Gaussian Chaos Noise: Variational Design of Noise Regularizers for Reliable Deep Learning

Ziran Liu
Shanghai Institute for Mathematics and Interdisciplinary Sciences,
Research Institute of Intelligent Complex Systems, Fudan University,
Shanghai 200433, China
zl1011@nyu.edu

Jinhao Wang
Dept. of Computer Science and Engineering, Santa Clara University,
Santa Clara, CA 95050
jwang11@scu.edu

Wei Wang
Futurewei Technologies
wei.wang@futurewei.com

Wei Jiang
Futurewei Technologies
wei.jiang@futurewei.com

Abstract

Noise regularizers inserted into deep networks are usually specified by templates such as dropout, spatial masks, hard masking, or additive noise and so on. We study how to design (according to some benefits that one needs) such a regularizer as a mechanism: what random field to sample, what spatial geometry it should follow, how it should be converted into a positive gate, where it should be inserted, and what relative structure of the representation it should preserve. Our framework, *Variational Kernel Design* (VKD), formulates these choices as a constrained variational problem over latent log-correlated fields. In a spatial quadratic instance, the optimizer is a Gaussian log-field with inverse-operator covariance; for a Dirichlet boundary grid operator this gives Green-kernel correlations and a positive mean-one gate, GCh, for which we call the Gaussian chaos noise. We then analyze the gate after deployment. We show that the sample-wise normalization yields exact Gaussian control of pairwise log-ratios, a margin-sensitive ranking law, and an intrinsic roughness budget, while hard masks destroy finite log-ratio geometry on positive evidence maps. Experiments on ImageNet, ImageNet-C, Swin-T, and Oxford-IIIT Pets show that GCh improves calibration and shift reliability at competitive accuracy, with diagnostics linking layer coherence, injection depth, and perturbation sensitivity.

1 Introduction

Noise regularizers are widely used to improve generalization, calibration, and robustness in deep networks. The usual choices–dropout, stochastic depth, spatial hard masks, token dropout, or additive noise–are useful, but they are typically specified as procedural templates [1, 23, 12, 7, 5, 15]. They answer how to inject randomness, but not how the geometry of that randomness should be chosen for the feature representation being perturbed.

This paper studies the design of a noise regularizer applied to the deep networks. We treat the regularizer as a complete mechanism: it samples a latent random field, assigns that field a spatial geometry, converts it into a positive gate, inserts the gate into a network layer, and should preserve

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

the relative evidence structure used by downstream computation. Our framework, *Variational Kernel Design* (VKD), turns these requirements into a variational problem over latent-field laws. The admissible class is determined by the needs of multiplicative feature perturbation: a latent log-field, centering, an operator budget for locality and coherence, a gauge convention, and a positive scale-preserving gate.

The resulting construction has two parts. The design part determines the latent geometry. In the spatial quadratic instance studied here, the optimizer has inverse-operator covariance; for the Dirichlet grid operator this gives Green-kernel correlations. The deployment part asks what the realized gate does to the representation. Later feature maps often behave as coherent evidence maps, where relative support between regions or tokens matters more than raw local variation. For such maps, pairwise log-ratios, ranking stability, and intrinsic roughness provide natural coordinates. We show that GCh gives finite, margin-aware deformations of these quantities, while hard binary masks can delete evidence and make finite relative log-geometry impossible.

The experiments follow this logic. Clean ImageNet isolates the mechanism and shows that correlation alone does not explain the calibration gain. Coherence diagnostics make the target representation variable measurable. Depth and strength sweeps identify when the mechanism is beneficial and when it fails. ImageNet-C provides the strongest evidence for reliability under corruption shift, and Swin-T and Oxford-IIIT Pets test whether the same design remains useful beyond the primary CNN setting.

Contributions. The main contributions are as follows.

1. **Variational design of a feature-noise regularizer.** We formulate a training-time noise regularizer as a mechanism whose latent law is determined by the requirements of multiplicative feature perturbation: log-field representation, centering, operator-level locality, gauge fixing, positive realization, and scale preservation.
2. **A solved spatial quadratic instance.** We solve the resulting variational problem for a spatial operator budget, obtaining an inverse-operator Gaussian log-field and, for the Dirichlet grid, Green-kernel perturbation geometry.
3. **A positive multiplicative noise structure.** We turn the latent field into GCh, a positive mean-one gate that combines spatial coherence, scale compatibility, and analyzable relative-evidence geometry.
4. **Representation-compatibility theory.** We prove pairwise log-ratio control, margin-sensitive ranking stability, and an intrinsic roughness budget for the deployed gate, and we show that hard masks fall outside finite pairwise log-ratio compatibility on positive evidence maps whenever they can zero a compared site.
5. **Reliability-focused empirical evidence.** We show that the mechanism primarily improves calibration and shift reliability at competitive accuracy. The experiments also reveal a stability pattern consistent with the compatibility theory: when hard masking is harmless, GCh is competitive, while representative multi-stage diagnostics show that GCh remains stable in settings where hard masks produce large swings in coherence. This can also been clearly seen in Figure 3d, as an evidence.

2 Main results and reading guide

The paper follows a simple chain: derive the geometry of a noise regularizer applied to feature maps, realize that geometry as a positive gate, deploy the gate inside a representation, and test whether the deployed perturbation preserves the relative structure predicted by the theory.

Contribution map. Table 1 summarizes where each component is established. The table is intended as a reading guide: the mathematical design, the deployed gate, the compatibility results, and the experimental evidence each play a distinct role.

Contribution	Main content	Location
VKD formulation	admissible laws induced by requirements of multiplicative feature noise	Sec. 3
Solved design instance	inverse-operator latent covariance; Green-kernel geometry for the Dirichlet grid	Sec. 3
Positive gate structure	spatially coherent, mean-one multiplicative realization	Secs. 3,4
Compatibility theory	finite log-ratios, margin-sensitive ranking, roughness budget; hard-mask boundary	Sec. 4
Empirical validation	mechanism isolation, regime diagnostics, depth/strength sensitivity, shift reliability, transfer	Sec. 5

Table 1: Main contributions and where they are established in the paper.

Design result. Let U be a finite perturbation domain and $Q \succ 0$ an operator encoding local smoothness. Section 3 defines an admissible class of centered latent log-field laws with a quadratic operator budget and finite entropy. The optimizer is

$$\psi \sim \mathcal{N}\left(0, \frac{2\varepsilon}{|U|} Q^{-1}\right),$$

with a KL entropy-gap identity certifying optimality inside this class. For $Q = L_U$, the Dirichlet Laplacian, the latent geometry is the Green kernel $G_U = L_U^{-1}$. The formulation is tied to the regularizer: the log-field, centering, operator budget, gauge convention, and positive realization are the mathematical form of the desired multiplicative feature perturbation.

Noise-structure advantage. The resulting gate is positive, spatially coherent, and multiplicative. Positivity avoids deletion and sign flips; inverse-operator geometry avoids independent sitewise fluctuations; multiplicative deployment respects feature scale and acts on relative evidence. These properties are not independent add-ons: the operator budget determines the latent covariance, and the realization map turns that latent field into a gate acting on feature maps.

Realization and deployed geometry. The exact variational gate is the Wick exponential

$$\xi_\gamma^{\mathrm{ex}}(x) = \exp\left(\gamma\psi(x) - \frac{\gamma^2}{2}\mathrm{Var}(\psi(x))\right).$$

The implementation used in the main experiments is the sample-wise mean-one realization. Its key property is that the sample-wise normalization cancels in pairwise log-ratios, giving exact Gaussian deformation of relative evidence. Section 4 also gives a Wick-samplewise deployed gate that combines local Wick correction with unit sample mean and preserves the exact pairwise geometry of the Wick gate.

Compatibility result. For a positive feature map h and the deployed sample-wise gate,

$$\log \frac{\widetilde{h}(x)}{\widetilde{h}(y)} - \log \frac{h(x)}{h(y)} = \gamma(\psi(x) - \psi(y)),$$

with variance determined by the Green effective resistance $R_G(x, y)$. This yields a margin-sensitive ranking law and an exact intrinsic roughness budget. A hard mask that can zero either compared site cannot preserve finite pairwise log-ratio geometry on positive fields; under inverted dropout, ranking preservation is margin-blind.

Empirical result. Clean ImageNet shows mechanism isolation: correlation alone does not reproduce the calibration gain. The selected ImageNet-C corruption slice gives the strongest current evidence for shift reliability, with improved ECE and NLL at competitive accuracy. Coherence diagnostics bridge theory and data by making intrinsic coherence measurable across layers and relating it to perturbation damage. Representative diagnostics further illustrate the structural advantage of the noise design: in harmless settings GCh is competitive with hard-mask baselines, while in a multi-stage Swin-T setting where hard masking produces large coherence swings, GCh remains comparatively stable (Appendix Fig. 12). Full protocols, ablations, and diagnostic figures are in the appendix.

3 Variational kernel design and Gaussian chaos noise

Variational design formulation. A noise regularizer applied to an intermediate feature map must specify four objects:

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

Here $\mathcal{F}$ is a family of latent-field laws on a perturbation domain U, K is the induced second-order geometry, ℓ converts a latent field into a positive gate, and $\mathcal{T}$ inserts that gate into the network. In this paper U is the $H \times W$ grid of the target feature map, and $\mathcal{T}$ is spatial multiplicative gating shared across channels.

The admissible class is determined by the intended use of the regularizer. We represent the perturbation by a latent log-field, center the field, constrain its spatial variation through a local operator budget, choose a gauge convention that makes the operator invertible, and realize the result as a positive mean-preserving gate. Thus the optimization variables are probability laws of latent fields, while the constraints encode the structure required before the perturbation can be used as a training-time gate.

Variational problem: mathematical formulation. Let $Q \succ 0$ be a symmetric positive definite operator on $\mathbb{R}^U$, let $n = |U|$, and let $\varepsilon > 0$. For a density p on $\mathbb{R}^U$, define $h(p) = -\int p(\psi) \log p(\psi) d\psi$ and

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

The resulting **mathematical problem** of the underlying variational design is to solve

$$\sup_{p \in \mathcal{A}(Q,\varepsilon)} h(p). \tag{2}$$

The operator Q encodes the local geometry on which latent perturbations may spend energy. Solving the design problem selects the inverse operator as the covariance geometry of the latent field. This inverse-operator language is familiar from Gaussian Markov random fields and Gaussian free fields, but here the operator is introduced by the requirements of a training-time multiplicative regularizer: it specifies the geometry of a perturbation that will later be realized and injected into a feature map [21, 22].

Theorem 3.1 (Quadratic VKD variational principle)**.** *Let $Q \succ 0$ be symmetric positive definite on $\mathbb{R}^U$, let $n = |U|$, and let $\varepsilon > 0$. The variational problem* (2) *has the unique optimizer*

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

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

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

Consequently, within this quadratic VKD class, the covariance geometry of the latent field is determined by the inverse operator Q^{-1}.

Corollary 3.2 (Dirichlet Green-kernel specialization)**.** *Taking $Q = L_U$, the Dirichlet graph Laplacian on the feature grid with auxiliary zero boundary, yields*

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

Thus the Dirichlet Green kernel is the latent geometry induced by the spatial quadratic design class.

From latent field to gates. Let $C = (\beta L_U)^{-1}$. The variational problem determines the latent field; the next step is to convert that field into a positive multiplicative gate. The exact Wick realization is

$$\xi^{\mathrm{ex}}_\gamma(x) = \exp\left(\gamma \psi(x) - \frac{\gamma^2}{2} C(x, x) \right).$$

Theorem 3.3 (Canonical exact GCh gate)**.** *For $\psi \sim \mathcal{N}(0, C)$, the gate ξ^{ex}_γ is positive and satisfies $\mathbb{E}[\xi^{\mathrm{ex}}_\gamma(x)] = 1$ for every x. Moreover, for any $x_1, \ldots, x_m \in U$,*

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

In particular, $\mathbb{E}[\xi^{\mathrm{ex}}_\gamma(x) \xi^{\mathrm{ex}}_\gamma(y)] = \exp(\gamma^2 C(x, y))$.

The sample-wise mean-one realization used in the experiments is

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

It is positive, preserves unit spatial average per sample, and admits exact pairwise log-ratio analysis. Section 4 also introduces a bridge realization that combines the local Wick correction with unit sample mean.

Target representation regime. The compatibility analysis is most relevant when feature maps behave as positive, coherent evidence maps: local regions or tokens carry semantic support, and downstream computation depends on relative comparisons among them. In this regime, the difference between smooth positive gating and hard deletion becomes geometric. The former changes evidence continuously in log-ratio coordinates; the latter may erase one side of a comparison.

4 Deployed compatibility and hard-mask mismatch

The design layer determines a latent geometry. The compatibility layer asks how a realized gate changes the feature map after it is inserted into the network. We analyze the sample-wise gate used in the reported experiments and a Wick-samplewise bridge realization that aligns a unit-mean deployed gate with the exact Wick geometry. Throughout this section, $h : U \to (0,\infty)$ denotes a positive feature map, ξ denotes a gate, and $\widetilde{h} = \xi \odot h$ denotes the perturbed map.

