Residual Transport in Batch-Normalization Adaptation: Mechanism and Parameter-Free Correction

Anonymous authors
Paper under double-blind review

Abstract

Replacing BatchNorm (BN) statistics with moments estimated from an unlabeled target domain is a useful test-time adaptation baseline, yet partial replacement can help, do nothing, or harm. We isolate one mechanism behind this variability. If a BN layer keeps its source mean but adopts a target variance, the scale change transports the BN output mean by a signed, channel-specific amount $q_c = \gamma_c(\mu_{t,c} - \mu_{s,c})(\sigma_{t,c}^{-1} - \sigma_{s,c}^{-1})$. We estimate this quantity on a calibration split and test it on disjoint evaluation images. Across CIFAR-10-C and a bounded external ImageNet-C blur-family replication with pretrained ResNets, the calibration-only prediction closely tracks the held-out observation. On fully trained models, injecting the predicted $+q$ into an otherwise frozen network degrades downstream performance, and exact-vector channel permutation substantially attenuates the effect, confirming that channel alignment matters. Relative to variance-only replacement, aligned residual compensation shows positive aggregate gains at both training durations and on both external ImageNet ResNet architectures, while permuted compensation yields a smaller gain.

The mechanism yields a direct algorithm: Residual-Balanced BN (RB-BN) keeps the target scale and installs a mean that preserves the frozen target expected BN output, exactly realizing $-q$ through BN statistics alone. In an all-layer study on CIFAR-10-C and CIFAR-100-C with two independently trained 100-epoch checkpoints per dataset, all 15 corruptions, two calibration budgets, and disjoint evaluation images, sequential RB-BN improves accuracy over frozen inference by 13.77 pp and 14.48 pp on CIFAR-10-C and CIFAR-100-C, respectively; sequential AdaBN improves by 13.61 pp and 14.66 pp. RB-BN uses no mixing weight. The evidence supports residual transport as a measurable, channel-aligned causal component of BN adaptation and provides a parameter-free correction derived from that mechanism.

1 Introduction

Models deployed outside their training distribution encounter changes in acquisition, weather, rendering, and noise. Test-time adaptation (TTA) attempts to respond using unlabeled target inputs and no retraining data. These shifts can be synthetic or natural, and robustness measured in one setting need not transfer to another (Taori et al., 2020; Cubuk et al., 2020; Koh et al., 2021; Subbaswamy et al., 2021; Zhang et al., 2023). Batch normalization (BN) is an attractive interface: its running moments are compact state variables, and replacing them with target estimates can improve robustness under covariate shift (Ioffe & Szegedy, 2015; Li et al., 2017; Schneider et al., 2020; Nado et al., 2020; Benz et al., 2021). The same simplicity creates an attribution problem. A BN layer standardizes with a mean and a scale at once. Updating only one of them does not isolate a "mean effect" or a "variance effect" because the two parameters interact in the output.

We study this interaction in a controlled CIFAR-10-C setting. Suppose the target activation mean differs from the source running mean. If the target variance replaces the source variance while the source mean is retained, the mean mismatch is rescaled. The induced

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

change in the BN output mean, after subtracting the learned affine bias, is

$$q_c = \gamma_c(\mu_{t,c} - \mu_{s,c}) \left(\frac{1}{\sigma_{t,c}} - \frac{1}{\sigma_{s,c}} \right). \tag{1}$$

The sign determines whether the target scale amplifies or attenuates the existing mean mismatch. We call this channel-wise change residual transport; "residual" refers to the BN output after subtracting its affine bias and does not refer to a skip connection.

Equation 1 is an algebraic identity under population moments. Evaluating it and the resulting residual on the same samples would therefore provide weak evidence. Our question is whether finite, unlabeled calibration estimates identify a transport pattern that (i) appears on disjoint target images, (ii) changes downstream predictions when directly realized at the predicted channels, and (iii) can be partly removed when the same target variance is used. We separate these questions with a measurement, intervention, and compensation sequence.

First, we estimate q on D_{cal} and compare it with the change measured on disjoint D_{eval}. Second, we add the predicted vector to the output of a fixed BN layer while all running statistics and learned parameters remain frozen. Third, we compare the aligned vector with its sign reversal and with exact-value channel permutations. Finally, we compare aligned and permuted compensation under matched target variances, and repeat the core study at two training durations and on a non-residual BN architecture.

This produces a compact causal chain. A corruption changes the incoming activation mean and scale; partial BN replacement converts their interaction into a predictable output transport; realizing that transport changes downstream loss; and destroying its channel alignment attenuates the effect. The chain is deliberately local. Joint moment replacement also changes centering and scaling throughout the network, so residual transport should explain a pathway through adaptation rather than serve as a complete performance surrogate.

Our contributions are:

- We formalize a signed residual transport induced by partial BN replacement and estimate it from unlabeled calibration images disjoint from evaluation.
- We establish downstream relevance with output-level direction and exact-vector location interventions, then test aligned compensation under a matched target scale across training durations and architectures.
- We derive Residual-Balanced BN, a fixed-target-scale mean correction that preserves the frozen target expected BN output and exactly implements residual compensation through BN statistics.
- We audit the local chain on CIFAR-10-C and the complete ImageNet-C blur family, and evaluate all-layer RB-BN on CIFAR-10-C, CIFAR-100-C, and a bounded external ImageNet-C stress test with matched controls.

2 Related work

Prediction-time normalization. BN learns affine parameters while maintaining running activation moments for inference (Ioffe & Szegedy, 2015; Santurkar et al., 2018; Bjorck et al., 2018). Re-estimating those moments on target data is a simple form of covariate-shift adaptation and has shown strong results on common-corruption benchmarks (Li et al., 2017; Schneider et al., 2020; Nado et al., 2020; Benz et al., 2021). Other normalization choices expose different trade-offs in how statistics are pooled or constrained (Ioffe, 2017; Wu & He, 2018). Recent work mixes source and target statistics, learns domain-sensitive interpolation, estimates moments online, accumulates information over time, or rectifies layers selectively (You et al., 2021; Lim et al., 2023; Hu et al., 2021; Su et al., 2024; Vianna et al., 2025). These methods ask how to obtain better normalization statistics. We hold the estimation protocol fixed and ask what one partial replacement does inside a layer.

Optimization-based TTA. Tent updates BN affine parameters by entropy minimization (Wang et al., 2021); EATA adds sample selection and anti-forgetting regularization (Niu

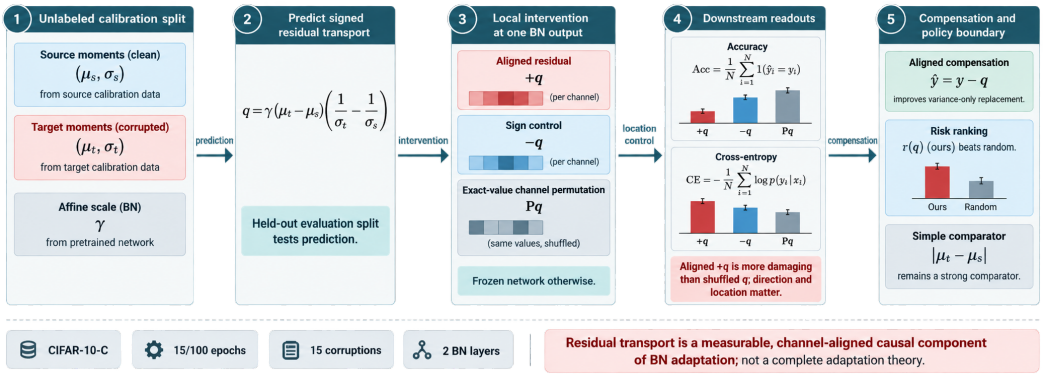

Figure 1: Overview of the evidence chain. Calibration-only moments predict a signed BN-output transport on held-out images; output-level interventions test direction and channel location; matched compensation tests whether removing the transport recovers variance-only performance. The final policy comparison is treated as an operational diagnostic rather than the mechanism claim itself.

et al., 2022). Other source-free and test-time methods use self-supervision, feature alignment, parameter-free output adjustment, contrastive learning, teacher–student updates, or stability controls (Sun et al., 2020; Liu et al., 2021; Liang et al., 2020; Boudiaf et al., 2022; Chen et al., 2022; Wang et al., 2022; Zhang et al., 2022; Niu et al., 2023; Yuan et al., 2023; Gong et al., 2022). These methods introduce optimization dynamics, stream-order effects, and possible degradation. Our principal experiment is calibration-only: it changes neither learned parameters nor source statistics permanently, and it uses an evaluation set disjoint from the calibration set.

Layer and channel selectivity. The behavior of normalization varies by layer, batch composition, and channel. TTN uses channel-wise domain-shift sensitivity when mixing moments (Lim et al., 2023); channel-selective normalization has also been used to limit failure under label shift (Vianna et al., 2025). More broadly, feature selectivity and representation bias can determine which shifts are consequential (Geirhos et al., 2019). Our result is complementary. We derive a signed intermediate, measure its realized value on held-out inputs, and test its location by permuting the complete intervention vector. A useful diagnostic need not be the best possible channel-selection policy.

3 Residual transport

3.1 Setting and notation

Let $z_c(x)$ denote the pre-BN activation in channel c. In inference mode,

$$y_c(x; \mu, \sigma) = \gamma_c \frac{z_c(x) - \mu_c}{\sigma_c} + \beta_c, \qquad \sigma_c = \sqrt{v_c + \epsilon}. \tag{2}$$

The source model stores (μ_s, σ_s). An unlabeled calibration set D_{cal} gives $(\widehat{\mu}_t, \widehat{\sigma}_t)$, while every reported evaluation metric is measured on disjoint D_{eval}. At one selected layer we compare frozen, mean-only, variance-only, and joint moment replacement.

