TRANSFERABLE RANKINGS, SHIFTED ACTION BOUNDARIES: THE VALUE OF BATCHNORM MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

BatchNorm (BN) running statistics retain information from earlier domains, but when should a model carry that state into a new domain? We measure its value through paired carry/reset interventions with frozen weights and identical target batches. Across 12 CIFAR-C receiver checkpoints and all 210 directed corruption transitions per checkpoint, an external effect map correlates with receiver effects ($r = 0.759$, $\rho = 0.721$), yet its zero-threshold rule loses 1.44 percentage points relative to resetting and is harmful at every checkpoint. On a locked 150/60 edge split, a single threshold learned from the other 11 receivers gains 0.83 points on held receivers and held edges, close to the 0.86 points from a receiver-specific affine fit; the zero threshold loses 1.02 points on the same edges. The transferred threshold also gains 0.93 points over resetting on 120 closed-loop streams with known domain identities and boundaries, without using held-receiver labels. State and layer interventions show that predictive utility is nonlinear even though BN-buffer differences decay exponentially under matched updates. These results identify a shared shift in the useful carry/reset threshold across the tested receivers: much of the source ranking remains useful once the threshold is calibrated.

1 INTRODUCTION

Batch normalization (BN; Ioffe & Szegedy, 2015) retains running means and variances from past inputs. In continual adaptation, these statistics can help normalize a new domain or introduce a mismatch. Resetting removes the inherited state but also discards information that may still be useful. Existing methods adapt, interpolate, retain, or restore normalization statistics (Schneider et al., 2020; Lim et al., 2023; Kang et al., 2024; Lim et al., 2026). We ask whether measurements of when BN memory helps in one trained system can guide carry/reset decisions in another.

A source score may rank transitions accurately while placing the zero of predicted benefit incorrectly. Its correlation with receiver effects would then overstate its value as a carry/reset rule. We evaluate both the transferred ranking and the decisions obtained by thresholding it, and test whether a threshold learned from other receivers can recover useful decisions.

We define the directed carry/reset effect

$$G(A \to B) = \mathrm{Acc}_B(\text{carry state from } A) - \mathrm{Acc}_B(\text{reset to source state}), \tag{1}$$

where both arms adapt their BN buffers on identical target batches and all network weights remain fixed. This paired intervention measures the finite-sample effect of the incoming BN state without weight updates, pseudo-labels, teachers, or replay buffers. Following the paired state-intervention framework of Sevetlidis & Pavlidis (2026), we compare these effects across independently trained receivers.

We measure all 210 directed transitions among 15 corruptions in 12 CIFAR-C receiver checkpoints. The source map is correlated with receiver effects, but its zero-threshold rule loses accuracy at every checkpoint. A threshold learned from the other 11 receivers produces positive gains on each held receiver's locked test edges, close to those of a receiver-specific affine fit. The same transferred threshold improves closed-loop accuracy by 0.93 pp when carry/reset decisions change subsequent states. Separate state interventions characterize the nonlinear response of predictions to BN buffers and test how donor acquisition reduces dependence on earlier history.

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.

The contribution is an empirical account of BN-memory transfer at three levels: effect rankings, carry/reset decisions on held receivers and edges, and decisions in streams with action-dependent state. The held-out comparisons establish that much of the decision gain can be recovered with a shared threshold, without fitting to labels from the receiver on which it is used.

2 Related work

Adapting normalization statistics. AdaBN and subsequent analyses establish that replacing source statistics can alleviate covariate shift (Li et al., 2018; Schneider et al., 2020), while feature alignment can also degrade predictions (Burns & Steinhardt, 2021). Source–target mixtures such as α-BN (within Core) and TTN trade adaptation against preservation of source statistics (You et al., 2021; Lim et al., 2023). DUA and MixNorm estimate normalization statistics online (Mirza et al., 2022; Hu et al., 2021); TEMA extends estimation beyond a single batch (Su et al., 2024); UnMix and MemBN retain multiple or historical statistics (Tomar et al., 2024; Kang et al., 2024). ResiTTA regularizes estimated BN statistics toward source values (Zhou et al., 2025), while ERSK tracks global BN statistics to detect domain changes and trigger source restoration (Niloy et al., 2024). These works design mechanisms for estimating, mixing, retaining, or restoring normalization state. We instead measure the paired utility of an incoming BN state and ask whether its ranking and induced action boundary transfer across receiver systems.

Continual adaptation and restoration. Tent adapts through entropy minimization (Wang et al., 2021); EATA, SAR, and RoTTA address efficiency, reliability, and challenging streams (Niu et al., 2022; 2023; Yuan et al., 2023). NOTE targets temporally correlated non-i.i.d. streams with instance-aware normalization (Gong et al., 2022). CoTTA combines teacher-based adaptation with stochastic restoration (Wang et al., 2022); RDumb highlights pretrained resets over long horizons (Press et al., 2023); AR-TTA combines replay with dynamic BN-statistics interpolation (Sójka et al., 2024), while ASR develops adaptive and selective reset strategies (Lim et al., 2026). MECTA combines memory-efficient continual adaptation with adaptive normalization (Hong et al., 2023). We use a frozen-weight intervention to isolate the contribution of BN state from the additional mechanisms present in full CTTA algorithms.

Evaluation, memory, and utility. TTAB documents model dependence and model-selection difficulties in test-time adaptation (Zhao et al., 2023). Recent work examines condition-dependent CIFAR-10-C outcomes (Guha Majumdar & Saha, 2026), accumulated teacher-state errors (Sójka et al., 2026), and latency-constrained adaptation utility (Sreeram et al., 2026). A training-memory survey formulates paired function-space state interventions and carry/reset primitives (Sevetlidis & Pavlidis, 2026). We apply this framework to incoming BN buffers and evaluate whether their measured effects support useful decisions across trained systems.

3 Measuring the value of incoming BN state

State intervention. Let θ denote frozen parameters, s_0 the pretrained BN state, and s_A the state after donor acquisition on domain A. For each target domain B, fork two copies with identical θ, initializing the carry arm with s_A and the reset arm with s_0. Both arms process the same target batches in the same order. For each batch, update BN running statistics in training mode, then predict that batch in evaluation mode; dropout remains disabled. Labels are used only to score outcomes. Effects are reported in percentage points (pp).

Data and receivers. We use the 15 common corruptions of CIFAR-10-C and CIFAR-100-C at severity 5 (Hendrycks & Dietterich, 2019). A receiver cell is one dataset, architecture, and training seed: two datasets, ResNet-20 and WRN-40-2, and three seeds yield 12 cells. Each cell measures all $15 \times 14 = 210$ directed non-self transitions under three donor acquisition orders. Averaging over orders gives 2,520 cell–edge effects from 7,560 raw measurements. Donor acquisition uses 3,200 images, batch size 64, and BN momentum $m = 0.1$; target updates use the same batch size and momentum. The primary outcome covers the first 128 target images (two updates), capturing early memory utility before further EMA attenuation. Secondary analyses use all 512 target images (eight updates; Section 6).

Table 1: Information available to each analysis. Receiver labels, when used, come from offline profiling; retrospective oracles additionally access the outcomes they score. Domain identities and boundaries are known; streams are introduced in Section 6.

