aware smooth multiplicative gate should remain compatible with those representations, whereas hard masking should become increasingly mismatched. Our experiments are designed to test exactly this distinction. On clean ImageNet, GCh improves calibration substantially, and on the selected 7-corruption ImageNet-C evaluation it improves both ECE and NLL while maintaining competitive accuracy. It also remains effective in late-stage injection settings where hard masking can degrade clean calibration.

Contributions. Our contributions are as follows.

- **A framework view of internal noise.** We formulate internal noise injection as a compositional design problem and introduce VKD, in which a mechanism is derived from learning-motivated constraints rather than selected from a fixed menu of perturbations.
- **A two-layer theory: design and compatibility.** We separate a mechanism-design layer from a representation-compatibility layer, making explicit the distinction between what is derived from first principles, what is realized in implementation, and what is subsequently measured on a target representation regime.
- **A solved quadratic MaxEnt design program.** We state the admissible class of centered log-field laws explicitly, solve the resulting finite-dimensional variational problem in closed form, and derive an entropy-gap identity certifying uniqueness of the optimizer.
- **Operator-forced kernel geometry.** For spatial log-fields with a Dirichlet-energy budget and gauge fixing, the optimizer is Gaussian with covariance proportional to the Dirichlet Green kernel. More generally, replacing the quadratic operator replaces the induced kernel by its inverse.
- **A canonical exact gate and an implementation-aware framework.** Exponentiating the MaxEnt log-field with Wick normalization yields GCh, a positive mean-one multiplicative gate with explicit multi-point moments; once the operator and budget are fixed, the exact gate becomes an effectively one-parameter family through $\tau = \gamma^2/\beta$. We also make explicit the split between the canonical exact gate and the sample-wise implementation used in practice.
- **Representation compatibility versus hard-mask mismatch.** For the sample-wise gate used in the experiments, we prove exact Gaussian control of pairwise log-ratios, margin-sensitive ranking stability, and an exact expected intrinsic roughness budget. For hard binary masking, we prove incompatibility with finite log-ratio geometry, a margin-blind ranking law for inverted dropout, an immediate loss-of-perfect-coherence result in expectation on perfectly coherent maps, and a late-stage mismatch theorem in the coherent-representation regime.
- **Empirical validation in the predicted late-stage regime.** On clean ImageNet, a selected 7-corruption ImageNet-C evaluation, Swin-T, and a fine-grained Oxford-IIIT Pets pilot, GCh improves calibration and, under shift, also improves NLL, all at competitive accuracy. Controlled ablations show the importance of correlation, positivity, and injection depth.

A practical way to read the paper. The variational results explain *where the kernel comes from.* The compatibility results explain *why the resulting implemented gate behaves differently from binary masking on semantic representations.* The experiments then test that distinction precisely in the late-stage regime where the mismatch should matter most.

Roadmap. Section 2 reviews stochastic regularization, calibration, and robustness under shift. Section 3 presents VKD as a compositional design system and situates the paper's solved instance inside that framework. Sections 4–5 develop the Dirichlet log-field construction, the quadratic MaxEnt theorem, and the exact and implemented GCh gates. Section 5.5 gives the representation-compatibility analysis, and Section 6 tests the resulting predictions empirically.

Paper in one sentence. We derive the noise geometry from first principles and then show that the resulting implemented smooth positive gate preserves finite, margin-aware relative geometry exactly in the regime where hard masking becomes singular or coherence-amplified.

What is classical and what is new. The isolated fact that quadratic maximum entropy yields a Gaussian law is classical. The contribution here is the *use* of that principle as an operator-level design map for training-time noise, together with the second layer of theory that is specific to this paper: exact representation-compatibility results for the implemented sample-wise gate and exact incompatibility results for hard binary masks on coherent positive semantic representations. Put differently, the variational theorem identifies the canonical kernel inside a chosen design class, and the later compatibility theorems explain why that designed mechanism behaves differently from masking in deep networks.

2. 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., 2020); ShakeDrop perturbs residual branches with randomized coefficients (Yamada et al., 2018). Spatial occlusion methods such as Cutout and DropBlock impose structured hard masking on feature maps (DeVries and Taylor, 2017; Ghiasi et al., 2018), while sample-level mixing methods such as Mixup and CutMix inject stochasticity at the data level (Zhang et al., 2018; Yun et al., 2019). In vision transformers, PatchDropout removes input patches and changes token topology (Liu et al., 2023). A common limitation is that the correlation structure of the noise is usually fixed a priori and often assumes spatial independence or hard discontinuities, which can mismatch late semantic representations.

Calibration and reliability under shift. Miscalibration is widespread in modern neural networks, and temperature scaling remains a strong post-hoc baseline (Guo et al., 2017). Nonparametric alternatives include BBQ (Naeini et al., 2015), while Dirichlet calibration extends beyond a single temperature parameter (Kull et al., 2019). Dropout admits an approximate Bayesian interpretation (Gal and Ghahramani, 2016), and deep ensembles remain a strong uncertainty baseline (Lakshminarayanan et al., 2017). Under distribution shift, calibration can deteriorate substantially (Ovadia et al., 2019), and recent work emphasizes that calibration depends strongly on architecture and training recipe (Minderer et al., 2021). Label smoothing can help but is context dependent (M"uller et al., 2019). These findings motivate methods that improve NLL and ECE directly during representation learning rather than relying only on post-hoc correction.

Robustness to corruptions and distribution shift. For worst-case robustness, adversarial training and TRADES formalize the robustness–accuracy trade-off (Madry et al., 2018; Zhang et al., 2019). For average-case corruptions, ImageNet-C/P provide standardized benchmarks (Hendrycks and Dietterich, 2019); subsequent work has also emphasized that performance on synthetic corruptions does not perfectly transfer to natural shifts (Taori et al., 2020), and broader OOD suites reveal substantial heterogeneity across shift types (Hendrycks et al., 2021). Simple augmentation 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 or speckle noise can also be effective (Rusak et al., 2020). Noisy Student further demonstrates the power of strong stochastic regularization in large-scale training (Xie et al., 2020). Our focus is complementary: rather than designing perturbations at the input level, we derive an *internal* spatial noise mechanism whose correlation structure follows from explicit desiderata.

3. Variational Kernel Design as a Compositional Design System

We treat internal noise not as a fixed perturbation template but as a mechanism to be derived from learning desiderata. The role of Variational Kernel Design (VKD) is to map a collection of task-level constraints to a stochastic mechanism and then to analyze how that mechanism acts on a

target representation regime. This viewpoint separates two layers that are often conflated in practice: a *mechanism-design layer*, which specifies what latent object is sampled, what geometry it must respect, and where it is injected, and a *compatibility layer*, which studies what geometric quantities the deployed perturbation preserves or distorts on the representations actually used by the network.

The benefit of this separation is conceptual as well as practical. It makes clear which parts of the construction are derived from first principles, which parts are implementation choices, and which parts are properties of the resulting perturbation on a given representation regime. In particular, VKD is not a menu of named noises; it is a compositional system for deriving, realizing, deploying, and analyzing an internal perturbation mechanism.

3.1. Mechanism space: VKD as a design system

Let Ω denote a perturbation domain and let $\mathcal{H}$ denote a feature space. A VKD mechanism is specified by a triple

$$N = (\mathcal{F}, K, \mathcal{T}),$$

whose three components encode complementary axes of design.

Definition 1 (VKD mechanism)**.** *A* VKD mechanism *on* $(\Omega, \mathcal{H})$ *is a triple*

$$N = (\mathcal{F}, K, \mathcal{T}),$$

where:

(i) $\mathcal{F}$ *is a family of laws on latent fields* $\psi \in \mathbb{R}^{\Omega}$*;*

(ii) K *is a positive semidefinite kernel on* $\Omega \times \Omega$ *encoding the intended second-order geometry;*

(iii) $\mathcal{T}$ *is an injection operator that deploys a realized perturbation inside the model.*

The three components play distinct roles. The family $\mathcal{F}$ determines what latent object is sampled; the kernel K encodes how that object is spatially correlated; and the operator $\mathcal{T}$ determines where and how the realized perturbation acts on the network. In this way, VKD separates *sampling*, *geometry*, and *deployment*.

To make the construction operational, we introduce a realization map

$$\ell : \mathbb{R}^{\Omega} \to (0, \infty)^{\Omega},$$

which turns a latent field ψ into a positive gate $\xi = \ell(\psi)$. The deployed perturbation is then

$$\widetilde{h} = \mathcal{T}(h; \xi), \qquad \psi \sim \mathcal{F}.$$

Thus the mechanism pipeline has the schematic form

$$(\mathcal{F}, K, \mathcal{T}) \quad \Longrightarrow \quad \psi \sim \mathcal{F} \quad \xrightarrow{\ell} \quad \xi \quad \xrightarrow{\mathcal{T}} \quad \widetilde{h}.$$

3.2. From desiderata to admissible mechanism classes

A central point of VKD is that the mechanism is not selected from a fixed heuristic menu. Instead, one starts from a collection of learning desiderata D—for example positivity, lack of systematic scale drift, locality, smoothness, or minimal extra information—and translates them into mathematical constraints on admissible mechanisms.

Accordingly, VKD should be read as a map

$$D \quad \longmapsto \quad \mathfrak{N}(D),$$

where $\mathfrak{N}(D)$ is an admissible class of mechanisms consistent with the desiderata. The design problem is then to derive a distinguished mechanism

$$N^\star \in \mathfrak{N}(D)$$

rather than choose one by convention.

This formulation is intentionally general. In some settings, the admissible class may leave several components independent. In other settings, the desiderata may couple the law and the geometry so strongly that the kernel is no longer a free modeling knob but a derived consequence of the design class itself.

3.3. A two-layer view: mechanism and compatibility

A VKD mechanism is only half of the story. Once a mechanism has been derived and realized, one must still ask how the deployed perturbation acts on the representations the network actually uses. We therefore separate a second object: a target representation regime $\mathcal{R}$ together with a collection of compatibility observables

$$\mathcal{O}(\widetilde{h}; \mathcal{R}),$$

such as pairwise log-ratio deformation, ranking stability, intrinsic roughness inflation, or topological stability.

The resulting conceptual split is:

- **Mechanism-design layer:** derive $(\mathcal{F}, K, \mathcal{T})$ and the realization map ℓ from desiderata;
- **Compatibility layer:** study the induced action of the deployed mechanism on observables relevant to a target representation regime $\mathcal{R}$.

This distinction is especially important in the present paper because the canonical object derived by the variational theory is an exact Wick-normalized gate, while the optimization-friendly implementation used in the main experiments is a sample-wise mean-one gate. The design layer tells us *what the canonical latent geometry is*; the compatibility layer tells us *what the implemented mechanism does once deployed*.

Remark 2 (VKD is compositional, not temporal)**.** *VKD should not be interpreted as a temporal dynamical system unless an explicit update rule is introduced. Its role here is compositional: desiderata define an admissible class, the variational principle derives a canonical latent law and geometry, the realization map produces a deployed gate, and the compatibility layer studies the induced action of that gate on a target representation regime.*

3.4. The solved instance studied in this paper

The present paper studies a solved quadratic VKD subfamily in which the latent object is a centered log-field and spatial coherence is imposed through a quadratic operator budget. In this subfamily, the law and the geometry are not independent design axes: once the operator Q and the energy budget ε are fixed, the unique maximum-entropy optimizer is Gaussian with covariance proportional to Q^{-1}. Thus the kernel is *operator-forced* rather than chosen heuristically.

In the main spatial construction, the perturbation domain is the feature grid U, the operator is the Dirichlet Laplacian $Q = L_U$, and the resulting latent law is a discrete Gaussian free field with covariance proportional to the Dirichlet Green kernel $G_U = L_U^{-1}$. A realization map then turns the latent log-field into either a canonical exact Wick-normalized gate or the sample-wise mean-one gate used in the experiments. The injected perturbation is studied precisely in the late-stage positive coherent regime where pairwise log-ratios, ranking preservation, and intrinsic roughness are the relevant observables.