Proposition 4.1 (Wick-samplewise deployed gate)**.** *Let $v(x) = C(x,x)$ and define*

$$\xi_\gamma^{\mathrm{ws}}(x) = \frac{\exp(\gamma\psi(x) - \frac{\gamma^2}{2}v(x))}{|U|^{-1}\sum_{z\in U}\exp(\gamma\psi(z) - \frac{\gamma^2}{2}v(z))}.$$

Then $\xi_\gamma^{\mathrm{ws}}(x) > 0$ and $|U|^{-1}\sum_x \xi_\gamma^{\mathrm{ws}}(x) = 1$ almost surely.

Theorem 4.2 (Exact-deployed pairwise bridge)**.** *For all $x, y \in U$,*

$$\log\frac{\xi_\gamma^{\mathrm{ws}}(x)}{\xi_\gamma^{\mathrm{ws}}(y)} = \log\frac{\xi_\gamma^{\mathrm{ex}}(x)}{\xi_\gamma^{\mathrm{ex}}(y)} = \gamma(\psi(x)-\psi(y)) - \frac{\gamma^2}{2}(v(x)-v(y)).$$

Moreover, for all x, y in any deployed region $D \subseteq U$,

$$\left|\log\frac{\xi_\gamma^{\mathrm{sw}}(x)}{\xi_\gamma^{\mathrm{sw}}(y)} - \log\frac{\xi_\gamma^{\mathrm{ws}}(x)}{\xi_\gamma^{\mathrm{ws}}(y)}\right| \le \frac{\gamma^2}{2}\operatorname{osc}_D(v),$$

where $\operatorname{osc}_D(v) = \max_{z\in D} v(z) - \min_{z\in D} v(z)$.

This result makes the relation between the variationally derived exact gate and the unit-mean deployed realizations explicit. The sample-wise gate used in the experiments removes global sample scale. The Wick-samplewise gate additionally aligns pairwise geometry with the exact Wick realization. The difference between the two deployed gates is not an uncontrolled implementation artifact; it is fully determined by the diagonal variance profile of the latent field.

Theorem 4.3 (Pairwise log-ratio stability)**.** *Let $h : U \to (0,\infty)$ and let $\widetilde{h} = \xi_\gamma^{\mathrm{sw}} \odot h$. For every $x, y \in U$,*

$$\log\frac{\widetilde{h}(x)}{\widetilde{h}(y)} - \log\frac{h(x)}{h(y)} = \gamma(\psi(x)-\psi(y)).$$

Consequently the deformation is centered Gaussian with variance

$$\tau R_G(x,y), \qquad R_G(x,y) = G_U(x,x) + G_U(y,y) - 2G_U(x,y), \quad \tau = \gamma^2/\beta.$$

Corollary 4.4 (Margin-sensitive ranking)**.** *Assume $x \neq y$ and $\tau R_G(x,y) > 0$. If $h(x) > h(y) > 0$ and $\delta_{xy}(h) = \log h(x) - \log h(y)$, then*

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

Large relative margins are preserved with high probability, while weak comparisons remain sensitive to perturbation.

Corollary 4.5 (Intrinsic roughness budget). *Let $h : U \to (0, \infty)$ and let $\mathcal{E}_{\mathrm{int}}$ be the intrinsic interior graph energy. For $\widetilde{h} = \xi_\gamma^{\mathrm{sw}} \odot h$,*

$$\mathbb{E}[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h})] = \mathcal{E}_{\mathrm{int}}(\log h) + \gamma^2 \varepsilon_{\mathrm{int}}, \qquad \varepsilon_{\mathrm{int}} = \mathbb{E}[\mathcal{E}_{\mathrm{int}}(\psi)].$$

Thus GCh adds a finite and quantified deformation in log-geometry rather than puncturing the map with zeros.

Proposition 4.6 (Boundary for finite pairwise log-geometry). *Let $h : U \to (0, \infty)$ and let $\xi : U \to [0, \infty)$ be a multiplicative gate. For a fixed pair (x, y), the deformation*

$$\log \frac{\xi(x)h(x)}{\xi(y)h(y)} - \log \frac{h(x)}{h(y)}$$

is almost surely finite if and only if $\xi(x) > 0$ and $\xi(y) > 0$ almost surely.

Theorem 4.7 (Hard-mask mismatch). *Let $h : U \to (0, \infty)$ and let $m : U \to \{0, a\}$ be a binary mask. If $\Pr(m(x) = 0$ or $m(y) = 0) > 0$ for a compared pair (x, y), then pairwise log-ratio geometry on that pair is not almost surely finite. For independent inverted dropout $m_q(z) = b(z)/q$ with $b(z) \sim \mathrm{Bernoulli}(q)$ and $q \in (0, 1]$, if $x \neq y$ and $h(x) > h(y) > 0$, then*

$$\Pr((m_q \odot h)(x) > (m_q \odot h)(y)) = q,$$

independent of the margin. Moreover, for the same inverted-dropout mask,

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

so the additive amplitude-roughness inflation is proportional to mass-weighted intrinsic degree. On coherent high-mass maps with small intrinsic energy, this term becomes large relative to the original roughness.

Observable	Positive multiplicative gate	Hard binary mask
Pairwise log-ratio	finite controlled deformation	non-finite if a compared site is zeroed
Ranking	margin-sensitive preservation	margin-blind under inverted dropout
Intrinsic roughness	log-domain additive budget	amplitude-domain coherence-amplified inflation
Late coherent regime	smooth evidence deformation	deletion or fragmentation of evidence

Table 2: Compatibility summary on positive feature maps, where pairwise log-ratios encode relative evidence. The sample-wise gate yields the centered Gaussian deformation in Theorem 4.3; the Wick-samplewise gate aligns pairwise geometry with the exact Wick realization up to the deterministic diagonal term in Theorem 4.2.

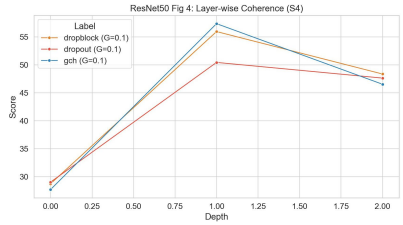
5 Experiments

The experiments test the claims developed above. Unless otherwise stated, GCh refers to the sample-wise mean-one realization used in training. We evaluate four questions: which ingredients matter beyond raw noise magnitude, whether coherence can be measured as a representation variable, where the gate is most useful in depth, and whether the effect transfers beyond the primary CNN setting.

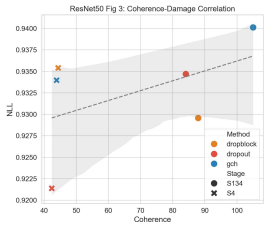
Mechanism isolation on clean ImageNet. Clean ImageNet isolates the mechanism rather than establishing a broad performance win. Table 3 shows that correlation alone is not sufficient: a correlated additive Gaussian baseline worsens ECE relative to no noise, while GCh substantially improves ECE at essentially unchanged accuracy. The calibration gain therefore depends on the positive mean-one multiplicative realization, not on spatial correlation alone.

Method	Top-1 ↑	NLL ↓	ECE ↓
None	0.765	0.931	0.030
Corr. Gaussian	0.765	0.944	0.037
GCh	0.764	0.934	**0.020**

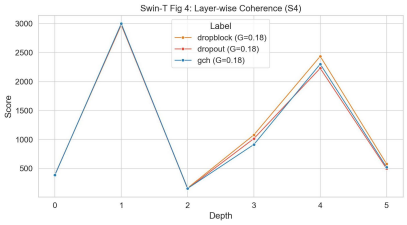
Table 3: Clean ImageNet mechanism isolation at late-stage injection. The gain is primarily calibration: positive mean-one multiplicative realization matters beyond correlation alone. Full multi-seed table is in Appendix A.



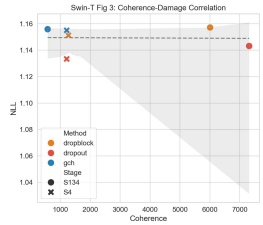
(a) ResNet, Stage 4



(b) ResNet coherence–damage

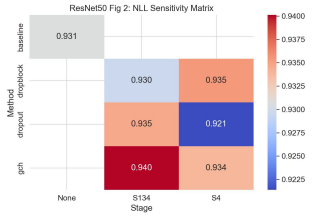


(c) Swin-T, Stage 4

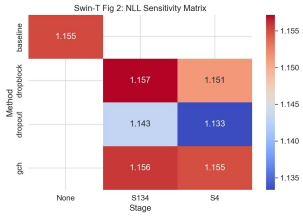


(d) Swin-T coherence–damage

Figure 1: Regime diagnostics. Layerwise coherence makes the target representation variable observable, while coherence–damage plots provide exploratory evidence that perturbation sensitivity depends on both representation geometry and deployed mechanism. Full Stage 134 diagnostics are in the appendix.



(a) ResNet-50



(b) Swin-T

Figure 2: NLL sensitivity across stages and perturbation configurations. This is a damage landscape, not a claim that any single depth is universally optimal. Full stage/gamma tables are in Appendix A.

Regime diagnostics. Before reporting shift metrics, we first check whether the representation quantities used by the theory can be measured in trained networks. Figure 1 shows intrinsic-coherence evolution and coherence–damage diagnostics. These plots are exploratory rather than causal: they make the target regime observable and show that perturbation damage varies with both representation geometry and the deployed perturbation. The point is not that coherence alone explains all damage; it is that coherence is a measurable variable that interacts with the noise mechanism.

Depth and strength sensitivity. Depth ablations show a clear trade-off. Early or middle injection can retain stronger accuracy or NLL, while late injection gives the strongest ECE. Thus the late-stage claim is a reliability claim, not a universal depth claim. Strength sweeps show the same pattern: moderate noise gives a usable operating range, while high strength leads to failure. Figure 2 summarizes NLL sensitivity across stages and perturbation configurations, while Table 4 records the clean depth trade-off and strength window.

Setting	Top-1 ↑	NLL ↓	ECE ↓
L2 early	0.767	**0.918**	0.031
L3 mid	0.765	0.925	0.029
L4 late	0.764	0.934	**0.020**
$\gamma = 0.03$	0.766	0.926	0.027
$\gamma = 0.10$	0.764	0.934	**0.020**
$\gamma = 0.35$	0.164	5.204	0.149

Table 4: Depth and strength summaries on clean ImageNet. Late injection is reliability-favored, while excessive strength leads to a failure mode. Full mean±std tables are in the appendix.

Corruption shift: ImageNet-C. The strongest empirical evidence is under corruption shift. On the selected seven-corruption slice of ImageNet-C, GCh improves both ECE and NLL at competitive accuracy, while correlated additive Gaussian does not reproduce the calibration gain. This is the primary benchmark support for the shift-reliability claim in the present experimental package; broader corruption coverage remains important for a fuller evaluation.

Method	Top-1 ↑	NLL ↓	ECE ↓
None	0.382	3.400	0.105
Corr. Gaussian	0.381	3.365	0.103
GCh	**0.383**	**3.287**	**0.056**

Table 5: Selected seven-corruption ImageNet-C summary, averaged over severities. This is the main shift-reliability evidence; full corruption-wise tables are in Appendix A.

Transfer evidence. On Swin-T, GCh improves Top-1 from 80.03% to 80.11%, NLL from 0.9213 to 0.9131, and ECE from 0.0762 to 0.0738 in the reported single-run setting. On Oxford-IIIT Pets, GCh gives the best NLL and ECE at competitive accuracy. We treat these as supporting transfer evidence rather than the empirical backbone of the paper. The appendix also includes representative multi-stage diagnostics. In the Swin-T multi-stage case, Appendix Fig. 12 shows that GCh remains stable in a configuration where hard masking produces large swings in coherence, while in milder configurations the methods are often close in aggregate performance.

Setting	Method	Top-1 ↑	NLL ↓	ECE ↓
Swin-T	None	80.03%	0.9213	0.0762
Swin-T	GCh	**80.11%**	**0.9131**	**0.0738**
Pets	None	0.9009	0.3669	0.0325
Pets	GCh	**0.9010**	**0.3627**	**0.0302**

Table 6: Transfer and secondary evidence. Swin-T is a single-run comparison; Pets is a 3-seed fine-grained evaluation summarized from the appendix.

Empirical synthesis. Taken together, the experiments support four conclusions. First, correlation alone does not explain the gains. Second, late-stage injection is best understood as reliability-favored, not universally best. Third, shift reliability is the strongest empirical signal. Fourth, coherence diagnostics provide a measurable bridge from the compatibility theory to observed perturbation damage, while remaining exploratory rather than causal. Table 7 summarizes the role of each evidence block.