Analysis	Receiver information	Held-out quantity
External source sign	None; frozen external effect map	Receiver outcomes
Receiver sign	Labeled calibration effects, all 210 edges	Evaluation identities
Affine magnitude, held edges	Labeled evaluation-role effects on 150 profile edges	60 edge identities
Global threshold	Calibration effects on 150 profile edges from 11 other cells	60 eval-role edges in held cell
Affine action, held edges	Calibration effects on 150 profile edges of held cell	60 eval-role edges
Affine, held cell	Labeled effects in 11 other cells	One checkpoint cell
Closed-loop profile rule	All-edge calibration map; 32 labeled profile edges per fit, 20 masks	Stream identities and trajectories

Identity separation and dependence. Within each experimental system, donor acquisition, an auxiliary geometry pool, target calibration, and target evaluation use disjoint underlying image identities, including across corruption views. The geometry pool supplies the earlier-domain history in Section 6 and is not used to define the external effect score. Calibration and evaluation each contain 512 images per domain. Acquisition seeds reorder fixed pools; edges share domains and images, and cells share benchmark definitions. We report pooled results descriptively and checkpoint-level dispersion across trained receivers. Source and receiver experiments are separate archived systems with their own partitions.

External source map. The frozen source score $S(A, B)$ is an archived WRN-40-2 CIFAR-10-C carry/reset effect map, fixed before the receiver analyses. Its source model used division by 255 for preprocessing; the receivers use dataset normalization and a different training recipe. Across the receiver collection, transfer also spans changes in architecture and dataset while retaining the same corruption taxonomy. This design measures transfer across systems but does not isolate which difference causes the boundary shift.

Separate prediction and decision tasks. For a score q, let $\pi_q(A, B) = \mathbf{1}\{q(A, B) > 0\}$, with ties resetting. Its mean gain over always resetting is

$$V(\pi_q) = \frac{1}{|\mathcal{E}|} \sum_{(A,B)\in\mathcal{E}} \pi_q(A, B) G_{\text{eval}}(A, B). \tag{2}$$

We compare the external source rule $q = S$ with the receiver-calibrated rule $q = G_{\text{cal}}$, which uses labeled receiver outcomes on separate image identities. The retrospective binary oracle uses $q = G_{\text{eval}}$. For magnitude prediction, we fit $\widehat{G} = a + bS$ on 150 receiver profile edges and test on 60 locked edges. The cross-role affine action analysis instead fits $a+bS$ to calibration-role effects on the held receiver's 150 profile edges, thresholds predictions at zero, and scores evaluation-role effects on its 60 held edges. To test transfer of the threshold itself, we fit the rule $S > t$ using calibration effects on the same 150 profile edges from the other 11 receivers, then score the twelfth receiver's 60 held edges. A further magnitude analysis fits on 11 cells and tests on all edges of the twelfth. The shared 150/60 split uses four cyclic offsets of a seeded domain permutation; Appendix D gives its construction and the profile-mask seeds. Table 1 lists the information available to each analysis.

4 Ranking transfers more readily than the decision boundary

4.1 A correlated source map can induce a harmful policy

The external map correlates with receiver evaluation effects (Pearson $r = 0.759$, Spearman $\rho = 0.721$; Figure 1). The association also holds within each checkpoint: Pearson ranges from 0.712 to 0.831 and Spearman from 0.680 to 0.766. Direct magnitude transfer is weaker, with $R^2 = 0.228$ and RMSE 14.36 pp. Receiver effects are predominantly negative: 364 positive, 2,146 negative,

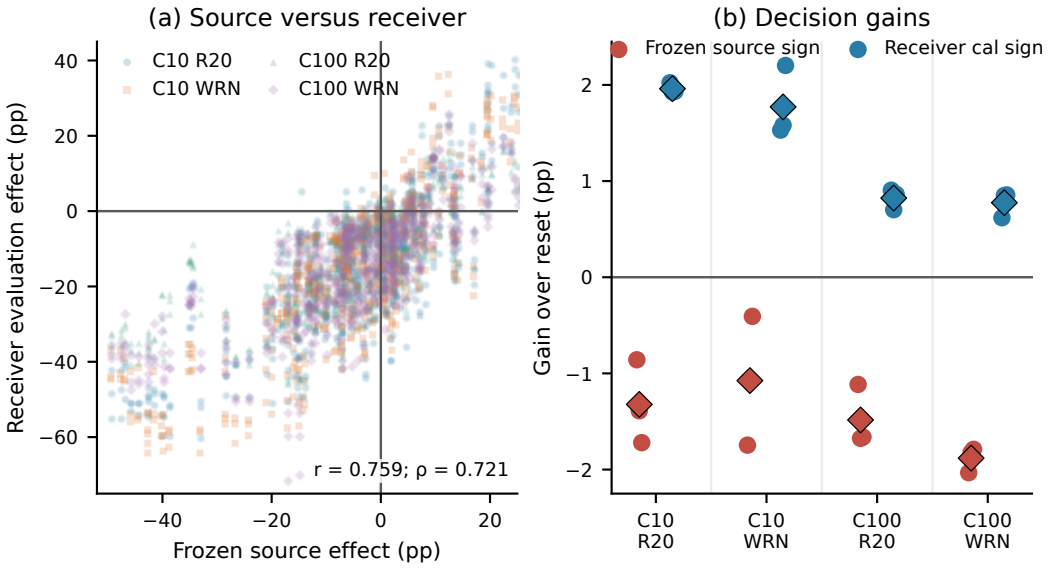


Figure 1: **Ranking transfer does not guarantee action transfer.** Left: the frozen external source effect versus the receiver effect on 2,520 cell–edges (three acquisition orders averaged). Right: gains relative to resetting for each receiver checkpoint; diamond markers denote the three-seed mean within each dataset/architecture group. All 12 source-threshold gains are negative; all 12 receiver-calibrated gains are positive. The receiver rule uses labeled calibration identities.

Table 2: Isolated first-128 decisions over 2,520 receiver cell–edges. Harmful fraction is measured among selected carry actions; carry-all and the retrospective oracle provide reference endpoints.

Rule	Gain (pp)	Carry count	Harmful carries	Useful recall
Always reset	0.00	0	—	0.0%
Always carry	−14.21	2,520	85.2%	100.0%
External source sign	−1.44	1,032	65.8%	94.2%
Receiver calibration sign	+1.33	423	22.7%	87.4%
Retrospective binary oracle	+1.50	364	0.0%	100.0%

and 10 zero cell–edges. Always carrying loses 14.21 pp relative to resetting, while the isolated binary oracle gains 1.50 pp. A useful carry rule must therefore select a small beneficial subset while avoiding large losses elsewhere.

The source sign rule carries on 1,032 cell–edges, capturing 94.2% of useful edges but also selecting 679 harmful ones. Thus 65.8% of its carry actions have negative evaluation effects. Its mean gain is −1.44 pp (checkpoint SD 0.48 pp), with losses at all 12 checkpoints. Among harmful selections, the median loss is 9.64 pp, the 95th percentile 24.74 pp, and the maximum 39.58 pp.

The receiver calibration sign rule carries on 423 cell–edges, of which 96 are harmful. The harmful fraction falls to 22.7%, while useful-edge recall remains 87.4% (Table 2). The rule gains 1.33 pp (checkpoint SD 0.59), is beneficial at every checkpoint, and recovers 89.0% of the isolated oracle gain. Losses among harmful selections are also smaller: median 2.34 pp, 95th percentile 8.72 pp, and maximum 12.76 pp.