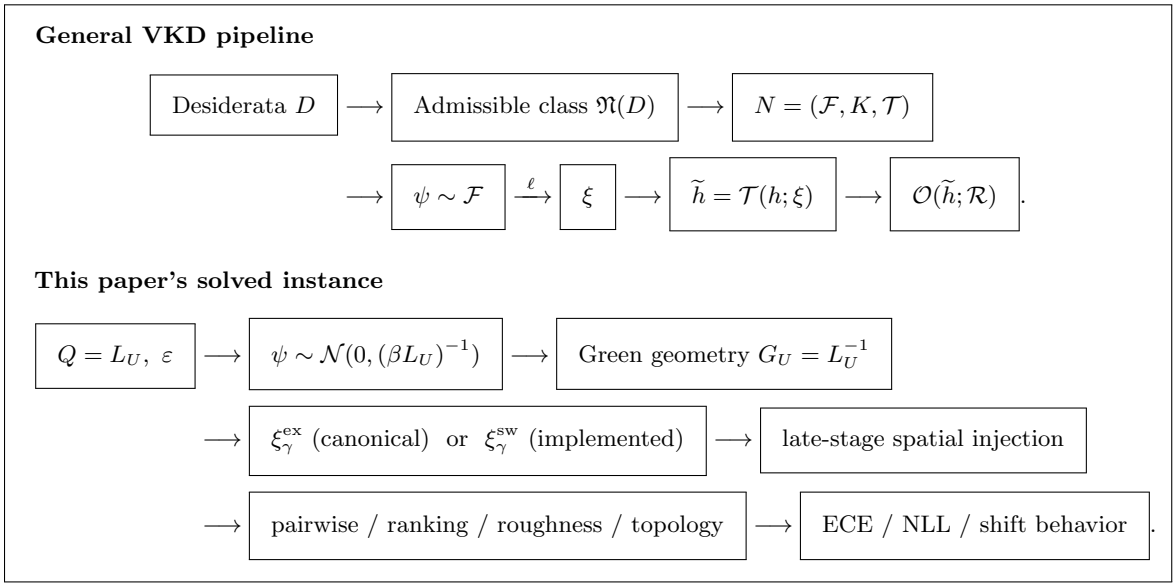

Figure 1. VKD as a compositional design system. The top row shows the general design logic: learning desiderata define an admissible mechanism class, from which a mechanism and realization map are derived and then evaluated through compatibility observables on a target representation regime. The bottom row shows the solved instance of this paper. A quadratic operator budget with $Q = L_U$ yields a Gaussian log-field with covariance proportional to the Dirichlet Green kernel. This latent field admits a canonical exact realization through Wick normalization and an optimization-friendly implemented realization through sample-wise mean-one normalization. The resulting deployed gate is analyzed through pairwise, ranking, roughness, and topological observables, and then tested empirically through calibration and reliability metrics.

Roadmap from framework to results. Section 4 fixes the perturbation domain and deployment axis for the spatial setting studied in this paper. Section 5 then solves the mechanism-design layer: desiderata define an admissible class, the quadratic MaxEnt principle derives the canonical latent law, and the realization map yields the exact GCH gate together with the implementation-aware variants used in practice. The later subsections of Section 5 begin the compatibility layer by analyzing the induced action of the implemented gate on the positive coherent regime relevant to late-stage representations.

4. A Solved Quadratic VKD Instance: Problem Setup and Discrete GFF Background

We now instantiate the general system of Section 3 in the spatial setting used throughout the paper. To keep the framework explicit, it is useful to separate what is fixed in this section from what is derived in Section 5. Here we fix the perturbation domain $\Omega = U$, the feature space $\mathcal{H} = \mathbb{R}^{C\times H\times W}$, and the deployment axis $\mathcal{T}$ as spatial multiplicative injection on a feature grid. In Section 5 we then derive the canonical latent law and its induced geometry from the learning desiderata. In the solved instance studied here, the latent object is a centered log-field, the operator budget is the Dirichlet energy, and the resulting canonical geometry is the Dirichlet Green kernel.

4.1. Injection site and spatial gating

Fix a 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, with channel index $c \in \{1, \dots, C\}$ and spatial location $x = (i, j) \in U$, where

$$U = \{1, \dots, H\} \times \{1, \dots, W\}.$$

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

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

and apply it identically across channels.

Injection operators. The basic multiplicative operator is

$$\mathcal{T}_\nu(h)(c, x) = h(c, x)\, \nu(x), \tag{1}$$

that is, pointwise multiplication with spatial broadcasting. For numerical stability or reduced perturbation strength, we may also use the residual form

$$\mathcal{T}_\nu^{\mathrm{res}}(h)(c, x) = h(c, x)\Big(1 + \alpha(\nu(x) - 1)\Big), \qquad \alpha \in (0, 1]. \tag{2}$$

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

Framework instantiation of the deployment axis. In the notation of Section 3, this subsection fixes the deployment part of the mechanism: the perturbation domain is the interior grid U, the feature space is $\mathcal{H} = \mathbb{R}^{C \times H \times W}$, and the admissible deployment operators are spatial multiplicative injections such as $\mathcal{T}_\nu$ and $\mathcal{T}_\nu^{\mathrm{res}}$. What remains open at this stage is the design layer: which latent law should be sampled, and what correlation geometry should it induce?

4.2. Discrete Gaussian free field on a rectangular grid

To make the implementation and spectral formulas consistent, we treat the feature grid itself as the interior domain and impose Dirichlet conditions on an *auxiliary outer boundary*. Fix integers $H, W \geq 1$ and define

$$U = \{1, \dots, H\} \times \{1, \dots, W\}, \qquad \bar{U} = \{0, \dots, H + 1\} \times \{0, \dots, W + 1\},$$

with auxiliary boundary

$$B = \bar{U} \setminus U.$$

Equip $\bar{U}$ with the nearest-neighbor undirected edge set

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

Optionally, allow positive symmetric edge weights $c_{xy} = c_{yx} > 0$ on $\{x, y\} \in E$; the unweighted case is $c_{xy} \equiv 1$.

A field is a function $\phi : U \to \mathbb{R}$. We extend it by zero to the auxiliary boundary:

$$\bar{\phi}(y) = \begin{cases} \phi(y), & y \in U, \\ 0, & y \in B. \end{cases}$$

Dirichlet Laplacian and energy. For $\phi : U \to \mathbb{R}$, define the Dirichlet Laplacian L_U by

$$(L_U \phi)(x) = \sum_{y : \{x, y\} \in E} c_{xy}\big(\phi(x) - \bar{\phi}(y)\big), \qquad x \in U. \tag{3}$$

Its quadratic form is the Dirichlet energy

$$\mathcal{E}(\phi) \coloneqq \frac{1}{2}\langle \phi, L_U \phi \rangle = \frac{1}{2} \sum_{\{x, y\} \in E} c_{xy}\big(\bar{\phi}(x) - \bar{\phi}(y)\big)^2. \tag{4}$$

Under Dirichlet boundary conditions, L_U is symmetric positive definite, so $\mathcal{E}(\phi) > 0$ for $\phi \neq 0$.

Discrete GFF. Fix an inverse-temperature parameter $\beta > 0$. The Dirichlet discrete Gaussian free field (GFF) on U is the centered Gaussian vector

$$\phi \sim \mathcal{N}\big(0, (\beta L_U)^{-1}\big). \tag{5}$$

Equivalently, its density on $\mathbb{R}^U$ is

$$p_\beta(\phi) = \frac{1}{Z_\beta} \exp\big(-\beta\, \mathcal{E}(\phi)\big) = \Big(\frac{\det(\beta L_U)}{(2\pi)^{|U|}}\Big)^{1/2} \exp\Big(-\tfrac{1}{2}\phi^\top(\beta L_U)\phi\Big), \tag{6}$$

with normalizing constant

$$Z_\beta = (2\pi)^{|U|/2} \det(\beta L_U)^{-1/2}. \tag{7}$$

Green kernel. Define the Dirichlet Green matrix

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

Then the covariance of the GFF is

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

Framework role of this subsection. At this point the state space of latent fields and the local operator have been fixed. The next section will solve the design layer inside this spatial VKD class: the variational principle will determine the canonical law family $\mathcal{F}^\star$, and the induced second-order geometry will appear as a consequence of the chosen operator rather than as an additional hyperparameter.

5. Solving the Design Layer: From Desiderata to Gaussian Chaos Noise

We now solve the mechanism-design layer of VKD for the spatial instantiation fixed in Section 4. The logical order is: specify learning desiderata, define the admissible class of latent laws, derive the canonical law and induced geometry, and only then choose a realization map that turns the latent object into a deployed gate. Read in this way, the section is not only about one new noise family; it is the full derivation of a solved VKD instance. The key mathematical point is that the optimization is performed over *laws of the latent log-field*; positivity and mean preservation are imposed afterwards at the realization stage through an exponential link and Wick normalization.

5.1. Design desiderata

Each desideratum constrains a different part of the framework: D1 selects the law inside an admissible class, D2–D3 constrain the realization map, D4 determines the operator geometry, and D5 ensures that the operator-level design problem is well posed.

D1 Least additional information (maximum entropy). Among all admissible laws satisfying the required constraints, choose the one with maximum differential entropy. Intuitively, the perturbation should avoid injecting unintended semantics.

D2 Positivity through an exponential link. The gate should modulate amplitude without introducing sign flips or hard artifact patterns. We therefore write

$$\xi = \exp(\zeta)$$

for a real-valued log-field $\zeta \in \mathbb{R}^U$.

D3 No systematic scale drift. The gate should not create a persistent gain shift. In the exact construction this is enforced by Wick normalization, giving $\mathbb{E}[\xi(x)] = 1$ for every site $x \in U$.

D4 Spatial coherence via a quadratic smoothness budget. The perturbation should be spatially coherent rather than pixelwise i.i.d. We encode this through a local quadratic budget on the log-field:

$$\mathbb{E}\left[\frac{1}{2}\langle \psi, Q\psi\rangle\right] = \varepsilon, \tag{9}$$

where $Q \succ 0$ is a symmetric positive definite operator on $\mathbb{R}^U$. In the canonical grid construction of this paper, $Q = L_U$ is the Dirichlet Laplacian.

D5 Well-posedness through gauge fixing. A gauge convention is required so that the quadratic operator is invertible. In the main text we impose auxiliary Dirichlet boundary conditions, which make $L_U \succ 0$.

Why separate ζ and ψ? For the variational problem, the object being optimized is the law of a centered log-field ψ. Positivity and mean preservation are then enforced *afterwards* by mapping ψ through a Wick-normalized exponential. This separation is useful because it makes clear which parts of the theory characterize the optimizer of the entropy problem and which parts define the final multiplicative gate.

5.2. A formal variational class

Fix an SPD operator Q on $\mathbb{R}^U$, an energy budget $\varepsilon > 0$, and let $n := |U|$. Define the admissible class

$$\mathcal{A}(Q,\varepsilon) := \left\{ p : \mathbb{R}^U \to [0,\infty) \;\middle|\; \begin{array}{l} \int_{\mathbb{R}^U} p(\psi)\, d\psi = 1, \\ \int_{\mathbb{R}^U} \psi\, p(\psi)\, d\psi = 0, \\ \int_{\mathbb{R}^U} \frac{1}{2}\langle \psi, Q\psi\rangle p(\psi)\, d\psi = \varepsilon, \\ h(p) > -\infty \end{array} \right\}, \tag{10}$$

where

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

is the differential entropy. The associated variational problem is

$$\sup_{p \in \mathcal{A}(Q,\varepsilon)} h(p). \tag{11}$$

This formulation clarifies the scope of the theory. The design class is determined by three ingredients only: (i) the state space $\mathbb{R}^U$ of log-fields, (ii) the centering and quadratic-budget constraints, and (iii) the choice of local operator Q. The role of the operator is especially important: once Q is fixed, the entropy maximizer—if it exists—must reveal the correlation geometry compatible with that operator.

