Testing the "Benign Running Statistics" Hypothesis: Class-Conditional Batch-Normalization Substitution Rejects the Distribution-Shift Account at $p < 10^{-19}$

Ali Saffarini [1]

Abstract

Batch-normalization (BN) running statistics are widely treated as a benign approximation to the true population mean/variance—a numerical convenience whose perturbation should affect accuracy in the same way as any other distribution shift. We pose this as a falsifiable null hypothesis and test it. H_0 **(distribution-shift account):** substituting any non-global statistics at inference should hurt accuracy roughly symmetrically; in particular, same-class statistics—the sample's own class-conditional statistics, the most accurate and least arbitrary substitution—should hurt no more than deliberately mismatched wrong-class statistics. We design a paired three-condition intervention—global / same-class / wrong-class—across four CNN architectures and three datasets, totalling 1,320 paired-trial comparisons over 22 trained models. The data reject H_0 sharply: same-class accuracy is 3.5% vs. wrong-class 65.4% on CIFAR-10/SmallResNet ($t(9)=-39.1$, $p < 10^{-10}$ paired; same-vs-global $t(9)= -374$, $p = 3.5\times10^{-20}$). We then run three pre-registered alternative-explanation tests: (i) a linear-probe test on the post-substitution features (probe accuracy $\geq 99.7\%$ under same-class; the encoder's class-discriminative structure survives the substitution rather than being destroyed—and because same-class selection uses the true label, we corroborate this with a label-independent *random*-class probe that also stays high, §4.2); (ii) a GroupNorm architectural-control test (no running statistics; unaffected, arguing against a "generic running-stat perturbation"); (iii) a temperature-scaling test (ECE halves from 0.394 to 0.151 but accuracy unchanged, arguing against a "logit-scale issue"). A within-architecture layer-wise ablation tests the multiplicative-compounding hypothesis: single-layer drops sum to 29.4% but the all-layer drop is 86.8%, a $2.95\times$ super-additive factor that replicates seed-to-seed (n=3). A per-statistic disentangling experiment tests a sub-hypothesis of the algebraic argument below: substituting only μ leaves 76.7% accuracy, only σ^2 leaves 27.2%, and the joint substitution reaches 2.8%—neither statistic alone reproduces the catastrophe, the two interact super-additively, and contrary to the one-layer linearization σ^2-substitution alone is more damaging than μ-substitution alone. We close with a one-line algebraic identity that restates the surviving asymmetry as an exact zero in a one-layer linearization. The paper is intended as both an empirical contribution and a worked example of intervention-design hypothesis testing as ML methodology.

[1]Harvard University. Correspondence to: Ali Saffarini <alisaffarini@college.harvard.edu>.

Accepted at the Workshop on Hypothesis Testing, ICML 2026. Non-archival.

1. Introduction

Batch normalization (Ioffe & Szegedy, 2015) stores per-layer running mean and variance vectors that are used at inference to standardize activations. The running statistics are typically thought of as a benign approximation to the true population mean and variance: a numerical convenience that affects optimization but not what the network represents.

We test that picture by perturbing the running statistics in a controlled way. For a network trained on C classes, we estimate one set of running statistics per class by passing the corresponding subset of the training data through the network in BN `train()` mode without updating any parameters. We then run inference under four conditions: *global* (the standard buffers), *same-class* (class-c statistics for class-c samples), *wrong-class* (class-$(c + \lfloor C/2 \rfloor) \bmod C$ statistics for class-c samples), and *random-class* (a random other class's statistics for each sample).

A skeptical reading of any single one of these conditions is straightforward: the classifier was trained against the

global statistics, so swapping in any other statistics is a distribution shift, and accuracy drops are unsurprising. What this paper documents is a pattern that the distribution-shift reading does not predict. Same-class statistics are not a mild refinement of global statistics; they are the *worst* of the three substitutions, and by a wide margin. On CIFAR-10 with a SmallResNet, same-class accuracy is 3.5%—below the 10% chance baseline—while wrong-class accuracy is 65.4% and random-class accuracy is 59.0%. The ordering "same $\ll$ random $<$ wrong $\ll$ global" replicates across four architectures and three datasets.

We frame the paper as a hypothesis-testing exercise. The default-assumed null—that BN running statistics are a benign approximation, and any non-global statistics produce a generic distribution-shift drop—makes a sharp prediction: same-class statistics, being the sample's *own* class-conditional statistics (the most accurate and least arbitrary substitution), should if anything be the *mildest*, not more damaging than the deliberately mismatched wrong-class. Our paired same-class / wrong-class / global / random-class design is constructed to test exactly this prediction. The data falsify it sharply (Section 4.1). We then run pre-specified tests of the three most natural alternative explanations, each of which is rejected (Section 4.3), and a within-architecture layer-wise ablation that tests a multiplicative-compounding hypothesis predicted by a simple algebraic argument (Section 4.5).

Contributions.

1. **A sharply rejected null.** We test a specific, falsifiable prediction of the distribution-shift account: that same-class statistics—the sample's *own* class-conditional statistics, the most accurate and least arbitrary substitution—should hurt *no more* than deliberately mismatched wrong-class. The data reverse it—same-class is uniquely catastrophic ($t(9) = -39.1$, $p < 10^{-10}$ paired against wrong-class; same-vs-global $t(9) = -374$, $p = 3.5 \times 10^{-20}$). The asymmetric ordering same $\ll$ random/wrong $\ll$ global replicates across four CNN architectures and three datasets, totaling 22 trained models and 1,320 paired-trial comparisons. (We formalize which distribution-shift baselines this does and does not adjudicate in §4.1 and §7.)
2. **Three pre-specified alternative-explanation tests.** (a) A linear-probe test shows the encoder's class-discriminative structure survives the substitution (probe accuracy $\geq 99.7\%$ under same-class; $\geq 90.8\%$ under the label-independent random-class control, §4.2)—evidence against "the encoder broke," with the caveat that the same-class probe is itself label-conditioned. (b) A GroupNorm architectural-control test argues against "generic running-statistic perturbation" (no running statistics; unaffected). (c) A temperature-scaling test argues against a "logit-scale issue" (ECE halves, accuracy unchanged).
3. **A multiplicative-compounding test that does *not* reject.** Single-layer drops sum to 29.4% but the all-layer drop is 86.8%—a $2.95\times$ super-additive factor (n=3 seeds, replicates seed-to-seed). The data are consistent with multiplicative cancellation predicted by the linearization in Section 5.
4. **A per-statistic sub-hypothesis test that partially rejects the linearization's prediction.** The single-layer linearization predicts μ-substitution alone should be sufficient to zero the class signal; empirically, μ-only substitution leaves 76.7% accuracy while σ^2-only substitution leaves 27.2% and the joint reaches 2.8%, with σ^2-substitution alone more damaging than μ-substitution alone and the two interacting super-additively (Section 4.6).
5. **An algebraic identity that survives the data.** A one-layer linearization shows the asymmetric ordering follows from a single equation: same-class statistics zero the class-c component of the head's input while wrong-class statistics only translate it. We treat this as a constraint on the surviving-hypothesis space, not as a derivation.
6. **A non-monotone probe trajectory under interpolation,** which the linearization does not capture and which we report as an open feature for future testing.