Averaged over all cell–edges, the source rule retains 1.478 pp of positive benefit but incurs 2.919 pp of harmful-carry cost. The receiver rule retains 1.454 pp while reducing that cost to 0.121 pp. Calibration therefore improves decisions mainly by avoiding harmful carries, rather than by finding additional beneficial edges (Appendix C).

Table 3: Source-score prediction of receiver effects. Pooled metrics are descriptive. Raw transfer uses no receiver labels; affine rows do. The eval-role row isolates held edge identities; the cal→eval row additionally fits on disjoint calibration image identities before evaluation. R^2 is predictive $1 - \mathrm{SSE}/\mathrm{SST}$, not squared correlation.

Prediction setting	Test rows	Pearson	Spearman	R^2	RMSE (pp)
Frozen source, uncalibrated	2,520	0.759	0.721	0.228	14.36
Affine, eval-role 150/60 split	720	0.777	0.710	0.602	10.05
Affine, cal→eval 150/60 split	720	0.775	0.705	0.592	10.17
Affine, one receiver held out	2,520	0.755	0.720	0.571	10.71

4.2 A SHIFTED BOUNDARY RECOVERS USEFUL DECISIONS

An affine fit on 150 receiver profile edges improves magnitude prediction on the remaining 60 edges per cell: pooled $R^2 = 0.602$, $r = 0.777$, and RMSE 10.05 pp (Table 3). Held-edge R^2 is positive in every cell, with mean 0.586 and SD 0.068. Two fitted parameters thus account for a substantial part of the source-to-receiver mismatch. This magnitude analysis uses labeled evaluation-role effects on the profile edges; the action comparisons below fit on calibration-role effects.

Can a threshold learned from other receivers recover this decision value? For each held receiver, we choose t from source-score values observed on the 150 profile edges, using only calibration-role effects from the other 11 receivers. We then apply $S > t$ to the held receiver's 60 locked edges. All 12 folds select $t = 7.031$ pp; each maximizes training gain on the same discrete score grid, with ties resolved by the smallest maximizing threshold (Appendix D). The rule gains $+0.829$ pp on the 720 held cell–edges and is beneficial at every receiver. The uncalibrated $S > 0$ rule loses 1.022 pp on those same edges. Using the midpoint of the discrete score gap, with ties resetting, gives $+0.703$ pp and remains beneficial in all 12 folds (Appendix D).

A receiver-specific affine rule gains $+0.857$ pp, only slightly more than the transferred threshold. It fits two coefficients to the held receiver's 150 calibration-role effects and is scored on disjoint evaluation-role effects for the 60 held edges. It also predicts effect magnitudes ($r = 0.775$, $\rho = 0.705$, $R^2 = 0.592$, RMSE 10.17 pp) and yields positive action gains at all 12 receivers (Appendix D). Within this receiver collection, a shared threshold therefore recovers most of the decision gain available from receiver-specific affine calibration.

For magnitude prediction with one receiver cell held out, the affine fit gives pooled $R^2 = 0.571$ and RMSE 10.71 pp. All 12 held cells have positive R^2 (mean 0.547, SD 0.078). This analysis fits labeled effects from all edges in the other 11 cells, testing transfer to a new checkpoint. The held-edge analysis instead tests new transition identities within a receiver.

The effect maps are stable across the three donor acquisition orders: within-cell pairwise Pearson correlation averages 0.990, with a minimum of 0.969. This stability holds for the fixed image pools and acquisition length used here. Appendix B shows the complete source and receiver maps and the per-checkpoint decisions.

5 WHY STATE PERSISTENCE IS NOT PREDICTIVE UTILITY

The matched-update identity. For fixed BN momentum m, two running-statistic states receiving identical batch statistics satisfy

$$s_{t+1} = (1-m)s_t + m\widehat{s}_t, \qquad \Delta s_t = (1-m)^t \Delta s_0. \tag{3}$$

In the frozen-weight protocol, all BN layers use batch statistics during the update pass, so both arms receive the same batch-statistic inputs. The EMA recurrence therefore describes buffer differences exactly. At $m = 0.1$, two target updates retain 81% of an incoming difference, and eight retain 43%. These percentages measure state persistence; they do not determine the resulting change in predictions.

Function-space response. BN applies channel scaling through $\left(v + \epsilon\right)^{-1/2}$, subsequent layers interact, and accuracy changes only when decisions cross class boundaries. For labeled target NLL

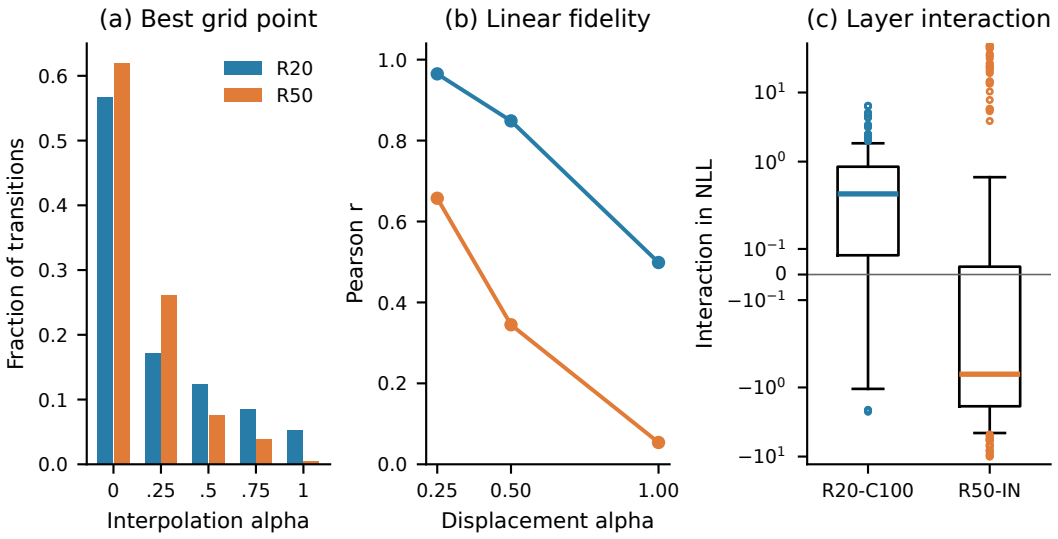


Figure 2: **Nonlinear utility of a linearly interpolated state.** Complete 210-edge sweeps for ResNet-20/CIFAR-100 and ResNet-50/ImageNet. Left: the NLL-minimizing point among five interpolation values. Middle: correlation of the Taylor diagnostic with measured NLL change as displacement grows. Right: all-layer NLL change minus the sum of single-group changes (symmetric-log scale). These are pure-prediction interventions, without target updates.

L_B, a local expansion along a state direction d gives

$$L_B(s_0 + \alpha d) - L_B(s_0) = \alpha \nabla L_B(s_0)^\top d + \tfrac{1}{2}\alpha^2 d^\top H_B(s_0) d + \mathcal{O}(\alpha^3). \tag{4}$$

The gradient, curvature, and decision margins depend on the receiving network and target. Exponential decay of the state difference therefore need not produce a proportional change in loss or accuracy.

Controlled state interpolation. We evaluate $s_\alpha = (1-\alpha)s_0 + \alpha s_A$ at $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, interpolating running means and raw running variances. The sweeps use pure prediction on the first 128 target images, without the target updates used in Section 3. They measure the response to a fixed state intervention. In complete ResNet-20/CIFAR-100 and ResNet-50/ImageNet sweeps, an interior value gives lower NLL than both endpoints in 80/210 (38.1%) and 79/210 (37.6%) transitions, respectively. Linear interpolation between endpoint losses would miss these interior optima. Labels are used here to characterize the response surface, not to select a deployable policy.