In VKD language, this subsection isolates the law-design part of the mechanism. The deployment axis has already been fixed in Section 4; the remaining task is to derive the canonical latent law and the geometry it induces.

5.3. Quadratic MaxEnt principle and operator-forced kernel geometry

The next theorem is the main design theorem for the solved quadratic VKD subfamily. It turns desiderata D1, D4, and D5 into a unique latent law, and it makes explicit how the operator budget fixes the scale of the optimizer. Relative to the earlier proof sketch, it yields the optimizer, its entropy value, the explicit scale, and an entropy-gap identity that certifies uniqueness.

Theorem 5.1 (Design theorem for the quadratic VKD subfamily)**.** *Let $Q \succ 0$ be symmetric positive definite on $\mathbb{R}^U$, let $n = |U|$, and let $\varepsilon > 0$. Then the variational problem (11) has a unique optimizer*

$$p^\star_{Q,\varepsilon} = \mathcal{N}\big(0, \Sigma_{Q,\varepsilon}\big), \qquad \Sigma_{Q,\varepsilon} = \frac{2\varepsilon}{n}\, Q^{-1}. \tag{12}$$

Equivalently,

$$p^\star_{Q,\varepsilon}(\psi) = \frac{1}{(2\pi)^{n/2}\det(\Sigma_{Q,\varepsilon})^{1/2}} \exp\Big(-\frac{1}{2}\psi^\top \Sigma^{-1}_{Q,\varepsilon}\psi\Big), \tag{13}$$

with precision matrix

$$\Sigma^{-1}_{Q,\varepsilon} = \frac{n}{2\varepsilon}\, Q.$$

Moreover, for every $p \in \mathcal{A}(Q,\varepsilon)$,

$$h(p^\star_{Q,\varepsilon}) - h(p) = \mathrm{KL}\big(p \,\|\, p^\star_{Q,\varepsilon}\big) \geq 0, \tag{14}$$

so the optimizer is unique. Its entropy is

$$h(p^\star_{Q,\varepsilon}) = \frac{1}{2}\log\Big((2\pi e)^n \det\big(\tfrac{2\varepsilon}{n}Q^{-1}\big)\Big). \tag{15}$$

Proof sketch. Let $p^\star$ denote the Gaussian density in (12). Since

$$\Sigma^{-1}_{Q,\varepsilon} = \frac{n}{2\varepsilon}Q,$$

the quadratic constraint implies

$$\mathbb{E}_{p^\star}\left[\frac{1}{2}\langle\psi, Q\psi\rangle\right] = \frac{1}{2}\mathrm{Tr}(Q\Sigma_{Q,\varepsilon}) = \frac{1}{2}\mathrm{Tr}\Big(Q\frac{2\varepsilon}{n}Q^{-1}\Big) = \varepsilon,$$

so $p^\star \in \mathcal{A}(Q,\varepsilon)$. For any feasible p,

$$\mathrm{KL}(p\|p^\star) = -h(p) - \int p(\psi)\log p^\star(\psi)\, d\psi.$$

Because $\log p^\star(\psi) = c - \frac{n}{4\varepsilon}\langle\psi, Q\psi\rangle$ for a constant c, and every feasible p has the same normalization, mean, and energy budget, the second term depends only on (Q,ε) and coincides with $-h(p^\star)$. Hence (14) holds. Uniqueness follows because $\mathrm{KL}(p\|p^\star) = 0$ iff $p = p^\star$ a.e. The entropy formula is the standard entropy of a centered Gaussian with covariance $\Sigma_{Q,\varepsilon}$. □

Corollary 2 (Operator-forced geometry in the Dirichlet instantiation)**.** *Taking $Q = L_U$ in Theorem 5.1 yields the unique entropy-maximizing log-field*

$$\psi \sim \mathcal{N}\big(0, (\beta L_U)^{-1}\big), \qquad \beta = \frac{n}{2\varepsilon}. \tag{16}$$

Its covariance is

$$\mathrm{Cov}(\psi) = \frac{2\varepsilon}{n}\, L_U^{-1} = \beta^{-1} G_U, \qquad G_U \coloneqq L_U^{-1}. \tag{17}$$

Thus, within the local quadratic design class determined by the Dirichlet energy, the correlation geometry is the Dirichlet Green kernel.

Remark 3 (What is and is not "forced")**.** *The theorem does* not *say that the Green kernel is universally optimal for every noise-design problem. It says something more precise: once the design class is fixed by a local quadratic budget with operator Q, the entropy maximizer has covariance proportional to Q^{-1}. The Green kernel is forced specifically because the operator chosen here is the Dirichlet Laplacian.*

Framework interpretation. Theorem 5.1 solves the latent-law axis of the VKD mechanism-design layer, and Corollary 2 shows that the second-order geometry is induced rather than tuned. What remains is the realization map ℓ: how to turn the derived latent log-field into a positive, mean-preserving gate that can actually be deployed.

5.4. From the MaxEnt log-field to the canonical realization map

At this point the latent law and induced geometry have been derived. The remaining step in the design layer is the realization map ℓ: how to turn the latent log-field into a positive gate satisfying D2–D3. The canonical answer in the solved VKD subfamily is the Wick-normalized exponential. Let

$$\psi \sim \mathcal{N}(0, C), \qquad C = (\beta L_U)^{-1} = \frac{2\varepsilon}{n} G_U. \tag{18}$$

For a strength parameter $\gamma \in \mathbb{R}$, define the exact Wick-normalized exponential

$$\xi_\gamma^{\mathrm{ex}}(x) \coloneqq :\exp(\gamma\psi(x)) := \exp\Big(\gamma\psi(x) - \tfrac{\gamma^2}{2} C(x,x)\Big), \qquad x \in U. \tag{19}$$

This is the canonical exact realization associated with the variationally derived log-field. Positivity comes from the exponential map; mean preservation comes from the Wick correction.

Theorem 5.4 (Canonical realization in the solved VKD subfamily)**.** *Under desiderata* D1–D5*, the canonical exact positive mean-one multiplicative gate is obtained by:*

1. *sampling the MaxEnt log-field*

$$\psi \sim \mathcal{N}\big(0, (\beta L_U)^{-1}\big), \qquad \beta = \frac{|U|}{2\varepsilon},$$

   and

2. *applying the Wick-normalized exponential* (19)*.*

For any sites $x_1, \dots, x_m \in U$,

$$\mathbb{E}\left[\prod_{r=1}^{m} \xi_\gamma^{\mathrm{ex}}(x_r)\right] = \exp\left(\gamma^2 \sum_{1 \le a < b \le m} C(x_a, x_b)\right). \tag{20}$$

In particular,

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\big] = 1, \tag{21}$$

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = \exp\big(\gamma^2 C(x,y)\big). \tag{22}$$

Hence the induced second-order gate kernel is

$$K_\gamma(x,y) \coloneqq \mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = \exp\big(\gamma^2 C(x,y)\big). \tag{23}$$

Proof. Because $(\psi(x_1), \dots, \psi(x_m))$ is jointly Gaussian,

$$\mathbb{E}\left[\exp\Big(\gamma \sum_{r=1}^{m} \psi(x_r)\Big)\right] = \exp\left(\frac{\gamma^2}{2} \sum_{a=1}^{m}\sum_{b=1}^{m} C(x_a, x_b)\right).$$

Multiplying by the Wick-normalization factor

$$\exp\left(-\frac{\gamma^2}{2} \sum_{r=1}^{m} C(x_r, x_r)\right)$$

leaves only the off-diagonal contribution, yielding (20). The one-point and two-point formulas are the cases $m = 1$ and $m = 2$. □

Proposition 5 (Effective one-parameter scaling)**.** *Define*

$$\tau := \frac{\gamma^2}{\beta} = \frac{2\varepsilon\,\gamma^2}{|U|}. \tag{24}$$

Then the exact gate law depends on (β, γ) *only through* τ*. Equivalently, if*

$$Y \sim \mathcal{N}(0, \tau G_U),$$

then

$$\xi_\gamma^{\mathrm{ex}}(x) \overset{d}{=} \exp\Big(Y(x) - \tfrac{1}{2}\mathrm{Var}(Y(x))\Big). \tag{25}$$

In particular,

$$K_\gamma(x,y) = \exp\big(\tau\, G_U(x,y)\big). \tag{26}$$

Proof. Since $\psi \sim \mathcal{N}(0, \beta^{-1}G_U)$, the rescaled field $Y := \gamma\psi$ is Gaussian with covariance

$$\mathrm{Cov}(Y) = \gamma^2\beta^{-1}G_U = \tau G_U.$$

The exact gate is precisely the Wick exponential of Y, so its law is determined by the law of Y, hence by τ alone. Equation (26) follows from (23). □

Proposition 6 (Small-strength expansion)**.** *For each fixed site* $x \in U$*,*

$$\xi_\gamma^{\mathrm{ex}}(x) = 1 + \gamma\psi(x) + \frac{\gamma^2}{2}\Big(\psi(x)^2 - C(x,x)\Big) + O_{L^2}(\gamma^3) \qquad (\gamma \to 0). \tag{27}$$

Moreover, for any $x, y \in U$*,*

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = 1 + \gamma^2 C(x,y) + O(\gamma^4), \tag{28}$$

$$\mathrm{Cov}\big(\xi_\gamma^{\mathrm{ex}}(x), \xi_\gamma^{\mathrm{ex}}(y)\big) = \gamma^2 C(x,y) + O(\gamma^4). \tag{29}$$

Proof. Expand

$$\exp\Big(\gamma\psi(x) - \tfrac{\gamma^2}{2}C(x,x)\Big)$$

in powers of γ and collect terms up to order γ^2, which gives (27). Equation (28) follows by expanding (23):

$$\exp\big(\gamma^2 C(x,y)\big) = 1 + \gamma^2 C(x,y) + O(\gamma^4).$$

Since $\mathbb{E}[\xi_\gamma^{\mathrm{ex}}(x)] = 1$, subtracting one yields (29). □

Interpretation of the small-γ regime. Proposition 6 shows that correlated additive Gaussian perturbation is only the *first-order proxy* of the exact gate. The full multiplicative construction retains positivity, exact mean preservation, and higher-order lognormal structure. This is one mathematically precise sense in which GCH is more than "correlated Gaussian noise with a different parameterization."

Exact theory versus implementation variant. The closed-form moment identities in Theorem 5.4 and Propositions 5 and 6 refer to the exact Wick-normalized gate (19). In practice, unless otherwise stated, our experiments use the sample-wise mean-one implementation variant

$$\xi_\gamma^{\mathrm{sw}}(x) = \frac{\exp(\gamma\psi(x))}{\frac{1}{|U|}\sum_{y\in U}\exp(\gamma\psi(y))}, \tag{30}$$

which is also positive and satisfies

$$\frac{1}{|U|}\sum_{x\in U}\xi_\gamma^{\mathrm{sw}}(x) = 1 \qquad \text{almost surely.} \tag{31}$$

However, (30) does *not* preserve the exact sitewise moment formulas of the Wick-normalized gate in general. Accordingly, the exact gate is the canonical object in the theory, while the sample-wise variant is the optimization-friendly implementation used in the main experiments. Appendix G.2 compares the two normalizations in more detail.

Framework interpretation. This separation is deliberate and belongs to the framework rather than only to this example. In VKD terms, ξ^{ex} is the canonical realization associated with the design desiderata, whereas ξ^{sw} is the deployed implementation used for training. The next subsection begins the compatibility layer for the latter.