Claim	Main evidence	Interpretation
Correlation alone is insufficient	clean ImageNet controls	positive multiplicative realization is needed for calibration gains
Late depth is reliability-favored	depth ablation and NLL matrix	late injection improves ECE most, while earlier depths may retain NLL/accuracy
Moderate strengths are usable	strength sweep	GCh has a stable operating range and high-strength failure mode
Shift reliability is the strongest evidence	selected ImageNet-C	ECE and NLL improve under corruption shift at competitive accuracy
Coherence is a measurable regime variable	coherence evolution and damage plots	diagnostics connect theory and experiments without asserting causality
Multi-stage stability	representative Swin-T diagnostic	GCh remains stable where hard masks show large coherence excursions

Table 7: Evidence map for the experimental section. The table clarifies what each empirical block supports and prevents overreading any single benchmark.

6 Discussion and limitations

The paper develops a noise regularizer for reliable deep learning through a solved spatial design instance. The variational mathematical structure determines a latent perturbation geometry; the realization fashion turns that field into a positive gate; the deployment step inserts the gate into feature maps; and the compatibility layer identifies which relative-geometry observables remain finite and margin-aware after perturbation. This mechanism point of view is the main message of the paper: **a noise regularizer should not be judged only by marginal variance or correlation structure, but by whether the deployed perturbation respects the representation geometry used by the network.**

The advantages of the proposed noise structure follow from this view. Green-kernel correlations encode spatial coherence, positivity avoids hard deletion and sign-flip artifacts, multiplicative deployment respects feature scale, and mean-one realization controls systematic gain shifts. More importantly, the deployed gates preserve pairwise relative evidence in log-geometry. These properties make the mechanism well matched to late coherent evidence maps, where downstream computation increasingly depends on relative support among semantic regions.

This also clarifies why the method should not be compared to hard masking only through aggregate accuracy. Hard masks and smooth positive gates can have similar nominal noise strength while acting on different geometric objects. A hard mask may be mean-preserving in amplitude, but on positive evidence maps it can make pairwise log-ratios infinite or undefined; a positive multiplicative gate keeps those quantities finite and admits margin-sensitive ranking laws. The central distinction is therefore not simply smooth versus nonsmooth noise, but whether the perturbation preserves the relative coordinates used by the representation.

Mechanism feature	Consequence for representations	Empirical role in the paper
Green-kernel geometry	coherent spatial perturbations rather than i.i.d. site noise	separates GCh from additive Gaussian controls
Positive gate	finite evidence ratios; no hard deletion or sign flip	matches the clean ImageNet mechanism-isolation result
Mean-one gate	controls global gain while leaving relative comparisons analyzable	supports training-time deployment
Late-stage use	acts where representations are more semantically aggregated and coherence diagnostics are meaningful	strongest ECE gains and shift-reliability signal
Hard-mask contrast	exposes the boundary where finite log-ratio geometry fails	distinguishes compatibility from generic regularization
Multi-stage stability	avoids large coherence swings in a representative case	supports deployment-side stability

Table 8: Mechanistic interpretation of the main claims. Each structural property has a representation-level consequence and a corresponding empirical role.

The empirical evidence is strongest for calibration and reliability under corruption shift. Clean accuracy is largely preserved but is not the primary win. The clean ImageNet controls show that correlated additive perturbations do not reproduce the ECE gain, while the positive multiplicative realization does. The ImageNet-C results provide the strongest benchmark evidence, because both ECE and NLL improve under corruption shift at competitive accuracy. The coherence diagnostics connect the target representation regime to observed perturbation damage and should be read as regime diagnostics rather than causal estimates. The Swin-T and Pets results provide supporting transfer evidence that the same mechanism is not confined to the primary ResNet setting. The representative multi-stage diagnostics further clarify the model advantage: GCh is competitive when hard masks behave harmlessly, and in the Swin-T multi-stage diagnostic it avoids the large coherence swings exhibited by hard masking.

For practical use, if the goal is calibration or robustness under shift, later semantic stages are a natural place to apply the gate; if the goal is clean accuracy or NLL alone, earlier or middle stages may remain preferable. The strength parameter should also stay in the moderate regime: the sweep shows both a usable window and a clear high-strength failure mode. This trade-off is part of the mechanism specification: GCh is a structured perturbation whose effect depends on where the representation has entered the coherent-evidence regime. The **scope of the main claim** is therefore deliberately

bounded: the paper establishes inverse-operator geometry in a solved spatial design class, finite relative-geometry results for positive coherent representations, and reliability gains on clean-isolation and selected shift settings. It does not claim universal optimality over all noise classes, arbitrary signed activations, or monotone depth superiority for every metric.

Broader shift coverage, multi-seed transformer studies, and architecture-adapted operators are natural extensions. Full proofs, ablations, diagnostic figures, and the mandatory checklist are provided in the appendix.

References

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

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Rémi Rhodes and Vincent Vargas. Gaussian multiplicative chaos and applications: A review. *Probability Surveys*, 11:315–392, 2014.

[21] Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005.

[22] Scott Sheffield. Gaussian free fields for mathematicians. *Probability Theory and Related Fields*, 139:521–541, 2007.

[23] 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.

[24] Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13888–13899, 2019.

[25] Deng-Bao Wang, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Min-Ling Zhang. On the pitfall of mixup for uncertainty calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7609–7618, 2023.

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

[27] 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.

[28] 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 Supplementary material

This appendix contains the material supporting the compressed main text: proofs of all theorem-style statements, extended framework details, complete experimental protocols and tables, supplementary theoretical consequences, diagnostic figures, and the checklist. The ordering mirrors the logic of the main paper: first proofs, then additional context, then experiments and diagnostics.

A.1 Proofs of the main-text results

The supplementary material is organized to support a first-pass reading of the main paper. We first collect the proofs of all theorem-style statements appearing in the main text. We then provide the extended VKD formulation, discrete GFF background, additional consequences of the exact gate, supplementary experimental protocols and tables, and diagnostic figures. The goal is to preserve the full technical record without interrupting the main-text argument.

A.1.1 Proof of Theorem 3.1

Let

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

and let $p^\star$ be the density of $\mathcal{N}(0, \Sigma_{Q,\varepsilon})$. Since $\Sigma_{Q,\varepsilon}^{-1} = (n/(2\varepsilon))Q$,

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

and the mean is zero, so $p^\star \in \mathcal{A}(Q,\varepsilon)$. For any $p \in \mathcal{A}(Q,\varepsilon)$,

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

Because

$$\log p^\star(\psi) = c - \frac{n}{4\varepsilon}\langle \psi, Q\psi\rangle$$

for a constant c, the second term is the same for all feasible p, since all feasible laws have the same energy budget. Evaluating it at $p^\star$ gives

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

Thus $h(p) \le h(p^\star)$, with equality iff $p = p^\star$ a.e. This proves uniqueness and the entropy-gap identity.

A.1.2 Proof of Corollary 3.2

Set $Q = L_U$ in Theorem 3.1. Then

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

This is the stated Dirichlet Green-kernel covariance.

A.1.3 Proof of Theorem 3.3

For each x,

$$\mathbb{E}[\xi_\gamma^{\mathrm{ex}}(x)] = \mathbb{E}\left[\exp\left(\gamma\psi(x) - \frac{\gamma^2}{2}C(x,x)\right)\right] = 1$$

by the moment-generating function of a centered Gaussian. For $x_1, \dots, x_m$, joint Gaussianity gives

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

Multiplying by the Wick correction

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

cancels the diagonal terms and leaves exactly the off-diagonal sum, proving the multipoint formula. The two-point kernel follows by taking $m = 2$.

A.1.4 Proof of Proposition 4.1

Positivity is immediate from the exponential form. The sample-mean-one identity follows by summing the definition:

$$\frac{1}{|U|}\sum_x \xi_\gamma^{\mathrm{ws}}(x) = \frac{|U|^{-1}\sum_x \exp(\gamma\psi(x) - \frac{\gamma^2}{2}v(x))}{|U|^{-1}\sum_z \exp(\gamma\psi(z) - \frac{\gamma^2}{2}v(z))} = 1.$$

A.1.5 Proof of Theorem 4.2

Let $D \subseteq U$ and $\mathrm{osc}_D(v) = \sup_{x,y\in D} |v(x) - v(y)|$. For the Wick-samplewise gate,

$$\log \xi_\gamma^{\mathrm{ws}}(x) = \gamma\psi(x) - \frac{\gamma^2}{2}v(x) - c_{\mathrm{ws}}(\psi),$$

where $c_{\mathrm{ws}}(\psi)$ is the logarithm of the sample average in the denominator. Subtracting the same identity at y removes $c_{\mathrm{ws}}(\psi)$, giving

$$\log \frac{\xi_\gamma^{\mathrm{ws}}(x)}{\xi_\gamma^{\mathrm{ws}}(y)} = \gamma(\psi(x) - \psi(y)) - \frac{\gamma^2}{2}(v(x) - v(y)).$$

The exact Wick gate has the same pairwise log-ratio by definition. The sample-wise mean-one gate without the Wick correction has pairwise log-ratio $\gamma(\psi(x) - \psi(y))$. Therefore the difference between the sample-wise mean-one and Wick-samplewise pairwise log-ratios is $(\gamma^2/2)(v(x) - v(y))$, whose absolute value is bounded by $(\gamma^2/2)\,\mathrm{osc}_D(v)$ on any deployed region D, where $\mathrm{osc}_D(v) = \max_{z\in D} v(z) - \min_{z\in D} v(z)$.

A.1.6 Proof of Theorem 4.3

Let $h : U \to (0, \infty)$. For the sample-wise gate,

$$\log \xi_\gamma^{\mathrm{sw}}(x) = \gamma\psi(x) - c_{\mathrm{sw}}(\psi),$$

where $c_{\mathrm{sw}}(\psi)$ is spatially constant. Hence the constant cancels in the difference:

$$\log \frac{\widetilde{h}(x)}{\widetilde{h}(y)} - \log \frac{h(x)}{h(y)} = \log \frac{\xi_\gamma^{\mathrm{sw}}(x)}{\xi_\gamma^{\mathrm{sw}}(y)} = \gamma(\psi(x) - \psi(y)).$$

Since ψ is Gaussian, this is centered Gaussian with variance

$$\gamma^2 \mathrm{Var}(\psi(x) - \psi(y)) = \gamma^2\beta^{-1}(G_U(x,x) + G_U(y,y) - 2G_U(x,y)) = \tau R_G(x,y).$$

A.1.7 Proof of Corollary 4.4

The event $\widetilde{h}(x) > \widetilde{h}(y)$ is equivalent to

$$\delta_{xy}(h) + \gamma(\psi(x) - \psi(y)) > 0.$$

By Theorem 4.3, the random term is $\mathcal{N}(0, \tau R_G(x,y))$. The assumption $\tau R_G(x,y) > 0$ makes the standard deviation nonzero, so the stated normal cdf formula follows.

A.1.8 Proof of Corollary 4.5

Let $h : U \to (0, \infty)$. In the log domain,

$$\log \widetilde{h} = \log h + \gamma\psi - c_{\mathrm{sw}}(\psi)\mathbf{1}.$$

The intrinsic Laplacian annihilates constants, so the constant term drops from $\mathcal{E}_{\mathrm{int}}$. Expanding the quadratic form,

$$\mathbb{E}[\mathcal{E}_{\mathrm{int}}(\log h + \gamma\psi)] = \mathcal{E}_{\mathrm{int}}(\log h) + \gamma^2\mathbb{E}[\mathcal{E}_{\mathrm{int}}(\psi)]$$

because $\mathbb{E}[\psi] = 0$, proving the result.

A.1.9 Proof of Proposition 4.6

Since $h(x), h(y) > 0$,

$$\log \frac{\xi(x)h(x)}{\xi(y)h(y)} - \log \frac{h(x)}{h(y)} = \log \xi(x) - \log \xi(y)$$

whenever $\xi(x), \xi(y) > 0$. Thus strict positivity on the pair implies finiteness. Conversely, if either gate value can be zero with positive probability, then the logarithm is either undefined or infinite with positive probability, so the pairwise log-ratio is not almost surely finite.

A.1.10 Proof of Theorem 4.7

If a binary mask zeros exactly one compared site, the log-ratio is $\pm\infty$; if it zeros both, the ratio is undefined. Thus finite pairwise log-geometry is not preserved whenever either site can be zeroed with positive probability. For independent inverted dropout $m_q(z) = b(z)/q$, with independent $b(z) \sim \text{Bernoulli}(q)$, if $x \neq y$ and $h(x) > h(y) > 0$, then the ordering is preserved iff $b(x) = 1$, independent of $b(y)$, hence the probability is q. For the intrinsic energy formula, for each interior edge $\{x, y\}$,

$$\mathbb{E}[(m_q(x)h(x) - m_q(y)h(y))^2] = (h(x) - h(y))^2 + \left(\frac{1}{q} - 1\right)(h(x)^2 + h(y)^2).$$

Multiplying by the edge weight and summing over intrinsic edges yields

$$\mathbb{E}[\mathcal{E}_{\text{int}}(m_q \odot h)] = \mathcal{E}_{\text{int}}(h) + \frac{1-q}{2q} \sum_x d_x^{\text{int}} h(x)^2.$$

This proves the theorem.

The formula is intentionally stated in amplitude geometry rather than log-geometry: after hard deletion, log-ratios may be non-finite, so amplitude roughness is the appropriate finite diagnostic for the corresponding distortion.