The first-order Taylor diagnostic becomes less predictive as displacement grows. Its correlation with measured NLL change at $\alpha = 0.25, 0.5, 1$ is $0.965, 0.849, 0.499$ for ResNet-20 and $0.657, 0.345, 0.054$ for ResNet-50. The functional implementation agrees with ordinary forward evaluation to maximum NLL discrepancies of 2.30×10^{-4} and 3.08×10^{-4}, respectively.

Layer interactions. We compare the NLL change from replacing all BN layers with the sum of changes from replacing the early, middle, and late layer groups separately. ResNet-20 has a median residual of $+0.341$, with 172/210 positive residuals. ResNet-50 has a median of -0.643 and 54/210 positive residuals; its positive mean of 1.798 is driven by the tail. The joint response thus differs from the sum of group responses, with different residual distributions across models. Because NLL is nonlinear, additivity serves as a descriptive reference. Appendix E reports the joint and single-group change scales.

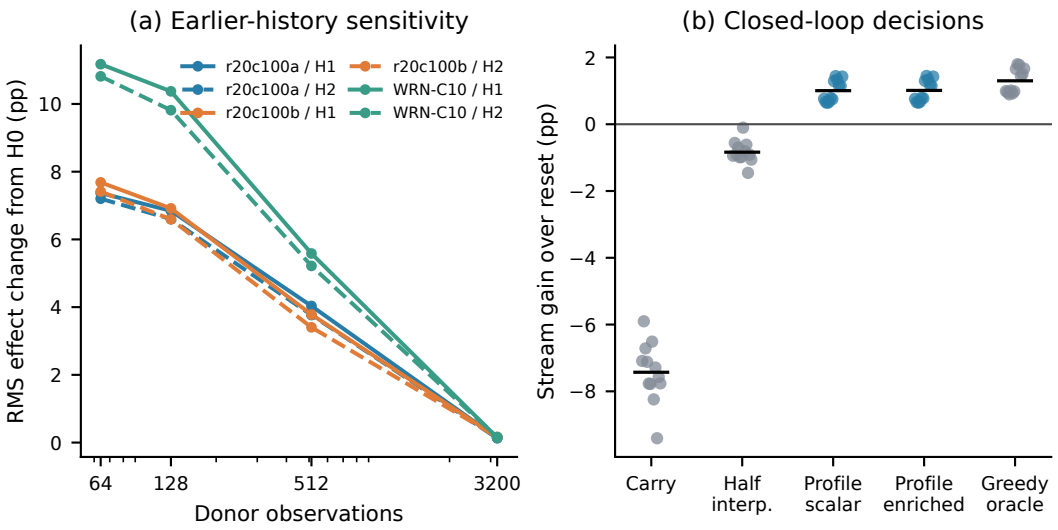


Figure 3: **History and operational consequences.** Left: RMS change in full-512 transition effects from adding earlier history, relative to acquisition from the source state; two histories for each of three receivers. Right: closed-loop accuracy gains over resetting, with each point averaging ten streams for one checkpoint. Policy actions alter subsequent states. The calibrated policies use labeled offline receiver profiles; the oracle uses current-segment labels.

6 From Isolated Edges to Action-Dependent Streams

6.1 Acquisition Length Controls Sensitivity to Earlier History

An effect map indexed only by donor A omits earlier history. On three CIFAR receiver checkpoints, we vary donor acquisition length after two different prior histories. Each history prepends a domain before A, using cyclic offsets 1 and 7 in the lexicographically sorted corruption identifiers and geometry-pool identities for the earlier domain (Appendix A). With 64 donor images, the RMS difference from edge maps acquired directly from the source state ranges from 7.20 to 11.17 pp. After 3,200 donor images, it falls to 0.128–0.166 pp (Figure 3, left). This measured convergence is consistent with EMA attenuation: 50 updates at $m = 0.1$ retain about 0.5% of a pre-donor buffer difference.

Long donor acquisition reduces the dependence of the isolated map on earlier domains. With short segments, that history still affects the incoming state. As a negative control, batch-statistics-only prediction yields zero carry/reset effect in all 630 tested transitions: the effect disappears when prediction does not use stored running buffers.

6.2 Transferred and Receiver-Profiled Decisions in Closed-Loop Evaluation

We evaluate 12 checkpoints under ten stream seeds, with 45 segments per stream (three cycles through 15 domains), 512 images per segment, and known domain identities and boundaries. Stream images are disjoint from receiver profiling pools. Each policy maintains its own state throughout the stream, so its carry/reset decisions affect the states entering later segments.

The receiver-profile experiment compares always carry, always reset, fixed half-interpolation between current and source BN buffers, and two profiled rules. The scalar rule uses the receiver's all-edge labeled calibration map as a feature and fits it to labeled effects on 32 profile edges per mask. Predictions are aggregated by the median over 20 predefined masks, whose union contains 147 distinct profile edges. Thus the information budget includes both the all-edge calibration map and the outcomes on those profile edges. The enriched rule adds the source score and donor/target indicators. Appendix D specifies the fits and masks. Neither rule uses current-stream labels. A greedy oracle provides a local reference: at each segment it forks carry/reset, selects the higher-

Table 4: Closed-loop full-segment accuracy, averaged over 120 streams. Gain SD is across 12 checkpoint means after averaging ten stream seeds per checkpoint. Gains use unrounded accuracies. The upper block is the receiver-profile experiment; the lower block reruns controls with the transferred threshold. Gains use each block's own reset baseline. The oracle uses current-segment labels.

Policy	Accuracy (%)	Gain over reset (pp)	Gain SD
Always carry	44.47	−7.43	0.90
Always reset	51.90	0.00	—
Fixed half-interpolation	51.06	−0.84	0.33
Receiver profile, scalar	52.90	+1.01	0.32
Receiver profile, enriched	52.91	+1.01	0.32
Greedy segment oracle	53.20	+1.30	0.37
Always carry (rerun)	44.47	−7.43	0.90
Always reset (rerun)	51.90	0.00	—
Transferred threshold	52.83	+0.93	0.31

accuracy branch using current-segment labels, and continues from that branch's final state. It does not optimize the full 45-segment trajectory.

Fixed half-interpolation loses 0.837 pp, while the scalar receiver-profiled rule gains 1.006 pp (Table 4); both directions hold at every checkpoint. The scalar rule recovers 77.3% of the greedy oracle's gain over reset. Adding the enriched features changes the mean gain by only 0.007 pp: 26 streams improve, ten decline, and 84 tie. Most of the measured gain is therefore already captured by the scalar calibration feature. Stream gains average all 512 images per segment, whereas isolated-edge gains use the first 128 target images.

Transferring the threshold into the stream. We freeze the leave-one-receiver-out threshold before rollout and carry at a known $A \to B$ boundary when the external score exceeds it. The rule acts on the realized incoming state without donor reacquisition and uses no labels or offline profile from the held receiver. On the same 120 stream configurations, with carry/reset controls rerun alongside it, the rule gains 0.931 pp over reset (checkpoint SD 0.307 pp). Gains are positive at all 12 checkpoints and in all 120 streams. The rule carries at 16.82% of post-initial boundaries. This stream test includes all visited transitions, including profile edges; it tests held-receiver transfer in action-dependent trajectories rather than held-edge generalization. Appendix G gives the accounting and rerun details.