The paper is intended as both an empirical contribution and a worked example of intervention-design hypothesis testing applied to a default ML assumption. Section 6 discusses what survives. All experiment scripts, raw JSON results, and reproduction instructions are available from the author on request.

2. Background and Related Work

Batch normalization. BN (Ioffe & Szegedy, 2015) normalizes activations using mini-batch statistics during training and exponentially-averaged *running* statistics $(\mu_\ell, \sigma_\ell^2)$ at inference, after which a learned affine transform γ_ℓ, β_ℓ is applied. Santurkar et al. (2018) and Bjorck et al. (2018) argue BN's main effect is to smooth the optimization landscape rather than to address "internal covariate shift". These works examine BN's training-time role; we examine the inference-time role of the running buffers specifically.

BN statistics as adaptable quantities. A number of methods treat BN statistics as something that can be re-estimated rather than treated as fixed. AdaBN (Li et al., 2018) replaces them with target-domain statistics for domain adaptation; TENT (Wang et al., 2021), Schneider

et al. (2020), and Benz et al. (2021) update them on test data for robustness to distribution shift. Frankle et al. (2020) show that training only the BN affine parameters of a randomly-initialized network can reach surprisingly high accuracy. These works modify BN parameters in directions that *help*; we instead document a regime in which a seemingly benign modification is sharply harmful.

Class-conditional normalization, when trained for. Conditional BN (de Vries et al., 2017), FiLM (Perez et al., 2018), and class-conditional BN in BigGAN (Brock et al., 2019) all use *per-class* normalization parameters successfully—but as part of training: the network learns per-class affine parameters (γ_c, β_c) alongside, so the head co-adapts to per-class statistics from the start. Our finding is not in tension with this. We substitute class-conditional (μ_c, σ_c) at inference while *holding the affine parameters* $(\gamma_\ell, \beta_\ell)$ *at their globally-trained values*; the cancellation in Eq. (2) is between class-mean activations and a head that is not expecting them. The CBN-trained networks above would not exhibit this cancellation precisely because their head has co-adapted. We flag the fixed-affine choice as limitation (v) and discuss it in Section 7.

Mean-ablation in mechanistic interpretability. Mean-ablation (Wang et al., 2023) replaces an activation with its dataset-mean to test what information it carries. Same-class BN substitution is the CNN analogue: it removes the class-mean component of the activation at every BN layer, and the probe-vs-classifier dissociation we report mirrors the standard mech-interp finding that ablating a direction can preserve linear decodability while breaking the original head—here in a normalization-layer rather than attention-head setting.

Linear probing. Linear probes (Alain & Bengio, 2016; Kornblith et al., 2019) measure whether class structure is linearly accessible in a representation. We use them in a counterfactual role: a network that mispredicts but whose features still admit a high-accuracy linear probe has a problem in the head, not in the encoder.

Calibration. Guo et al. (2017) introduced expected calibration error (ECE) and temperature scaling. We use both as diagnostics. Temperature scaling is monotone in logits and so cannot change the argmax; if it nonetheless recovered accuracy, the same-class failure would be a mere overconfidence problem, which it is not.

Normalization variants. GroupNorm (Wu & He, 2018) computes statistics over feature-map groups within each example and stores no running averages; we use it as an architectural control. Normalization-free architectures (Brock et al., 2021) replace BN entirely.

3. Methodology

Architectures. We train four image classifiers on CIFAR-10 with standard augmentation (random crop pad 4, horizontal flip), SGD with momentum 0.9, weight decay 5×10^{-4}, initial LR 0.1, cosine annealing. **SmallResNet**: a four-stage residual net with widths $[32, 64, 128, 256]$ and 20 BN layers (10 seeds, 20 epochs); a five-stage **SmallResNet64** variant with widths $[32, 64, 128, 256, 512]$ is used for 64×64 inputs. **VGG-11-BN** (8 BN layers, 3 seeds), **SimpleCNN** (4-layer plain CNN, 4 BN layers, 3 seeds), **WideResNet-28-10** (Zagoruyko & Komodakis, 2016) (depth 28, widen factor 10, dropout p=0.3, 3 seeds). We additionally train SmallResNet on CIFAR-100 (5 seeds, 25 epochs) and SmallResNet64 on Tiny-ImageNet-200 (3 seeds, 25 epochs). Full hyperparameters are in Appendix A.

Class-conditional statistics. Let f_θ be a trained CNN with BN layers $\{B_\ell\}$ carrying running means μ_ℓ and running variances σ_ℓ^2. For class c we reset the BN buffers, place the model in `train()` mode (so BN updates the buffers via its standard EMA), iterate in `torch.no_grad()` over the class-c training subset, and snapshot the resulting buffers as $(\mu_\ell^{(c)}, (\sigma_\ell^{(c)})^2)$. *No model parameters are updated.* The original buffers are restored before the next class's pass.

Inference conditions. For test sample x with label c we evaluate four conditions: *global* (load original buffers), *same-class* (load $(\mu_\ell^{(c)}, (\sigma_\ell^{(c)})^2)$), *wrong-class* (load class-$(c + \lfloor C/2 \rfloor) \bmod C$), and *random-class* (load a uniformly random class $\neq c$, drawn per-sample). *We substitute both μ_ℓ and σ_ℓ^2 jointly* (the standard same-class condition above); a per-statistic disentangling experiment that substitutes only $\mu_\ell^{(c)}$ or only $(\sigma_\ell^{(c)})^2$ tests which of the two carries the effect (Section 4.6). For each condition we record top-1 accuracy, mean true-class softmax confidence, and the penultimate-layer features.