5.5. Compatibility layer for the implemented mechanism

At this point the design layer is complete: the perturbation domain, deployment axis, latent law, induced geometry, and realization map have all been specified. We now turn to the second layer of VKD, namely compatibility on a target representation regime. In the language of Section 3, the regime of interest here is the class of positive coherent late-semantic maps, and the main observables are pairwise log-ratio deformation, ranking stability, intrinsic roughness, and topology.

The exact Wick gate is the canonical object of the variational theory, but the main experiments use the sample-wise gate (30). That implementation admits a sharp geometric description because it differs from the unnormalized exponential by a *spatially constant* correction in the log domain. This makes it possible to derive exact compatibility statements for the deployed mechanism rather than only for the canonical realization.

Framework reading guide. The next results should be read as properties of the induced action of the implemented mechanism on the observables above. The pairwise log-ratio theorem gives the local relative-geometry law; the ranking corollary converts it into a probability of evidence preservation; the intrinsic-energy corollary gives the whole-map roughness budget; and the hard-mask results identify the failure mode of discontinuous deletion in the same representation regime. In one sentence: the implemented GCh gate yields a *finite, margin-aware Gaussian deformation* of relative geometry, whereas hard masking yields a *singular, margin-blind deformation* whose relative damage worsens as representations become more coherent.

Assumption-to-practice map. Three mathematical objects are worth translating immediately into deep-learning language. Positivity models post-ReLU, attention-weighted, or saliency-like activations that encode evidence by magnitude. Pairwise log-ratios model *relative evidence*: how much more strongly one region or token is supported than another. The intrinsic energy $\mathcal{E}_{\mathrm{int}}$ measures edgewise variation after removing global scale, so low intrinsic energy corresponds to a map that is spatially coherent or low-frequency. With that translation in mind, the next results can be read as statements about evidence preservation, ranking preservation, and geometric distortion of semantic feature maps.

Positivity domain of the log-geometry statements. All results in this subsection that involve $\log h$ or pairwise log-ratios are stated for strictly positive fields. This is deliberate: the mathematical object under study is the geometry of *positive evidence maps*. In practice, the statements apply exactly on positive-support channels or regions, and one may also work with an ε-lifted field $h + \varepsilon$ if a numerical implementation needs to avoid exact zeros. The paper does not claim these log-geometry theorems for arbitrary signed activations.

Theorem 5.7 (Pairwise log-ratio stability of the implemented GCh gate)**.** *Let $h : U \to (0, \infty)$ be a fixed positive field and define*

$$\widetilde{h} := \xi_\gamma^{\mathrm{sw}} \odot h,$$

where ξ_γ^{sw} is given by (30). For every $x, y \in U$,

$$\Delta_{xy}^{\mathrm{sw}}(h) := \log \frac{\widetilde{h}(x)}{\widetilde{h}(y)} - \log \frac{h(x)}{h(y)} = \gamma\big(\psi(x) - \psi(y)\big). \tag{32}$$

Consequently, $\Delta_{xy}^{\mathrm{sw}}(h)$ is centered Gaussian with variance

$$\mathrm{Var}\big(\Delta_{xy}^{\mathrm{sw}}(h)\big) = \gamma^2\big(C(x,x) + C(y,y) - 2C(x,y)\big) = \tau R_G(x,y), \tag{33}$$

where $\tau = \gamma^2/\beta$ *and*

$$R_G(x,y) := G_U(x,x) + G_U(y,y) - 2G_U(x,y). \tag{34}$$

More generally, for any collection of pairs $\{(x_r, y_r)\}_{r=1}^m$*, the vector*

$$\big(\Delta^{\mathrm{sw}}_{x_r y_r}(h)\big)_{r=1}^m$$

is jointly Gaussian. In particular, $R_G(x,y) \geq 0$ *for all* $x, y \in U$ *because it is the variance proxy of a Gaussian difference field.*

Proof. Write

$$c(\psi) := \log\Big(\frac{1}{|U|}\sum_{z\in U} e^{\gamma\psi(z)}\Big).$$

Then

$$\log \widetilde{h}(x) = \log h(x) + \gamma\psi(x) - c(\psi) \qquad \text{for every } x \in U.$$

Subtracting the same identity at y gives (32). Since ψ is Gaussian and $\Delta^{\mathrm{sw}}_{xy}(h)$ is a linear functional of ψ, it is centered Gaussian with variance

$$\gamma^2 \mathrm{Var}\big(\psi(x) - \psi(y)\big) = \gamma^2\big(C(x,x) + C(y,y) - 2C(x,y)\big).$$

Using $C = \beta^{-1}G_U$ yields the final expression $\tau R_G(x,y)$. Joint Gaussianity for finitely many pairs is immediate for the same reason. □

Corollary 8 (Margin-sensitive ranking stability under the implemented GCh gate)**.** *Assume* $x \neq y$*,* $h(x) > h(y) > 0$*, and* $\tau R_G(x,y) > 0$*, and define the log-margin*

$$\delta_{xy}(h) := \log h(x) - \log h(y) > 0. \tag{35}$$

Then under the implemented sample-wise gate,

$$\Pr\big(\widetilde{h}(x) > \widetilde{h}(y)\big) = \Phi\left(\frac{\delta_{xy}(h)}{\sqrt{\tau R_G(x,y)}}\right), \tag{36}$$

where Φ *is the standard Gaussian cdf. Equivalently,*

$$\Pr\big(\widetilde{h}(x) \leq \widetilde{h}(y)\big) = \Phi\left(-\frac{\delta_{xy}(h)}{\sqrt{\tau R_G(x,y)}}\right) \leq \exp\left(-\frac{\delta_{xy}(h)^2}{2\tau R_G(x,y)}\right). \tag{37}$$

Proof. By Theorem 5.7,

$$\log\frac{\widetilde{h}(x)}{\widetilde{h}(y)} = \log\frac{h(x)}{h(y)} + \Delta^{\mathrm{sw}}_{xy}(h) = \delta_{xy}(h) + \Delta^{\mathrm{sw}}_{xy}(h),$$

where $\Delta^{\mathrm{sw}}_{xy}(h) \sim \mathcal{N}(0, \tau R_G(x,y))$. Therefore

$$\Pr\big(\widetilde{h}(x) > \widetilde{h}(y)\big) = \Pr\big(\delta_{xy}(h) + \Delta^{\mathrm{sw}}_{xy}(h) > 0\big) = \Phi\left(\frac{\delta_{xy}(h)}{\sqrt{\tau R_G(x,y)}}\right),$$

which is (36). The tail bound follows from the standard Gaussian bound $\Phi(-u) \leq e^{-u^2/2}$ for $u > 0$. □

Deep learning interpretation. Pairwise log-ratios are a natural coordinate system for relative evidence: how much stronger one region, token, or semantic part is than another. Corollary 8 says that the implemented GCH gate is *margin-sensitive*: if a feature comparison already has a large semantic log-margin, then the probability of preserving that ordering is exponentially close to one. This is the kind of behavior one wants from a late-stage regularizer—strong semantic contrasts become more, not less, stable. The mild condition $\tau R_G(x,y) > 0$ simply excludes the degenerate zero-variance case; on a connected Dirichlet grid it is automatic whenever $x \neq y$ and $\gamma \neq 0$.

To aggregate pairwise distortions over the grid, define the *intrinsic* interior edge set

$$E_{\mathrm{int}} \coloneqq \{\{x,y\} \in E : \ x,y \in U\}$$

and the associated intrinsic graph energy

$$\mathcal{E}_{\mathrm{int}}(f) \coloneqq \frac{1}{2} \sum_{\{x,y\} \in E_{\mathrm{int}}} c_{xy}\big(f(x) - f(y)\big)^2 = \frac{1}{2}\langle f, L_{\mathrm{int}} f\rangle, \tag{38}$$

where L_{int} is the interior graph Laplacian on U with *no* auxiliary boundary term. Unlike the Dirichlet energy used in the variational design, $\mathcal{E}_{\mathrm{int}}$ is invariant under adding spatial constants, so it measures *relative* geometry.

Corollary 9 (Exact expected intrinsic roughness budget under the implemented gate)**.** *Let $h : U \to (0,\infty)$ and $\widetilde{h} = \xi_\gamma^{\mathrm{sw}} \odot h$. Then*

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h})\big] = \mathcal{E}_{\mathrm{int}}(\log h) + \gamma^2 \varepsilon_{\mathrm{int}}, \qquad \varepsilon_{\mathrm{int}} \coloneqq \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\psi)\big] = \tfrac{1}{2}\mathrm{Tr}(L_{\mathrm{int}} C). \tag{39}$$

Proof. From the proof of Theorem 5.7,

$$\log \widetilde{h} = \log h + \gamma\psi - c(\psi)\mathbf{1},$$

where $\mathbf{1}$ is the all-ones vector on U. Because $L_{\mathrm{int}}\mathbf{1} = 0$, the constant term drops out of $\mathcal{E}_{\mathrm{int}}$. Hence

$$\mathcal{E}_{\mathrm{int}}(\log \widetilde{h}) = \mathcal{E}_{\mathrm{int}}(\log h + \gamma\psi).$$

Expanding the quadratic form and taking expectation gives

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h})\big] &= \mathcal{E}_{\mathrm{int}}(\log h) + \gamma\, \mathbb{E}\big[\langle \log h, L_{\mathrm{int}}\psi\rangle\big] + \gamma^2 \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\psi)\big] \\ &= \mathcal{E}_{\mathrm{int}}(\log h) + \gamma^2 \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\psi)\big],\end{aligned}$$

where the cross term vanishes because $\mathbb{E}[\psi] = 0$. Finally,

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\psi)\big] = \frac{1}{2}\mathbb{E}[\psi^\top L_{\mathrm{int}}\psi] = \frac{1}{2}\mathrm{Tr}(L_{\mathrm{int}} C).$$

□

Deep learning interpretation. Corollary 9 is the whole-map counterpart of the pairwise result. In the intrinsic log-geometry of a positive feature map, the implemented GCH gate adds an *exactly quantified expected* amount of roughness. It deforms the representation by a finite random field rather than puncturing it with hard zeros. For practitioners, this is the rigorous version of the intuition that GCH injects controlled uncertainty rather than discontinuous semantic damage.

Corollary 10 (Scale compatibility of the implemented GCH gate)**.** *For any $a > 0$ and any positive field $h : U \to (0,\infty)$, let $\widetilde{h}_a \coloneqq \xi_\gamma^{\mathrm{sw}} \odot (ah)$. Then for every $x, y \in U$,*

$$\Delta_{xy}^{\mathrm{sw}}(ah) = \Delta_{xy}^{\mathrm{sw}}(h), \tag{40}$$

and

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h}_a)\big] - \mathcal{E}_{\mathrm{int}}(\log(ah)) = \gamma^2 \varepsilon_{\mathrm{int}}. \tag{41}$$

Thus the pairwise deformation law and the added intrinsic roughness budget are invariant under global amplitude rescaling.

Proof. Because $\log(ah) = \log h + (\log a)\mathbf{1}$, global rescaling adds only a spatial constant in the log domain. Both Theorem 5.7 and the intrinsic energy $\mathcal{E}_{\text{int}}$ are invariant under such constants, which yields (40) and (41). □

Deep learning interpretation. This is a concrete advantage of working in multiplicative log-geometry. If the same semantic feature map is globally rescaled—for example by a change in channel gain, normalization, or overall confidence level—the geometric effect of the implemented GCH gate does not change. The perturbation tracks relative structure rather than absolute amplitude.