For a distribution P, define the affine residual

$$r_c(P; \mu, \sigma) = \mathbb{E}_{x \sim P}[y_c(x; \mu, \sigma)] - \beta_c. \tag{3}$$

If the target pre-BN mean is μ_t, the frozen and variance-only residuals are

$$r_c^F = \gamma_c \frac{\mu_{t,c} - \mu_{s,c}}{\sigma_{s,c}}, \qquad r_c^V = \gamma_c \frac{\mu_{t,c} - \mu_{s,c}}{\sigma_{t,c}}. \tag{4}$$

Their difference is Equation 1. The calibration-only estimate is

$$\widehat{q}_c = \gamma_c(\widehat{\mu}_{t,c} - \mu_{s,c}) \left(\frac{1}{\widehat{\sigma}_{t,c}} - \frac{1}{\sigma_{s,c}} \right). \tag{5}$$

We compare it with the held-out observation

$$q_c^{\text{obs}} = \mathbb{E}_{x \in D_{\text{eval}}}[y_c(x; \mu_s, \widehat{\sigma}_t)] - \mathbb{E}_{x \in D_{\text{eval}}}[y_c(x; \mu_s, \sigma_s)]. \tag{6}$$

The discrepancy includes calibration sampling error and any difference between calibration and evaluation activation means.

3.2 Interventions and estimands

High correlation between Equations 5 and 6 establishes measurement fidelity, not downstream causality. We therefore intervene at the output of a fixed BN layer while the network otherwise uses frozen source statistics. This treatment–contrast framing follows the narrower intervention logic of randomized causal analysis, while making no claim that the full deployment shift is identified by this one manipulation (Rubin, 1974; Pearl, 2009):

$$\widetilde{y}_c(x) = y_c(x; \mu_s, \sigma_s) + a_c. \tag{7}$$

The aligned adverse-direction intervention uses $a = \widehat{q}$. The sign control uses $a = -\widehat{q}$. The location control uses $a = P\widehat{q}$, where P is an exact channel permutation. It preserves the complete multiset of values and hence the signs and both ℓ_1 and ℓ_2 norms. A difference between $\widehat{q}$ and $P\widehat{q}$ therefore identifies channel alignment rather than intervention magnitude.

The estimand is the paired change in accuracy and cross-entropy relative to frozen inference on the same evaluation images. Because later nonlinearities need not respond linearly or symmetrically, $-\widehat{q}$ is a directional control rather than an assumed inverse treatment. For compensation, we use the same target variance as variance-only replacement and compare an aligned residual correction with a permutation of the same correction vector.

3.3 Residual-Balanced BatchNorm

The intervention also has a direct, parameter-free BN implementation. Fix a target scale σ_t and seek a mean that preserves the expected frozen BN output on the target activation distribution. For channels with $\gamma \neq 0$, the unique solution is

$$\mu_{\text{RB}} = \mu_s + \left(1 - \frac{\sigma_t}{\sigma_s} \right) (\mu_t - \mu_s). \tag{8}$$

Indeed, if $\gamma \neq 0$ and $\sigma_t > 0$,

$$\mathbb{E}_t[y(z; \mu_{\text{RB}}, \sigma_t)] = \mathbb{E}_t[y(z; \mu_s, \sigma_s)]. \tag{9}$$

The result follows by substituting $\mathbb{E}_t[z] = \mu_t$ and solving the resulting linear equation in μ; the nonzero coefficient gives uniqueness. Moreover, it realizes the output-level compensation exactly for every activation z:

$$y(z; \mu_{\text{RB}}, \sigma_t) = y(z; \mu_s, \sigma_t) - q. \tag{10}$$

We call the resulting rule Residual-Balanced BN (RB-BN). In practice, target moments in Equation 8 are estimated from D_{cal}.

RB-BN is related to source–target interpolation, but is derived from a different constraint. Its mean can be rewritten as $\mu_{\text{RB}} = (\sigma_t/\sigma_s)\mu_s + (1 - \sigma_t/\sigma_s)\mu_t$, which is a channel-dependent interpolation or extrapolation coefficient. α-BN instead uses a chosen source weight to mix both the mean and standard deviation (You et al., 2021), while TTN obtains interpolation weights through a post-training sensitivity procedure (Lim et al., 2023). RB-BN fixes the target scale and derives that mean analytically; its effective coefficient is not a learned or validation-tuned hyperparameter, and it uses no gradient step, target label, or source-data replay.

We evaluate two estimators. Simultaneous RB-BN collects every target moment under the frozen source checkpoint and installs all RB-BN statistics in one pass. This is fast, but

changing an early BN layer changes later-layer inputs. Sequential RB-BN therefore collects each later layer's target moment after preceding layers have received their RB-BN statistics. Equation 9 is a one-layer identity for a fixed input distribution; it does not assert that the simultaneous all-layer update preserves every later activation expectation.

Budgeted ablation. If only K BN layers may change, we rank the simultaneous estimates by $E_\ell = \|q_\ell\|_2^2$ and install RB-BN only in the top-K layers. This exactly maximizes the declared additive local surrogate $\sum_{\ell \in S} E_\ell$ subject to $|S| \leq K$, by a direct exchange argument. It is not an accuracy-optimality theorem; we compare it with matched mean-shift, scale-shift, and random selection controls.

4 Experimental design

Core protocol. The local study uses all 15 CIFAR-10-C corruptions (Hendrycks & Dietterich, 2019), severities 1, 3, and 5, with disjoint 1,000-image evaluation splits and unlabeled calibration pools. It spans two fully trained 100-epoch CIFAR ResNets, two 15-epoch ResNets, and two PlainBN networks without shortcuts; the local intervention uses two prespecified early BN layers.

Local and all-layer matrices. The short and fully trained local matrices contain 480 and 240 model–corruption–budget–repeat conditions, respectively. We report paired accuracy and cross-entropy with corruption-cluster bootstrap intervals; target labels never select a layer, budget, or method. For all-layer RB-BN, we additionally use CIFAR-10-C and CIFAR-100-C at severity 3, two independently trained 100-epoch checkpoints per dataset, all 15 corruptions, budgets 32 and 256, two calibration repeats, and disjoint 1,000-image evaluation splits: 120 base conditions per dataset. Every one of the 15 BN layers is updated. We compare frozen inference, simultaneous and sequential AdaBN and RB-BN, and fixed-α interpolation at source weights 0.7 and 0.9, without validation tuning, target labels, gradients, or source-data replay. A separate finite-calibration matrix extends the same two checkpoints to budgets 8–256, three calibration draws, and confidence-gated controls; a third independently trained checkpoint tests a theory-motivated scale-gated variant (Appendix E).

External mechanism replication. We use ImageNet-C's complete blur family (four corruptions, five severities), torchvision pretrained ResNet-18 and ResNet-50, and a prespecified BN layer. Each condition uses 10,000 calibration and 10,000 disjoint evaluation images. Complete split construction, hook-direction audit, model details, and the statistical aggregation unit are in Appendix B.

5 Results

5.1 Calibration moments predict a held-out intermediate

The mean within-condition signed Spearman correlation is 0.966 for the short matrix and 0.969 for the fully trained matrix; the medians are above 0.99 (detailed rows in Appendix C). The short matrix includes both residual and non-residual architectures, while the fully trained matrix uses two independent ResNets. The result is therefore a transfer of an intermediate measured on D_{eval}, not a same-sample algebraic check.

The result establishes a reliable measurement of the local intermediate. It does not by itself imply that the intermediate matters for classification. The next experiment changes only that intermediate.

5.2 Direct intervention establishes downstream relevance and channel alignment

On fully trained CIFAR-10-C, adding $+\widehat{q}$ to the frozen BN output decreases accuracy by 2.64 pp and increases cross-entropy by 0.261 (detailed rows in Appendix C). The realized output shift agrees with the requested vector to numerical precision. The same signed vector, randomly permuted across channels, changes accuracy by only -0.87 pp and cross-entropy

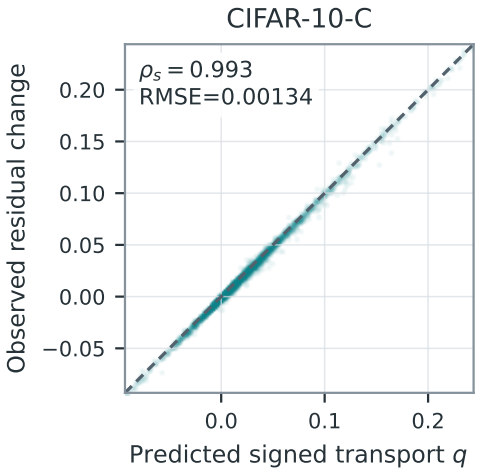

Figure 2: Calibration-only prediction of signed residual transport on disjoint evaluation images. Each point is a channel observation; the dashed line is identity. The relationship remains close to identity for both the short residual/non-residual matrix and the fully trained ResNet matrix.

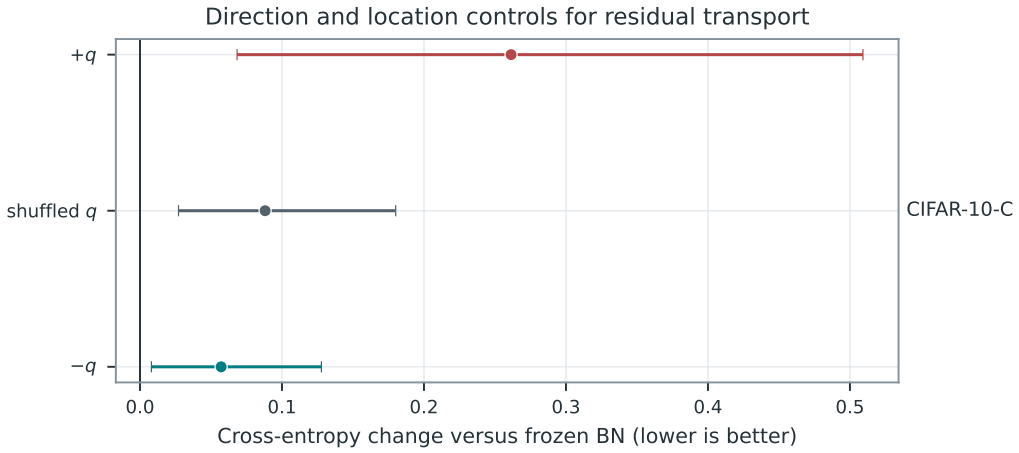

Figure 3: The complete local intervention chain. The top row shows direct $+\widehat{q}$, exact-value permutation, and sign reversal on fully trained models. The bottom row compares variance-only, aligned compensation, permuted compensation, and joint replacement at 15 and 100 epochs. Accuracy changes are in percentage points and cross-entropy changes are in nats per example; error bars are corruption-cluster 95% intervals.

by +0.088. The paired aligned-minus-shuffled difference is −1.99 pp in accuracy and +0.167 in cross-entropy, with corruption-cluster intervals excluding zero.