Linear probe. Penultimate features (post global average pooling) are split 80/20; we fit ridge regression (λ=10^{-3}, one-hot targets, closed form) for CIFAR-10, and multinomial logistic regression (L-BFGS, C=1.0) for $\geq$ 100-class settings, and report held-out accuracy.

Interpolation. For $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ we set $\mu_\alpha = \alpha\mu_{\text{global}} + (1-\alpha)\mu_{\text{same}}$ and $\sigma_\alpha^2 = \alpha\sigma_{\text{global}}^2 + (1-\alpha)\sigma_{\text{same}}^2$ and evaluate.

Calibration controls. ECE uses 15 equal-width bins. Temperature scaling (Guo et al., 2017) optimizes a single scalar T on a held-out validation set under NLL, then divides logits by T. Because $T > 0$ preserves the argmax,

Table 1. SmallResNet on CIFAR-10 across the four BN-statistic conditions. Mean $\pm$ std over 10 seeds.

Condition	Acc. (%)	Conf. (%)	Probe (%)
Global (baseline)	91.08 ± 0.12	88.88 ± 0.15	90.86 ± 0.55
Same-class	$\mathbf{3.55 \pm 0.69}$	4.80 ± 0.52	99.88 ± 0.09
Wrong-class	65.37 ± 4.69	50.64 ± 3.22	100.00 ± 0.00
Random-class	58.99 ± 5.84	45.47 ± 5.19	97.63 ± 1.11

Table 2. Cross-architecture results on CIFAR-10 (mean $\pm$ std). The ratio same/wrong is < 0.18 in every architecture.

Arch.	BN	Global	Same	Wrong	S/W
SmallResNet	20	91.08 ± 0.12	3.55 ± 0.69	65.37 ± 4.69	0.054
VGG-11-BN	8	85.44 ± 1.28	10.88 ± 1.91	62.96 ± 1.25	0.173
SimpleCNN	4	83.17 ± 0.23	26.33 ± 0.48	78.10 ± 0.46	0.337
WRN-28-10	many	90.61 ± 0.51	11.12 ± 0.90	64.06 ± 3.14	0.174

any accuracy change under temperature scaling is by construction zero; the diagnostic value lies in comparing ECE before and after.

GroupNorm control. We retrain SmallResNet with all BatchNorm layers replaced by GroupNorm (8 groups, fewer when channels are smaller). GroupNorm has no running statistics, so the substitution paradigm is inapplicable and the relevant question is whether the architecture suffers any analogous loss.

Statistical protocol. SmallResNet/CIFAR-10 uses 10 seeds (42–51); SmallResNet/CIFAR-100 uses 5 (42–46); all other configurations use 3 (42–44). We report mean $\pm$ std and paired t-tests where stated.

4. Results

4.1. The asymmetric substitution ordering

Table 1 shows the four-condition comparison on our most-replicated configuration. Three observations follow.

First, every non-global condition lies strictly below global accuracy, as a generic distribution-shift account would predict. Second, the ordering among the three substitution conditions is the opposite of what such an account would predict: same-class statistics—the sample's *own* class-conditional statistics, the most accurate and least arbitrary of the three substitutions—are the most damaging. Same-class accuracy (3.5%) lies below the 10% chance baseline; wrong- and random-class accuracies (65.4% and 59.0%) sit much higher. Same-vs-global accuracy is $t(9) = -374.1, p = 3.5 \times 10^{-20}$; same-vs-wrong is $t(9) = -39.1, p < 10^{-10}$.

Third, per-class deltas are uniformly large and negative across all 10 CIFAR-10 classes (Appendix B), ranging from -0.97 for *automobile* to -0.72 for *cat*; the collapse is not driven by one or two pathological classes.

What the null does and does not claim. The distribution-shift account admits several operationalizations, and we test one specific, falsifiable version rather than a strawman. On the benign-approximation reading, the running statistics are just a numerical convenience, so every non-global substitution is a generic distribution shift and the three conditions should hurt comparably; if anything, same-class statistics—the sample's *own* class-conditional statistics, the most accurate and least arbitrary of the three substitutions—should be the mildest, not the most damaging. The data reverse this: same-class is uniquely catastrophic, more than $18\times$ worse than the deliberately mismatched wrong-class. We do *not* claim to have ruled out every distribution-shift baseline: perturbed-global statistics, statistics estimated from corrupted inputs, or statistics borrowed from another dataset would each probe a different off-distribution direction, and characterizing them is future work (Limitation vii). What the random-class condition already contributes is one non-class-matched off-distribution baseline: it patterns with wrong-class (59.0% vs. 65.4%) and sits far above same-class (3.5%), so the catastrophe is specific to statistics matching the sample's own class, not to off-distributionness per se.

Cross-architecture replication. Table 2 shows the ordering replicates across all four architectures we tested. In every case, same-class accuracy is well below wrong-class accuracy, and the ratio is at least $4\times$ for the BN-heavy networks.

The same-class accuracy correlates roughly with the number of BN layers: SmallResNet (20 layers) drops the most, SimpleCNN (4 layers) the least. WideResNet-28-10 (a deep BN-heavy network with dropout p=0.3) shows a same-class accuracy of 11.1%, somewhat higher than SmallResNet's 3.5% despite many more BN layers; the dropout regularization may be partially buffering the head against activation perturbation.

Cross-dataset replication. Table 3 extends the experiment to CIFAR-100 and Tiny-ImageNet-200 with SmallResNet/SmallResNet64. With $C = 100$ or $C = 200$, all three substitution conditions collapse to floor and the same-vs-wrong gap shrinks below the seed-to-seed standard deviation. The accuracy values ($0.60 \pm 0.49\%$ on CIFAR-100, $0.17 \pm 0.24\%$ on Tiny-IN-200) are statistically indistinguishable from $1/C$ chance, so we read these datasets as showing only that the substitution effect persists at scale, not that it produces below-chance accuracy.

Table 3. Cross-dataset results with SmallResNet (CIFAR-10/100) and SmallResNet64 (Tiny-ImageNet-200). Chance is $1/C$. Same-class accuracy collapses to floor on both larger datasets.