A.2 Appendix organization

The main text is deliberately compressed. To make the supplementary material easier to navigate, we avoid repeating the main-paper introduction and instead preserve only the extra material needed for verification. Section A.1 proves the theorem-style statements from the main text. Section A.3 records related-work details that would otherwise interrupt the main line. The remaining sections give the extended VKD construction, additional theoretical consequences, full experimental protocols and tables, and diagnostic figures.

A.3 Additional related work

Noise injection and structured regularization. Training with noise has long been connected to regularization [1]. Modern deep networks use a broad family of stochastic regularizers, including dropout [23], stochastic depth [12], LayerDrop [5], ShakeDrop [26], spatial occlusion methods such as Cutout and DropBlock [4, 7], and token/patch dropout in vision transformers [15]. These methods specify useful stochastic operations, but the perturbation geometry is typically procedural: it is chosen through a masking, dropping, or augmentation rule. VKD instead derives a latent geometry from an operator-level design constraint and then studies how the realized gate acts on representation geometry.

Calibration, augmentation, and reliability under shift. Modern networks can be accurate while miscalibrated [8], and uncertainty quality can degrade under distribution shift [18, 17]. Post-hoc calibration, Dirichlet calibration, ensembles, and Bayesian interpretations of dropout provide important baselines [13, 14, 6]. Data mixing and augmentation methods such as Mixup, CutMix, RandAugment, and AugMix also affect reliability and robustness [28, 27, 2, 11]; the calibration effect of Mixup itself has been studied and qualified in later work [24, 25]. The present work is not a post-hoc calibration method and not an input augmentation policy. It studies a training-time internal perturbation whose strongest empirical signal is reliability under corruption shift.

Gaussian fields and multiplicative gates. Gaussian random fields, sparse precision matrices, and Green-kernel covariance structures are standard objects in spatial statistics and probability [21, 22]. Gaussian multiplicative chaos provides a mathematical language for exponentiating Gaussian fields [20]. In the present paper these objects enter through a design problem for a complete noise mechanism whose admissible class is induced by the requirements of multiplicative training noise applied to feature maps. The realization and compatibility analyses are therefore part of the learned perturbation mechanism, not a separate probabilistic decoration.

Scope table. For completeness, Table 9 records the boundary of the main claims in tabular form.

Question	Established here	Outside the main claim
Noise geometry	mechanism-induced inverse-operator covariance	universal optimality over all noise classes
Compatibility	finite log-ratios, ranking law, roughness budget	arbitrary signed activations or all layers
Depth	strongest reliability signal in late stages	monotone superiority for every metric
Evidence	clean isolation and selected ImageNet-C shift gains	full ImageNet-C suite and multi-seed transformer coverage

Table 9: Scope of the main claims (appendix version). The table separates the contribution established in this paper from broader claims not made here.

A.4 Extended VKD framework and spatial construction

High-level view. We view a noise mechanism not as a member of a fixed menu of perturbations, but as a structured object to be *designed*. The central question is: given a learning context and a small set of desired properties, what stochastic mechanism should be constructed so that those properties hold by design?

Mechanism components. We parameterize a spatial noise mechanism by

$$\boxed{\mathsf{N} = \big(\mathcal{F},\, K,\, \ell,\, \mathcal{T}\big)}, \tag{3}$$

where the four components separate sampling, geometry, realization, and deployment:

(i) Distribution family $\mathcal{F}$ (what is sampled). $\mathcal{F}$ is a family of laws over latent fields on a domain Ω:

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

It specifies the latent random object before exponentiation or deployment.

(ii) Kernel K (how the latent field correlates). K is a positive semidefinite kernel on $\Omega \times \Omega$ encoding the intended second-order geometry: locality, smoothness, anisotropy, and scale. In the Gaussian designs studied here, K is the covariance induced by the inverse operator.

(iii) Realization map ℓ (how the latent field becomes a gate). The realization map sends a latent field to a positive gate, for example through an exponential link with either Wick normalization, sample-wise normalization, or the Wick-samplewise realization used to bridge the two.

(iv) Injection operator $\mathcal{T}$ (where and how the gate acts). $\mathcal{T}$ inserts the realized gate into the model. Given a feature tensor F and gate $\xi = \ell(\psi)$, it produces a perturbed tensor

$$\widetilde{F} = \mathcal{T}(F; \xi), \tag{4}$$

covering multiplicative gating, additive perturbation, channel-wise injection, block-wise injection, and multi-scale variants.

For the spatial multiplicative case used in this paper, if $F \in \mathbb{R}^{C\times H\times W}$ and $\xi \in (0,\infty)^{H\times W}$, then

$$(\widetilde{F})_{c,i,j} = F_{c,i,j}\, \xi_{i,j}, \qquad c = 1,\dots,C, \quad i = 1,\dots,H, \quad j = 1,\dots,W, \tag{5}$$

where the same realized spatial gate is broadcast across channels.

From desiderata to mechanism. VKD converts learning-level desiderata into a stochastic design. We begin with requirements such as minimal extra information, positivity, lack of systematic scale drift, locality, and smoothness. These are translated into mathematical constraints on admissible mechanisms, after which one solves for the mechanism consistent with those constraints. In this sense, VKD is a blueprint for deriving a noise mechanism from inductive bias rather than selecting one by convention.

How this paper instantiates VKD. In the spatial setting studied here, VKD yields a fully explicit design: a Gaussian log-field with Green-kernel correlations and a positive multiplicative gate obtained by exact Wick normalization. This construction is easy to sample on a rectangular grid and can be inserted as a structured training-time perturbation in convolutional or grid-based transformer pipelines.

A.5 Discrete GFF setup and sampling background

A.5.1 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

$$\mathcal{T}_\nu(h)(c,x) = h(c,x)\,\nu(x), \tag{6}$$

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

$$\mathcal{T}_\nu^{\mathrm{res}}(h)(c,x) = h(c,x)\Big(1+\alpha(\nu(x)-1)\Big), \qquad \alpha\in(0,1]. \tag{7}$$

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

A.5.2 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 $H,W\ge 1$ 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

$$(L_U\phi)(x) = \sum_{y:\{x,y\}\in E} c_{xy}\big(\phi(x) - \bar{\phi}(y)\big), \qquad x \in U. \tag{8}$$

Its quadratic form is the Dirichlet energy

$$\mathcal{E}(\phi) \coloneqq \frac{1}{2}\langle \phi, L_U\phi\rangle = \frac{1}{2}\sum_{\{x,y\}\in E} c_{xy}\big(\bar{\phi}(x) - \bar{\phi}(y)\big)^2. \tag{9}$$

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

$$\phi \sim \mathcal{N}\big(0, (\beta L_U)^{-1}\big). \tag{10}$$

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

$$p_\beta(\phi) = \frac{1}{Z_\beta}\exp\big(-\beta\,\mathcal{E}(\phi)\big) = \Big(\frac{\det(\beta L_U)}{(2\pi)^{|U|}}\Big)^{1/2}\exp\Big(-\tfrac{1}{2}\phi^\top(\beta L_U)\phi\Big), \tag{11}$$

with normalizing constant

$$Z_\beta = (2\pi)^{|U|/2}\det(\beta L_U)^{-1/2}. \tag{12}$$

Green kernel. Define the Dirichlet Green matrix

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

Then the covariance of the GFF is

$$\mathrm{Cov}\big(\phi(x), \phi(y)\big) = \beta^{-1}G_U(x,y), \qquad x, y \in U. \tag{13}$$

A.6 Extended derivation of Gaussian Chaos Noise

We now instantiate VKD for spatial multiplicative noise. The aim is to design a positive gate $\xi : U \to (0,\infty)$ that perturbs intermediate representations while encoding only the structure demanded by the learning problem. The main mathematical refinement in this section is to make the variational problem fully explicit: the optimization is performed over *laws of the log-field*, the operator-level smoothness constraint is stated as a quadratic budget, and positivity/mean preservation are imposed through the exponential link and Wick normalization.

A.6.1 Design desiderata

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:

$$\mathbb{E}\left[\frac{1}{2}\langle \psi, Q\psi\rangle\right] = \varepsilon, \tag{14}$$

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 $L_U \succ 0$.

Latent law versus gate realization. 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.

A.6.2 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

$$\mathcal{A}(Q,\varepsilon) := \left\{ p : \mathbb{R}^U \to [0,\infty) \;\middle|\; \begin{array}{l} \int_{\mathbb{R}^U} p(\psi)\,d\psi = 1, \\ \int_{\mathbb{R}^U} \psi\, p(\psi)\,d\psi = 0, \\ \int_{\mathbb{R}^U} \frac{1}{2}\langle \psi, Q\psi\rangle p(\psi)\,d\psi = \varepsilon, \\ h(p) > -\infty \end{array} \right\}, \tag{15}$$

where

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

is the differential entropy. The associated variational problem is

$$\sup_{p\in\mathcal{A}(Q,\varepsilon)} h(p). \tag{16}$$

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.

A.6.3 Quadratic VKD variational principle and operator-determined kernel geometry

The next theorem gives the exact solution of the VKD variational problem in (16). It identifies the optimizer, its entropy value, the explicit scale, and an entropy-gap identity that certifies uniqueness.

Theorem A.1 (Quadratic VKD variational principle)**.** *Let $Q \succ 0$ be symmetric positive definite on $\mathbb{R}^U$, let $n = |U|$, and let $\varepsilon > 0$. Then the variational problem* (16) *has a unique optimizer*

$$p^\star_{Q,\varepsilon} = \mathcal{N}\big(0, \Sigma_{Q,\varepsilon}\big), \qquad \Sigma_{Q,\varepsilon} = \frac{2\varepsilon}{n}\, Q^{-1}. \tag{17}$$

Equivalently,

$$p^\star_{Q,\varepsilon}(\psi) = \frac{1}{(2\pi)^{n/2}\det(\Sigma_{Q,\varepsilon})^{1/2}} \exp\Big(-\frac{1}{2}\psi^\top \Sigma^{-1}_{Q,\varepsilon}\psi\Big), \tag{18}$$

with precision matrix

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

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

$$h(p^\star_{Q,\varepsilon}) - h(p) = \mathrm{KL}\big(p \,\|\, p^\star_{Q,\varepsilon}\big) \geq 0, \tag{19}$$

so the optimizer is unique. Its entropy is

$$h(p^\star_{Q,\varepsilon}) = \frac{1}{2}\log\Big((2\pi e)^n \det\big(\tfrac{2\varepsilon}{n}Q^{-1}\big)\Big). \tag{20}$$

Proof sketch. Let $p^\star$ denote the Gaussian density in (17). 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 (19) 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 A.2 (Dirichlet specialization and Green-kernel specialization)**.** *Taking $Q = L_U$ in Theorem A.1 yields the unique entropy-optimal log-field*

$$\psi \sim \mathcal{N}\big(0, (\beta L_U)^{-1}\big), \qquad \beta = \frac{n}{2\varepsilon}. \tag{21}$$

Its covariance is

$$\mathrm{Cov}(\psi) = \frac{2\varepsilon}{n}\,L_U^{-1} = \beta^{-1}G_U, \qquad G_U \coloneqq L_U^{-1}. \tag{22}$$

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

Remark A.3 (What is and is not determined by the design class"). The theorem does *not* assert universal optimality of the Green kernel over all noise-design problems. It states a sharper design-level fact: once the admissible class is fixed by a local quadratic budget with operator Q, the optimizing latent law has covariance proportional to Q^{-1}. The Green kernel appears in this paper because the design operator is the Dirichlet Laplacian.