The sign control is direction-sensitive: $-\widehat{q}$ gives −0.70 pp accuracy and +0.057 cross-entropy. It does not perfectly undo $+\widehat{q}$, which is expected because the downstream network is non-linear and examples occupy different decision margins. The three-way comparison isolates two facts at once: the predicted transport can affect the network, and its effect is tied to the channels where it is realized.

5.3 Compensation recovers part of the variance-only effect

Variance-only replacement is a useful bridge between the local intervention and ordinary BN adaptation. Relative to variance-only, aligned compensation improves accuracy by 2.54 pp at 15 epochs and 3.43 pp at 100 epochs. The corresponding cross-entropy changes are

-0.149 and -0.272. Permuting the correction vector reduces the gain to 0.80 pp and 0.43 pp, respectively. The sign is the same at both training durations after correcting the hook conventions, which rules out the earlier apparent training-duration sign reversal.

Joint replacement is stronger than either local compensation reference: it improves accuracy by 6.01 pp at 15 epochs and 6.23 pp at 100 epochs relative to frozen inference. This comparison places the mechanism in the right causal role. Residual compensation addresses one transport induced by the variance update; joint adaptation also changes centering, scale, and downstream states throughout the network.

5.4 The mechanism survives architecture and corruption changes

The short matrix includes a plain convolutional BN network to test whether the phenomenon is an artifact of residual shortcuts. Both architectures have high transport correlations and positive joint and compensation effects (Appendix C). Effects vary by corruption, as expected for a product of mean shift, scale shift, affine gain, and downstream margin geometry rather than a corruption-independent constant.

5.5 The local mechanism replicates on ImageNet-C

The external replication extends the mechanism beyond CIFAR-scale models without changing the local estimand. Across the four blur types and five severities, the calibration-only transport prediction reaches mean Spearman 0.99974 for both ResNet-18 and ResNet-50. The result is measured on 10,000 held-out images per condition, with no overlap between calibration and evaluation identities; detailed values are in Appendix C.

The downstream direction is also consistent. Relative to variance-only replacement, aligned $-q$ compensation improves accuracy by 4.80 pp for ResNet-18 and 4.39 pp for ResNet-50, while the matched exact-vector shuffle improves accuracy by only 0.45 pp and decreases it by 0.65 pp, respectively. The aligned-minus-shuffled contrasts are $+4.36$ pp and $+5.03$ pp; their cluster-bootstrap intervals exclude zero. The reverse $+q$ injection decreases accuracy by 4.71 pp and 7.10 pp, and increases cross-entropy in both architectures. Joint replacement remains a strong reference, but the local compensation result shows that the channel-aligned transport is itself downstream-relevant in standard ImageNet models.

5.6 RB-BN turns the local mechanism into an all-layer rule

Figure 4 evaluates the rule derived in Section 3.3 as an all-layer adaptation algorithm. Sequential RB-BN improves over frozen inference by 13.77 pp on CIFAR-10-C and 14.48 pp on CIFAR-100-C. These gains are close to sequential AdaBN (13.61 pp and 14.66 pp), with paired method differences whose corruption-cluster intervals include zero. Thus the parameter-free rule is performance-competitive with sequential moment replacement, rather than a claim of uniform superiority. Its aggregate accuracy is descriptively higher than the two fixed-α controls in this study, without choosing an interpolation weight; this is not a tuning sweep over the α-BN family. Exact accuracies are in Appendix D.

The simultaneous comparison is an essential boundary check. RB-BN substantially reduces the failure of simultaneous AdaBN, but remains below frozen inference (-17.06 pp and -14.56 pp), whereas simultaneous α-BN is positive. This does not contradict the one-layer identity: an early simultaneous update changes the input distribution at every later layer, so the independent frozen-checkpoint moment estimates no longer describe those inputs. Sequential collection restores the distributional premise layer by layer.

Across 3,600 layer-condition evaluations, the calibration-set expectation identity holds to numerical precision; under frozen upstream statistics, its discrepancy on disjoint evaluation images decreases as the calibration budget grows (Appendix D). This audit is a statistic-level check, not an accuracy theorem.

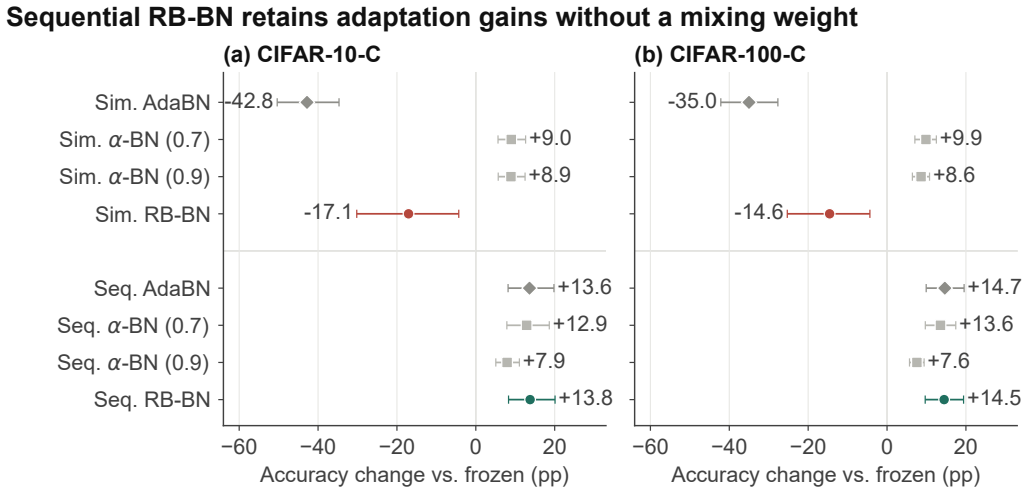

Figure 4: All-layer RB-BN comparison on held-out CIFAR-C images. Sequential RB-BN retains the adaptation gain of sequential AdaBN without selecting a source–target mixing weight. The poor simultaneous RB-BN result exposes the layer-composition boundary of the one-layer identity; fixed-α interpolation is a strong one-pass control. Points are aggregate paired accuracy changes from frozen inference and horizontal intervals are corruption-cluster 95% intervals.

5.7 An external all-layer stress test bounds the algorithm claim

At severity 3, we also compare all-layer sequential RB-BN, AdaBN, and frozen inference over the four ImageNet-C blur families. Each complete eight-cell grid uses one pretrained weight variant per architecture, 10 held-out evaluation images per class, and either 2 or 10 calibration images per class. RB-BN improves over frozen inference in six of eight cells at each budget. At 2/10 calibration images per class, its mean gains are +2.37 pp for ResNet-18 and +0.33 pp for ResNet-50; at 10/10, they are +2.35 pp and +0.54 pp, respectively. The four-family bootstrap intervals for these means include zero, and defocus blur is adverse at both budgets. RB-BN trails sequential AdaBN in all eight cells at each budget: the corresponding mean gaps are −5.97 pp and −5.64 pp at 2/10, and −6.06 pp and −5.94 pp at 10/10, for ResNet-18 and ResNet-50, respectively. Thus the fixed-layer mechanism replicates externally, while the CIFAR all-layer competitiveness does not extend to this pretrained blur setting. The two budgets have different evaluation identities, so their numerical difference is descriptive. Complete cell values and intervals are in Appendix C.

5.8 Finite-calibration controls separate residual correction from moment adaptation

At fixed severity 3, a confidence-gated scale lets us compare residual-balanced centering with a source-mean variance-only control (CVO) and a confidence-gated joint-moment control (Confidence-AdaBN). Across two independently trained checkpoints, 15 corruptions, five budgets, and three calibration draws, sequential Confidence-RB-BN (CRB-BN) exceeds CVO by 0.19–0.23 pp on CIFAR-10-C and 0.31–0.36 pp on CIFAR-100-C; cross-entropy decreases at every budget. Confidence-AdaBN remains stronger than CRB-BN. On a third, held-out training seed, retaining the raw mean with the gated scale improves over CRB-BN at most budgets, but again does not surpass Confidence-AdaBN. The local same-scale correction is algebraic; sequential policies collect later-layer moments under different upstream states, so their accuracy differences are paired policy outcomes rather than an exact whole-network decomposition. Full intervals and audits are in Appendix E.

6 Discussion: what the experiments identify

The evidence follows one chain: calibration-only moments predict signed transport on held-out images; direct $+q$ changes downstream loss; exact-vector permutation attenuates that effect; and aligned compensation recovers more variance-only performance than a matched permutation. The same fixed-layer pattern holds in pretrained ImageNet ResNets. RB-BN implements the correction through BN statistics and is competitive with sequential AdaBN on both CIFAR-C datasets. Its two complete external all-layer grids improve over frozen inference in six of eight cells but trail AdaBN throughout. The simultaneous ablation shows why the layer-local identity need not compose when earlier updates change later inputs. Confidence controls support residual relevance at small budgets, while joint adaptation remains a stronger performance reference. The claim is a local mechanism and design principle, without a global mediation percentage or network-wide identity.

7 Limitations

The local CIFAR mechanism study uses small models and early BN layers. The ImageNet-C replication and severity-3 all-layer stress test cover only blur corruptions and one pretrained weight variant per ResNet architecture, not independent ImageNet training seeds. The confidence variant adds one training seed per CIFAR dataset. Intervals resample observed corruption families, not model populations, and the calibration pools range from 8–256 CIFAR images to 10,000 ImageNet images; non-i.i.d. streams, label shift, and continual target changes may alter moment quality (Gong et al., 2022; Yuan et al., 2023). Accuracy and loss are reported together because they capture different outcomes. Simultaneous all-layer RB-BN and transport-energy channel ranking remain scope checks; the sequential algorithm and local mechanism claims do not rely on either producing the best downstream score.