7 Scope, Implications, and Limitations

For the tested receivers, useful decisions do not require an exact prediction of effect magnitude: a shared threshold on the source score recovers nearly as much held-edge gain as receiver-specific affine calibration. This makes threshold transfer a useful baseline for BN-memory selection. Its simplicity concerns the number of fitted parameters, however, not the amount of calibration data. Each held-receiver fit uses 150 labeled profile-edge effects from each of 11 other receivers.

The results concern frozen-weight convolutional models under the shared CIFAR-C corruption taxonomy. Source and receiver systems differ in architecture, dataset, preprocessing, and training recipe, so the experiments do not attribute the boundary shift to a single factor. Each dataset–architecture group has three training seeds, and repeated edges share benchmark structure. The ImageNet state sweeps characterize predictive responses under pure prediction, a different protocol from the carry/reset transfer analysis. The corrected ImageNet decision check covers nine distinct transitions per model and mostly negative carry effects (Appendix F).

The receiver-sign and profiled stream rules require offline receiver labels. The transferred threshold requires labels only from other receivers, but still assumes known domain identities and boundaries. Its stream evaluation includes profile edges as well as held edges. Transfer to unseen corruption types, natural shifts, unknown-domain streams, or temporally class-correlated inputs has not been established. Weight-updating CTTA also changes the state-to-prediction map, and LayerNorm lacks the across-batch running buffers studied here. At a fixed target horizon, changing BN momentum or

batch size changes memory attenuation; threshold transfer under mismatched EMA settings remains untested. The conclusions apply to the source map and receiver family evaluated here.

8 CONCLUSION

Across 12 CIFAR-C receivers, an external BN-memory effect map preserves useful ranking information but produces harmful decisions when thresholded at zero. A single threshold learned from other receivers gains 0.83 pp on held edges, close to the 0.86 pp from receiver-specific affine calibration. It also gains 0.93 pp in closed-loop streams without held-receiver labels. These findings show that the carry/reset threshold can transfer across the tested receivers even when the source threshold fails. Evaluations of memory transfer should therefore report decision gains alongside ranking metrics and specify which calibration information those decisions require.

REPRODUCIBILITY STATEMENT

The accompanying artifact contains the experimental result files, input hashes, and scripts used to generate the reported tables and figures. CPU reanalysis recomputes effect aggregation, affine calibration, the leave-one-receiver-out threshold, decision gains, harm tails, and stream summaries. A separate script checks the transferred-threshold rollout, its action rule, and paired controls. The package supports reanalysis of saved results; benchmark images and model checkpoints needed to rerun the experiments are not included. Appendix A specifies the intervention and evaluation roles, and the artifact README lists the reanalysis commands.

AI USE STATEMENT

AI tools assisted with literature discovery and metadata checking, code and result auditing, numerical reanalysis, manuscript drafting, figure generation, and LaTeX preparation. All reported numerical results come from experimental outputs and their CPU reanalysis.

REFERENCES

Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2103.05898.

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://arxiv.org/abs/2208.05117.

Sreeja Guha Majumdar and Aratrika Saha. When test-time adaptation helps, harms, or becomes inactive: A condition-level study on CIFAR-10-C. *arXiv preprint arXiv:2608.22233*, 2026. URL https://arxiv.org/abs/2608.22233.

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.

Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. MECTA: Memory-economic continual test-time model adaptation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=N92hjSf5NNh.

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

Juwon Kang, Nayeong Kim, Jungseul Ok, and Suha Kwak. MemBN: Robust test-time adaptation via batch norm with statistics memory. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/04143.pdf.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. doi: 10.1016/j.patcog.2018.03.005. URL https://doi.org/10.1016/j.patcog.2018.03.005.

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.

Taejun Lim, Joong-Won Hwang, and Kibok Lee. When and where to reset matters for long-term test-time adaptation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.03796.

M. Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.00463.

Fahim Faisal Niloy, Sk Miraj Ahmed, Dripta S. Raychaudhuri, Samet Oymak, and Amit K. Roy-Chowdhury. Effective restoration of source knowledge in continual test time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2091–2100, 2024. doi: 10.1109/WACV57701.2024.00209. URL https://doi.org/10.1109/WACV57701.2024.00209.

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*, Proceedings of Machine Learning Research. PMLR, 2022. URL https://arxiv.org/abs/2204.02610.

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://arxiv.org/abs/2302.12400.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05401.

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://arxiv.org/abs/2006.16971.

Vasileios Sevetlidis and George Pavlidis. Training memory in deep neural networks: Mechanisms, evidence, and measurement gaps. *arXiv preprint arXiv:2601.21624*, 2026. URL https://arxiv.org/abs/2601.21624.

Damian Sójka, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. AR-TTA: A simple method for real-world continual test-time adaptation. In *British Machine Vision Conference*, 2024. URL https://bmva-archive.org.uk/bmvc/2024/papers/Paper_25/paper.pdf.

Damian Sójka, Marc Masana, Bartłomiej Twardowski, and Sebastian Cygert. Rethinking the teacher-student framework for test-time adaptation. In *Conference on Lifelong Learning Agents*, 2026. URL https://arxiv.org/abs/2609.02507.

Sudarshan Sreeram, Young D. Kwon, and Cecilia Mascolo. Tempora: Characterising the time-contingent utility of online test-time adaptation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.06136.

Zixian Su, Jingwei Guo, Kai Yao, Xi Yang, Qiufeng Wang, and Kaizhu Huang. Unraveling batch normalization for realistic test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2312.09486.

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Un-mixing test-time normalization statistics: Combatting label temporal correlation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.08328.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno 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*, 2022. URL https://arxiv.org/abs/2203.13591.

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*, 2023. URL https://arxiv.org/abs/2303.13899.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Xingzhi Zhou, Boyang Zhang, Zhiliang Tian, Yibo Zhang, Xin Niu, Ka Chun Cheung, Simon See, and Nevin L. Zhang. Resilient test-time adaptation by mitigating batch-normalization overfitting. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10888647. URL https://doi.org/10.1109/ICASSP49660.2025.10888647.

A Protocol details and interpretation

A.1 The experimental unit

The receiver matrix contains CIFAR-10-C and CIFAR-100-C, ResNet-20 and WRN-40-2, and three training seeds per dataset–architecture pair. For each cell and each directed non-self corruption edge, three acquisition-order measurements are averaged before fitting the cross-system predictors or evaluating the isolated sign rules. This averaging produces the 2,520 rows in the main analysis. The original 7,560 measurements remain available in the artifact.

Training seed identifiers in the supplied experiment are 2026092101, 2026092102, and 2026092103. They are integer RNG identifiers, not dates of execution. The normalized receiver recipe uses 200 training epochs, SGD with momentum 0.9, weight decay 5×10^{-4}, batch size 128, initial learning rate 0.1, and milestones at epochs 60, 120, and 160. The archived source map belongs to a different preprocessing/training system. Keeping these systems distinct is essential to interpreting transfer.

A.2 Target timing and buffer handling

Every paired measurement begins from the same frozen checkpoint parameters. Donor acquisition changes running means and variances; the carry and reset arms differ in their incoming stored state. During each target update, BN modules are in training mode while dropout remains disabled. A subsequent evaluation-mode pass scores the same batch using the updated running statistics. Using current target inputs to update normalization before scoring is part of the declared protocol. This is neither predict-before-update evaluation nor pure batch-statistics inference.