A.6.4 From the variationally derived log-field to the canonical exact gate

We now pass from the centered log-field to the positive multiplicative gate. Let

$$\psi \sim \mathcal{N}(0, C), \qquad C = (\beta L_U)^{-1} = \frac{2\varepsilon}{n}G_U. \tag{23}$$

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

$$\xi_\gamma^{\mathrm{ex}}(x) \coloneqq :\exp(\gamma\psi(x)):= \exp\Big(\gamma\psi(x) - \tfrac{\gamma^2}{2}C(x,x)\Big), \qquad x \in U. \tag{24}$$

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

Theorem A.4 (Canonical exact GCh gate)**.** *Under desiderata* D1–D5*, the canonical exact positive mean-one multiplicative gate is obtained by:*

1. *sampling the variationally derived 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* (24)*.*

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

$$\mathbb{E}\left[\prod_{r=1}^{m}\xi_\gamma^{\mathrm{ex}}(x_r)\right] = \exp\left(\gamma^2 \sum_{1\le a<b\le m} C(x_a,x_b)\right). \tag{25}$$

In particular,

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\big] = 1, \tag{26}$$

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = \exp\big(\gamma^2 C(x,y)\big). \tag{27}$$

Hence the induced second-order gate kernel is

$$K_\gamma(x,y) \coloneqq \mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = \exp\big(\gamma^2 C(x,y)\big). \tag{28}$$

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 (25). The one-point and two-point formulas are the cases $m = 1$ and $m = 2$. □

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

$$\tau := \frac{\gamma^2}{\beta} = \frac{2\varepsilon\,\gamma^2}{|U|}. \tag{29}$$

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

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

then

$$\xi_\gamma^{\mathrm{ex}}(x) \overset{d}{=} \exp\Big(Y(x) - \tfrac{1}{2}\mathrm{Var}(Y(x))\Big). \tag{30}$$

In particular,

$$K_\gamma(x, y) = \exp\big(\tau\, G_U(x, y)\big). \tag{31}$$

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 (31) follows from (28). □

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

$$\xi_\gamma^{\mathrm{ex}}(x) = 1 + \gamma\psi(x) + \frac{\gamma^2}{2}\Big(\psi(x)^2 - C(x, x)\Big) + O_{L^2}(\gamma^3) \qquad (\gamma \to 0). \tag{32}$$

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

$$\mathbb{E}\big[\xi_\gamma^{\mathrm{ex}}(x)\xi_\gamma^{\mathrm{ex}}(y)\big] = 1 + \gamma^2 C(x, y) + O(\gamma^4), \tag{33}$$

$$\mathrm{Cov}\big(\xi_\gamma^{\mathrm{ex}}(x), \xi_\gamma^{\mathrm{ex}}(y)\big) = \gamma^2 C(x, y) + O(\gamma^4). \tag{34}$$

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 (32). Equation (33) follows by expanding (28):

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

Since $\mathbb{E}[\xi_\gamma^{\mathrm{ex}}(x)] = 1$, subtracting one yields (34). □

Interpretation of the small-γ regime. Theorem A.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 variational gate and deployed sample-wise gate. The closed-form moment identities in Theorems A.4 to A.6 refer to the exact Wick-normalized gate (24). This is the variationally derived object associated with the variationally derived log-field. In practice, unless otherwise stated, our experiments use the sample-wise mean-one realization

$$\xi_\gamma^{\mathrm{sw}}(x) = \frac{\exp(\gamma\psi(x))}{\frac{1}{|U|}\sum_{y\in U}\exp(\gamma\psi(y))}, \tag{35}$$

which is also positive and satisfies

$$\frac{1}{|U|}\sum_{x\in U}\xi_\gamma^{\mathrm{sw}}(x) = 1 \qquad \text{almost surely.} \tag{36}$$

The two realizations should not be conflated: the exact gate preserves sitewise Wick moments, while the sample-wise gate enforces a per-sample spatial mean. The reason the deployed gate can nevertheless be analyzed sharply is that its normalization is a spatial constant in the log domain. It cancels from pairwise log-ratios, so the compatibility observables below can be studied exactly for the mechanism used in training. Appendix A.16.2 compares the two normalizations in more detail.

A.6.5 Representation-compatibility theory: relative geometry under deployed noise

The design layer determines a latent field and a variationally derived exact gate. The question in this subsection is different: once a gate is actually deployed inside a network, what part of the representation geometry does it preserve? We answer this question for positive evidence maps using observables that are intrinsic to relative geometry: pairwise log-ratios, ranking stability, and intrinsic roughness.

Choice of observables. Pairwise log-ratios are the primitive coordinates of relative evidence: they record how much more strongly one region or token is supported than another. Ranking stability is the decision-level consequence of perturbing those coordinates. Intrinsic roughness aggregates the effect over the whole spatial graph after removing global scale. These observables are therefore not chosen merely because they make the algebra convenient; they are the quantities through which a positive semantic map expresses relative evidence, spatial coherence, and downstream comparisons.

Deployed-gate payoff. The sample-wise gate used in training admits a sharp analysis because its normalization is a spatial constant in the log domain. As a result, it cancels from pairwise log-ratios. The next theorem is therefore not a statement about an idealized object detached from the experiments; it is the primitive compatibility statement for the deployed mechanism itself. The hard-mask results then evaluate binary masking on the same observables, which is what makes the comparison structural rather than purely empirical.

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 A.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 (35)*. For every $x, y \in U$,*

$$\Delta_{xy}^{\mathrm{sw}}(h) := \log\frac{\widetilde{h}(x)}{\widetilde{h}(y)} - \log\frac{h(x)}{h(y)} = \gamma\big(\psi(x)-\psi(y)\big). \tag{37}$$

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

$$\mathrm{Var}\big(\Delta_{xy}^{\mathrm{sw}}(h)\big) = \gamma^2\big(C(x,x)+C(y,y)-2C(x,y)\big) = \tau R_G(x,y), \tag{38}$$

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

$$R_G(x,y) \coloneqq G_U(x,x) + G_U(y,y) - 2G_U(x,y). \tag{39}$$

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

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

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 (37). 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. □

Consequence for representation geometry. Theorem A.7 is the basic bridge from the deployed gate to representation geometry. It says that the effect of the sample-wise gate on relative evidence is a centered Gaussian difference field whose variance is determined by the Green geometry. In particular, the per-sample normalization does not introduce an uncontrolled distortion of relative comparisons; it disappears exactly from the pairwise log domain.

Corollary A.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*

$$\delta_{xy}(h) \coloneqq \log h(x) - \log h(y) > 0. \tag{40}$$

Then under the implemented sample-wise gate,

$$\Pr\big(\widetilde{h}(x) > \widetilde{h}(y)\big) = \Phi\left(\frac{\delta_{xy}(h)}{\sqrt{\tau R_G(x,y)}}\right), \tag{41}$$

where Φ is the standard Gaussian cdf. Equivalently,

$$\Pr\big(\widetilde{h}(x) \leq \widetilde{h}(y)\big) = \Phi\left(-\frac{\delta_{xy}(h)}{\sqrt{\tau R_G(x,y)}}\right) \leq \exp\left(-\frac{\delta_{xy}(h)^2}{2\tau R_G(x,y)}\right). \tag{42}$$

Proof. By Theorem A.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 (41). The tail bound follows from the standard Gaussian bound $\Phi(-u) \leq e^{-u^2/2}$ for $u > 0$. □

Consequence for representation geometry. Theorem A.8 turns the pairwise geometry theorem into a decision-level statement. If a representation already separates two regions by a large log-margin, the deployed gate preserves that ordering with high probability; if the margin is small, the perturbation is allowed to express uncertainty. This is exactly the reliability behavior targeted by late-stage training-time perturbation: strong relative evidence should be stable, while weak comparisons may remain uncertain. 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

$$\mathcal{E}_{\mathrm{int}}(f) \coloneqq \frac{1}{2} \sum_{\{x,y\} \in E_{\mathrm{int}}} c_{xy}\big(f(x) - f(y)\big)^2 = \frac{1}{2}\langle f, L_{\mathrm{int}} f\rangle, \tag{43}$$

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 A.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*

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h})\big] = \mathcal{E}_{\mathrm{int}}(\log h) + \gamma^2 \varepsilon_{\mathrm{int}}, \qquad \varepsilon_{\mathrm{int}} \coloneqq \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\psi)\big] = \tfrac{1}{2}\mathrm{Tr}(L_{\mathrm{int}} C). \tag{44}$$

Proof. From the proof of Theorem A.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,

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

□

Consequence for representation geometry. Theorem A.9 is the global counterpart of the pairwise theorem. It shows that the deployed gate does not merely preserve individual comparisons in distribution; it also adds a finite and exactly quantified amount of intrinsic log-roughness to the entire positive map. This is the aggregate sense in which GCh behaves as controlled stochastic deformation rather than discontinuous semantic deletion.

Corollary A.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 $x, y \in U$,*

$$\Delta_{xy}^{\mathrm{sw}}(ah) = \Delta_{xy}^{\mathrm{sw}}(h), \tag{45}$$

and

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h}_a)\big] - \mathcal{E}_{\mathrm{int}}(\log(ah)) = \gamma^2 \varepsilon_{\mathrm{int}}. \tag{46}$$

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 A.7 and the intrinsic energy $\mathcal{E}_{\mathrm{int}}$ are invariant under such constants, which yields (45) and (46). □

Interpretation for representation learning. 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 A.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 $\widetilde{h} = \xi_\gamma^{\mathrm{sw}} \odot h$,*

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(\log \widetilde{h})\big] = \gamma^2 \varepsilon_{\mathrm{int}}. \tag{47}$$

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 $\mathcal{E}_{\mathrm{int}}(\log h) = 0$. The claim follows immediately from Theorem A.9. □

Interpretation for representation learning. 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. Theorem A.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.

We now evaluate hard binary masks on the same relative-geometry observables. This is important: the comparison is not between a smooth method and a hard method under different criteria, but between two deployed perturbations measured by the same positive-field geometry. The next theorem states the boundary case. Whenever a mask can zero one member of a compared pair with positive probability, finite pairwise log-ratio geometry is no longer available.

Theorem A.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*

$$\Pr\big(m(x) = 0 \text{ or } m(y) = 0\big) > 0, \tag{48}$$

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 A.7 can hold for such a mask. For inverted dropout at distinct *compared sites $x \neq y$,*

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

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 (48) 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 A.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*

$$\Pr\big((m_q \odot h)(x) > (m_q \odot h)(y)\big) = q. \tag{49}$$

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 $(m_q \odot h)(x) = 0 \leq (m_q \odot h)(y)$. Therefore the ordering is preserved if and only if $b(x) = 1$, which occurs with probability q. □

Consequence for representation geometry. Theorem A.13 isolates the decision-level failure mode of inverted dropout on positive evidence maps. Even if one site has an arbitrarily larger semantic margin than another, the ordering is preserved only with probability q. The probability does not improve with the margin. This contrasts directly with Theorem A.8, where the preservation probability increases with the log-margin.

Proposition A.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 $h : U \to \mathbb{R}$,*

$$\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big] = \mathcal{E}_{\mathrm{int}}(h) + \frac{1-q}{2q} \sum_{x \in U} d_x^{\mathrm{int}} h(x)^2, \tag{50}$$

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 (50). □

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

$$\kappa(h) := \frac{\sum_{x \in U} d_x^{\mathrm{int}} h(x)^2}{2\mathcal{E}_{\mathrm{int}}(h)}. \tag{51}$$

Then inverted dropout satisfies

$$\frac{\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big]}{\mathcal{E}_{\mathrm{int}}(h)} = 1 + \frac{1-q}{q}\,\kappa(h). \tag{52}$$

Proof. Divide both sides of (50) by $\mathcal{E}_{\mathrm{int}}(h) > 0$ and rearrange. □

Consequence for representation geometry. The scalar $\kappa(h)$ is a mismatch factor for coherent maps: it is large when a feature map carries nontrivial activation mass but varies only weakly across space. Theorem A.15 therefore identifies the regime in which hard deletion becomes disproportionately damaging in relative geometric terms. This is the theorem-level reason the experiments later track intrinsic coherence as a diagnostic variable.

Corollary A.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*

$$\mathcal{E}_{\mathrm{int}}(h) = 0, \qquad \mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q \odot h)\big] = \frac{1-q}{2q} c^2 \sum_{x \in U} d_x^{\mathrm{int}} > 0. \tag{53}$$

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 Theorem A.14. □

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

$$\inf_{\ell\geq 1}\sum_{x\in U} d_x^{\mathrm{int}} h_\ell(x)^2 > 0, \qquad \mathcal{E}_{\mathrm{int}}(h_\ell) > 0 \text{ for every } \ell, \qquad \mathcal{E}_{\mathrm{int}}(h_\ell)\to 0. \tag{54}$$

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

$$\frac{\mathbb{E}\big[\mathcal{E}_{\mathrm{int}}(m_q\odot h_\ell)\big]-\mathcal{E}_{\mathrm{int}}(h_\ell)}{\mathcal{E}_{\mathrm{int}}(h_\ell)} \longrightarrow \infty. \tag{55}$$

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

Proof. By Theorem A.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 A.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

$$\frac{\delta_\ell}{\sqrt{\tau R_G(x,y)}} \longrightarrow \infty, \tag{56}$$

then under the implemented sample-wise GCh gate,

$$\Pr\big(\widetilde{h}_\ell(x) > \widetilde{h}_\ell(y)\big) \longrightarrow 1. \tag{57}$$

Under inverted dropout with keep probability q, however,

$$\Pr\big((m_q\odot h_\ell)(x) > (m_q\odot h_\ell)(y)\big) = q \qquad \text{for every } \ell. \tag{58}$$

Proof. Equation (57) follows immediately from Theorem A.8 and the assumption (56). Equation (58) is exactly Theorem A.13. □

Interpretation for representation learning. Theorem A.18 is a margin-level statement about the late-layer representation regime considered in the paper. 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 A.19 (Representation-compatibility dichotomy)**.** *Under the hypotheses of Theorems A.7, A.9, A.12 and A.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* (54)*.*

The assumptions in Theorem A.17 abstract the late-semantic regime targeted by the paper: the representation retains nontrivial mass but becomes increasingly low-frequency or spatially coherent. In that regime, hard masking becomes more and more mismatched because its relative roughness inflation is measured against a vanishing intrinsic baseline. By contrast, Theorems A.7 to A.9, A.18 and A.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.

Takeaway. If a layer encodes positive region-level evidence or token-level saliency, then the relevant question is not merely whether noise is mean-preserving in expectation, but whether it preserves the relative comparisons that downstream computation 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 A.13: 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.