8 Conclusion

Partial BN adaptation couples target mean shift and scale replacement. The signed residual transport is label-free, appears on disjoint images, changes predictions at its channel locations, and is partly removed by aligned compensation. RB-BN turns this correction into a parameter-free sequential rule that retains the adaptation gain of sequential AdaBN on two CIFAR-C datasets. Residual transport is therefore a reproducible local causal component and a concrete design principle for BN adaptation.

AI assistance disclosure

Generative AI tools assisted with hypothesis refinement, experimental design, mathematical exposition and derivation checking, code and figure preparation, literature organization, and English prose. The authors reviewed and verified all AI-assisted material, made the final scientific decisions, and take responsibility for the paper's claims, experiments, and artifacts.

Reproducibility statement

The project archives fixed calibration and evaluation indices, checkpoint hashes, condition-level completion markers, channel statistics, per-image losses and predictions, and the actual injected vectors. The protocol4 short matrix contains 480 completed conditions and the fully trained matrix contains 240. The all-layer RB-BN study has 120 completed base conditions per CIFAR-C dataset, each with paired method predictions; the identity audit covers 3,600 layer-condition evaluations. The confidence-gated two-checkpoint matrix validates 900/900 complete conditions with 11 paired methods; the separate seed-2 Scale-RB-BN matrix validates 450/450 with 13 methods. Both audit per-image predictions, losses, split disjointness, script and checkpoint hashes, and numerical identities, and resample corruption families rather than images. The ImageNet-C fixed-layer mechanism replication

contains 40 completed conditions across two pretrained ResNets, four blur types, and five severities, with 400,000 paired evaluation images in total. The two external severity-3 all-layer grids each audit eight complete paired cells and 10,000 evaluation images per cell, with 10,000 corruption-family bootstrap draws per grid. Protocols and generated audit reports are archived in paper/ and results/; the numerical experiment code is in scripts/. The anonymous code supplement contains first-party scripts and protocol documents; datasets, model weights, checkpoints, and large per-image outputs are excluded and must be obtained or regenerated to reproduce every numerical result.

References

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. In IEEE Winter Conference on Applications of Computer Vision, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Benz_Revisiting_Batch_Normalization_for_Improving_Corruption_Robustness_WACV_2021_paper.html.

Johan Bjorck, Carla Gomes, Bart Selman, and Kilian Q. Weinberger. Understanding batch normalization. arXiv preprint arXiv:1806.02375, 2018. URL https://arxiv.org/abs/1806.02375.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8344–8353, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Boudiaf_Parameter-Free_Online_Test-Time_Adaptation_CVPR_2022_paper.html.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 295–305, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Contrastive_Test-Time_Adaptation_CVPR_2022_paper.html.

Ekin D. Cubuk, Norman Mu, Dan Hendrycks, Balaji Lakshminarayanan, Barret Zoph, and Justin Gilmer. Augmix: A simple data processing method to improve robustness and uncertainty. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap. Chapman and Hall/CRC, 1993. doi: 10.1007/978-1-4899-4541-9.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In Advances in Neural Information Processing Systems, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/ae6c7dbd9429b3a75c41b5fb47e57c9e-Abstract-Conference.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations, 2019. URL https://arxiv.org/abs/1903.12261.

Xuefeng Hu, Gokhan Uzunbas, Sirius Chen, Rui Wang, Ashish Shah, Ram Nevatia, and Ser-Nam Lim. Mixnorm: Test-time adaptation through online normalization estimation. arXiv preprint arXiv:2110.11478, 2021. URL https://arxiv.org/abs/2110.11478.

Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In Advances in Neural Information Processing Systems, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/c54e7837e0cd0ced286cb5995327d1ab-Abstract.html.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning, pp. 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In ICLR 2017 Workshop Track Proceedings, 2017. URL https://openreview.net/forum?id=Hk6dkJQFx.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 6028–6039. PMLR, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In International Conference on Learning Representations, 2023. URL https://arxiv.org/abs/2302.05155.

Yuejiang Liu, Parth Kothari, Bastien Germain van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In Advances in Neural Information Processing Systems, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/b618c3210e934362ac261db280128c22-Abstract.html.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963, 2020. URL https://arxiv.org/abs/2006.10963.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Proceedings of the 39th International Conference on Machine Learning, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In Advances in Neural Information Processing Systems, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/905056c1ac1dad141560467e0a99e1cf-Abstract.html.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In Advances in Neural Information Processing Systems, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/85690f81aadc1749175c187784afc9ee-Abstract.html.

Zixian Su, Jingwei Guo, Kai Yao, Xi Yang, Qiufeng Wang, and Kaizhu Huang. Unraveling batch normalization for realistic test-time adaptation. Proceedings of the AAAI Conference on Artificial Intelligence, 2024. doi: 10.1609/AAAI.V38I13.29436. URL https://ojs.aaai.org/index.php/AAAI/article/view/29436.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pp. 2611–2619. PMLR, 2021. URL https://proceedings.mlr.press/v130/subbaswamy21a.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 9229–9248. PMLR, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In Advances in Neural Information Processing Systems, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

Pedro Vianna, Muawiz Sajjad Chaudhary, Paria Mehrbod, An Tang, Guy Cloutier, Guy Wolf, Michael Eickenberg, and Eugene Belilovsky. Channel-selective normalization for label-shift robust test-time adaptation. In Proceedings of The 3rd Conference on Lifelong Learning Agents, volume 274 of Proceedings of Machine Learning Research, pp. 514–533. PMLR, 2025. URL https://proceedings.mlr.press/v274/vianna25a.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In International Conference on Learning Representations, 2021. URL https://arxiv.org/abs/2006.10726.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Yuxin Wu and Kaiming He. Group normalization. In Proceedings of the European Conference on Computer Vision, pp. 3–19, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Yuxin_Wu_Group_Normalization_ECCV_2018_paper.html.

Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. arXiv preprint arXiv:2110.04065, 2021. URL https://arxiv.org/abs/2110.04065.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15922–15932, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yuan_Robust_Test-Time_Adaptation_in_Dynamic_Scenarios_CVPR_2023_paper.html.

Haoran Zhang, Harvineet Singh, Marzyeh Ghassemi, and Shalmali Joshi. Why did the model fail? attributing model performance changes to distribution shifts. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 41550–41578. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23ai.html.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In Advances in Neural Information Processing Systems, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

A Additional implementation details

Architecture. The small ResNet uses a 32-channel stem followed by three stages with two basic blocks each and widths 32, 64, and 128. Downsampling occurs in the first block of stages two and three. Each basic block has two 3×3 convolutions with BN; shortcuts use a 1×1 convolution and BN when shape changes. The PlainBN model has six convolution–BN–ReLU blocks with widths $32, 32, 64, 64, 128, 128$, max pooling after blocks two and four, global average pooling, and no skip connections.

Target moments. Moments are collected from normalized calibration images in batches of 128. Source moments and affine parameters are restored after every condition. The selected BN layer is adapted while the rest of the network remains frozen for local experiments. The compensation vector is broadcast over examples and spatial positions.

Ranking. For severity 3, we evaluate budgets 32 and 256 and retain fractions 25% and 50%. Three deterministic random subsets are averaged before paired comparison. Risk, absolute mean deviation, standardized mean deviation, affine-change, and standard-deviation shrinkage are computed from calibration statistics. The primary random control uses the same correction values and changes only the selected channel indices.

B Full experimental protocol

Local CIFAR-C matrices. For each CIFAR-10-C corruption, 1,000 images form the fixed evaluation split and the remaining 9,000 form the calibration pool; no calibration identity occurs in evaluation. Severity 1 and 5 use 256 calibration images, while severity 3 uses 32 and 256. We draw two calibration repeats. The short matrix has two 15-epoch ResNets and two independently trained six-convolution PlainBN networks; the fully trained matrix has two 100-epoch ResNets. We preselect two early BN layers per architecture. Each condition archives moment estimates, channel statistics, intervention vectors, logits, predictions, and per-image losses.

All-layer algorithm matrix. At severity 3, each CIFAR-10-C and CIFAR-100-C base condition combines two trained checkpoints, 15 corruptions, two budgets, and two calibration repeats. All 15 BN layers receive each method's statistics. The fixed-α control mixes source and target means and standard deviations at source weights 0.7 and 0.9. Simultaneous methods estimate every layer under the frozen checkpoint; sequential methods collect a later layer only after prior layers have received their method-specific moments.

Hook and aggregation audit. The short-study hook subtracts its offset input, whereas the fully trained hook adds its shift input. Thus both use the signed value that realizes the same $-\widehat{q}$ compensation. Random ranking uses the identical signed correction vector as risk ranking. Completion markers and realized output means verify the direction; the 100-epoch audit found 240/240 completed conditions, no issues, and maximum injection-fidelity RMSE 4.22×10^{-9}. Repeats and random permutations are averaged within model, corruption, severity, budget, and layer before summary. The local protocol uses 5,000-draw corruption-cluster bootstrap intervals, whereas the all-layer RB-BN figures use 10,000 percentile resamples of the 15 corruption-family means (Efron & Tibshirani, 1993); these intervals quantify variation across the observed corruption set, not uncertainty over arbitrary model populations.

External ImageNet-C replication. The official ImageNet-C blur archive contains defocus, glass, motion, and zoom blur at severities 1–5. We use torchvision ImageNet-1K pretrained

Table 1: Calibration-only prediction of signed residual transport on disjoint evaluation images. Entries are means and medians of within-condition Spearman correlations; RMSE is in BN-output units.

Matrix	n	mean ρ_s	median ρ_s	mean RMSE
15-epoch, ResNet + PlainBN	960	0.966	0.993	0.0016
100-epoch, CIFAR-10 ResNet	240	0.969	0.993	0.0011

Table 2: Output-level interventions on fully trained CIFAR-10-C. Values are changes relative to frozen inference with corruption-cluster 95% bootstrap intervals; accuracy is in percentage points and cross-entropy is in nats per example.