Corollary 11 (Finite expected intrinsic roughness for a perfectly coherent positive map under the implemented gate)**.** *Let $h : U \to (0, \infty)$ satisfy $\log h(x) \equiv c$ on U for some constant $c \in \mathbb{R}$. Then for $\widetilde{h} = \xi_\gamma^{\text{sw}} \odot h$,*

$$\mathbb{E}\big[\mathcal{E}_{\text{int}}(\log \widetilde{h})\big] = \gamma^2 \varepsilon_{\text{int}}. \tag{42}$$

In particular, a perfectly coherent positive map acquires a finite and explicitly budgeted expected *intrinsic roughness under the implemented* GCH *gate.*

Proof. If $\log h$ is constant on U, then $\mathcal{E}_{\text{int}}(\log h) = 0$. The claim follows immediately from Corollary 9. □

Deep learning interpretation. A late-stage representation is often close to piecewise coherent in log-amplitude: within a semantically consistent region, the main issue is not whether the feature is exactly constant, but whether the perturbation preserves the region as a coherent object. Corollary 11 gives an expectation-level statement: starting from zero intrinsic roughness, the implemented GCH gate produces a finite and explicitly budgeted expected roughness level rather than a singular or uncontrolled distortion.

The next result formalizes the opposite behavior of hard binary masks. The singular-ratio statement applies whenever a compared pair can be zeroed with positive probability, and therefore covers dropout, DropBlock, and related hard-masking mechanisms in their natural nontrivial regime.

Theorem 5.12 (Binary masks are incompatible with finite log-ratio geometry)**.** *Let $h : U \to (0, \infty)$, let $a > 0$, and let $m : U \to \{0, a\}$ be any random binary mask. Define $\widetilde{h}^m := m \odot h$. If there exist $x, y \in U$ such that*

$$\Pr\big(m(x) = 0 \text{ or } m(y) = 0\big) > 0, \tag{43}$$

then

$$\log \frac{\widetilde{h}^m(x)}{\widetilde{h}^m(y)}$$

fails to be an almost surely finite real-valued random variable. In particular, no finite-variance analog of Theorem 5.7 can hold for such a mask. For inverted dropout at distinct *compared sites $x \neq y$,*

$$m_q(z) = \frac{b(z)}{q}, \qquad b(z) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(q),$$

the total probability of a zero event at the compared pair is $1 - q^2$, with asymmetric singular events of total probability $2q(1-q)$ and a joint-erasure event of probability $(1-q)^2$.

Proof. If $m(x) = 0$ and $m(y) = a$, then $\widetilde{h}^m(x) = 0 < \widetilde{h}^m(y)$ and the log-ratio equals $-\infty$. If $m(x) = a$ and $m(y) = 0$, then the log-ratio equals $+\infty$. If $m(x) = m(y) = 0$, then both numerator and denominator vanish and the log-ratio is undefined, hence not a finite real number. Therefore the log-ratio fails to be almost surely finite whenever (43) holds. For inverted dropout at distinct sites $x \neq y$, independence gives

$$\Pr\big(m_q(x) = 0 \text{ or } m_q(y) = 0\big) = 1 - q^2,$$

while the asymmetric events have total probability $2q(1-q)$ and the joint-erasure event has probability $(1-q)^2$. □

Corollary 13 (Margin-blind ranking under inverted dropout)**.** *Assume $h(x) > h(y) > 0$ and let m_q be inverted dropout with keep probability $q \in (0, 1]$. Then*

$$\Pr\big((m_q \odot h)(x) > (m_q \odot h)(y)\big) = q. \tag{44}$$

In particular, the probability of preserving the ordering is independent of the magnitude of the underlying feature margin.

Proof. Write $m_q(z) = b(z)/q$ with $b(z) \in \{0, 1\}$. If $b(x) = 1$, then $(m_q \odot h)(x) = h(x)/q$ and regardless of whether $b(y) = 0$ or 1, one has $(m_q \odot h)(x) > (m_q \odot h)(y)$ because $h(x) > h(y) > 0$. If $b(x) = 0$, then $(m_q \odot h)(x) = 0 \le (m_q \odot h)(y)$. Therefore the ordering is preserved if and only if $b(x) = 1$, which occurs with probability q. □

Deep learning interpretation. This corollary is intentionally blunt: even if one activation is *arbitrarily* more semantically decisive than another, inverted dropout preserves that ordering with probability exactly q and destroys or erases it with probability $1 - q$. In that sense hard masking is *margin-blind*. By comparison, Corollary 8 shows that the implemented GCh gate becomes more stable as the semantic margin increases.

Proposition 14 (Exact intrinsic energy inflation under inverted dropout)**.** *Let $m_q(x) = b(x)/q$ with i.i.d. $b(x) \sim \mathrm{Bernoulli}(q)$ and $q \in (0, 1]$. Then for every deterministic field $h : U \to \mathbb{R}$,*

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big] = \mathcal{E}_{\mathrm{int}}(h) + \frac{1-q}{2q} \sum_{x \in U} d_x^{\mathrm{int}} h(x)^2, \tag{45}$$

where

$$d_x^{\mathrm{int}} := \sum_{y : \{x,y\} \in E_{\mathrm{int}}} c_{xy}$$

is the intrinsic weighted degree of x.

Proof. Fix an interior edge $\{x, y\} \in E_{\mathrm{int}}$. Since $m_q(x)$ and $m_q(y)$ are independent and $\mathbb{E}[m_q(x)] = 1$, $\mathbb{E}[m_q(x)^2] = 1/q$, we have

$$\begin{aligned}\mathbb{E}\big[(m_q(x)h(x) - m_q(y)h(y))^2\big] &= \frac{1}{q}h(x)^2 + \frac{1}{q}h(y)^2 - 2h(x)h(y) \\ &= \big(h(x) - h(y)\big)^2 + \Big(\frac{1}{q} - 1\Big)\big(h(x)^2 + h(y)^2\big).\end{aligned}$$

Multiply by $c_{xy}/2$ and sum over E_{int}. The first term sums to $\mathcal{E}_{\mathrm{int}}(h)$, while the second becomes

$$\frac{1-q}{2q} \sum_{x \in U} d_x^{\mathrm{int}} h(x)^2.$$

This is exactly (45). □

Corollary 15 (Coherence amplification factor for inverted dropout)**.** *Assume $\mathcal{E}_{\mathrm{int}}(h) > 0$ and define the coherence score*

$$\kappa(h) := \frac{\sum_{x \in U} d_x^{\mathrm{int}} h(x)^2}{2\mathcal{E}_{\mathrm{int}}(h)}. \tag{46}$$

Then inverted dropout satisfies

$$\frac{\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big]}{\mathcal{E}_{\mathrm{int}}(h)} = 1 + \frac{1-q}{q}\,\kappa(h). \tag{47}$$

Proof. Divide both sides of (45) by $\mathcal{E}_{\mathrm{int}}(h) > 0$ and rearrange. □

Deep learning interpretation. The scalar $\kappa(h)$ is an interpretable mismatch factor: it is large when a feature map carries nontrivial activation mass but varies only weakly across space, i.e. when the representation is coherent. Corollary 15 therefore says that hard masking damages coherent representations more severely in relative terms, and it does so by a completely explicit amplification factor.

Corollary 16 (Immediate loss of perfect coherence under inverted dropout in expectation)**.** *Assume $q \in (0,1)$ and that the interior graph has at least one edge. Let $h(x) \equiv c$ on U for some constant $c \neq 0$. Then*

$$\mathcal{E}_{\mathrm{int}}(h) = 0, \qquad \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big] = \frac{1-q}{2q} c^2 \sum_{x \in U} d_x^{\mathrm{int}} > 0. \tag{48}$$

Thus perfect coherence is not preserved by a single masking step: the post-mask field has strictly positive expected *intrinsic roughness.*

Proof. A constant field has zero intrinsic energy, so the claim follows immediately from Proposition 14. □

Deep learning interpretation. This is the cleanest possible statement of hard-mask mismatch. Even if a feature map is spatially perfectly coherent before perturbation, binary masking does not preserve that zero-roughness state in any controlled relative sense. After one masking step the representation acquires strictly positive *expected* edgewise roughness, reflecting the discontinuities introduced by hard deletion.

Corollary 17 (Late-stage mismatch of inverted dropout under coherence)**.** *Let $(h_\ell)_{\ell \geq 1}$ be deterministic fields on U such that*

$$\inf_{\ell \geq 1} \sum_{x \in U} d_x^{\mathrm{int}} h_\ell(x)^2 > 0, \qquad \mathcal{E}_{\mathrm{int}}(h_\ell) > 0 \text{ for every } \ell, \qquad \mathcal{E}_{\mathrm{int}}(h_\ell) \to 0. \tag{49}$$

Then for every fixed $q \in (0,1)$,

$$\frac{\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h_\ell)\big] - \mathcal{E}_{\mathrm{int}}(h_\ell)}{\mathcal{E}_{\mathrm{int}}(h_\ell)} \longrightarrow \infty. \tag{50}$$

Thus, as the representation becomes more spatially coherent, the relative geometric distortion induced by binary masking diverges.

Proof. By Proposition 14,

$$\frac{\mathbb{E}[\mathcal{E}_{\mathrm{int}}(m_q \odot h_\ell)] - \mathcal{E}_{\mathrm{int}}(h_\ell)}{\mathcal{E}_{\mathrm{int}}(h_\ell)} = \frac{1-q}{2q} \cdot \frac{\sum_{x \in U} d_x^{\mathrm{int}} h_\ell(x)^2}{\mathcal{E}_{\mathrm{int}}(h_\ell)}.$$

The numerator is bounded below by assumption, whereas the denominator tends to zero, so the ratio diverges to $+\infty$. □

Corollary 18 (Margin-growth regime: GCH strengthens while dropout saturates)**.** *Fix distinct $x, y \in U$ and assume $\tau R_G(x,y) > 0$. Let $(h_\ell)_{\ell \geq 1}$ be positive fields with $h_\ell(x) > h_\ell(y)$ for every ℓ. Define*

$$\delta_\ell := \log h_\ell(x) - \log h_\ell(y).$$

If

$$\frac{\delta_\ell}{\sqrt{\tau R_G(x,y)}} \longrightarrow \infty, \tag{51}$$

then under the implemented sample-wise GCH *gate,*

$$\Pr\big(\widetilde{h}_\ell(x) > \widetilde{h}_\ell(y)\big) \longrightarrow 1. \tag{52}$$

Under inverted dropout with keep probability q, however,

$$\Pr\big((m_q \odot h_\ell)(x) > (m_q \odot h_\ell)(y)\big) = q \qquad \text{for every } \ell. \tag{53}$$

Proof. Equation (52) follows immediately from Corollary 8 and the assumption (51). Equation (53) is exactly Corollary 13. □

Deep learning interpretation. Corollary 18 is the mathematically clean version of the informal slogan that GCH becomes more compatible with later, sharper semantic representations. It does *not* claim that depth is always beneficial in every model. Instead it says: whenever late-stage representations become more decisively separated in their relative log-margins, the implemented GCH gate respects those rankings with probability tending to one, while hard masking stays stuck at the same keep-probability ceiling.

Corollary 19 (Representation-compatibility dichotomy)**.** *Under the hypotheses of Theorems 5.7 and 5.12 and Corollaries 9 and 17, the implemented* GCH *gate and hard binary masks exhibit qualitatively different behavior on positive coherent representations:*

1. *the implemented* GCH *gate preserves a finite relative log-geometry, with exact Gaussian pairwise deformations, margin-sensitive ranking stability, and an exact additive intrinsic roughness budget;*
2. *any hard binary mask that can zero one or both members of a compared pair with positive probability fails to preserve finite log-ratio geometry, and inverted dropout preserves pairwise ranking only with the margin-blind probability q; and*
3. *for inverted dropout, the relative intrinsic distortion diverges along coherent representation sequences satisfying* (49)*.*

The assumptions in Corollary 17 are a clean mathematical abstraction of the late-semantic regime: the representation retains nontrivial mass but becomes increasingly low-frequency or spatially coherent. In that regime, binary masking becomes more and more mismatched. By contrast, Theorem 5.7 and Corollaries 8, 9, 18 and 19 show that the implemented GCH gate continues to produce a finite Gaussian deformation whose pairwise, ranking, and aggregate effects are controlled by the Green geometry.