Dataset	C	Chance	Global	Same	Wrong	Probe
CIFAR-10	10	10.00%	91.08	3.55	65.37	99.88
CIFAR-100	100	1.00%	70.11	0.60	1.00	100.00
Tiny-IN-200	200	0.50%	60.77	0.17	0.50	100.00

4.2. The features survive: a probe-vs-classifier dissociation

The right-most columns of Tables 1 and 3 report the linear probe accuracy under same-class statistics. Despite classification accuracy of 3.5% on CIFAR-10 and 0.17% on Tiny-ImageNet-200, a linear probe trained on the same penultimate features achieves 99.9% and 100.0% respectively—in the latter case, with zero variance across seeds.

This dissociation matters for interpreting Section 4.1: whatever the same-class substitution is doing, it is not destroying the encoder's class-discriminative information. A linear refit of the head on the same features is essentially perfect. The classifier's inability to recover the class is therefore a property of the *specific learned head*, not of the representation it operates on. Notably the probe accuracy under same-class often *exceeds* the global-statistic probe (99.9% vs 90.9% on CIFAR-10), suggesting that class-conditional normalization removes within-class variation in a way that makes the resulting features more linearly separable.

The same-class probe is somewhat lower for SimpleCNN (54.7%; see Appendix C), the smallest architecture; on the BN-heavy networks (SmallResNet, VGG, WRN) the probe sits at $\geq 98.4\%$.

Does the probe merely read an injected label? Same-class substitution is a *label-conditioned* intervention: choosing which statistics to load requires the true label c. As the linearization in Section 5 makes explicit, the class identity is partly re-encoded in the *choice* of (μ_c, σ_c), so a probe on same-class features could in part be reading the label back out of the normalization; the same-class probe *alone* therefore does not establish that the encoder is intact. The clean control is the *random-class* condition, which normalizes each sample by a randomly chosen *other* class, re-drawn per sample. There is then no statistic that consistently tags a sample with its own class—two class-c samples receive different, non-matching normalizations—so the choice of statistics carries no recoverable true-label information (the only residual label dependence, that the drawn class differs from the true one, excludes a single class and cannot produce $> 90\%$ accuracy). Yet the random-class linear probe still reaches 97.6% (SmallResNet), 96.6% (WRN-28-10), and 90.8% (SimpleCNN) (Tables 1, 6). Because it cannot be reading a class-matched normalization, this accuracy must reflect class structure retained in the encoder. We therefore rest the "encoder survives" conclusion on the random-class probe and report the same-class probe only alongside it. A fully label-agnostic probe—trained on global-statistic features and evaluated on substituted features—would be a still cleaner single test, and we flag it as the preferred protocol for future work.

4.3. Two ruled-out explanations

Generic normalization-swap effect. A natural alternative explanation is that the asymmetry is an artifact of *any* deep architecture with running statistics that are perturbed at inference. We test this by training SmallResNet with all BatchNorm layers replaced by GroupNorm (Wu & He, 2018), which computes statistics from the current input and has no running buffers (3 seeds, 20 epochs). The GroupNorm model achieves $81.65 \pm 1.94\%$ accuracy and a $81.08 \pm 1.64\%$ linear probe under standard inference; there are no running statistics to substitute, so the asymmetry is inapplicable to it by construction. The effect is thus specific to the running-statistic mechanism in BN.

Logit-calibration issue. A second alternative is that same-class statistics merely sharpen or rescale the logits in a way that softmax confidence and ECE pick up but argmax does not. We test this with temperature scaling (Guo et al., 2017): optimize a single scalar T on validation NLL under same-class statistics, then divide all logits by T at test time.

Fitting T on a held-out validation split (excised from training, so the same-class accuracy here is 2.87% rather than Table 1's 3.55%), the optimum is $T \approx 3.43$: same-class logits are substantially over-scaled. Temperature scaling cuts ECE from 0.394 to 0.151 but recovers 0% of the lost accuracy (by construction, since monotone-in-logits transforms preserve argmax); the diagnostic content is the residual ECE—even after optimal rescaling the probabilities remain poorly calibrated, so the logit *ranking*, not just its scale, has been disrupted. The same-class collapse is not a temperature problem.

4.4. Asymmetry magnitude scales with BN depth

Across the CIFAR-10 architectures the same-class drop grows monotonically with the number of BN layers in the main path: $\log(\text{global}/\text{same}) = 1.15$ for SimpleCNN (4 BN), 2.06 for VGG-11-BN (8 BN), 3.25 for SmallResNet (20 BN). WRN-28-10 (BN-heavy with dropout p=0.3) sits at 2.10, which we attribute to dropout buffering the head against activation perturbation (Section 4.1). Four architectures are too few to fit a quantitative trend, so we read this cross-architectural ordering as suggestive and reserve the strong evidence for the within-architecture layer-wise ablation in Section 4.5, which shows single-layer drops com-

Table 4. Layer-wise BN-statistic replacement on SmallResNet/CIFAR-10 (mean $\pm$ std over 3 seeds). Full per-layer numbers in Appendix E.

Replacement	# layers	Acc. (%)
Global baseline	0	91.11 ± 0.14
Single layer (mean over 20)	1	89.59 ± 0.21
Single layer (max drop, layer4.1.bn1)	1	86.20 ± 1.01
Cumul. shallow→deep, k=5	5	85.46 ± 1.43
Cumul. shallow→deep, k=10	10	76.10 ± 0.33
Cumul. shallow→deep, k=15	15	46.96 ± 0.88
Cumul. deep→shallow, k=5	5	59.65 ± 6.41
Cumul. deep→shallow, k=10	10	23.74 ± 3.25
Cumul. deep→shallow, k=15	15	11.14 ± 0.48
All 20 layers (full SC)	20	4.33 ± 1.32

pounding super-additively into the full effect.

4.5. Within-architecture: which BN layers carry the effect?

Section 4.4 documents that the same-class drop scales with the *number* of BN layers *across* architectures. We localize it *within* a single architecture by training SmallResNet on CIFAR-10 (3 seeds, 20 epochs) and substituting class-conditional statistics at controlled subsets of its 20 BN layers, leaving other layers at their global values. The trained models reach $91.11 \pm 0.14\%$ global / $4.33 \pm 1.32\%$ all-layer-same-class, replicating the multi-seed numbers in Section 4.1.