Intervention	Baseline	Accuracy change (pp)	Cross-entropy change (nats)
$+\widehat{q}$	frozen	−2.64 [-5.03, -0.72]	0.26 [0.07, 0.51]
shuffled $\widehat{q}$	frozen	−0.87 [-1.67, -0.31]	0.09 [0.03, 0.18]
$-\widehat{q}$	frozen	−0.70 [-1.67, -0.03]	0.06 [0.01, 0.13]
variance-only	frozen	−0.03 [-2.31, 2.06]	0.11 [-0.08, 0.34]
joint	frozen	6.23 [4.11, 8.43]	−0.31 [-0.42, -0.20]

ResNet-18 and ResNet-50, fixing the selected BN layer before inspecting outcomes. Ten images per WNID form calibration and another disjoint ten form evaluation, yielding 10,000 images in each split per condition. Labels are used only for scoring. The external study is a mechanism replication over one corruption family, rather than a complete ImageNet-C algorithm benchmark or an independent ImageNet training-seed study.

External all-layer stress test. At severity 3, separate complete grids use 2 or 10 calibration images per WNID and 10 disjoint evaluation images per WNID. All 1,000 classes and four blur types are included. Frozen inference, sequential AdaBN, and sequential RB-BN share evaluation identities within each cell; 10,000-draw intervals resample the four blur-family clusters. Each architecture has one pretrained weight variant. The evaluation block follows the calibration block in a deterministic image ordering, so the two budgets have different evaluation identities and are not a controlled calibration-budget intervention.

C Detailed local and external results

D All-layer RB-BN audit and budgeted selection

Identity audit. Table 10 isolates the exact statistic-level claim from classification. For every all-layer condition, it evaluates the expected-output identity with the calibration moments used to construct the update, then repeats the calculation on disjoint evaluation images while upstream BN statistics remain frozen. The calibration and algebraic errors are at numerical precision. The held-out discrepancy is finite-sample estimation error and declines with the calibration budget.

Budgeted layer selection. For a fixed budget of one, two, or four layers, we compare transport energy with affine mean-shift energy, log-scale-shift energy, and four deterministic random selections. All selected layers receive exactly the same RB-BN update; only the rule selecting their locations changes. Figure 9 reports paired accuracy change from frozen inference. Transport energy often helps over frozen inference, but it does not consistently exceed the simpler controls. This is why Proposition 2 is stated for the declared transport-energy surrogate rather than downstream accuracy.

Table 3: Moment replacement and residual compensation across training duration. Accuracy changes are in percentage points and cross-entropy changes are in nats per example. The aligned and permuted compensation rows are compared with variance-only replacement; all other rows use frozen inference.

Epochs	Intervention	Baseline	Accuracy change (pp)	Cross-entropy change (nats)
15	variance-only	frozen	−0.15 [-1.45, 1.09]	0.05 [-0.04, 0.17]
15	aligned compensation	variance_only	2.54 [0.78, 4.63]	−0.15 [-0.29, -0.04]
15	permuted compensation	variance_only	0.80 [0.22, 1.46]	−0.05 [-0.10, -0.01]
15	joint	frozen	6.01 [3.93, 8.18]	−0.25 [-0.34, -0.17]
100	variance-only	frozen	−0.03 [-2.26, 2.08]	0.11 [-0.08, 0.34]
100	aligned compensation	variance_only	3.43 [0.65, 6.54]	−0.27 [-0.53, -0.05]
100	permuted compensation	variance_only	0.43 [-0.04, 1.06]	−0.04 [-0.08, -0.00]
100	joint	frozen	6.23 [4.08, 8.35]	−0.31 [-0.41, -0.20]

Table 4: Short-training architecture check. Joint gain is relative to frozen inference; aligned compensation gain is relative to variance-only replacement.

Architecture	n	mean ρ_s	joint gain (pp)	compensation gain (pp)
plain	480	0.969	7.80	3.82
resnet	480	0.963	6.01	2.54

E Finite-calibration controls and held-out scale-gated ablation

Confidence-gated construction. For each BN channel, calibration images are the independent sampling units for the spatial first and second moments. Let $\widehat{\mu}$, $\widehat{\sigma}$, se_μ, and se_σ denote the estimated moments and image-cluster delta-method standard errors. With the fixed two-sided normal quantile $z = 1.9599639845400536$, define

$$\begin{aligned} d_\mu &= \mathrm{sign}(\widehat{\mu} - \mu_s)[|\widehat{\mu} - \mu_s| - z\,\mathrm{se}_\mu]_+, \\ d_\sigma &= \mathrm{sign}(\widehat{\sigma} - \sigma_s)[|\widehat{\sigma} - \sigma_s| - z\,\mathrm{se}_\sigma]_+, \\ \mu_C &= \mu_s + d_\mu, \qquad s_C = \max(\sigma_s + d_\sigma, \sqrt{\epsilon}), \\ v_C &= \max(s_C^2 - \epsilon, 10^{-5}), \qquad \widetilde{\sigma}_C = \sqrt{v_C + \epsilon}. \end{aligned} \tag{11}$$

The 95% confidence level is fixed before evaluating the held-out labels and is an approximate marginal delta-method choice, not simultaneous coverage over channels. Confidence-variance-only BN (CVO) installs (μ_s, v_C); Confidence-AdaBN installs (μ_C, v_C). Confidence-RB-BN (CRB-BN) installs

$$\mu_{\mathrm{CRB}} = \mu_s + (1 - s_C/\sigma_s)(\mu_C - \mu_s), \qquad v_{\mathrm{CRB}} = v_C. \tag{12}$$

When the variance floor is inactive, the same-moment local contrast between CRB-BN and CVO is exactly $-\gamma(\mu_C - \mu_s)(s_C^{-1} - \sigma_s^{-1})$ for every input activation. The floor introduces only the archived numerical discrepancy. If the image-cluster event $|\widehat{\mu} - \mu_t| \le z\,\mathrm{se}_\mu$ holds, $|\mu_t - \mu_C| \le 2z\,\mathrm{se}_\mu$, giving a conditional local residual bound of $2|\gamma| z\,\mathrm{se}_\mu\,|s_C^{-1} - \sigma_s^{-1}|$. This is an approximation under the sampling assumptions, not a bound on classification accuracy.

Two-checkpoint mechanism matrix. The formal v5 matrix uses two independent 100-epoch checkpoints per CIFAR-C dataset, all 15 standard corruptions at severity 3, budgets $B \in \{8, 16, 32, 64, 256\}$, three paired calibration draws, and 1,000 disjoint evaluation images. It contains 900/900 valid condition directories, each with 11 paired methods. The audit recomputes accuracy and cross-entropy from archived per-image predictions and losses, checks calibration/evaluation separation and provenance hashes, and finds no exclusions. Repeats are averaged within checkpoint and corruption; checkpoints are then averaged within corruption; the 95% percentile bootstrap resamples 15 corruption-family means. These intervals describe variation across the observed corruption set, not a model population.

Table 12 reports CRB-BN minus CVO on paired evaluation images. The accuracy improvement is positive and the cross-entropy difference is negative in every dataset–budget cell. Confidence-AdaBN is nevertheless a stronger joint-proxy reference, as Table 11 shows. The

Table 5: External ImageNet-C blur-family replication on pretrained ResNets. Each row aggregates 20 corruption–severity conditions. Accuracy changes are percentage points relative to variance-only replacement; cross-entropy changes are in nats. Corruption-cluster intervals are shown in Figure 5 and archived with the condition-level results.

Model	ρ_s	$-q$ Δacc.	$-q$ ΔCE	shuffle Δacc.	aligned–shuffle Δacc.	$+q$ Δacc.
ResNet-18	0.99974	+4.80	−0.969	+0.45	+4.36	−4.71
ResNet-50	0.99974	+4.39	−0.275	−0.65	+5.03	−7.10

Table 6: ImageNet-C blur-family results at severity 3 for pretrained ResNet-18 and ResNet-50. Entries are absolute top-1 accuracy (%); the mean is the unweighted average across the four blur families. The final column gives the paired RB–AdaBN mean accuracy difference in percentage points with the audit's 95% percentile corruption-family cluster-bootstrap interval (10,000 resamples).

Model	Sequential method	Defocus	Glass	Motion	Zoom	Mean	RB–AdaBN paired mean Δ accuracy (pp), 95% CI
ResNet-18	Frozen	25.18	14.22	25.87	27.84	23.28	
	Sequential AdaBN	23.74	23.30	38.91	40.50	31.61	
	Sequential RB–BN	16.57	17.67	32.76	35.58	25.64	−5.9675 [−6.7850, −5.2275]
ResNet-50	Frozen	43.28	17.31	44.27	42.95	36.95	
	Sequential AdaBN	34.22	31.56	53.28	52.60	42.91	
	Sequential RB–BN	26.34	26.22	48.93	47.62	37.28	−5.6375 [−7.1550, −4.5975]

Notes. Calibration uses 2 images per class; one pretrained weight variant is used per architecture. The complete audit grid has 8 valid model–blur cells. Intervals resample the four corruption-family clusters ($n = 4$); they are not seed confidence intervals. Paired mean RB–AdaBN cross-entropy differences (nats; lower is better), with the same 95% family-cluster bootstrap intervals, are ResNet-18 +0.6070 [+0.3793, +0.8347] and ResNet-50 −0.0464 [−0.1041, +0.0374].

sequential methods recollect later-layer moments after their own previous updates. Thus the table is a comparison of deployable policies; the exact same-scale $-q$ identity applies locally when the incoming moments are held common, and it is not an identity between full networks.

Held-out Scale-RB-BN ablation. The local residual bound suggests retaining the raw calibration mean while gating only the scale. To match the effective denominator after the variance floor, Scale-RB-BN installs

$$\mu_{\mathrm{SRB}} = \mu_s + (1 - \widetilde{\sigma}_C/\sigma_s)(\widehat{\mu} - \mu_s), \qquad v_{\mathrm{SRB}} = v_C. \tag{13}$$

At one layer and a fixed incoming distribution, its expected BN output equals frozen inference when the mean is the calibration proxy $\widehat{\mu}$. For the unknown target mean μ_t, the exact residual is $\gamma(\mu_t - \widehat{\mu})(\widetilde{\sigma}_C^{-1} - \sigma_s^{-1})$; on the same approximate image-cluster event, its absolute value is at most $|\gamma| z\,\mathrm{se}_\mu\,|\widetilde{\sigma}_C^{-1} - \sigma_s^{-1}|$. This one-standard-error expression versus CRB-BN's two-standard-error expression motivates the variant but gives no global performance ordering. Both the identity and its $-q$ compensation were checked numerically in all completed seed-2 conditions; the largest archived local discrepancy is below 5×10^{-7}.