Scope of the compatibility results. 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.

A.6.6 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:

$$\widetilde{F}_c(x) = F_c(x)\,\xi_\gamma(x), \qquad x \in U. \tag{59}$$

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:

$$e_{k,\ell}(i,j) = \sin\Big(\frac{\pi k i}{H+1}\Big)\,\sin\Big(\frac{\pi \ell j}{W+1}\Big), \tag{60}$$

$$\lambda_{k,\ell} = 4\sin^2\Big(\frac{\pi k}{2(H+1)}\Big) + 4\sin^2\Big(\frac{\pi \ell}{2(W+1)}\Big), \tag{61}$$

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 (61); 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 $\widetilde{F}$

A.7 Additional experimental protocols and full results

The experiments are organized to test the mechanism predicted by the theory rather than to serve as an unrelated benchmark suite. The empirical program has four roles: isolate which ingredients matter beyond raw noise magnitude, check whether the coherence-related regime targeted by the compatibility theory is measurable in trained networks, characterize the depth/strength trade-off, and test whether the effect transfers beyond the primary CNN setting. Detailed protocols are provided in Appendix A.7.

A.7.1 Experimental goals and setup

Our primary backbone is ResNet-50 trained on ImageNet-1k. The main controlled comparisons use three seeds and hold the training recipe fixed while varying the perturbation mechanism. We report Top-1 accuracy, negative log-likelihood (NLL), and expected calibration error (ECE), separating predictive performance from probabilistic reliability. The clean-data experiments are used mainly for mechanism isolation and calibration; the corruption-shift experiments are the strongest evidence for reliability under distribution shift.

A.7.2 Mechanism isolation on clean ImageNet

We first ask whether the observed gains can be explained by raw noise magnitude or spatial correlation alone. To isolate the mechanism, we compare GCh with Dropout, DropBlock, i.i.d. additive Gaussian noise, and correlated additive Gaussian noise under matched controlled settings. For compactness, we use a unified strength knob g: for GCh, $g = \gamma$; for Gaussian baselines, $g = \sigma$; for Dropout and DropBlock, $g = p$; and for the no-noise baseline, $g = 0$.

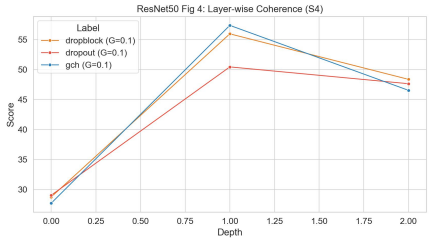
Method	g	Top-1↑	NLL↓	ECE↓
None	0	0.765±0.001	0.931±0.004	0.030±0.001
Dropout	0.1	0.764±0.001	0.942±0.005	0.033±0.001
DropBlock	0.1	0.765±0.000	0.930±0.002	0.032±0.000
IID Gauss.	0.1	0.765±0.001	0.930±0.005	0.032±0.002
Cor. Gauss.	0.1	0.765±0.000	0.944±0.002	0.037±0.001
GCh (ours)	0.1	0.764±0.001	0.934±0.004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

Table 10: 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.

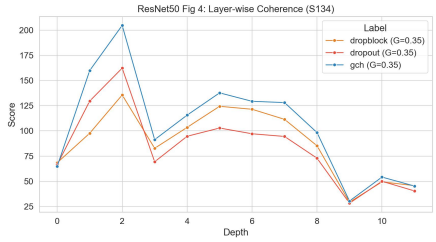
Interpretation. Table 10 is best understood as a mechanism-isolation result. On clean ImageNet, GCh mainly improves calibration rather than accuracy or NLL: the Top-1 accuracy remains essentially unchanged, while ECE drops substantially. The correlated additive Gaussian baseline does not reproduce this effect and can even worsen ECE. Thus the clean-data evidence supports a specific conclusion: correlation alone is not sufficient; the benefit appears when spatial correlation is coupled with a positive, mean-one multiplicative realization.

A.7.3 Regime diagnostics: coherence and perturbation damage

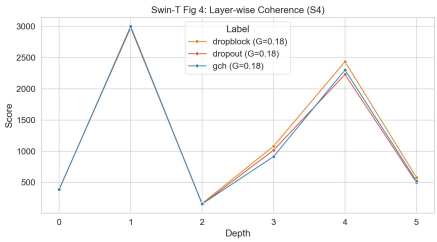
The compatibility theory is conditional on a target regime: positive, coherent, late-semantic representations. Before turning to benchmark metrics, we therefore ask whether geometry-related diagnostics are measurable in trained networks. The diagnostics below are not used as causal estimates; they operationalize the regime hypothesis and connect the paper's observables to network behavior.



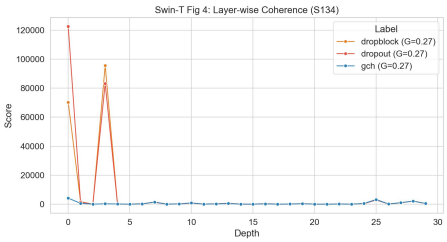
(a) ResNet-50, stage-4 injection.



(b) ResNet-50, multi-stage injection.

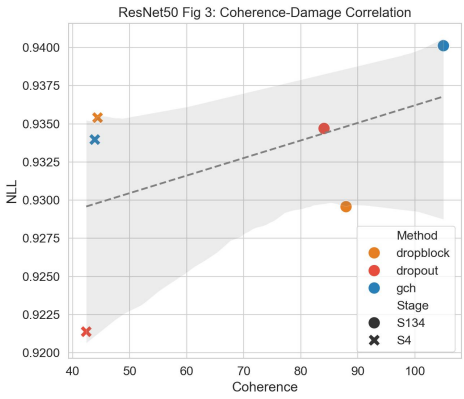


(c) Swin-T, stage-4 injection.

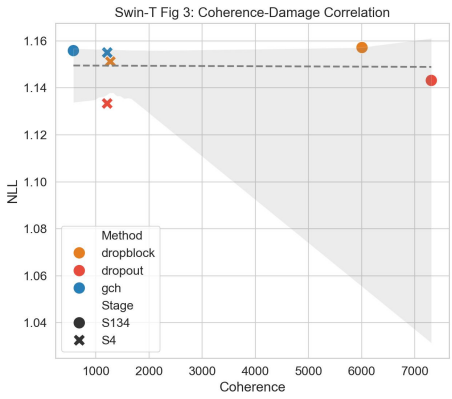


(d) Swin-T, multi-stage injection.

Figure 3: Layerwise intrinsic-coherence diagnostics. Intrinsic-coherence statistics across depth and perturbation settings for ResNet-50 and Swin-T. The figure operationalizes one component of the target regime in the compatibility theory: coherence is treated as a measurable stage-dependent variable, not as a direct performance metric or causal explanation.



(a) ResNet-50.



(b) Swin-T.

Figure 4: Exploratory relation between intrinsic coherence and perturbation-induced NLL damage. Each point corresponds to an evaluated configuration. The figure is a diagnostic consistency check rather than a causal claim: coherence helps expose the target regime, while damage also depends on the deployed perturbation mechanism and stage.

Interpretation. Figure 3 begins to make the target regime observable: intrinsic coherence varies with depth, architecture, and perturbation setting. Figure 4 further suggests that coherence is relevant to perturbation damage, but not sufficient by itself to explain the outcome. This is consistent with the central claim of the paper: representation compatibility depends jointly on the representation regime and on the deployed realization of the perturbation.

A.7.4 Depth and strength sensitivity

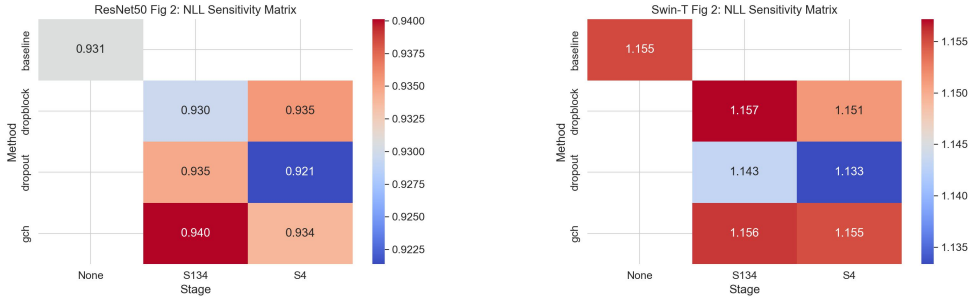
The theory predicts that the distinction between smooth positive gating and hard deletion should be most visible when representations are coherent and semantically sharp. We therefore study injection depth and perturbation strength separately. These ablations identify a reliability-favored operating regime; they do not claim that later or stronger injection is universally better.

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

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



(a) ResNet-50.

(b) Swin-T.

Figure 5: NLL sensitivity across perturbation stages and mechanisms. The heatmaps summarize how NLL responds to injection stage and perturbation choice. They are depth-sensitivity diagnostics rather than the primary aggregate performance table; absolute values should be compared within the same panel and protocol.

Interpretation. The depth ablation shows a reliability–performance trade-off: early or middle injection can preserve stronger Top-1/NLL, while late injection gives the strongest ECE gain. The strength sweep shows a usable moderate range and a clear failure mode at overly large strengths. Thus the empirical conclusion is not that deeper or stronger injection is always better, but that GCh has a reliability-favored operating regime that can be identified through depth and strength diagnostics.

A.7.5 Corruption shift: reliability on ImageNet-C

The strongest empirical evidence for the paper is under corruption shift. We evaluate on a selected subset of seven ImageNet-C corruption types, each averaged across severities 1–5. This is not the full ImageNet-C suite, and we state this explicitly. Within this evaluated shift slice, the question is whether the mechanism improves reliability while maintaining competitive accuracy.

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.

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

Main robustness takeaway. Table 13 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 14: 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 14 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 15: 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 15 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-wise breakdown.

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

Which corruptions benefit most. Table 16 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.

Interpretation. The ImageNet-C results provide the clearest empirical spine of the paper. GCh substantially reduces ECE and improves NLL relative to the no-noise baseline at competitive accuracy. The correlated additive Gaussian baseline remains close to the no-noise baseline in ECE, reinforcing the mechanism-isolation message from clean ImageNet: spatial correlation by itself is not sufficient. The stage-wise table shows a reliability-targeted depth trade-off: late-stage injection gives the strongest calibration gain, while mid-stage injection retains stronger Top-1/NLL on this selected corruption slice. The strength sweep identifies a moderate operating range and a clear collapse regime at excessive perturbation strength. Finally, the corruption-wise breakdown shows that reliability gains are strongest for noise-type corruptions and compression/pixelation, and weaker for defocus blur; this heterogeneity is informative and argues against presenting the method as uniformly robust to every shift type.

A.7.6 Transfer beyond the primary CNN setting

We next ask whether the effect is specific to the primary ResNet-50 setting. The Swin-T result is reported as preliminary transfer evidence because it is a single-run full-recipe comparison. The Oxford-IIIT Pets experiment, reported in Appendix A.15, provides an additional fine-grained test where coherent part structure is relevant.

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

Table 17: Swin-T transfer evidence. Full-recipe training, best checkpoint, single run. The result is reported as preliminary architecture-transfer evidence rather than as a multi-seed aggregate.

Interpretation. The Swin-T and Pets results suggest that the mechanism is not restricted to the primary CNN setting, but they play a secondary role. The current strongest evidence remains the controlled ResNet-50 and ImageNet-C package. The transfer results are useful because they show directionally consistent reliability gains beyond the main architecture and dataset, but they should not be overstated until multi-seed transformer experiments are added.

A.7.7 Empirical synthesis

The empirical results support the paper at four distinct levels. First, the clean ImageNet controlled study isolates the mechanism: correlation alone does not explain the calibration gain; the positive mean-one multiplicative realization is essential. Second, the coherence diagnostics begin to make the target representation regime measurable. They do not prove a causal law, but they show that the geometric variables used by the theory are observable in trained networks and relevant to perturbation damage. Third, the depth and strength ablations identify a reliability-favored operating regime rather than a monotone law that later or stronger injection is always better. Fourth, the selected ImageNet-C evaluation gives the strongest reliability evidence: GCh improves both ECE and NLL under shift at competitive accuracy. Together, these results support the paper's central claim at the right level of strength: GCh is a principled training-time noise regularizer whose benefits are most visible when reliability, calibration, and compatibility with coherent positive representations are the target.

A.8 Extended concluding remarks

We introduced a design-oriented view of noise injection in which the mechanism itself is derived from learning-relevant constraints rather than selected from a fixed heuristic menu. In the resulting VKD framework, a spatial noise mechanism is specified by its distribution family, correlation kernel, and injection operator. On the design side, we formulated the log-field construction problem as a finite-dimensional variational problem over admissible latent laws shaped by the intended multiplicative perturbation: centering, a quadratic operator budget, gauge fixing, and later positive mean-preserving realization. Solving this problem yields a uniquely selected Gaussian log-field whose covariance is the inverse of the chosen operator. 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

secondary evaluation. 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 two-step recipe for future work: first choose the operator that encodes the geometry one wants the noise to respect; then ask whether the resulting implemented mechanism is actually compatible with the representations a deep network uses. That perspective opens the door to principled variants based on massive, anisotropic, graph-adapted, or architecture-specific operators while preserving the same mathematical blueprint.