Three observations follow. First, the per-layer single-layer drops sum to 29.4%, but the full-replacement drop is 86.8%: the compounding is super-additive by a factor of $2.95\times$, replicating across all three seeds. This is the multiplicative-cancellation prediction of Section 5 made concrete. Second, deep BN layers carry the bulk of the effect: replacing only the last 5 BN layers (cumul. deep→shallow, k=5) costs 31.5pp of accuracy, while replacing only the first 5 (cumul. shallow→deep, k=5) costs only 5.7pp—a roughly $5\times$ asymmetry between deep and shallow halves of the same architecture. Third, the largest single-layer drops are the first BN layer (bn1, 4.0%) and the second-to-last residual BN layer (layer4.1.bn1, 4.9%), while residual-shortcut BN layers (active only at stride changes) contribute the least (full per-layer numbers in Appendix E).

4.6. Disentangling μ vs σ^2: which statistic carries the effect?

The same-class condition substitutes both class-conditional running means $\mu_\ell^{(c)}$ and class-conditional variances $(\sigma_\ell^{(c)})^2$ jointly. To test which of the two carries the effect, we re-evaluate SmallResNet on CIFAR-10 (91.25% global) under two further conditions: *μ-only* (load $\mu_\ell^{(c)}$, keep σ_ℓ^2 global) yields 76.66%; *σ^2-only* (load $(\sigma_\ell^{(c)})^2$, keep μ_ℓ global) yields 27.18%; the joint same-class condition yields 2.80%.

Three observations follow. *(i)* Neither statistic alone reproduces the catastrophe. *(ii)* The contributions are super-additive: an independent-multiplicative model on the accuracy ratio predicts $0.7666 \times 0.2718/0.9125 = 0.228$, an order of magnitude above the observed 0.028. *(iii)* σ^2-substitution alone is more damaging than μ-substitution alone—which the single-layer linearization of Section 5 does *not* predict (it attributes the effect to μ-cancellation since $a^{(c)} - \mu_c \approx 0$). The disagreement is consistent with the multi-layer compounding picture of Section 4.5: at a single layer μ-cancellation suffices to zero the class signal, but stacked through 20 BN layers the per-class $\sigma_\ell^{(c)}$ rescaling produces an independent class-dependent reweighting of feature magnitudes that the globally-trained head also cannot read. The two statistics together produce a class-dependent affine map of activation space that is far more damaging than either piece in isolation.

4.7. A non-monotone interpolation between class-conditional and global

We linearly interpolate the BN buffers between same-class ($\alpha = 0$) and global ($\alpha = 1$) and evaluate at each step (full 11-point sweep in Appendix F, Table 11). Classifier accuracy rises smoothly and monotonically from 3.4% to 91.0% (no phase transition), but the linear probe is *non-monotone*: high at both the same-class endpoint (99.9%) and global (90.9%), yet dropping to a minimum of 74.65% near $\alpha \approx 0.6$. That both endpoints admit high-quality linear separability while an intermediate point does not is constraining for any explanation of the effect: it suggests $\alpha = 0$ and $\alpha = 1$ are two qualitatively different geometric regimes rather than a uniform off-distribution move (Section 6).

5. An algebraic observation: where the class signal goes

The cross-condition ordering admits a one-line algebraic restatement in a one-layer linearization of the encoder-to-head interface. The argument below makes the asymmetric same $\ll$ wrong ordering an identity rather than a coincidence, in a single equation; the gap between this argument and a predictive model of deep nonlinear networks is examined at the end of this section.

Setup. Fix a single BN layer immediately before a linear head. Let $f(x) \in \mathbb{R}^d$ denote the pre-BN penultimate features; for a class-c input write $f(x) = a^{(c)} + \varepsilon(x)$ where $a^{(c)} = \mathbb{E}[f(x) \mid y = c]$ is the class-c centroid and ε has

class-conditional mean zero. Assume balanced classes so the global mean is $\mu_g = \frac{1}{C}\sum_c a^{(c)}$, and write σ_g for the global per-channel std (similarly μ_c, σ_c for class c, where $\mu_c \approx a^{(c)}$ in the limit of large class-c samples). The head produces logits $z_{c'}(x) = w_{c'} \cdot \hat{h}(x) + b_{c'}$ where $\hat{h}$ is the BN output. The head is trained against μ_g, σ_g and so satisfies, in expectation over class-c samples,

$$(w_c - w_{c'}) \cdot \frac{a^{(c)} - \mu_g}{\sigma_g} + (b_c - b_{c'}) > 0 \tag{1}$$

for all $c' \neq c$ (otherwise the network mispredicts on its training class).

Effect of substitution on the class margin. Under each substitution, the expected logit margin on class-c samples becomes:

$$\begin{aligned}
\Delta^{\text{global}}_{c\to c'} &= (w_c - w_{c'}) \cdot \frac{a^{(c)} - \mu_g}{\sigma_g} + (b_c - b_{c'}) \\
\Delta^{\text{wrong}(d)}_{c\to c'} &= (w_c - w_{c'}) \cdot \frac{a^{(c)} - \mu_d}{\sigma_d} + (b_c - b_{c'}) \\
\Delta^{\text{same}}_{c\to c'} &= (w_c - w_{c'}) \cdot \frac{a^{(c)} - \mu_c}{\sigma_c} + (b_c - b_{c'}) \\
&\approx 0 + (b_c - b_{c'}).
\end{aligned} \tag{2}$$

The first term in Δ^{same} vanishes because $\mu_c \approx a^{(c)}$ by construction: same-class statistics centre the class-c activation at zero, and any vector $w_c - w_{c'}$ projected onto 0 is 0. The trained head's correctness inequality (1) is therefore reduced to a pure bias inequality $b_c > b_{c'}$, which softmax-trained heads do not in general satisfy.

Three empirical patterns the identity captures.