With fixed momentum, the running-statistic interpolation weight is determined by the EMA update rather than the BN batch counter. We therefore analyze the intervention through running means and variances. Both arms process the same target order and use the same labels solely for evaluation. The first-128 metric scores two 64-image batches; the full-512 metric scores eight. Changing these horizons changes the amount of donor-state attenuation and the set of predictions averaged.

A.3 Identity roles

Underlying clean-image identities, not merely corrupted file paths, define separation. An identity used for donor acquisition is excluded from geometry, calibration, and evaluation roles in that system, including alternative corruption views. Calibration and evaluation effects are measured on different identities. This separation lets the labeled calibration rule be evaluated on image identities distinct from those used to fit it.

The source-map experiment and receiver experiments are separate archived systems with their own partitions. The transfer analysis compares measured effects across these systems under a shared corruption taxonomy; its generalization axis is the receiver system rather than a new corruption vocabulary. In the stream experiment, stream identities are checked against all four receiver roles, and the three visits to a domain use 1,536 distinct identities in total.

A.4 Corruption taxonomy

The 15 nodes are gaussian noise, shot noise, impulse noise; defocus blur, glass blur, motion blur, zoom blur; snow, frost, fog, brightness; and contrast, elastic transform, pixelate, JPEG compression. These are grouped as noise, blur, weather, and digital in the visualizations. Every off-diagonal ordered pair is measured. The visual family grouping is descriptive; edge dependence still follows from shared donors and targets. The lexicographic order used for history offsets is brightness, contrast, defocus_blur, elastic_transform, fog, frost, gaussian_noise, glass_blur, impulse_noise, jpeg_compression, motion_blur, pixelate, shot_noise, snow, zoom_blur; offsets wrap modulo 15.

A.5 Metric definitions

Predictive R^2 is $1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$. Correlation measures association and can be high even when raw R^2 is low because the scale or offset is incorrect. For a sign policy, useful-edge recall is the number

of selected edges with $G_{\text{eval}} > 0$ divided by all positive evaluation edges. Harmful-carry fraction is the number of selected edges with $G_{\text{eval}} < 0$ divided by all selected edges. Ties select reset. Harmful regret statistics summarize $-G_{\text{eval}}$ among harmful selected actions.

The isolated oracle gain is $|\mathcal{E}|^{-1} \sum_e \max(G_e, 0) = 1.4974$ pp. The receiver sign rule's 1.3326 pp gain divided by this quantity yields 88.99%. This upper reference is valid for the isolated binary interventions being evaluated, not for all possible state mixtures, future sequences, or full adaptation algorithms.

B Complete maps and checkpoint-level decisions

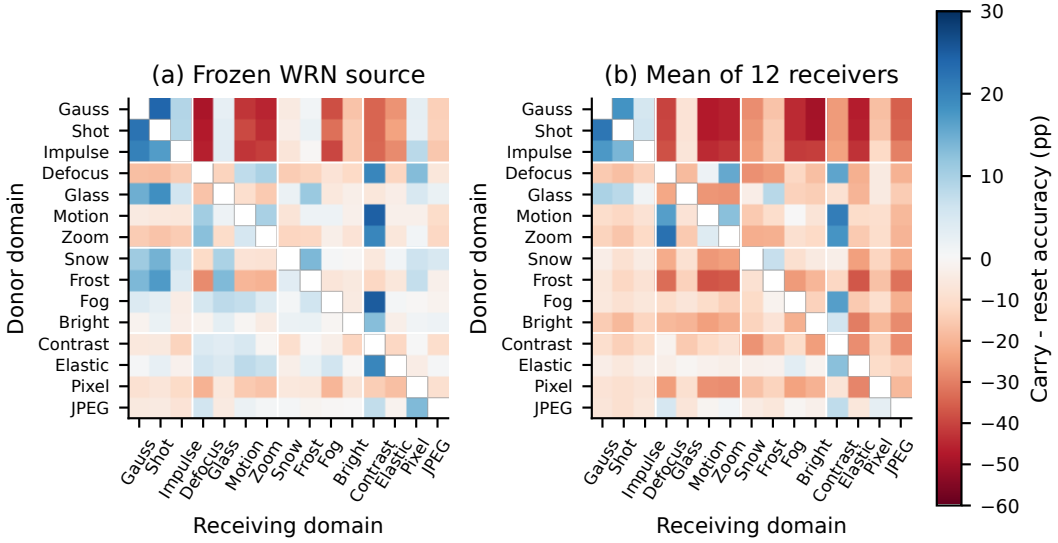


Figure 4: Frozen external source map and mean receiver map on a shared color scale. Rows are donors and columns are targets; diagonal entries are unmeasured. The receiver map averages 12 checkpoints. Similar large-scale structure can coexist with a different location of zero. The pooled metrics additionally retain checkpoint-level variation.

Table 5: Per-checkpoint isolated decision gains (pp). Seeds 1–3 follow the ordered identifiers in Appendix A.

Dataset / architecture	Seed	External source sign	Receiver sign
C100 / ResNet-20	1	-1.115	+0.905
C100 / ResNet-20	2	-1.675	+0.702
C100 / ResNet-20	3	-1.662	+0.863
C100 / WRN-40-2	1	-2.032	+0.619
C100 / WRN-40-2	2	-1.822	+0.851
C100 / WRN-40-2	3	-1.789	+0.856
C10 / ResNet-20	1	-0.858	+2.023
C10 / ResNet-20	2	-1.389	+1.926
C10 / ResNet-20	3	-1.720	+1.933
C10 / WRN-40-2	1	-1.746	+1.530
C10 / WRN-40-2	2	-1.071	+1.581
C10 / WRN-40-2	3	-0.407	+2.202

C WHERE DECISION VALUE IS LOST

For a binary carry policy π_e and isolated effect G_e, define $G_e^+ = \max(G_e, 0)$ and $G_e^- = \max(-G_e, 0)$. Averaging over the fixed evaluation edge set gives the accounting identities

$$V(\pi) = \underbrace{\mathbb{E}[\pi G^+]}_{\text{retained benefit}} - \underbrace{\mathbb{E}[\pi G^-]}_{\text{harmful-carry cost}}, \tag{5}$$

$$V(\pi^*) - V(\pi) = \underbrace{\mathbb{E}[(1-\pi)G^+]}_{\text{missed benefit}} + \underbrace{\mathbb{E}[\pi G^-]}_{\text{harmful-carry cost}}. \tag{6}$$

Here $\mathbb{E}$ denotes the finite arithmetic mean over the evaluated edge set. The decomposition separates missed opportunities from damaging actions, two quantities that correlation and unweighted sign accuracy combine. The external rule misses just 0.019 pp of available positive benefit; its oracle regret is dominated by 2.919 pp of harmful-carry cost. Receiver calibration slightly increases missed benefit to 0.044 pp while reducing harmful cost to 0.121 pp. The net improvement therefore reflects better avoidance of expensive mistakes. This decomposition uses the already-scored evaluation set descriptively; it supplies no additional policy-training information or independent test set.

Table 6: Decision-value decomposition, averaged over all 2,520 cell–edges (pp). Costs are unconditional averages, unlike the conditional harmful-loss quantiles in the main text.