The Scale-RB formula and comparison plan were fixed in the theory audit before inspecting the independent seed-2 outcomes. One newly trained 100-epoch checkpoint per CIFAR-C dataset was then evaluated over the same 15 corruptions, five budgets, three draws, and 1,000-image disjoint evaluation split, for 450/450 complete conditions with 13 paired methods and no exclusions. This is a separate confirmation matrix and is not pooled into the two-checkpoint v5 intervals. Table 13 and Figures 12–13 show that Scale-RB-BN improves over CVO and generally over CRB-BN, but Confidence-AdaBN retains the stronger classification result, especially on CIFAR-100-C. It therefore serves as a theory ablation; sequential RB-BN remains the main parameter-free algorithm.

Table 7: ImageNet-C blur-family results at severity 3 for pretrained ResNet-18 and ResNet-50, using 10 calibration images per class and 10 evaluation images per class. One pretrained weight variant is used per architecture. Entries are absolute top-1 accuracy (%); the mean is the unweighted average over the four blur families. The paired RB–BN mean accuracy effects versus Frozen and AdaBN are reported in percentage points with 95% percentile intervals from resampling the four corruption-family clusters ($n = 4$; 10,000 bootstrap draws). These estimates describe the 10/10 protocol; comparison with the 2/10 run does not identify a causal calibration-budget effect because evaluation identities differ across budgets.

Model	Sequential method	Defocus	Glass	Motion	Zoom	Mean	RB–BN paired Δ accuracy (pp) estimate [95% family-cluster CI]
ResNet-18	Frozen	26.46	14.94	25.95	28.06	23.85	
	Sequential AdaBN	24.30	24.15	39.41	41.18	32.26	
							vs Frozen: +2.3500 [−5.1400, +7.8300]
	Sequential RB–BN	17.09	18.05	33.50	36.17	26.20	vs AdaBN: −6.0575 [−6.8850, −5.2825]
ResNet-50	Frozen	43.03	18.19	44.22	42.11	36.89	
	Sequential AdaBN	34.65	32.17	53.39	53.27	43.37	
							vs Frozen: +0.5425 [−11.0475, +7.1950]
	Sequential RB–BN	26.20	26.23	48.88	48.41	37.43	vs AdaBN: −5.9400 [−7.5525, −4.6850]

Notes. Each method is evaluated on the same audited identities within a cell. The family-cluster intervals summarize variation over four blur families, not independent model seeds; they do not support a causal comparison of calibration budgets.

Table 8: Channel selection is compared at matched layer, budget, fraction, and condition. Accuracy differences are in percentage points and cross-entropy differences are in nats per example; the intervals are corruption-cluster 95% bootstrap intervals. Risk beats random, but its paired margin over the simpler absolute mean-shift score is small and its accuracy interval includes zero.

Epochs	Contrast	Budget fraction	Accuracy difference (pp)	CE difference (nats)
15	risk vs random	25%	2.09 [1.03, 3.38]	-0.151
15	risk vs random	50%	2.09 [1.11, 3.24]	-0.139
15	risk vs $\|\Delta\mu\|$	25%	0.17 [-0.07, 0.52]	-0.012
15	risk vs $\|\Delta\mu\|$	50%	0.25 [-0.03, 0.69]	-0.017
15	$\|\Delta\mu\|$ vs random	25%	1.92 [0.71, 3.32]	-0.139
15	$\|\Delta\mu\|$ vs random	50%	1.84 [0.69, 3.14]	-0.122
100	risk vs random	25%	2.34 [0.75, 4.36]	–
100	risk vs random	50%	1.76 [0.62, 3.24]	–
100	risk vs $\|\Delta\mu\|$	25/50%	0.20 [-0.02, 0.44]	-0.011

F Evaluation protocol and audit

The short matrix has four models, 15 corruptions, four severity–budget combinations, two repeats, and two selected layers: $4 \times 15 \times 4 \times 2 \times 2 = 960$ layer-condition groups. The fully trained matrix has $2 \times 15 \times 4 \times 2 \times 2 = 480$ layer-condition groups. Each condition includes frozen and moment-replacement baselines, direct interventions, matched permutations, and ranking rows. Accuracy is reported in percentage points and cross-entropy is the mean per-image negative log likelihood.

For a method contrast, repeats are averaged at fixed model, corruption, severity, budget, and layer. The bootstrap resamples corruption names with replacement; all other rows belonging to a sampled corruption travel together. This preserves the dependence induced by evaluating multiple layers and severities of the same corruption.

G Generated artifacts

All main numerical tables are generated from the audited protocol4 directories by scripts/build_protocol4_artifacts.py. Figures are generated by scripts/plot_protocol4_figures.py; the external ImageNet-C figure is generated by scripts/plot_imagenetc_blur_resnets.py. Each is written as vector PDF and 300-dpi PNG. The figure set includes transport prediction, direct intervention and compensation, ranking

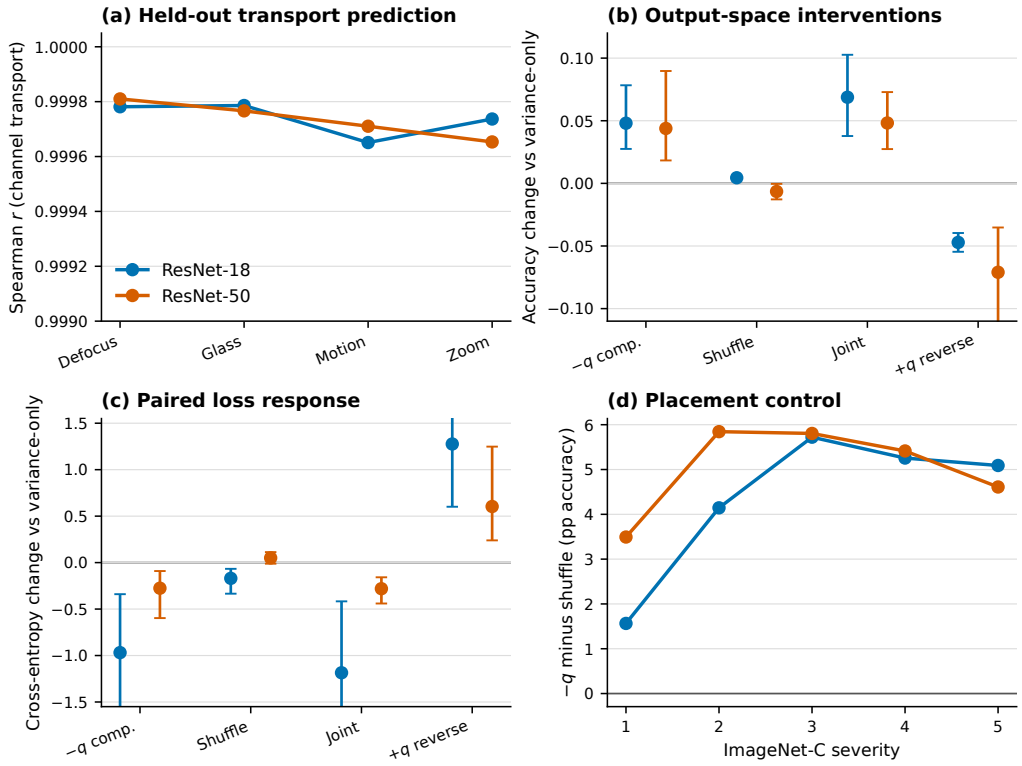

Figure 5: External ImageNet-C blur-family replication on pretrained ResNet-18 and ResNet-50. (a) Calibration-only channel transport prediction on held-out images, in additive BN-output units. (b–c) Accuracy changes (percentage points) and cross-entropy changes (nats per example) relative to variance-only replacement, with corruption-cluster 95% intervals. (d) The aligned $-q$ correction exceeds the exact-vector shuffle at every severity on average.

Table 9: RB-BN algorithm comparison on held-out CIFAR-C images. Accuracy is in percent; Δ is the paired change from frozen inference in percentage points. Each entry averages two independently trained checkpoints, 15 corruptions, two calibration budgets, and two calibration repeats.

	CIFAR-10-C		CIFAR-100-C	
Method	Accuracy	Δ	Accuracy	Δ
Frozen	71.1	+0.0	43.0	+0.0
Sim. AdaBN	28.3	-42.8	8.0	-35.0
Sim. α-BN (0.7)	80.1	+9.0	52.8	+9.9
Sim. α-BN (0.9)	80.0	+8.9	51.6	+8.6
Sim. RB-BN	54.0	-17.1	28.4	-14.6
Seq. AdaBN	84.7	+13.6	57.6	+14.7
Seq. α-BN (0.7)	84.0	+12.9	56.5	+13.6
Seq. α-BN (0.9)	79.0	+7.9	50.5	+7.6
Seq. RB-BN	84.9	+13.8	57.5	+14.5

controls, corruption heterogeneity, training duration, and the ImageNet-C/ResNet external replication.