Engineering takeaway. If a layer encodes positive region-level evidence or token-level saliency, then the mathematically relevant question is not merely whether noise is mean-preserving in expectation, but whether it preserves *relative comparisons* that the downstream model relies on. The results above say that GCH perturbs those comparisons through a finite, margin-aware Gaussian deformation, whereas hard binary masks can delete them outright and become especially mismatched when the representation is coherent and semantically sharp.

A topological complement is given in Appendix E: positive multiplicative gates perturb superlevel sets only through a multiplicative threshold band, whereas hard Bernoulli masking destroys loop-type excursion topology with probability $1 - q^n$ on an n-cycle.

What these theorems do and do not claim. They do *not* prove that one should always inject noise deeper, nor that every masking strategy is inferior in every possible regime. What they prove is a sharper and more defensible statement: once a layer behaves like a positive coherent evidence map, there is a mathematically meaningful comparison to make. In that regime, the implemented GCH gate preserves finite relative geometry, ranking information, and an explicit global roughness budget, while hard binary masking either makes those quantities singular or amplifies their distortion by an explicit coherence factor. That is exactly the regime targeted by the late-stage experiments in this paper.

5.6. Implementation and efficient sampling

Injecting the gate. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, we inject the spatial gate multiplicatively:

$$\widetilde{F}_c(x) = F_c(x)\,\xi_\gamma(x), \qquad x \in U. \tag{54}$$

In the experiments, β is fixed once the grid, operator, and normalization convention are chosen; γ is the reported strength knob.

FFT/DST sampling of the GFF log-field. For the unweighted four-neighbor Dirichlet Laplacian on the $H \times W$ interior grid U, the eigenbasis is the 2D sine basis:

$$e_{k,\ell}(i,j) = \sin\Big(\frac{\pi k i}{H+1}\Big)\,\sin\Big(\frac{\pi \ell j}{W+1}\Big), \tag{55}$$

$$\lambda_{k,\ell} = 4\sin^2\Big(\frac{\pi k}{2(H+1)}\Big) + 4\sin^2\Big(\frac{\pi \ell}{2(W+1)}\Big), \tag{56}$$

for $1 \le k \le H$ and $1 \le \ell \le W$. 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 compute $\psi = \mathrm{IDST2}(A)$ using an orthonormal inverse discrete sine transform. Fast DST implementations rely on FFT internally, giving near-linear complexity in the number of spatial sites.

Algorithm 1 GCh on an $H \times W$ grid (Dirichlet; FFT/DST implementation)

Input: grid size (H, W), parameters $\beta > 0$, $\gamma \in \mathbb{R}$, feature map $F \in \mathbb{R}^{C\times H\times W}$
Precompute once: eigenvalues $\lambda_{k,\ell}$ in (56); choose a DST convention; optionally precompute the variance map $v(x) = C(x,x)$
Sample spectral coefficients: draw i.i.d. $Z_{k,\ell} \sim \mathcal{N}(0,1)$
Scale by the Laplacian spectrum: set $A_{k,\ell} \leftarrow Z_{k,\ell}/\sqrt{\beta\,\lambda_{k,\ell}}$
Inverse transform: $\psi \leftarrow \mathrm{IDST2}(A)$ (so $\psi \sim \mathcal{N}(0, (\beta L_U)^{-1})$)
Exponentiate: $G(x) \leftarrow \exp(\gamma\psi(x))$ for all $x \in U$
Normalize (choose one):
  Exact Wick: $\xi(x) \leftarrow \exp(\gamma\psi(x) - \frac{\gamma^2}{2} v(x))$
  Sample-wise mean-one: $\xi(x) \leftarrow G(x) \Big/ \Big(\frac{1}{|U|}\sum_{y\in U} G(y)\Big)$
Inject into features: $\widetilde{F}_c(x) \leftarrow F_c(x)\,\xi(x)$ for all channels c and sites $x \in U$
Output: noised feature map $\widetilde{F}$

6. Experiments

We evaluate whether the design principles behind GCh translate into practical gains. Our empirical questions are: (i) which ingredients matter beyond raw noise magnitude, (ii) where in network depth is the mechanism most effective, and (iii) whether the effect transfers beyond the primary CNN setting. Detailed protocols are provided in Appendix F.

Theory-to-experiment map. The representation-compatibility results make four concrete empirical predictions. Pairwise log-ratio stability and the margin-sensitive ranking law predict that when late-stage representations encode decisive relative evidence, GCh should preserve that evidence better than hard masking. The intrinsic roughness budget predicts a broad non-destructive regime of stochasticity rather than abrupt fragmentation. The immediate loss-of-perfect-coherence and coherence-mismatch results predict that once a representation becomes spatially coherent, binary masking should incur disproportionate damage, especially at late stages. Finally, the topological appendix is most relevant to the fine-grained Pets pilot, where preserving coherent part structure matters most directly.

6.1. Controlled multi-baseline ImageNet study

To isolate the source of the gains, we run a controlled 3-seed comparison that separates noise magnitude, spatial correlation, and positivity/mean-one multiplicative gating. We compare GCh against Dropout and DropBlock, together with additive Gaussian baselines (i.i.d. and correlated) whose injected strength is energy-matched. Table 1 reports mean±std.

Unless otherwise stated, the GCh experiments use the sample-wise mean-one normalization of Algorithm 1, which is the implementation variant used throughout the main body.

Unified strength knob. To keep tables compact, we write g for the method-specific strength parameter. For GCh, $g \equiv \gamma$; for Gaussian baselines, $g \equiv \sigma$; for Dropout and DropBlock, $g \equiv p$; and for the no-noise baseline, $g = 0$.

Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.765±0.001	0.931±0.004	0.030±0.001
Dropout	0.1	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. Transfer to a transformer backbone: Swin-T

We additionally evaluate GCh on Swin-T under the same full-recipe training setup. Direct Dropout/DropBlock analogues are less aligned with transformer pipelines because token-based representations and attention updates no longer correspond to contiguous suppression on a convolutional feature grid, and standard transformer regularization usually acts on different objects (e.g., stochastic depth, attention dropout, or MLP dropout). We therefore report the clean full-recipe baseline 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. What the evidence shows

The ImageNet controlled study supports three main conclusions.

First, *correlation alone is not enough.* In Table 1, the correlated additive Gaussian baseline can worsen calibration relative to the no-noise baseline at matched strength. The strongest ECE improvements appear only when correlation is combined with *positive mean-one multiplicative gating*, namely in GCh.

Second, *depth matters*. Injection depth induces a clear accuracy–calibration trade-off (Appendix Table 5): moving from earlier to later stages substantially improves calibration while changing accuracy only modestly. In the selected 7-corruption ImageNet-C evaluation, the late-stage setting reduces ECE by 46% and improves NLL by 3.3% relative to the no-noise baseline (Appendix Table 7), with corruption-wise details in Table 10.

Third, *there is a stable operating regime*. A strength sweep reveals a broad useful range around $g \approx 0.07$–0.18. When g becomes too large, accuracy collapses and NLL rises sharply; ECE can also become misleadingly small under severe underconfidence, which we treat as a failure mode rather than a favorable outcome (Appendix Table 9). This empirical pattern is consistent with Theorem 5.7 and Corollaries 8, 9, 11, 13, 16 and 17: the implemented GCh gate induces a finite Gaussian deformation in relative log-geometry together with an exact expected intrinsic roughness budget, while also becoming more stable when semantic margins are sharper; hard masking, by contrast, is margin-blind and incurs a coherence-sensitive geometric penalty whose relative size grows in the coherent-representation regime.

Finally, the Swin-T result provides preliminary transfer evidence beyond the primary ResNet-50 setting, and the Oxford-IIIT Pets pilot in the appendix supports the claim that spatially coherent positive gating is especially compatible with fine-grained structure-sensitive recognition.

7. Conclusion

We proposed Variational Kernel Design as a compositional framework for internal noise in deep learning. In VKD, a stochastic mechanism is not chosen from a fixed heuristic menu; it is derived from learning-relevant constraints and then analyzed on the representation regime where it is actually deployed. This two-layer viewpoint—mechanism design followed by compatibility analysis—is the main conceptual contribution of the paper.

Within the solved quadratic VKD subfamily studied here, we formulated the log-field construction problem as an explicit finite-dimensional maximum-entropy program under a quadratic operator budget, solved that program in closed form, and obtained an entropy-gap identity showing that the optimizer is uniquely Gaussian. For the Dirichlet operator, this makes the Green kernel emerge as the induced correlation geometry; after Wick normalization, it yields the canonical exact GCh gate. Once the operator and energy budget are fixed, the exact gate becomes an effectively one-parameter family through $\tau = \gamma^2/\beta$.

The more distinctive message of the paper is the representation-compatibility layer that sits on top of this variational design. For the sample-wise gate actually used in the experiments, we established exact Gaussian control of pairwise log-ratios, margin-sensitive ranking stability, and an exact expected intrinsic roughness budget. For hard binary masking, we proved the opposite kind of statement: incompatibility with finite log-ratio geometry, margin-blind ranking under inverted dropout, immediate loss of perfect coherence in expectation on perfectly coherent maps, and a relative distortion term that diverges in the coherent-representation regime. The central contrast is therefore not merely *Gaussian versus Bernoulli*; it is *finite, margin-aware deformation versus singular or coherence-amplified deletion*.

These theorems are intentionally conditional rather than universal. They do not claim that every deeper layer in every architecture will automatically favor GCh. They claim something more precise and, for practice, more useful: whenever positive semantic representations become coherent and their relative evidence sharpens, smooth multiplicative gating preserves those comparisons in a way that hard deletion cannot. That conditional form is exactly what allows the theory to speak directly to the late-stage regime without pretending to replace empirical evaluation.

Empirically, GCh improves calibration on clean ImageNet, improves both ECE and NLL on a selected 7-corruption ImageNet-C evaluation, remains effective in late semantic stages where hard masking can degrade clean calibration, and shows encouraging transfer to Swin-T and a fine-grained

pilot. The practical takeaway is simple: if a layer carries positive, coherent, region-level evidence, then the right question is not merely whether noise is unbiased, but whether it perturbs relative evidence smoothly or deletes it abruptly, and whether that perturbation respects the comparisons the downstream network actually uses. Our theory says that GCh does the former, whereas canonical hard binary masks such as dropout and DropBlock-type deletion mechanisms tend to do the latter in the coherent late-stage regime.

More broadly, the paper suggests a reusable recipe for future work. First choose the operator that encodes the geometry one wants the noise to respect. Then derive the corresponding latent law and realization. Finally, ask whether the implemented mechanism is compatible with the representation regime of interest. That perspective opens the door to principled variants based on massive, anisotropic, graph-adapted, or architecture-specific operators while preserving the same mathematical blueprint.

References

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

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 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)*, pages 702–703, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations (ICLR)*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10727–10737, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision (ECCV)*, pages 646–661, 2016.

Meelis Kull, Miquel Perelló Nieto, Markus K"angsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12316–12326, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6402–6413, 2017.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. PatchDropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4917–4926, 2023.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15682–15694, 2021.

Rafael M"uller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4696–4705, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2901–2907, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13991–14002, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012.

Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision (ECCV)*, pages 53–69, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18583–18599, 2020.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020.