Rule	Retained benefit	Harm cost	Missed benefit	Net gain
External source sign	1.478	2.919	0.019	-1.441
Receiver calibration sign	1.454	0.121	0.044	+1.333

D AFFINE CALIBRATION AND SPLIT BOUNDARIES

The locked edge split consists of 150 profile edges and 60 test edges per cell. Sort the 15 corruption identifiers lexicographically and permute them with NumPy `RandomState(44001)`. If the resulting order is $p_0, \ldots, p_{14}$, the test edges are $(p_i, p_{(i+d) \bmod 15})$ for every i and $d \in \{1, 4, 7, 11\}$. The 150 remaining directed non-self edges form the lexicographically sorted profile list. This outcome-independent construction gives each node four incoming and four outgoing test edges; it is not a corruption-family holdout. The exact lists are in `evidence/q2/config_locked.json`. For the original magnitude analysis, ordinary least squares fits an intercept and slope on the receiver evaluation-role first-128 effects of the 150 profile edges, averaged across acquisition orders. The 60 scored edges are excluded from that fit. Because several edges can share a target domain, this split isolates edge identities; image-role separation is handled separately by the cross-role analysis.

The cross-role action analysis uses the same pre-existing split but fits the two affine coefficients to *calibration-role* effects on the held receiver's 150 profile edges. It then applies $\mathbf{1}\{a + bS > 0\}$ to the 60 locked edges and scores only their evaluation-role effects. Calibration and evaluation identities are disjoint by construction, so both the scored edge outcomes and scored image identities are absent from the fit. On the 720 held cell–edges, the receiver-specific affine action gains 0.857 pp versus reset while the frozen source-sign action loses 1.022 pp.

The global-threshold baseline asks whether the boundary shift transfers without any labels from the held receiver. In each of 12 leave-one-receiver-out folds, we pool the 150 profile-edge calibration effects from the other 11 receivers. We search thresholds t only over source-score values observed in this training set and choose the one maximizing mean calibration gain for the policy $\mathbf{1}\{S > t\}$; ties at the threshold reset, matching the main policy convention. We then freeze t and score the 60 locked evaluation-role edges of the held receiver. Every fold selects $t = 7.03125$ pp and the mean held-edge gain is 0.8286 pp, positive in all 12 receivers (Table 8). The best training decisions are unchanged for any threshold between the adjacent observed profile scores 7.03125 and 7.55208 pp. To expose the resulting discrete-grid ambiguity, we also evaluate their exact midpoint, 7.29167 pp, with ties resetting; it gains 0.7028 pp and remains positive in all 12 receivers. The conclusion that a shared shifted boundary is useful therefore does not depend on carrying the held edge whose source score lands exactly at that midpoint.

For leave-one-cell-out magnitude prediction, the same two affine coefficients are fitted on all 2,310 cell–edges from the other 11 receivers and evaluated on the 210 edges of the held receiver. All 210 corruption pairs remain represented among the training receivers. This measures transfer to a held checkpoint within the receiver collection and is distinct from the one-parameter threshold test, which holds out both the receiver and the 60 scored edge identities.

The coefficients in Table 7 belong to the original evaluation-role magnitude fit and make the location of its implied source-score zero explicit. The cross-role affine action uses separate coefficients from calibration-role effects. The global threshold, affine coefficients, per-checkpoint gains, and midpoint sensitivity are all regenerated by the artifact's CPU analysis. The larger affine fit and the stream's 32-edge per-mask fitting procedure remain different analyses with different information costs.

Table 7: Receiver-specific affine fits on 150 profile edges. The implied zero is $-a/b$, in source-effect pp.

Dataset / architecture	Seed	Intercept a	Slope b	Implied zero
C100 / ResNet-20	1	-8.47	0.647	13.09
C100 / ResNet-20	2	-10.13	0.647	15.65
C100 / ResNet-20	3	-9.62	0.655	14.69
C100 / WRN-40-2	1	-11.25	0.644	17.46
C100 / WRN-40-2	2	-10.75	0.762	14.11
C100 / WRN-40-2	3	-11.27	0.789	14.29
C10 / ResNet-20	1	-8.33	0.996	8.37
C10 / ResNet-20	2	-10.82	1.017	10.64
C10 / ResNet-20	3	-10.67	0.944	11.31
C10 / WRN-40-2	1	-11.23	1.061	10.59
C10 / WRN-40-2	2	-10.23	1.065	9.61
C10 / WRN-40-2	3	-8.61	0.997	8.63

Table 8: Held-edge action gains (pp) on the locked 60-edge evaluation set per receiver. The global threshold is learned from calibration-role effects on the 150 profile edges of the other 11 receivers; the affine rule is fitted to the held receiver's 150 calibration-role profile edges. All rules are scored on the same 60 evaluation-role edges.

Dataset / architecture	Seed	Source $S > 0$	Global $S > t$	Receiver affine
C100 / ResNet-20	1	-0.655	+0.525	+0.582
C100 / ResNet-20	2	-1.276	+0.352	+0.430
C100 / ResNet-20	3	-0.898	+0.538	+0.599
C100 / WRN-40-2	1	-1.059	+0.512	+0.577
C100 / WRN-40-2	2	-1.076	+0.781	+0.586
C100 / WRN-40-2	3	-1.059	+0.642	+0.538
C10 / ResNet-20	1	-0.781	+1.046	+0.885
C10 / ResNet-20	2	-1.050	+1.285	+1.081
C10 / ResNet-20	3	-2.010	+0.634	+0.890
C10 / WRN-40-2	1	-1.463	+0.716	+1.042
C10 / WRN-40-2	2	-0.907	+1.128	+1.194
C10 / WRN-40-2	3	-0.030	+1.784	+1.884
Mean	–	-1.022	+0.829	+0.857

Stream profile masks and fitted rules. For each seed $s \in \{99000, \ldots, 99019\}$, apply `RandomState(s).permutation(150)`, retain the first 32 indices, and sort those indices into the profile list above. Sampling is without replacement within a mask; masks may overlap and their union contains 147 edges. These fixed masks are shared across receivers. The artifact stores the masks as profile-list indices; the verification script exports their resolved edge lists and checks exact reconstruction.

The scalar stream rule fits an intercept and nonnegative slope from the receiver calibration-map feature to evaluation-role first-128 effects on each mask's 32 offline profile edges. The enriched rule fits

ridge regression with an intercept, the calibration feature, donor/target indicators, and the standardized frozen source score. The intercept is unpenalized. Its penalty is selected from $\{0.1, 1, 10, 100\}$ by node-held-out validation within the profile mask, minimizing mean fold RMSE and breaking ties toward the smaller penalty. Both rules take the median prediction over masks and carry only if it is positive. The fitting outcomes are offline receiver measurements, disjoint from stream identities; "evaluation-role" here does not mean current-stream labels. The original result files call these rules M1 and M4, respectively.

E EMA IDENTITY AND INTERVENTION BOUNDARIES

Let s_t^c and s_t^r be the carry and reset buffer vectors immediately before a matched update. Subtracting their recurrences gives

$$s_{t+1}^c - s_{t+1}^r = (1-m)(s_t^c - s_t^r) + m(\widehat{s}_t^c - \widehat{s}_t^r).$$

When batch statistics match, the last term vanishes. Induction yields Equation 3. Under the all-BN-training update pass with fixed parameters and identical inputs, current activations use batch statistics, so differing running buffers do not induce a different forward path for computing those statistics. The matched-update identity relies on fixed parameters, matched randomness, and the all-BN-training update pass; changing those conditions changes the recurrence being analyzed.

The mechanism experiment instead fixes a state s_α and performs pure prediction, with no target-statistic update. Running variance is interpolated directly as a stored buffer, which defines the state path used in the experiment. Labels provide NLL and the retrospective best interpolation value, allowing the sweep to characterize the response surface along that path.