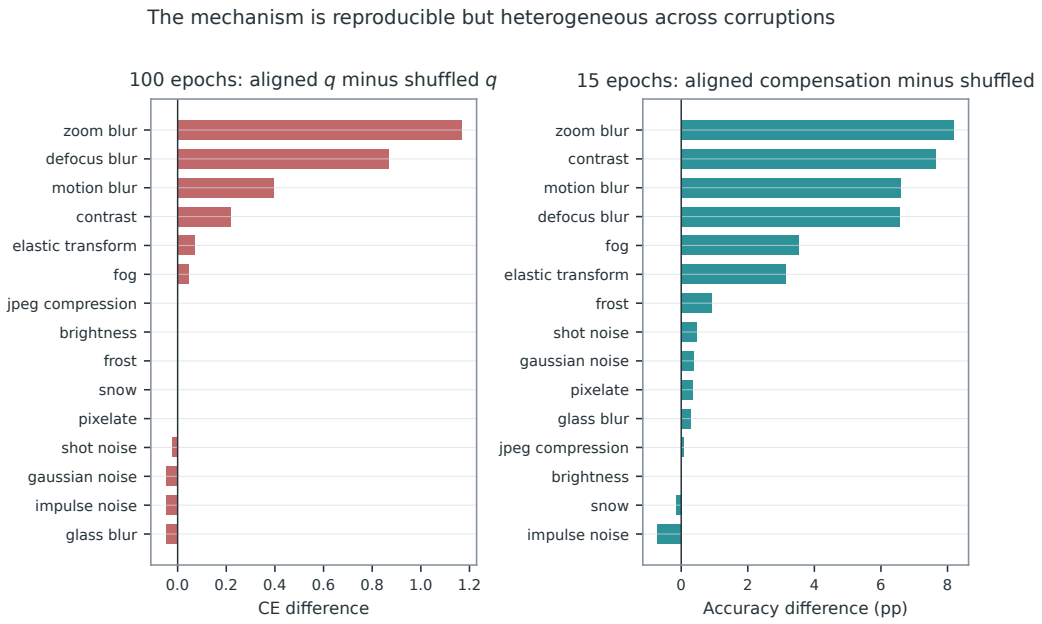

Figure 6: Per-corruption heterogeneity. Left: cross-entropy excess (nats per example) of aligned $+q$ over its exact-value permutation for the fully trained matrix. Right: accuracy gain (percentage points) of aligned compensation over a permuted correction in the short matrix. Values aggregate over models, severities, budgets, repeats, and selected layers.

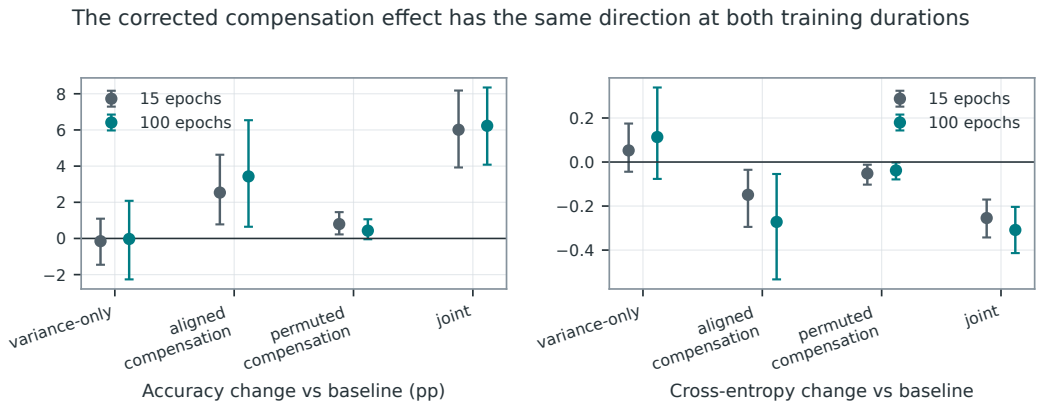

Figure 7: Training-duration comparison after sign correction. Accuracy changes are in percentage points and cross-entropy changes are in nats per example. The aligned compensation effect has the same direction at 15 and 100 epochs; the joint reference remains larger. Error bars are corruption-cluster 95% intervals.

Table 10: Audit of the RB-BN expectation-preservation identity. Calibration values use the moments that construct RB-BN; held-out values use a disjoint 1,000-image evaluation split under frozen upstream statistics. Values are mean absolute BN-output differences except the two maximum columns.

Dataset	$\lvert D_{\text{cal}}\rvert$	Calibration mean	Held-out mean	Held-out max	Algebraic max
CIFAR-10-C	32	4.9e-09	2.1e-03	1.84e-01	1.0e-07
CIFAR-10-C	256	4.8e-09	7.5e-04	5.49e-02	1.2e-07
CIFAR-100-C	32	7.0e-09	2.8e-03	3.99e-01	2.3e-07
CIFAR-100-C	256	7.0e-09	9.7e-04	1.14e-01	2.1e-07

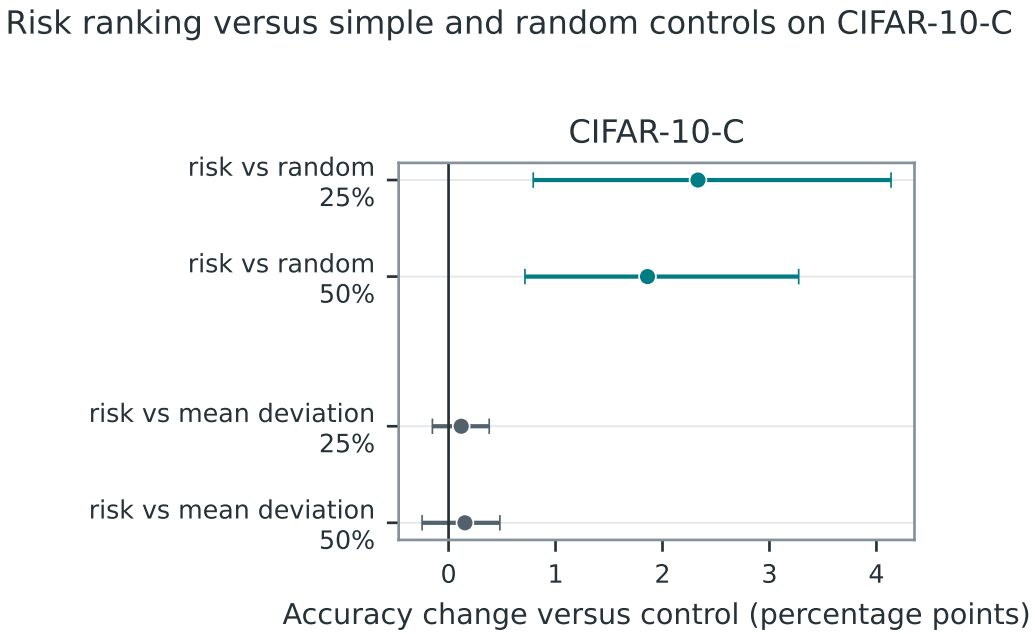

Figure 8: Fixed-layer risk ranking compared with matched controls. Accuracy differences are in percentage points and intervals are corruption-cluster 95% intervals. The random control receives the same signed correction values and changes only their channel locations.

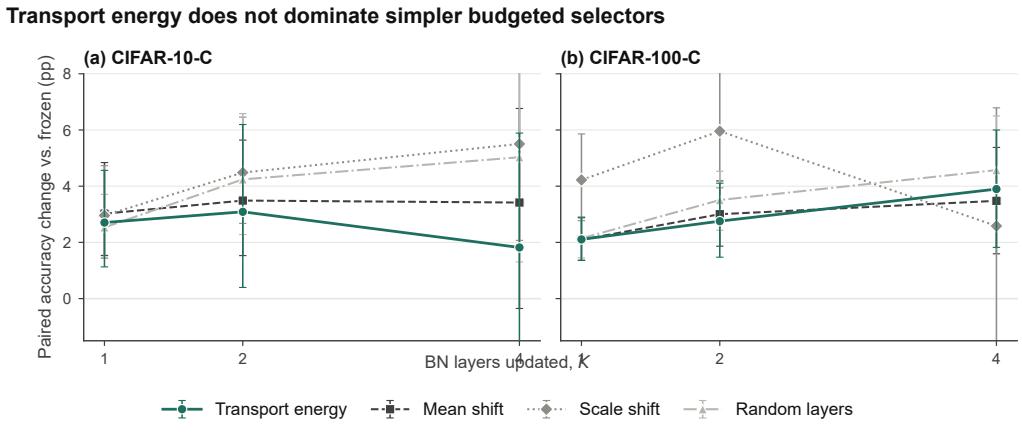

Figure 9: Budgeted RB-BN selection ablation. Every selector installs the same RB-BN statistics at its selected layers; only the label-free location rule differs. Although transport-energy selection improves over frozen inference in many settings, it has no consistent independent accuracy advantage over mean-shift, scale-shift, or repeated random selection. Points are aggregate paired changes from frozen inference and intervals are corruption-cluster 95% intervals.

Table 11: Confidence-aware sequential normalization on held-out CIFAR-C images. Entries are top-1 accuracy (%). Each cell averages three calibration draws within each of two independently trained checkpoints and then averages the 15 corruption-family means. CRB-BN is the residual-balanced correction at a confidence-gated scale; CVO holds the source mean at that same style of scale. Family-bootstrap intervals are shown in Figure 10.

Dataset	B	Frozen	AdaBN	RB-BN	CVO	Confidence-AdaBN	CRB-BN
CIFAR-10-C	8	71.1	78.9	80.8	81.6	82.1	81.8
	16	71.1	82.1	83.0	82.5	83.5	82.7
	32	71.1	83.8	84.2	83.5	84.4	83.7
	64	71.1	84.7	84.9	84.1	85.0	84.3
	256	71.1	85.5	85.4	84.9	85.5	85.1
CIFAR-100-C	8	43.0	49.6	51.6	53.4	54.8	53.7
	16	43.0	53.6	54.7	55.0	56.6	55.3
	32	43.0	56.2	56.5	56.0	57.5	56.3
	64	43.0	57.8	57.7	56.8	58.3	57.1
	256	43.0	58.9	58.3	57.6	58.9	57.9

Table 12: Confidence-gated residual correction versus the scale-only CVO construction. Entries are sequential policy outcomes; each policy recollects downstream moments after its own preceding updates. The final columns are paired CRB-BN minus CVO effects with 95% percentile bootstrap intervals over 15 corruption-family averages. Positive accuracy and negative cross-entropy favor CRB-BN.