Yoshihiro Yamada, Masakazu Iwamura, Takuya Akiba, and Koichi Kise. ShakeDrop regularization for deep residual learning. *arXiv preprint arXiv:1802.02375*, 2018.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

A. Notation and Terminology (Glossary)

- U: the $H \times W$ feature grid on which the gate is sampled and applied.
- B: the auxiliary Dirichlet boundary outside U; $\bar{U} = U \cup B$.
- L_U: Dirichlet Laplacian on U; $G_U = L_U^{-1}$: Dirichlet Green kernel.
- $\mathcal{F}$: law family of latent log-fields in the VKD mechanism.
- K: intended second-order geometry / kernel in the VKD mechanism.
- $\mathcal{T}$: injection operator in the VKD mechanism.
- ℓ: realization map from latent log-field to positive gate.
- $\mathcal{R}$: target representation regime for compatibility analysis.
- ψ: log-field; ξ: positive multiplicative gate; γ: GCh strength parameter.
- g: unified strength knob in the experimental tables ($g = \gamma/\sigma/p$ depending on the method).
- GCh: Gaussian Chaos Noise / gate (ours).
- **IID/Corr. Gaussian**: additive Gaussian baselines with matched injected energy.

B. Full variational derivation for Theorem 5.1

This appendix gives a fuller proof of the quadratic MaxEnt principle, including an entropy-gap identity and, for completeness, the corresponding Euler–Lagrange stationarity calculation.

Let $Q \succ 0$ be symmetric positive definite on $\mathbb{R}^U$, let $n = |U|$, and let $\varepsilon > 0$. Recall the admissible class

$$\mathcal{A}(Q,\varepsilon) = \left\{ p : \mathbb{R}^U \to [0,\infty) \ : \ \int p = 1, \ \int \psi\, p(\psi)\, d\psi = 0, \ \int \frac{1}{2}\langle \psi, Q\psi\rangle p(\psi)\, d\psi = \varepsilon, \ h(p) > -\infty \right\}.$$

B.1. Entropy-gap proof of optimality and uniqueness

Define

$$\Sigma_{Q,\varepsilon} \coloneqq \frac{2\varepsilon}{n} Q^{-1}, \qquad p^\star(\psi) \coloneqq \frac{1}{(2\pi)^{n/2} \det(\Sigma_{Q,\varepsilon})^{1/2}} \exp\Big(-\frac{1}{2} \psi^\top \Sigma_{Q,\varepsilon}^{-1} \psi \Big).$$

Since $\Sigma_{Q,\varepsilon}^{-1} = \frac{n}{2\varepsilon} Q$, we have

$$\mathbb{E}_{p^\star}\left[\frac{1}{2} \langle \psi, Q\psi \rangle \right] = \frac{1}{2} \mathrm{Tr}(Q \Sigma_{Q,\varepsilon}) = \frac{1}{2} \mathrm{Tr}\Big(Q \frac{2\varepsilon}{n} Q^{-1} \Big) = \varepsilon,$$

and clearly $\mathbb{E}_{p^\star}[\psi] = 0$, so $p^\star \in \mathcal{A}(Q, \varepsilon)$.

Now fix any $p \in \mathcal{A}(Q, \varepsilon)$. Using the definition of KL divergence,

$$\begin{aligned} \mathrm{KL}(p \| p^\star) &= \int p(\psi) \log \frac{p(\psi)}{p^\star(\psi)} \, d\psi \\ &= -h(p) - \int p(\psi) \log p^\star(\psi) \, d\psi. \end{aligned} \tag{57}$$

Since

$$\log p^\star(\psi) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log \det(\Sigma_{Q,\varepsilon}) - \frac{1}{2} \psi^\top \Sigma_{Q,\varepsilon}^{-1} \psi,$$

and $\Sigma_{Q,\varepsilon}^{-1} = \frac{n}{2\varepsilon} Q$, the energy constraint gives

$$\begin{aligned} -\int p(\psi) \log p^\star(\psi) \, d\psi &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma_{Q,\varepsilon}) + \frac{n}{2\varepsilon} \int \frac{1}{2} \langle \psi, Q\psi \rangle p(\psi) \, d\psi \\ &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma_{Q,\varepsilon}) + \frac{n}{2}. \end{aligned} \tag{58}$$

But the right-hand side is exactly the entropy of $p^\star$:

$$h(p^\star) = \frac{1}{2} \log\Big((2\pi e)^n \det(\Sigma_{Q,\varepsilon}) \Big).$$

Therefore (57) and (58) imply

$$\mathrm{KL}(p \| p^\star) = h(p^\star) - h(p).$$

Because $\mathrm{KL}(p \| p^\star) \geq 0$, we obtain

$$h(p) \leq h(p^\star),$$

with equality iff $\mathrm{KL}(p \| p^\star) = 0$, i.e. iff $p = p^\star$ almost everywhere. This proves both optimality and uniqueness.

B.2. Euler–Lagrange derivation (for completeness)

The same optimizer can be recovered by stationarity. Introduce Lagrange multipliers $\lambda_0 \in \mathbb{R}$, $\lambda \in \mathbb{R}^U$, and $\beta \in \mathbb{R}$, and define

$$\begin{aligned} \mathcal{L}(p) = &-\int p(\psi) \log p(\psi) \, d\psi + \lambda_0 \Big(\int p(\psi) \, d\psi - 1 \Big) \\ &+ \Big\langle \lambda, \int \psi \, p(\psi) \, d\psi \Big\rangle - \beta \Big(\int \frac{1}{2} \langle \psi, Q\psi \rangle p(\psi) \, d\psi - \varepsilon \Big). \end{aligned} \tag{59}$$

For an interior optimum, the first variation in a direction δp gives

$$\int \Big(-\log p(\psi) - 1 + \lambda_0 + \langle \lambda, \psi \rangle - \beta \tfrac{1}{2} \langle \psi, Q\psi \rangle \Big) \delta p(\psi) \, d\psi = 0.$$

Hence the Euler–Lagrange equation is

$$-\log p(\psi) - 1 + \lambda_0 + \langle \lambda, \psi \rangle - \beta \tfrac{1}{2} \langle \psi, Q\psi \rangle = 0,$$

so

$$p(\psi) \propto \exp(\langle \lambda, \psi \rangle) \exp\Big(- \beta \tfrac{1}{2} \langle \psi, Q\psi \rangle \Big).$$

The centering constraint forces $\lambda = 0$, and integrability requires $\beta > 0$ because $Q \succ 0$. Thus

$$p(\psi) \propto \exp\Big(- \beta \tfrac{1}{2} \langle \psi, Q\psi \rangle \Big) = \exp\Big(- \tfrac{1}{2} \psi^\top (\beta Q) \psi \Big),$$

which is the centered Gaussian $\mathcal{N}(0, (\beta Q)^{-1})$. Matching the energy budget yields

$$\varepsilon = \frac{1}{2} \mathrm{Tr}\big(Q(\beta Q)^{-1}\big) = \frac{n}{2\beta}, \qquad \text{so} \qquad \beta = \frac{n}{2\varepsilon}.$$

This reproduces

$$(\beta Q)^{-1} = \frac{2\varepsilon}{n} Q^{-1} = \Sigma_{Q,\varepsilon}.$$

B.3. Dirichlet specialization

Setting $Q = L_U$ gives the optimizer used in the main text:

$$p^\star_{L_U,\varepsilon} = \mathcal{N}\big(0, (\beta L_U)^{-1}\big), \qquad \beta = \frac{|U|}{2\varepsilon}.$$

Its covariance is

$$\mathrm{Cov}(\psi) = \frac{2\varepsilon}{|U|} L_U^{-1} = \beta^{-1} G_U.$$

Other boundary conditions. If one uses periodic or Neumann boundary conditions on a connected finite graph, the Laplacian has a constant nullspace, so the corresponding field must be defined after gauge fixing, for example by pinning one site or imposing zero spatial mean and using the Moore–Penrose pseudoinverse. Under the auxiliary Dirichlet boundary used in the main text, $L_U \succ 0$ and no additional gauge fixing is needed.

Massive variant. A regularized or *massive* variant replaces L_U by $L_U + \mu I$ for $\mu > 0$:

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

This corresponds to the quadratic energy

$$\mathcal{E}_\mu(\psi) = \frac{1}{2} \psi^\top (L_U + \mu I) \psi$$

and yields a better conditioned covariance with shorter-range correlations.

C. Further properties of the exact Gaussian-chaos gate

This appendix collects simple but useful consequences of the exact construction.

C.1. All-order moment formula

Let ξ^{ex}_γ be defined by (19). For any $x_1, \dots, x_m \in U$,

$$\mathbb{E}\left[\prod_{r=1}^{m} \xi^{\mathrm{ex}}_\gamma(x_r) \right] = \exp\left(\gamma^2 \sum_{1 \le a < b \le m} C(x_a, x_b) \right).$$

The proof is the same Gaussian moment-generating calculation used in Theorem 5.4.

C.2. Effective one-parameter scaling

Writing $\tau = \gamma^2/\beta$, the exact gate can be rewritten as the Wick exponential of a Gaussian field $Y \sim \mathcal{N}(0, \tau G_U)$:

$$\xi_\gamma^{\mathrm{ex}}(x) \overset{d}{=} \exp\Big(Y(x) - \tfrac{1}{2}\mathrm{Var}(Y(x))\Big).$$

Hence the exact gate law depends on (γ, β) only through τ. This is the precise sense in which, once the operator and energy budget are fixed, the exact mechanism becomes an effectively one-parameter family.

C.3. Small-strength regime

Expanding the exact gate at small γ gives

$$\xi_\gamma^{\mathrm{ex}}(x) = 1 + \gamma\psi(x) + \frac{\gamma^2}{2}\big(\psi(x)^2 - C(x,x)\big) + O_{L^2}(\gamma^3).$$

Consequently,

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = 1 + \gamma^2 C(x,y) + O(\gamma^4), \qquad \mathrm{Cov}\big(\xi_\gamma^{\mathrm{ex}}(x), \xi_\gamma^{\mathrm{ex}}(y)\big) = \gamma^2 C(x,y) + O(\gamma^4).$$

This makes explicit that additive correlated Gaussian noise is a first-order approximation of the exact gate but does not preserve the positivity or higher-order structure of the multiplicative mechanism.

D. Why the Green kernel emerges in this design class

The purpose of this appendix is to state the precise structural lesson of the variational analysis. The Green kernel is not an extra hypothesis layered on top of the model. It appears because the design class is built from a *quadratic operator budget*, and the entropy maximizer in such a class always has covariance equal to the inverse of that operator.

General operator principle. Let $Q \succ 0$ be any symmetric positive definite operator on $\mathbb{R}^U$ and consider the admissible class $\mathcal{A}(Q, \varepsilon)$ from (10). By Theorem 5.1, the unique entropy maximizer is

$$\mathcal{N}\big(0, \tfrac{2\varepsilon}{|U|}Q^{-1}\big).$$

Hence the covariance kernel is forced to be proportional to Q^{-1}.

Dirichlet specialization. In the main text, the operator in the budget is the Dirichlet Laplacian $Q = L_U$, so the covariance becomes

$$\mathrm{Cov}(\psi) = \frac{2\varepsilon}{|U|}L_U^{-1} = \beta^{-1}G_U.$$

This is the exact sense in which the Dirichlet Green kernel is *forced*: it is the inverse operator corresponding to the chosen local smoothness budget.

Operator substitution principle. The same reasoning immediately yields a family of designed noises.

Corollary 1 (Replacing the operator replaces the kernel)**.** *Fix any SPD operator Q on $\mathbb{R}^U$ and define the quadratic budget*

$$\mathbb{E}\left[\frac{1}{2}\langle\psi, Q\psi\rangle\right] = \varepsilon.$$

Then the unique maximum-entropy log-field is Gaussian with covariance

$$\mathrm{Cov}(\psi) = \frac{2\varepsilon}{|U|} Q^{-1}.$$

After Wick normalization, the exact multiplicative gate has kernel

$$K_\gamma(x, y) = \exp\Big(\gamma^2 \frac{2\varepsilon}{|U|} Q^{-1}(x, y)\Big).$$

This corollary is useful conceptually. It shows that VKD is not tied to one operator or one architecture. Choosing $Q = L_U$ gives the massless Dirichlet construction of the main paper; choosing $Q = L_U + \mu I$ gives a massive variant with shorter-range correlations; choosing an anisotropic or graph-adapted operator would produce the corresponding inverse-kernel geometry. The core variational logic remains unchanged.