A.9 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.
- 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.

A.10 Full variational derivation for Theorem A.1

This appendix gives a fuller proof of the quadratic VKD variational 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

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

A.10.1 Entropy-gap proof of optimality and uniqueness

Define

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

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,

$$\begin{aligned} \mathrm{KL}(p \| p^\star) &= \int p(\psi) \log \frac{p(\psi)}{p^\star(\psi)}\, d\psi \\ &= -h(p) - \int p(\psi) \log p^\star(\psi)\, d\psi. \end{aligned} \tag{62}$$

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

$$\begin{aligned} -\int p(\psi) \log p^\star(\psi)\, d\psi &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma_{Q,\varepsilon}) + \frac{n}{2\varepsilon} \int \frac{1}{2} \langle \psi, Q\psi \rangle p(\psi)\, d\psi \\ &= \frac{n}{2} \log(2\pi) + \frac{1}{2} \log \det(\Sigma_{Q,\varepsilon}) + \frac{n}{2}. \end{aligned} \tag{63}$$

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 (62) and (63) 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.

A.10.2 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

$$\begin{aligned}\mathcal{L}(p) = &- \int p(\psi)\log p(\psi)\,d\psi + \lambda_0\Big(\int p(\psi)\,d\psi - 1\Big) \\ &+ \Big\langle \lambda, \int \psi\, p(\psi)\,d\psi\Big\rangle - \beta\Big(\int \frac{1}{2}\langle\psi, Q\psi\rangle p(\psi)\,d\psi - \varepsilon\Big).\end{aligned} \tag{64}$$

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

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

A.10.3 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.

A.11 Further properties of the exact Gaussian-chaos gate

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

A.11.1 All-order moment formula

Let ξ_γ^{ex} be defined by (24). For any $x_1, \dots, x_m \in U$,

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

The proof is the same Gaussian moment-generating calculation used in Theorem A.4.

A.11.2 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.

A.11.3 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.

A.12 Why the Green kernel is induced in this design class

The purpose of this appendix is to state the precise structural lesson of the VKD variational analysis. The Green kernel is not an extra hypothesis layered on top of the model. It is the covariance geometry induced by the quadratic operator budget in this design class.

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 (15). By Theorem A.1, the unique variational optimizer is

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

Hence the covariance kernel is determined 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 *induced*: 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 A.20 (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 variationally selected 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.

A.13 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 A.21 (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 $t > 0$,*

$$S_{te^\eta}(h) \subseteq S_t(\xi \odot h) \subseteq S_{te^{-\eta}}(h). \tag{65}$$

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 $[te^{-\eta}, te^{\eta}]$.

Interpretation for representation learning. The proposition shows that positive multiplicative gating changes superlevel geometry only through a controlled multiplicative threshold band. Thus coherent superlevel regions are deformed through threshold shifts rather than punctured by exact 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 A.22 (Sample-wise GCh obeys a random sandwich width)**.** *For the implemented gate* ξ_γ^{sw} *in* (35)*,*

$$\|\log \xi_\gamma^{\mathrm{sw}}\|_\infty \le |\gamma|\ \mathrm{osc}(\psi), \qquad \mathrm{osc}(\psi) \coloneqq \max_{x\in U}\psi(x) - \min_{x\in U}\psi(x). \tag{66}$$

Hence Theorem A.21 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 (66). □

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 A.23 (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*

$$\Pr\big(\beta_1(S_t(\widetilde{h})) = 1\big) = q^n, \qquad \Pr\big(\beta_1(S_t(\widetilde{h})) = 0\big) = 1 - q^n. \tag{67}$$

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. □

Interpretation for representation learning. Closed contours, ring-like activation patterns, and loop-shaped superlevel regions are idealized but meaningful models of semantic geometry. Theorem A.23 shows that hard deletion is topologically brittle in this model: a single dropped segment breaks the loop. This provides a complementary topological view of the same finite-geometry mismatch studied in the main text.

Remark A.24 (Relation to DropBlock). DropBlock changes the spatial correlation of the zero set but retains the hard-deletion mechanism. Any block pattern that removes a connected arc from a loop-like superlevel set also destroys its cycle rank. The theorem above isolates the corresponding topological failure mode of binary deletion.

A.14 Additional experimental details and results

A.14.1 Experimental setup

Datasets. We evaluate on ImageNet-1k [3] (1.28M training images, 50k validation images, 1000 classes). To measure robustness under common corruptions, we additionally use ImageNet-C [10]. 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 compact fine-grained secondary evaluation, we also evaluate on Oxford-IIIT Pet [19], a 37-class benchmark whose labels are sensitive to shape cues.

Architectures and injection sites. Our primary backbone is ResNet-50 [9]. 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 [16] 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 secondary evaluation. For the fine-grained secondary evaluation, 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$\pm$std over 3 seeds.

Baselines. We compare GCh against Dropout [23], DropBlock [7], 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.

A.14.2 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 18: 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 19: 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.

A.14.3 Injection depth (L2/L3/L4)

We apply the same GCh mechanism at different residual stages under the controlled 3-seed protocol. The table is reported in the consolidated depth/strength diagnostic appendix above; the takeaway is

that 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.

A.14.4 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. The table is reported in the consolidated depth/strength diagnostic appendix above. The takeaway is that there is a robust small-to-moderate regime in which GCh preserves accuracy and improves reliability, while very large γ values cause the expected breakdown from excessive multiplicative perturbation.

A.14.5 ImageNet-C detailed 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$\pm$std over three seeds.

Reading the tables. Table 13 gives the main late-stage comparison at $g = 0.1$. Table 14 isolates the effect of injection depth under shift. Table 15 reports strength sensitivity under shift. Table 16 provides the corruption-wise breakdown.

The detailed ImageNet-C tables are reported once in Section A.7.5. We avoid duplicating them here to keep table numbering and interpretation unambiguous.

A.15 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 20: 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 21: 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.

A.16 Normalization and implementation details

A.16.1 Interpretation of Theorem A.4

Theorem A.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.

A.16.2 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.

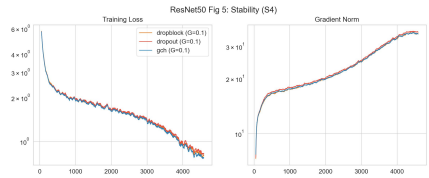
A.16.3 Implementation notes

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 $\xi \equiv 1$.

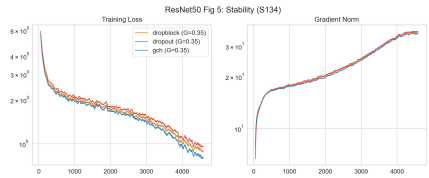
A.17 Additional diagnostic figures

The figures in this section support the experimental interpretation but are not used as primary aggregate evidence. Optimization-stability plots check for large-scale training or signal-flow pathologies, and representative configuration figures provide qualitative single-run inspections while keeping the main text focused on the central theory-to-experiment line.

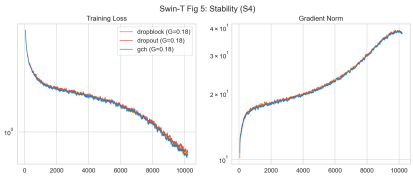
A.17.1 Optimization stability and signal-flow diagnostics



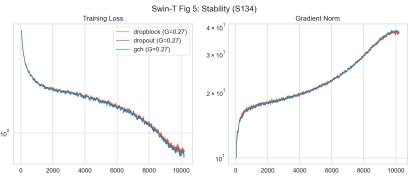
(a) ResNet-50, late-stage injection.



(b) ResNet-50, multi-stage injection.

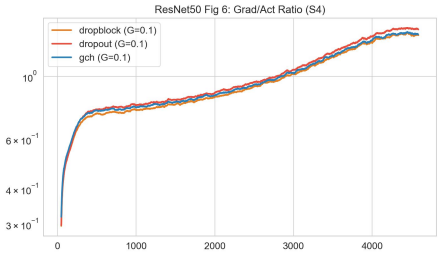


(c) Swin-T, late-stage injection.

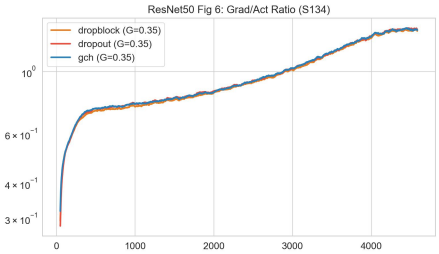


(d) Swin-T, multi-stage injection.

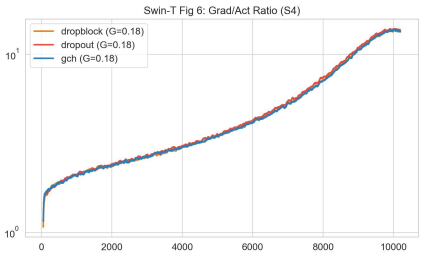
Figure 6: Training stability diagnostics. Training-loss and gradient-norm trajectories for representative settings. These plots are used to check for large-scale optimization instability and are not used as primary performance evidence.



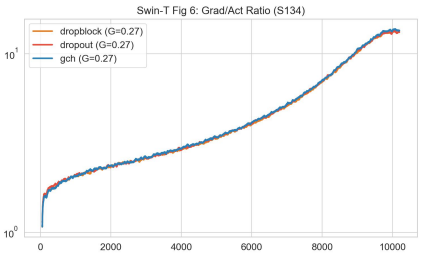
(a) ResNet-50, late-stage injection.



(b) ResNet-50, multi-stage injection.



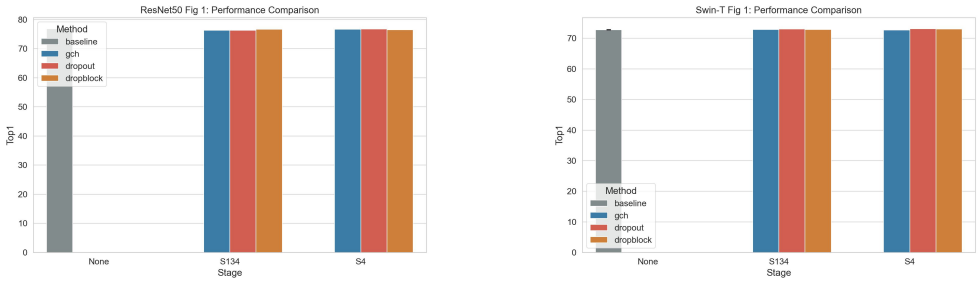
(c) Swin-T, late-stage injection.



(d) Swin-T, multi-stage injection.

Figure 7: Gradient-to-activation norm ratio. Signal-flow diagnostics for representative settings. These plots are used to check whether the method introduces obvious gradient-to-activation pathologies.

A.17.2 Additional performance visualizations



(a) ResNet-50. **(b)** Swin-T.

Figure 8: Additional clean-performance visualization. Accuracy and NLL across representative injection settings. The main text emphasizes calibration and shift reliability; this figure is included as a visual summary of clean-performance behavior.

A.17.3 Representative configuration diagnostics

The following figures are representative single-run diagnostics. They are included for qualitative inspection of internal trajectories and should not be read as aggregate evidence. They are useful for showing the deployment-side distinction suggested by the theory: in mild configurations, hard masks and GCh may be close in aggregate metrics, whereas in more incompatible multi-stage settings hard masking can produce large coherence excursions while GCh remains comparatively stable.