Dataset	B	CVO acc.	CRB acc.	Δ acc. (pp)	Δ CE (nats)
CIFAR-10-C	8	81.61	81.82	+0.21 [+0.05, +0.41]	-0.013 [-0.027, -0.002]
	16	82.53	82.74	+0.21 [+0.03, +0.43]	-0.012 [-0.025, -0.003]
	32	83.47	83.67	+0.20 [+0.06, +0.39]	-0.012 [-0.024, -0.002]
	64	84.10	84.33	+0.23 [+0.02, +0.52]	-0.011 [-0.023, -0.003]
	256	84.90	85.09	+0.19 [+0.04, +0.39]	-0.011 [-0.022, -0.002]
CIFAR-100-C	8	53.38	53.73	+0.35 [+0.12, +0.65]	-0.033 [-0.068, -0.007]
	16	54.96	55.27	+0.31 [+0.09, +0.60]	-0.030 [-0.060, -0.007]
	32	56.00	56.32	+0.31 [+0.06, +0.67]	-0.028 [-0.057, -0.006]
	64	56.75	57.11	+0.36 [+0.09, +0.68]	-0.027 [-0.053, -0.006]
	256	57.59	57.90	+0.31 [+0.03, +0.66]	-0.028 [-0.055, -0.007]

Table 13: Held-out seed-2 Scale-RB-BN ablation on CIFAR-C. The two absolute columns are top-1 accuracy (%). The final columns are paired Scale-RB-BN effects with 95% percentile bootstrap intervals over 15 corruption-family means after three calibration draws are averaged. Confidence-AdaBN remains the stronger joint-proxy reference. Absolute CVO and CRB-BN curves are shown in Figure 12.

Dataset	B	Scale-RB acc.	Conf.-AdaBN acc.	Δ Scale–CVO (pp)	Δ Scale–CRB (pp)
CIFAR-10-C	8	82.44	82.54	+0.46 [+0.19, +0.84]	+0.22 [+0.08, +0.36]
	16	83.49	83.92	+0.52 [+0.17, +0.96]	+0.23 [+0.07, +0.43]
	32	84.20	84.64	+0.50 [+0.20, +0.89]	+0.17 [+0.08, +0.27]
	64	84.76	85.20	+0.47 [+0.19, +0.81]	+0.15 [+0.09, +0.22]
	256	85.38	85.68	+0.27 [+0.06, +0.55]	+0.01 [-0.03, +0.05]
CIFAR-100-C	8	53.34	54.25	+0.79 [+0.31, +1.38]	+0.40 [+0.21, +0.62]
	16	54.87	56.10	+0.60 [+0.26, +1.03]	+0.32 [+0.18, +0.48]
	32	55.89	57.12	+0.55 [+0.19, +1.06]	+0.26 [+0.13, +0.45]
	64	56.57	57.81	+0.45 [+0.14, +0.89]	+0.13 [+0.04, +0.22]
	256	57.43	58.44	+0.39 [+0.12, +0.72]	+0.09 [+0.04, +0.14]

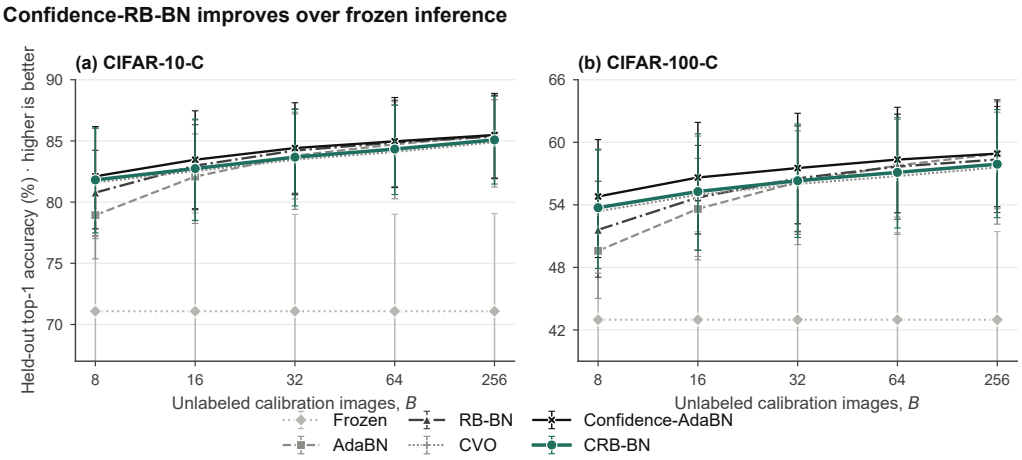

Figure 10: Sequential confidence-aware normalization on CIFAR-C at severity 3. Each point averages three calibration draws within each of two independently trained checkpoints and then averages 15 corruption families. Error bars are 95% percentile intervals over the family means. The comparison includes frozen inference, AdaBN, RB-BN, matched scale-only CVO, Confidence-AdaBN, and CRB-BN.

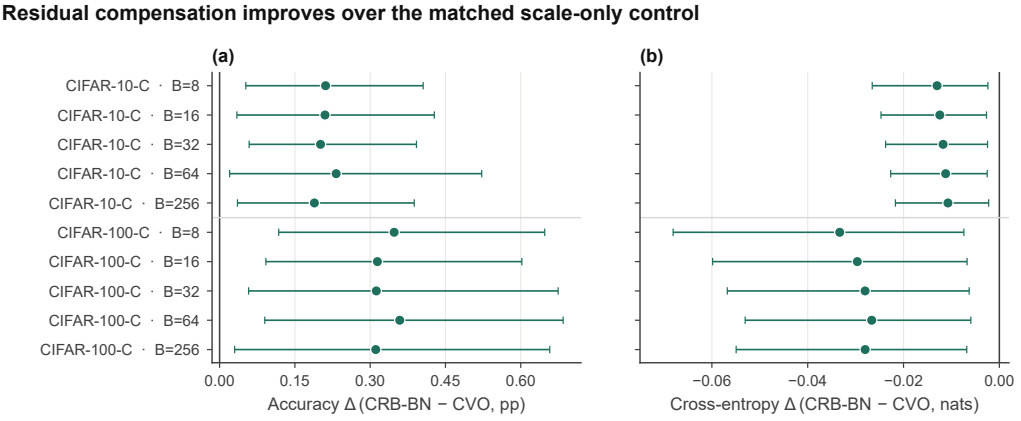

Figure 11: Paired CRB-BN minus CVO effects for accuracy and cross-entropy. Positive accuracy and negative cross-entropy favor CRB-BN. Points are means and whiskers are 95% percentile bootstrap intervals over 15 corruption-family averages, after averaging calibration draws and checkpoints within each family.

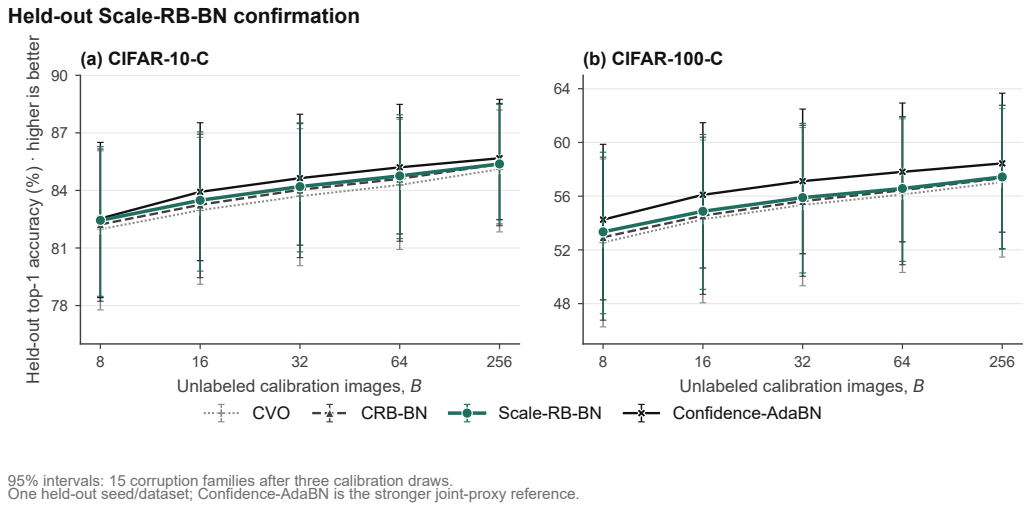

Figure 12: Held-out seed-2 Scale-RB-BN ablation on CIFAR-C at severity 3. Curves show sequential accuracy after three calibration draws are averaged within each corruption; error bars resample 15 corruption families. The one-checkpoint-per-dataset design is a held-out algorithm check and does not support training-seed population inference.

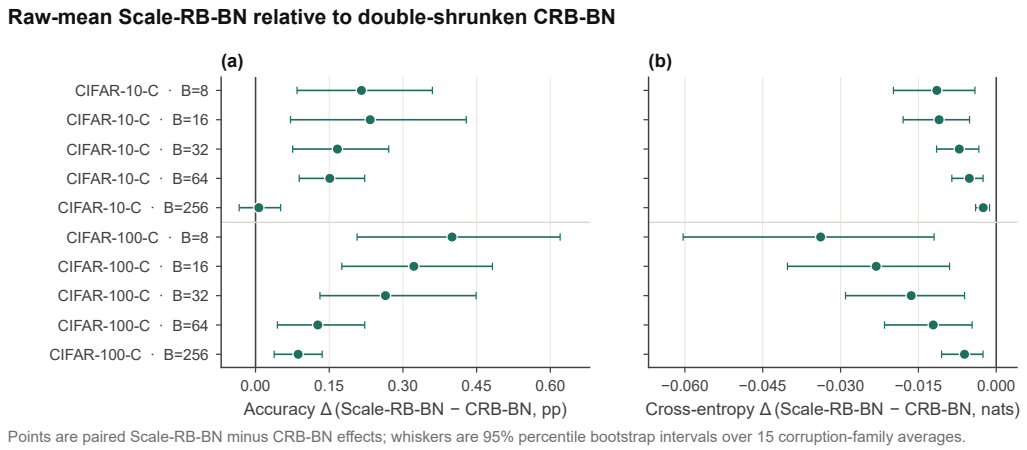

Figure 13: Paired Scale-RB-BN minus CRB-BN accuracy and cross-entropy on the independent seed-2 checkpoint. The whiskers are 95% percentile intervals over 15 corruption-family means after three calibration draws are averaged. The scale-only and joint-moment controls remain in Table 13.