1. *Same-class is uniquely damaging.* The substitution $\mu_g \to \mu_c$ erases the between-class component of the input to the head; $\mu_g \to \mu_d$ for $d \neq c$ only translates it. So $\Delta^{\text{same}} \approx b_c - b_{c'}$ while $\Delta^{\text{wrong}(d)}$ retains $(w_c - w_{c'}) \cdot (a^{(c)} - \mu_d)/\sigma_d$, a generically nonzero class signal. The ordering same $\ll$ wrong (Tables 1, 2) follows immediately.
2. *Floor-level accuracy.* The remaining bias term $b_c - b_{c'}$ reflects class-frequency calibration of a softmax head; in a balanced training set it is small in magnitude and need not be of the right sign. With C classes, predictions are dominated by within-class noise $w_{c'} \cdot \varepsilon/\sigma_c$, which is uncorrelated with the true class. The expected accuracy is at floor (chance or below it within seed noise), as observed: 3.5% vs $1/C{=}10\%$ on CIFAR-10—a clean below-chance result—and indistinguishable from $1/C$ on CIFAR-100 and Tiny-ImageNet-200 once seed noise is accounted for.
3. *Probe survives, often* exceeds *the global probe.* Under same-class, $\hat{h}(x) = \varepsilon(x)/\sigma_c$ with σ_c chosen per class. The class identity is fully encoded in the choice of σ_c (which is observed by the probe at training time but not by the original head's fixed weights). Moreover within-class variance is normalised away, leaving features that are *more* linearly separable than global features (where within-class variance is normalised by the class-blind σ_g). This matches the data: probe(same) $>$ probe(global) on SmallResNet (99.9% vs 90.9%), VGG-11 (98.4% vs 85.0%), and WRN-28-10 (99.8% vs 91.0%).

Multi-layer compounding. The same argument applied at L stacked BN layers gives an iterated cancellation: each layer's class-c statistics centre that layer's class-c mean activation. Architectures with more BN layers therefore lose the class signal at more depths, predicting that the same-class drop should grow with BN-layer count. The cross-architectural ordering reported in Section 4.4 (log-ratios of 1.15, 2.06, 3.25 for 4, 8, 20 BN layers) is consistent with this picture. The within-architecture layer-wise ablation (Section 4.5) is the strong evidence: across 3 seeds, single-layer drops sum to 29.4% while the all-layer drop is 86.8%—a $2.95\times$ super-additive compounding factor that replicates seed-to-seed—and deep BN layers contribute disproportionately, exactly as a multiplicative-cancellation account predicts.

What this argument does *not* do. The argument is not predictive: it restates the empirical ordering as an identity rather than deriving it. Three gaps are worth naming. First, the non-monotone probe trajectory (Section 4.7) is not captured by (2), which predicts a smooth, not U-shaped, dependence on α; the minimum near $\alpha \approx 0.6$ likely reflects nonlinear interactions with the fixed affine parameters γ, β. Second, the argument assumes one BN layer; the multi-layer evidence (Sections 4.4, 4.5) is consistent with multiplicative compounding but its quantitative form is left open. Third, the linearization attributes the drop to μ-cancellation alone, yet the disentangling experiment (Section 4.6) shows σ^2-substitution alone is also strongly damaging ($91\% \to 27\%$) and interacts super-additively with μ. We therefore use the identity as a constraint on the space of explanations, not a first-principles derivation of the asymmetry.

6. Discussion

The findings of Section 4 constrain how the same-class effect can be explained. Section 4.3 argues against two natural alternatives—the effect is not a generic running-statistic perturbation (GroupNorm, no running statistics, is unaffected) nor a logit-scale issue (temperature scaling halves ECE without recovering accuracy)—and Section 5 restates the pattern as a single equation, (2), in which the class-mean signal the head depends on is exactly cancelled by

same-class statistics and only translated by wrong-class statistics; that asymmetry is the source of the ordering.

An OOD-input reading is not enough. As Section 4.1 argues, a generic-OOD reading predicts the three substitutions hurt comparably—and if anything predicts same-class, the most accurate class-matched estimate, to be the mildest; instead same-class is the *most* damaging, by $> 18\times$ over wrong-class on CIFAR-10. Section 5 shows why: same-class is the unique direction in which the between-class activation component collapses to zero, not a generic OOD shift.

BN statistics as an implicit calibrator. A useful restatement: BN running statistics couple the encoder to the head, and same-class statistics maximally decouple them—zeroing the projection of $a^{(c)} - \mu_g$ onto the head's class-c direction w_c without changing the encoder. A linear refit (the probe) absorbs the residual structure, which is why probe accuracy stays high and, on BN-heavy networks, rises above the global baseline.

The non-monotone probe. The interpolation curve (Section 4.7) is where the linearization is incomplete: it predicts smooth, not U-shaped, probe accuracy in α. The empirical minimum near $\alpha \approx 0.6$ suggests the same-class and global endpoints are two qualitatively different organisations of activation space—at same-class, within-class variation is suppressed and features are unusually separable—with intermediate statistics interfering destructively rather than interpolating smoothly.

Connection to prior empirical work. The interpretation also suggests why AdaBN (Li et al., 2018) and TENT (Wang et al., 2021) work: they re-estimate BN statistics on the data the head must work with, recalibrating the encoder-to-head coupling rather than learning new features, so their benefit should be largely orthogonal to representation quality. The same lens flags a practical caution: post-hoc compression that quantizes or prunes BN buffers can have outsized, asymmetric accuracy effects without altering the underlying representation, and should be evaluated for it.

Methodological implication. A network at 0.17% accuracy on Tiny-ImageNet-200 may have a 100%-linearly-separable representation. Reading classification accuracy as a proxy for representation quality is unsafe in any setting in which the encoder-to-head alignment can be perturbed independently of the encoder—domain shift, adversarial inputs, normalization-layer interventions. Diagnosing such failures benefits from separately measuring representation quality (e.g., via a linear probe) and head-alignment (e.g., via ECE plus a temperature-scaling check).

7. Limitations and Conclusion

Limitations. (i) All experiments are on natural-image classification at scales up to Tiny-ImageNet-200. Whether the asymmetry holds on ImageNet-1k, fine-grained recognition, or non-vision domains is open. (ii) Architectures with non-BN normalization (LayerNorm, RMSNorm, transformers) carry no running statistics and are not directly addressed. (iii) Wrong-class is operationalized via the fixed mapping $c \mapsto (c + \lfloor C/2 \rfloor) \bmod C$; characterizing the full $C \times C$ surface of stat-class pairings would be a richer analysis. (iv) The layer-wise sensitivity (Section 4.5) is reported on SmallResNet only; whether the same depth-asymmetry pattern holds for VGG, WRN, and other BN-heavy architectures is left to future work. (v) When we replace running statistics we leave the affine parameters $(\gamma_\ell, \beta_\ell)$ unchanged; cleanly disentangling their contribution is left to future work. (vi) ECE with fixed binning has known sensitivities (Nixon et al., 2019); we use 15 bins following standard practice. (vii) We falsify one operationalization of the distribution-shift null—distance-monotone damage, contradicted by same $\ll$ wrong—but do not adjudicate every off-distribution baseline; perturbed-global, corrupted-input, and cross-dataset statistics probe different directions and are left to future work (the random-class condition supplies one non-class-structured baseline). (viii) The same-class intervention, and the same-class probe in particular, are conditioned on the true label; we therefore rely on the label-independent random-class probe (§4.2) to establish that the encoder survives, and we recommend a fully label-agnostic probe (trained on global-statistic features, evaluated on substituted features) as the cleaner protocol.