E. Topological stability of positive gates and fracture under hard masks

For a positive field $f : U \to (0, \infty)$ and a threshold $t > 0$, define the superlevel set

$$S_t(f) \coloneqq \{x \in U : f(x) \geq t\},$$

and view it as the induced subgraph of the underlying adjacency graph on U.

Proposition 1 (Threshold-band stability under positive multiplicative gating)**.** *Let $h : U \to (0, \infty)$, let $\xi : U \to (0, \infty)$, and assume $\|\log \xi\|_\infty \leq \eta$. Then for every $t > 0$,*

$$S_{te^\eta}(h) \subseteq S_t(\xi \odot h) \subseteq S_{te^{-\eta}}(h). \tag{60}$$

In particular, the superlevel topology of $\xi \odot h$ at level t can differ from that of h only through threshold events already present in the band $[te^{-\eta}, te^\eta]$.

Deep learning interpretation. The proposition says that a positive multiplicative gate does not tear the superlevel geometry apart arbitrarily. It only moves the effective threshold within a multiplicative band. For representation learning, this is a rigorous way to say that coherent regions can shift smoothly under GCh rather than being punctured by hard zeros.

Proof. From $\|\log \xi\|_\infty \leq \eta$ we have $e^{-\eta} \leq \xi(x) \leq e^\eta$ for every $x \in U$. If $h(x) \geq te^\eta$, then

$$\xi(x)h(x) \geq e^{-\eta}\, te^\eta = t,$$

so $x \in S_t(\xi \odot h)$. Conversely, if $\xi(x)h(x) \geq t$, then

$$h(x) \geq \frac{t}{\xi(x)} \geq te^{-\eta},$$

so $x \in S_{te^{-\eta}}(h)$. □

Proposition 2 (Sample-wise GCh obeys a random sandwich width)**.** *For the implemented gate ξ_γ^{sw} in (30),*

$$\|\log \xi_\gamma^{\mathrm{sw}}\|_\infty \leq |\gamma|\ \mathrm{osc}(\psi), \qquad \mathrm{osc}(\psi) \coloneqq \max_{x \in U} \psi(x) - \min_{x \in U} \psi(x). \tag{61}$$

Hence Proposition 1 applies with $\eta = |\gamma|\, \mathrm{osc}(\psi)$.

Proof. Write

$$\log \xi_\gamma^{\mathrm{sw}}(x) = \gamma\psi(x) - c(\psi), \qquad c(\psi) = \log\Big(\frac{1}{|U|}\sum_{y\in U} e^{\gamma\psi(y)}\Big).$$

Since the logarithm of an average of exponentials lies between the minimum and maximum exponent,

$$\min_{y\in U} \gamma\psi(y) \le c(\psi) \le \max_{y\in U} \gamma\psi(y).$$

Therefore each quantity $\gamma\psi(x) - c(\psi)$ lies in the interval

$$[-|\gamma|\,\mathrm{osc}(\psi),\ |\gamma|\,\mathrm{osc}(\psi)],$$

which is exactly (61). □

To contrast this with hard masking, recall that for any finite graph H the first Betti number equals the cycle rank

$$\beta_1(H) = |E(H)| - |V(H)| + \beta_0(H).$$

Theorem E.3 (Cycle-topology fracture under inverted dropout)**.** *Let the underlying graph be the cycle C_n, let $q \in (0,1]$, let $h \equiv c > 0$ on its vertices, and choose a threshold $t \in (0, c/q)$. Under inverted dropout,*

$$m_q = \frac{b}{q}, \qquad b(v) \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(q),$$

define $\widetilde{h} \coloneqq m_q \odot h$. Then $S_t(\widetilde{h})$ is exactly the induced subgraph on the kept vertices, and

$$\Pr\big(\beta_1(S_t(\widetilde{h})) = 1\big) = q^n, \qquad \Pr\big(\beta_1(S_t(\widetilde{h})) = 0\big) = 1 - q^n. \tag{62}$$

Equivalently, the loop topology is destroyed with probability $1 - q^n$.

Proof. Because $t < c/q$, a vertex belongs to $S_t(\widetilde{h})$ if and only if it is kept. Thus $S_t(\widetilde{h})$ is the induced subgraph on the kept vertices. If all n vertices are kept, this induced subgraph is the full cycle C_n, so $\beta_1 = 1$. If at least one vertex is dropped, the induced subgraph is a disjoint union of paths, hence acyclic and therefore has $\beta_1 = 0$. The all-kept event has probability q^n. □

Deep learning interpretation. Closed contours, ring-like activation patterns, and loop-shaped superlevel regions are idealized but meaningful models of semantic geometry. Theorem E.3 says that hard deletion is topologically brittle: a single dropped segment breaks the loop. This makes precise the intuition that binary masks can fracture spatial semantics rather than perturb them smoothly.

Remark 4 (Why this is relevant to DropBlock)**.** *DropBlock changes the spatial correlation of the zero set, not the basic hard-mask mechanism. Any block pattern that removes a connected arc from a loop-like superlevel set also destroys its cycle rank. The theorem above therefore isolates the essential topological failure mode already present in the binary-masking mechanism itself.*

F. Additional experimental details and results

F.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, we additionally use ImageNet-C (Hendrycks and Dietterich, 2019). Our main corruption-shift analysis reports averages over a selected subset of 7 corruption types, each averaged across severities 1–5. To complement the large-scale setting with a fast fine-grained pilot, we also evaluate on Oxford-IIIT Pet (Parkhi et al., 2012), a 37-class benchmark whose labels are sensitive to shape cues.

Architectures and injection sites. Our primary backbone is ResNet-50 (He et al., 2016). We inject the spatial gate at selected residual stages (L2/L3/L4) to study depth-dependent effects. ResNet-50 is also the fairest setting for comparisons to Dropout and DropBlock because those methods are naturally defined on convolutional feature grids. Since GCh acts on a 2D grid wherever such a representation exists, we further evaluate on Swin-T (Liu et al., 2021) to test transfer beyond the primary CNN regime.

Training protocols and reproducibility. **Main ImageNet protocol.** Unless otherwise specified, ImageNet models are trained from scratch for 270 epochs using SGD with momentum 0.9 and weight decay 10^{-4}, with learning-rate schedules held fixed across methods. Clean ImageNet metrics are reported on the standard validation set, and ImageNet-C metrics are computed from the corresponding trained checkpoints.
Controlled ablation protocol. For extensive multi-seed comparisons and strength sweeps, we also use a shorter matched-budget protocol described in the table captions. Within each controlled study, all hyperparameters aside from the noise mechanism are held fixed.
Oxford-IIIT Pets pilot. For the fine-grained pilot, we train a ResNet-18 from scratch for 40 epochs using Adam (lr $= 10^{-3}$), 224×224 inputs, and standard normalization. Results are reported as mean±std over 3 seeds.

Baselines. We compare GCh against Dropout (Srivastava et al., 2014), DropBlock (Ghiasi et al., 2018), additive i.i.d. Gaussian noise, and additive correlated Gaussian noise. The Gaussian baselines are energy-matched to GCh to separate the effect of structure from the effect of raw magnitude.

Metrics. We report Top-1 accuracy, negative log-likelihood (NLL), and expected calibration error (ECE). These metrics capture both predictive performance and probabilistic reliability.

F.2. Best vs. latest checkpoint on clean ImageNet

We compare late-stage (L4) injection at two evaluation points: the best checkpoint observed during training and the final checkpoint.

Protocol note. These tables come from the full-recipe single-run checkpoint protocol and are therefore complementary to, rather than numerically comparable with, the 3-seed controlled table in the main text. They summarize a separate evaluation slice of the same late-stage setting, whereas the main-text causal-control table reports the matched 3-seed protocol used for mechanism isolation. They are included to show that the late-stage reliability pattern is not specific to one checkpointing convention.

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 (clean), 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 (clean), latest checkpoint / final epoch, L4 injection.

Takeaway. Across both checkpoints, GCh improves reliability relative to both the no-noise baseline and DropBlock while remaining close to the baseline in Top-1 accuracy. The pattern is especially informative at the final epoch, where DropBlock shows a pronounced late-stage degradation whereas GCh does not.

F.3. Injection depth (L2/L3/L4)

We apply the same GCh mechanism at different residual stages under the controlled 3-seed protocol.

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.

Takeaway. Depth induces a clear trade-off: earlier injection favors Top-1 and NLL, while later injection gives the strongest ECE gains. This is why the main paper emphasizes the late-stage regime when discussing reliability.

F.4. Strength sensitivity

We sweep $\gamma \in \{0.03, 0.07, 0.10, 0.18, 0.27, 0.35\}$ at L4 injection under the controlled protocol.

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).

Takeaway. There is a robust small-to-moderate regime in which GCh preserves accuracy and improves reliability. Very large γ values cause the expected breakdown from excessive multiplicative perturbation.

F.5. ImageNet-C full results

Evaluation protocol and aggregation. We report Top-1 accuracy, NLL, and ECE on a selected 7-corruption subset of ImageNet-C. For each corruption type, metrics are averaged across severities 1–5; the reported aggregate numbers then average across the selected corruption types. All ImageNet-C metrics are computed from the same checkpoints used in the clean ImageNet tables, and we report mean±std over three seeds.

Reading the tables. Table 7 gives the main late-stage comparison at $g = 0.1$. Table 8 isolates the effect of injection depth under shift. Table 9 reports strength sensitivity under shift. Table 10 provides the corruption-wise breakdown.

Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.382±0.003	3.400±0.030	0.105±0.002
Dropout	0.1	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.

F.6. 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.

G. Additional theory details

G.1. Operational meaning of Theorem 5.4

Theorem 5.4 gives the canonical exact construction:

1. sample a GFF log-field ψ with covariance $(\beta L_U)^{-1}$;
2. exponentiate with exact Wick normalization to obtain a positive mean-one multiplicative gate.

Once the operator, gauge convention, and energy budget are fixed, the remaining reported strength parameter in the experiments is γ.

G.2. Mean-one normalization choices

The exact mean-one gate requires the variance map $v(x) = \mathrm{Var}(\psi(x)) = C(x, x)$. On a finite Dirichlet grid, $v(x)$ is not spatially constant. Two practical normalization choices are standard.

1. **Exact Wick normalization.** Precompute
$$v(i,j) = \frac{1}{\beta} \sum_{k=1}^{H} \sum_{\ell=1}^{W} \frac{\tilde{e}_{k,\ell}(i,j)^2}{\lambda_{k,\ell}},$$
where $\tilde{e}_{k,\ell}$ denotes the orthonormal sine basis. Then use
$$\xi_\gamma^{\mathrm{ex}}(x) = \exp\Big(\gamma\psi(x) - \frac{\gamma^2}{2} v(x)\Big).$$
This is the exact object in the theory and preserves $\mathbb{E}[\xi_\gamma^{\mathrm{ex}}(x)] = 1$ sitewise.
2. **Sample-wise mean-one normalization.** Compute $G(x) = \exp(\gamma\psi(x))$ and normalize by the spatial mean:
$$\xi_\gamma^{\mathrm{sw}}(x) = \frac{G(x)}{\frac{1}{|U|}\sum_{y\in U} G(y)}.$$
This guarantees unit spatial average per sample and is often convenient in optimization. It is the implementation used in the main experiments unless otherwise noted.

G.3. Implementation notes

1. A single gate may be shared across channels, or independent gates may be sampled channel-wise.
2. In multi-resolution architectures, the gate can be sampled directly at the feature resolution of the target layer or sampled at a base resolution and then resized.
3. At inference time, noise can be disabled by setting $\xi \equiv 1$.