The layer interaction residual is

$$I = \big[L(s_A) - L(s_0)\big] - \sum_{g\in\{\text{early,mid,late}\}} \big[L(s_0^{(g\leftarrow A)}) - L(s_0)\big],$$

where $s_0^{(g\leftarrow A)}$ replaces only group g's buffers. The residual measures functional non-additivity across the three layer groups; its sign describes whether the joint state change exceeds or falls below the sum of the corresponding single-group changes.

Table 9: NLL-change scales across 210 edges per model on the first 128 target images, without target updates. J is the joint all-layer change, A the sum of the three single-group changes, and $I = J - A$ the interaction residual. Entries are medians of edge-wise absolute magnitudes.

Model	median$\lvert J\rvert$	median$\lvert A\rvert$	median$\lvert I\rvert$
ResNet-20/CIFAR-100	1.503	1.039	0.452
ResNet-50/ImageNet	3.108	3.913	1.136

Table 9 contextualizes the residual against the joint and summed single-group NLL changes on the same 210 edges per model. Absolute magnitudes avoid cancellation across positive and negative changes. These are descriptive summaries, not tests of statistical independence or evidence that non-additivity uniquely arises in BN.

Only complete 210-edge state sweeps enter the main mechanism figures. Additional local-update logs contain unequal windows of 1, 1, 2, 4, and 8 batches. Since position and adaptation dose vary together in those windows, they are not used to claim that a particular time interval uniquely causes early benefit or late harm.

F CORRECTED IMAGENET CALIBRATION CHECK

The corrected ImageNet check contains nine distinct directed transitions per model and three realizations per transition, giving 27 measurements for ResNet-18 and 27 for ResNet-50. These are 54 measurements, not 54 independent transition types. Calibration-to-evaluation Pearson correlation is 0.942 for ResNet-18 and 0.987 for ResNet-50; RMSE is 5.24 and 2.86 pp. Strict three-way sign agreement (positive, negative, zero) is $24/27$ and $27/27$, respectively.

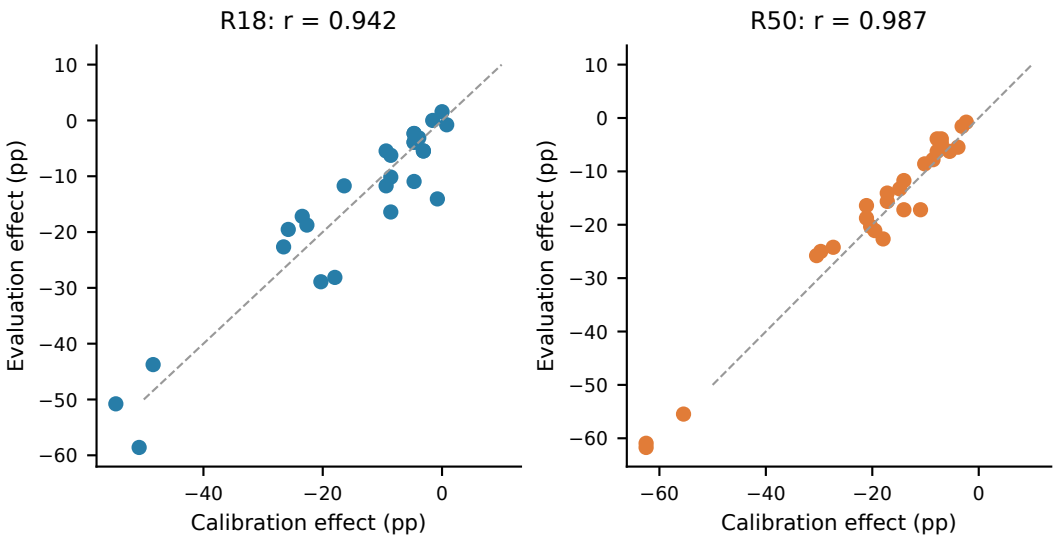


Figure 5: ImageNet calibration and evaluation effects on the corrected, limited transition set. High agreement primarily concerns negative carry effects. ResNet-50 has no positive evaluation effect in these 27 measurements; an always-reset classifier already identifies every nonpositive outcome.

ResNet-18 has one positive, 25 negative, and one zero evaluation effect. ResNet-50 has 27 negative effects. These measurements therefore show strong calibration-to-evaluation consistency on ImageNet, concentrated largely in the negative-carry regime. The complete ResNet-50 alpha sweep addresses the complementary question of how loss varies along a fixed interpolated-state path.

G Stream accounting and evidence provenance

The original receiver-profile experiment has 120 stream files containing 45 segments for each of seven policies, totaling 37,800 segment–policy rows. Main-table accuracy averages all segment predictions, then stream seeds, with equal segment lengths. Checkpoint dispersion averages the ten stream seeds within a checkpoint before computing standard deviations across the 12 checkpoints. The stream count reflects repeated trajectories over the same checkpoint collection and benchmark domains; checkpoint-level summaries provide the across-model dispersion.

The source-only inference control leaves target statistics untouched, whereas reset adapts them from the source state. The greedy oracle selects the better carry/reset branch by current-segment accuracy and continues from the winning state, providing a local reference at its own visited states. Since actions alter future states, this is not an upper bound on complete-trajectory accuracy for every policy.

Transferred-threshold rollout. The additional run contains $12 \times 10 \times 45 \times 3 = 16{,}200$ segment–policy rows: transferred threshold, always carry, and always reset. It uses the same 12 receiver identifiers, ten stream seeds (2026092601–2026092610), three 15-domain cycles, and 512 images per segment, with batch size 64, BN momentum 0.1, frozen weights, and update-then-predict evaluation. Each fold fits on 150 profile edges from each of the other 11 receivers (1,650 rows); all folds select $t = 7.03125$ pp under Appendix D's grid and tie rule. The threshold is fixed throughout the held receiver's streams. Held-receiver labels are used only for scoring.

The first segment initializes from source buffers and has no boundary decision. Subsequent decisions use the previous and current domain identifiers and the frozen empirical source map; the incoming BN state is the actual previous segment's output. No fresh 3,200-image donor acquisition is performed. The 120 streams visit 192 distinct directed transitions, including profile and held edges. There are 888 carry actions among $120 \times 44 = 5{,}280$ post-initial boundaries (16.82%); this is a boundary-weighted rate, not an average over edge types.

Mean accuracies are 44.467484% (carry), 51.897786% (reset), and 52.829246% (transferred threshold). The paired gain is 0.931460 pp; checkpoint-mean gains range from 0.571181 to 1.320747 pp, with sample SD 0.307429 pp. All 120 individual stream gains are positive. These repeated streams share checkpoints and corruption domains and are not 120 independent model replications.

The new run records an RTX 3080 Ti, whereas the receiver-profile run records an RTX 3090. Compared with the original controls on matched segments, rerun mean accuracy changes by -0.000687 pp for carry and $+0.001302$ pp for reset; the largest segment discrepancy is two correct predictions out of 512. We compute every new gain against its same-run reset control. Comparisons to the receiver-profiled rules are descriptive across runs and calibration budgets, not equivalence tests. The artifact includes raw segment counts, fitting rows, configuration, and checkpoint receipts. Receipts report matching checkpoint hashes, but checkpoint binaries are outside this compact artifact; the original lock file contains stale retraining language, documented in the artifact provenance note.