Conclusion. Substituting class-conditional BN running statistics at inference produces a sharp asymmetric ordering—same $\ll$ random/wrong $\ll$ global—that holds across four architectures and three datasets. The encoder remains class-discriminative under the substitution—linear probe $\geq 99.7\%$ under same-class, and $\geq 90.8\%$ under the label-independent random-class control that is not subject to label leakage through the choice of statistics (§4.2)—and the failure is not explained by a generic normalization-swap effect (GroupNorm models are unaffected) nor by a logit-scale issue (temperature scaling halves ECE but recovers 0% accuracy). A within-architecture layer-wise ablation shows the effect compounds super-additively at $2.95\times$ across BN layers, and a per-statistic disentangling shows neither μ nor σ^2 alone reproduces the catastrophe (μ-only 76.7%, σ^2-only 27.2%, joint 2.8%). We treat the asymmetry as a forensic observation about a specific intervention rather than a sweeping claim about the role of BN statistics; the data constrain—rather than determine—a mechanistic story, and we hope the asymmetry itself becomes a useful target for theoretical work on BN.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Benz, P., Zhang, C., Karjauv, A., and Kweon, I. S. Revisiting batch normalization for improving corruption robustness. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. Understanding batch normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Brock, A., De, S., Smith, S. L., and Simonyan, K. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning (ICML)*, 2021.

de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Frankle, J., Schwab, D. J., and Morcos, A. S. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020.

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

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. Revisiting batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Wu, Y. and He, K. Group normalization. In *European Conference on Computer Vision (ECCV)*, 2018.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

A. Experimental configuration

Table 5. Full experimental configuration.

Parameter	Value
SmallResNet	BasicBlock, $[2, 2, 2, 2]$, channels $[32, 64, 128, 256]$, 20 BN layers
SmallResNet64	BasicBlock, 5-stage, channels $[32, 64, 128, 256, 512]$, for 64×64
VGG-11-BN	Standard VGG-11 with BN, 8 BN layers
SimpleCNN	4-layer CNN, 4 BN layers
WideResNet-28-10	depth=28, widen_factor=10, dropout=0.3
Training epochs	20 (CIFAR-10), 25 (CIFAR-100, Tiny-ImageNet-200)
Batch size	128
Optimizer	SGD (momentum=0.9, weight_decay=5×10^{-4})
Learning rate	0.1, cosine annealing
Augmentation	RandomCrop(32, pad=4), RandomHorizontalFlip (CIFAR)
Normalization (input)	$\mu=(0.4914, 0.4822, 0.4465)$, $\sigma=(0.2470, 0.2435, 0.2616)$
Linear probe	Ridge ($\lambda=10^{-3}$, CIFAR-10) or LogisticRegression ($\geq$ 100-class), 80/20 split
ECE	15 equal-width bins
Temperature scaling	NLL-optimized scalar T on validation set
Seeds	42–51 (SR/CIFAR-10), 42–46 (SR/CIFAR-100), 42–44 (others)
Device	Apple Silicon (MPS backend)

B. Per-class accuracy deltas

On CIFAR-10/SmallResNet, per-class same-minus-global accuracy deltas (averaged over 10 seeds) range from -0.97 (*automobile*) to -0.72 (*cat*). All ten classes show deltas ≤ -0.72, ruling out the hypothesis that the collapse is concentrated in a small subset of classes.

C. Cross-architecture results in detail

Table 6. Cross-architecture results on CIFAR-10 with all four conditions and the linear probe (mean $\pm$ std, n=3 seeds for VGG/SimpleCNN/WRN).

Architecture	Condition	Acc. (%)	Conf. (%)	Probe (%)
VGG-11-BN	Global	85.44 ± 1.28	80.47 ± 1.78	85.00 ± 1.80
	Same-class	10.88 ± 1.91	10.29 ± 1.23	98.37 ± 0.63
	Wrong-class	62.96 ± 1.25	58.61 ± 1.31	99.57 ± 0.27
	Random-class	57.50 ± 3.03	52.29 ± 3.12	95.28 ± 0.78
SimpleCNN	Global	83.17 ± 0.23	75.68 ± 0.11	80.53 ± 1.44
	Same-class	26.33 ± 0.48	20.71 ± 0.18	54.68 ± 0.74
	Wrong-class	78.10 ± 0.46	65.79 ± 0.59	93.67 ± 0.27
	Random-class	72.11 ± 0.36	59.20 ± 0.33	90.82 ± 0.52
WRN-28-10	Global	90.61 ± 0.51	89.36 ± 0.43	91.02 ± 0.28
	Same-class	11.12 ± 0.90	11.04 ± 0.73	99.78 ± 0.09
	Wrong-class	64.06 ± 3.14	52.64 ± 3.51	99.60 ± 0.07
	Random-class	62.41 ± 1.76	51.21 ± 2.25	96.63 ± 0.44

D. Cross-dataset results in detail

Table 7. Full four-condition results on CIFAR-100 (SmallResNet, n=5 seeds, logistic-regression probe).

Condition	Acc. (%)	Conf. (%)	Probe (%)
Global	70.11 ± 0.57	63.98 ± 0.73	64.23 ± 1.34
Same-class	0.60 ± 0.49	0.94 ± 0.00	100.00 ± 0.00
Wrong-class	1.00 ± 0.00	1.00 ± 0.00	100.00 ± 0.00
Random-class	1.00 ± 0.63	1.00 ± 0.00	65.49 ± 8.75

Table 8. Full four-condition results on Tiny-ImageNet-200 (SmallResNet64, n=3 seeds, logistic-regression probe).

Condition	Acc. (%)	Conf. (%)	Probe (%)
Global	60.77 ± 0.30	53.75 ± 0.24	51.83 ± 0.55
Same-class	0.17 ± 0.24	0.49 ± 0.00	100.00 ± 0.00
Wrong-class	0.50 ± 0.00	0.50 ± 0.00	100.00 ± 0.00
Random-class	0.50 ± 0.00	0.50 ± 0.00	57.02 ± 3.41

E. Full layer-wise replacement

Table 9. Single-layer BN-statistic replacement on SmallResNet/CIFAR-10 (mean $\pm$ std, n=3 seeds). Each row: replace only the named BN layer with its class-conditional statistics; all other layers retain global statistics.