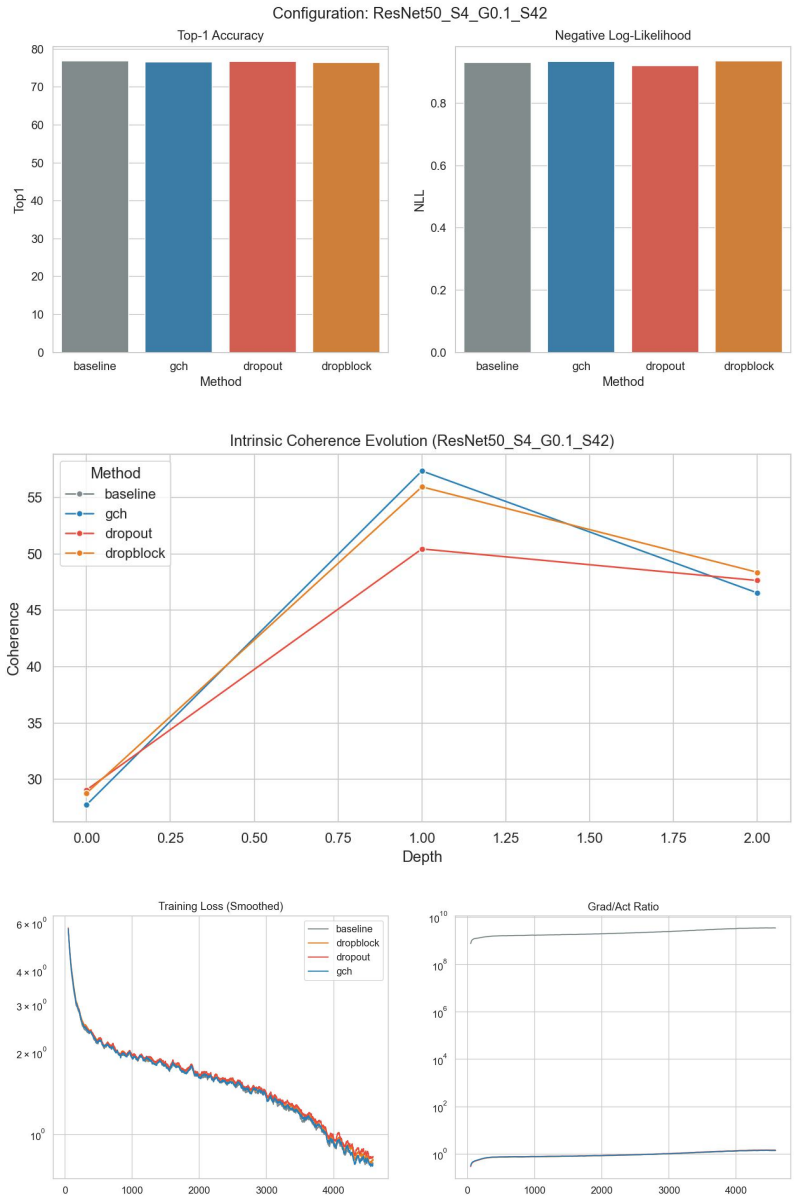


Figure 9: Representative diagnostics for ResNet-50, late-stage injection, $\gamma = 0.1$, seed 42. Single-run diagnostic only.

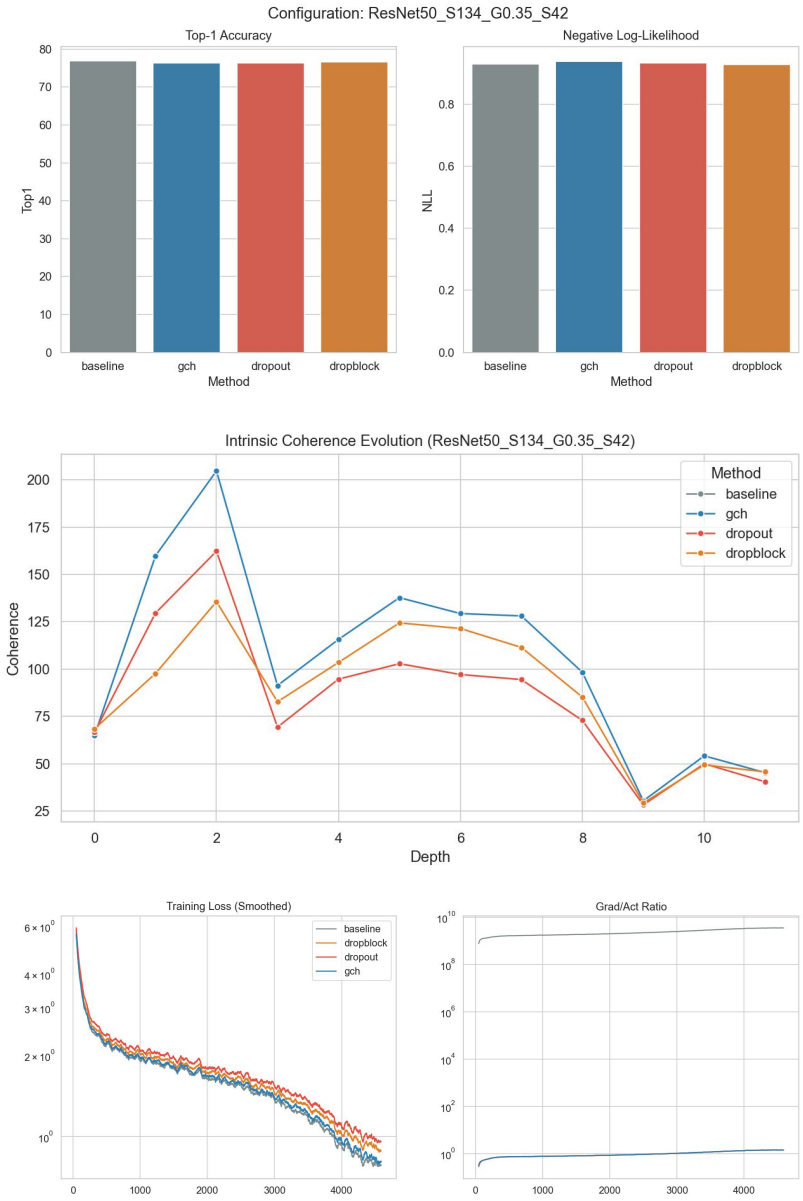


Figure 10: Representative diagnostics for ResNet-50, multi-stage injection, $\gamma = 0.35$, seed 42. Single-run diagnostic only.

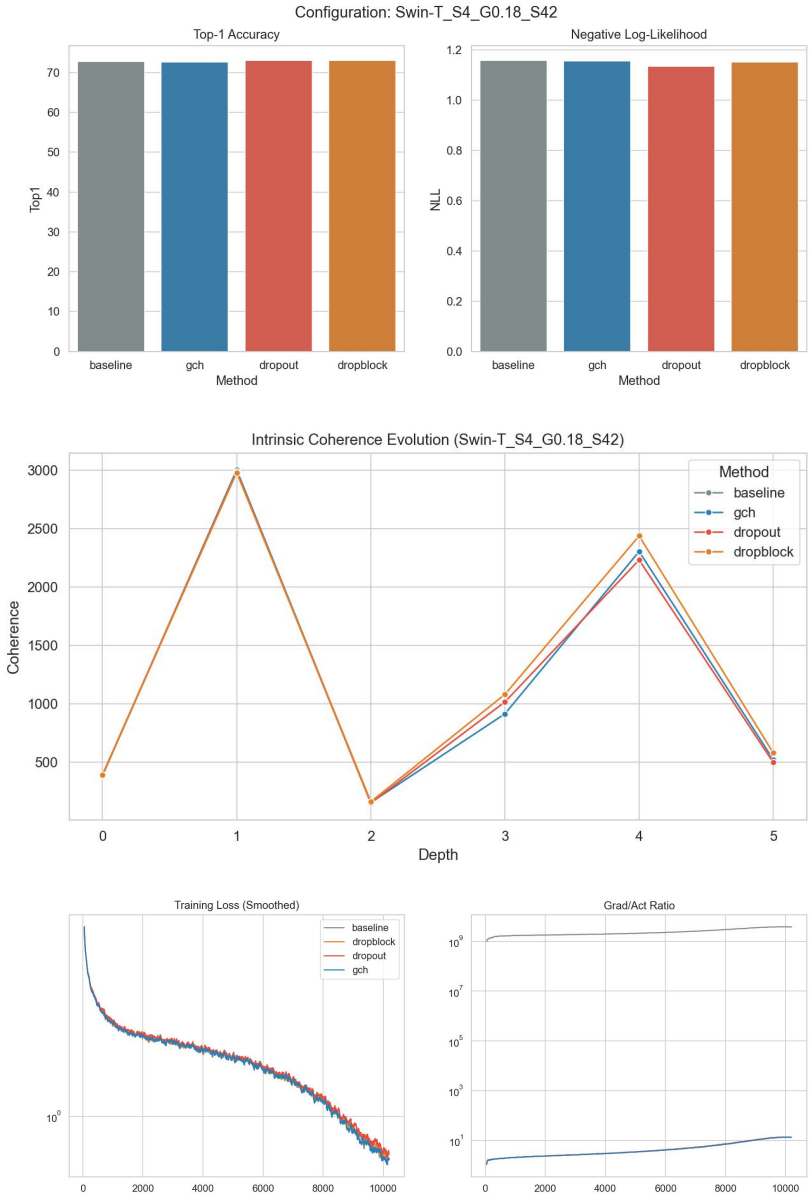


Figure 11: Representative diagnostics for Swin-T, late-stage injection, $\gamma = 0.18$, seed 42. Single-run diagnostic only.

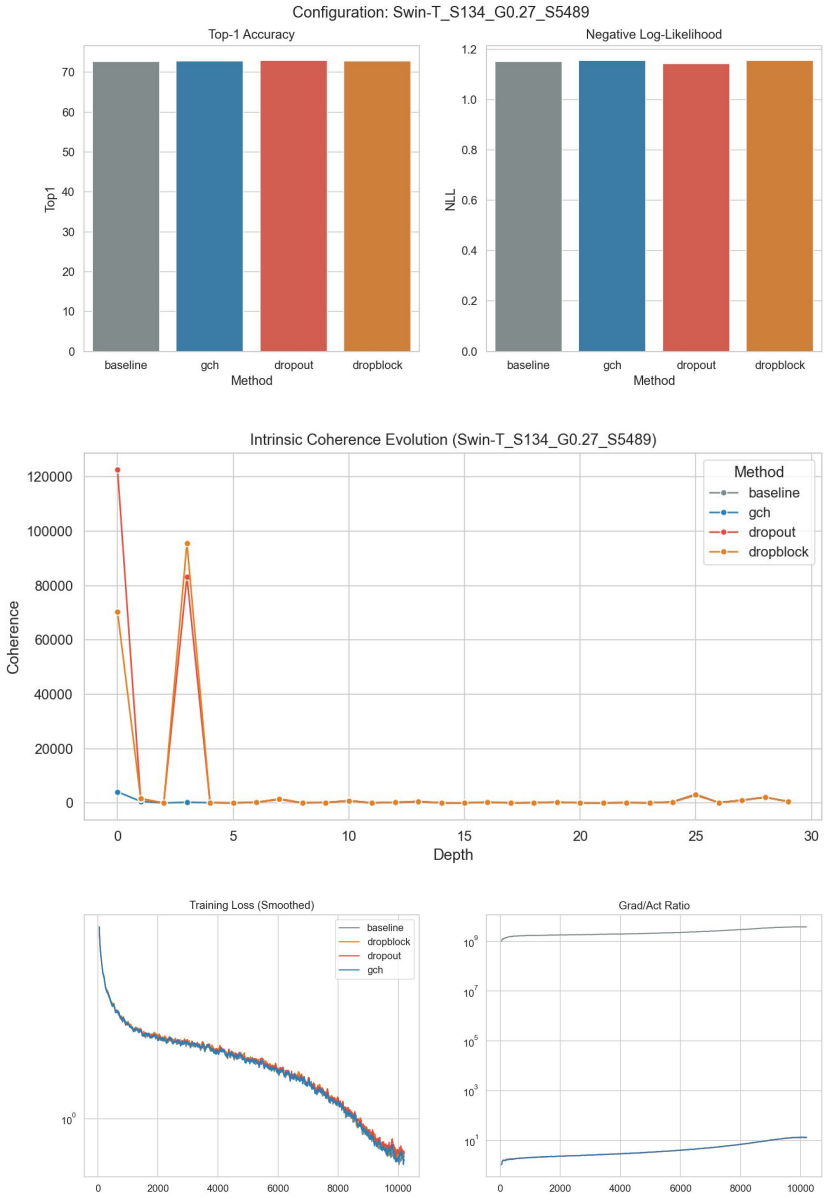


Figure 12: Representative diagnostics for Swin-T, multi-stage injection, $\gamma = 0.27$, seed 5489. Single-run diagnostic only. This configuration illustrates the deployment-side stability advantage: hard masking exhibits large coherence excursions under multi-stage perturbation, while GCh remains comparatively stable in the diagnostic trajectories.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the variational design problem for a complete noise mechanism, the compatibility results, the empirical scope, and the main limitations.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The discussion section states that the current evidence is strongest for calibration and selected corruption-shift reliability, and identifies broader corruption coverage and multi-seed transformer studies as limitations.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All main theoretical results are numbered in the main text; complete proofs are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The experimental setup, metrics, datasets, training protocols, baselines, and full tables are described in the appendix.

   Guidelines:

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

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

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

   Answer: [No]

   Justification: The submission does not include an anonymized code release at this stage; the method is specified algorithmically and implementation details are provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The main text and appendix specify datasets, models, metrics, injection stages, strength parameters, and training protocols.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Main controlled CNN experiments report multi-seed summaries; single-run transfer results are explicitly identified as preliminary.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The appendix reports training protocols and model settings; detailed hardware/accounting can be added if required by the final submission system.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses standard image-classification benchmarks and does not involve human-subject data collection or safety-critical deployment.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The broader-impact discussion is limited because the work is a foundational reliability method; potential benefits and limitations are discussed in the paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, or assets requiring special safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All datasets and baselines used in the experiments are cited; dataset licenses should be checked and added to the final camera-ready metadata if required.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset or released model asset.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The experiments do not involve crowdsourcing or human-subject studies.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The experiments do not involve human-subject studies requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not part of the proposed method or experimental pipeline.

Guidelines:

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