Idx	Layer	Drop (%)
0	bn1	4.03 ± 0.27
1	layer1.0.bn1	0.29 ± 0.28
2	layer1.0.bn2	0.81 ± 0.46
3	layer1.1.bn1	0.26 ± 0.28
4	layer1.1.bn2	0.41 ± 0.14
5	layer2.0.bn1	1.04 ± 0.29
6	layer2.0.bn2	1.00 ± 0.35
7	layer2.0.shortcut	0.32 ± 0.21
8	layer2.1.bn1	0.72 ± 0.19
9	layer2.1.bn2	0.70 ± 0.05
10	layer3.0.bn1	1.75 ± 0.05
11	layer3.0.bn2	1.90 ± 0.34
12	layer3.0.shortcut	0.21 ± 0.13
13	layer3.1.bn1	1.57 ± 0.48
14	layer3.1.bn2	1.86 ± 0.14
15	layer4.0.bn1	3.19 ± 0.40
16	layer4.0.bn2	2.72 ± 0.04
17	layer4.0.shortcut	0.46 ± 0.05
18	layer4.1.bn1	$\mathbf{4.91 \pm 1.01}$
19	layer4.1.bn2	1.27 ± 0.38
Sum of mean single-layer drops		29.40
Mean all-layer drop		86.78
Super-additive factor		$2.95\times$

Table 10. Cumulative replacement: shallow→deep replaces the first k BN layers (in forward order); deep→shallow replaces the last k. Mean $\pm$ std over 3 seeds, SmallResNet/CIFAR-10.

k	Shallow→Deep (%)	Deep→Shallow (%)
1	87.08 ± 0.36	89.84 ± 0.24
5	85.46 ± 1.43	59.65 ± 6.41
10	76.10 ± 0.33	23.74 ± 3.25
15	46.96 ± 0.88	11.14 ± 0.48
20	4.33 ± 1.32	4.33 ± 1.32

F. Full interpolation sweep

G. Detailed ECE and temperature-scaling sweep

Note that wrong-class statistics produce moderate ECE (0.064) despite substantially reduced accuracy: the network's confidence and accuracy remain roughly aligned, just at a lower level. Same-class statistics produce severely miscalibrated outputs (ECE 0.394) and accuracy near floor. The optimal temperature $T \approx 3.43$ indicates the same-class logits are substantially over-scaled.

Table 11. 11-point interpolation from same-class ($\alpha = 0$) to global ($\alpha = 1$) statistics on SmallResNet/CIFAR-10.

α	Accuracy (%)	Linear Probe (%)
0.0	3.44	99.90
0.1	7.13	99.70
0.2	13.41	98.67
0.3	23.14	94.85
0.4	36.47	87.10
0.5	51.61	77.60
0.6	65.17	74.65
0.7	75.26	77.88
0.8	82.62	83.00
0.9	87.64	87.53
1.0	91.01	90.85

Table 12. Expected Calibration Error by condition (SmallResNet, CIFAR-10, n=3 seeds).

Condition	Accuracy (%)	ECE
Global (baseline)	91.33 ± 0.12	0.025 ± 0.001
Same-class	2.87 ± 0.49	0.394 ± 0.013
Wrong-class	63.40 ± 4.26	0.064 ± 0.013

H. Per-seed tables

Table 13. Per-seed temperature-scaling on the same-class condition.

Seed	Glob. ECE	SC ECE	SC ECE (post-T)	T
42	0.026	0.382	0.150	3.456
43	0.024	0.412	0.148	3.428
44	0.026	0.388	0.154	3.416

Table 14. SmallResNet CIFAR-10 – Classification accuracy by seed.

Seed	Global	Same-Class	Wrong-Class	Random-Class
42	90.96	3.85	60.19	55.26
43	91.05	4.02	70.70	67.73
44	91.01	2.46	69.93	55.45
45	91.07	4.06	64.59	56.74
46	91.26	3.84	57.08	56.07
47	91.30	3.37	60.49	57.01
48	90.95	2.39	67.19	50.82
49	91.19	2.90	63.57	56.01
50	90.93	4.22	69.00	66.09
51	91.04	4.35	70.98	68.72

Table 15. SmallResNet CIFAR-10 – Linear probe accuracy by seed.

Seed	Global	Same-Class	Wrong-Class	Random-Class
42	90.30	99.90	100.00	96.30
43	90.65	99.80	100.00	96.85
44	91.10	99.90	100.00	97.05
45	89.65	99.90	100.00	96.90
46	90.80	100.00	100.00	98.30
47	91.80	99.95	100.00	99.35
48	90.80	99.90	100.00	96.80
49	91.15	99.90	100.00	98.90
50	91.00	99.65	100.00	96.65
51	91.30	99.90	100.00	99.15

Table 16. SmallResNet CIFAR-100 – Per-seed results (n=5).

Seed	Glob. Acc	SC Acc	Glob. Probe	SC Probe
42	70.95	1.00	65.65	100.00
43	69.68	1.00	64.85	100.00
44	70.21	0.00	65.00	100.00
45	69.31	1.00	61.80	100.00
46	70.42	0.00	63.85	100.00

Table 17. WideResNet-28-10 CIFAR-10 – Per-seed accuracy (n=3).

Seed	Global	Same-Class	Wrong-Class	Random-Class
42	90.17	10.44	64.21	62.49
43	90.35	12.40	67.83	64.53
44	91.32	10.53	60.15	60.22

Table 18. WideResNet-28-10 CIFAR-10 – Per-seed linear probe (n=3).

Seed	Global	Same-Class	Wrong-Class	Random-Class
42	91.40	99.85	99.70	96.30
43	90.90	99.85	99.55	96.35
44	90.75	99.65	99.55	97.25

Table 19. SmallResNet64 Tiny-ImageNet-200 – Per-seed results (n=3).

Seed	Glob. Acc	Glob. Conf	Glob. Probe	SC Acc	SC Probe
42	60.88	53.78	51.55	0.00	100.00
43	61.08	54.04	52.60	0.50	100.00
44	60.36	53.44	51.35	0.00	100.00