INTERVENTION-DEPENDENT SELECTION CREDIT IN STATEFUL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective test-time adaptation (TTA) methods choose which test samples drive each update. Their gains are often credited to reliable samples, but selection also changes sample counts and loss weights, while each update changes the model that scores later samples. Standard ablations conflate these effects. We measure the value of sample identity by recording native per-stage counts and rerunning each stream with those counts, choosing samples by each run's own native ranking or at random. Crossing these choices at the two stages of DeYO, EATA, and SAR yields a factorial that splits identity value between stages and measures their interaction. Native selection has positive identity value in 164 of 180 ImageNet-C cells. Yet stage credit changes with the comparison: on a five-corruption panel with 10 seeds, first-stage dominance is resolved in 31 of its 35 cells under the original comparison; matching mean loss weights leaves 19 of 40 cells first-stage dominant, with 8 resolved reversals (9 with joint resampling), and removes most of the ViT-B/16 identity value. In 8 of 12 candidate SAR cells, 20-seed intervals show that stages helping individually can hurt together. Waterbirds further separates identity-driven worst-group gains from harms that weaken under uniform loss weights. Selection credit is therefore intervention-dependent: it describes a stage's marginal effect within a specified stateful comparison, rather than an intrinsic property of its criterion.

1 INTRODUCTION

Deployed networks face distribution shifts (Hendrycks & Dietterich, 2019; Koh et al., 2021; Hendrycks et al., 2021). Test-time adaptation (TTA) updates models from unlabeled test data (Liang et al., 2025; Xiao & Snoek, 2024) through entropy minimization (Wang et al., 2021), normalization statistics (Schneider et al., 2020; Nado et al., 2020), or self-supervised and augmentation-based objectives (Sun et al., 2020; Liu et al., 2021; Zhang et al., 2022).

Each update changes how later samples are predicted and selected. This feedback can reinforce errors and cause collapse (Niu et al., 2022; 2023; Wang et al., 2022; Press et al., 2023; Hoang et al., 2024). Selective TTA filters candidates by entropy, prediction diversity, or prediction change (Niu et al., 2022; 2023; Lee et al., 2024), or uses confidence, class balance, or memory consistency (Gong et al., 2022; 2023; Yuan et al., 2023), before weighting survivors and updating the model.

Gains over all-sample adaptation conflate sample identities, counts, loss weights, and the model state that scores later samples. Component ablations mix the same effects. This feedback also arises in curriculum learning, online batch selection, and confidence-thresholded self-training (Bengio et al., 2009; Kumar et al., 2010; Mindermann et al., 2022; Sohn et al., 2020); data valuation instead scores training points under a specified learning procedure (Ghorbani & Zou, 2019; Kwon & Zou, 2022). Consequently, the apparent value of a selection stage need not be an intrinsic property of its criterion; it is a marginal effect under a particular counterfactual intervention. Therefore we ask: at fixed per-stage counts and weighting/update rules, how much does *sample identity* matter, and which stage contributes its value?

We record native selection counts, then rerun each stream with those counts and weighting/update rules, filling slots by own-state native ranking or uniform random choice. Every run follows its own trajectory. Crossing these choices at two stages yields total identity value, two-player Shapley stage credits (Shapley, 1953), and Shapley interaction (Grabisch & Roubens, 1999; Sundararajan

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.

et al., 2020). A performance-difference identity (Kakade & Langford, 2002) explains why these trajectory-level effects cannot be inferred from one update.

Applying this factorial to 15 ImageNet-C corruptions, two model families, and two stream permutations yields three findings. *First, identity matters*: native selection outperforms exact-count random selection in 164 of 180 cells, so the benefit of selective TTA is not merely an effect of adapting on fewer samples. *Second, how this value divides between stages depends on the comparison.* The original factorial gives more credit to the first, entropy-based stage in 151 of 180 cells, but that stage also has more room to change the kept samples, and in DeYO and EATA native samples carry 1.1–3.2$\times$ the loss weight of random ones. On the five-corruption panel, the original stage-1 ordering is resolved in 31 of 35 dominant cells. Weight matching leaves stage 1 dominant in 19 of 40 cells (10 seeds), with 17 positive-to-negative reversals (8 resolved under independent resampling; 9 under joint resampling) and one reversal the other way. It leaves ResNet-50 identity value intact or larger and removes 76–91% of it for ViT-B/16. *Third, stages that help individually can hurt jointly*: given a native first stage, SAR's second stage underperforms random choice in 56 of 60 cells, and with 20 seeds, 8 of 12 candidate cells support a reversal in which each stage helps alone but both together do worse than random selection at both stages. A Waterbirds study (Sagawa et al., 2020) shows the same entanglement at the level of populations: native selection's worst-group gain comes from which samples it keeps, while its harms to other groups largely weaken under uniform loss weights.

We contribute stateful selection interventions, a complete factorial across three methods and two model families, and controls showing that stage credit is a context-specific marginal rather than an intrinsic criterion score.

2 Related Work

Test-time adaptation and sample selection. TTA updates a source model on unlabeled test data, for example by minimizing prediction entropy (Wang et al., 2021), re-estimating normalization statistics (Schneider et al., 2020; Nado et al., 2020), or optimizing a self-supervised task (Sun et al., 2020; Liu et al., 2021). Continual and long-horizon evaluations show that such updates can accumulate errors and collapse (Wang et al., 2022; Press et al., 2023), which motivates selecting the samples that drive each update. Selection criteria include low entropy and prediction diversity (Niu et al., 2022), low entropy after a sharpness-aware perturbation (Niu et al., 2023), prediction change under object destruction (Lee et al., 2024), and confidence or class balance in memory-based methods (Gong et al., 2022; 2023; Yuan et al., 2023); DualTTA instead assigns different objectives to likely correct and incorrect samples (Phuong et al., 2026). These methods are evaluated against adaptation on all samples and through component ablations (Lee et al., 2024; Mounsaveng et al., 2024), both of which change how many samples adapt together with which ones. We keep a method's per-stage counts fixed and vary only which samples fill them.

Controls for which samples adapt. Random controls already appear in several studies. SNAP compares entropy- and representation-based memories with random memories under sparse updates (Cha et al., 2025), GoTTA compares memory policies across host adapters (Alhuwaider et al., 2026), CAS compares adaptation with random skipping at controlled rates (Jiang et al., 2026), and RECAST changes which buffered samples are used while keeping the host objective (Jo et al., 2026). Our controls differ in three ways: they match the native count at every stage and batch; they keep the native weighting, update, and recovery rules; and they cross native and random choice across stages, so that each stage's contribution and the interaction between stages can be measured along independently evolving runs.

Valuing data and components. Data valuation and training-data attribution score individual training points, by Shapley-style averages over subsets or by their estimated influence on a trained model (Ghorbani & Zou, 2019; Kwon & Zou, 2022; Koh & Liang, 2017; Pruthi et al., 2020; Ilyas et al., 2022; Park et al., 2023). They vary data under a fixed learning procedure, whereas in TTA the selection rule is part of the procedure and reshapes every later update. Online data selection during training, such as curriculum and self-paced learning, loss-based batch selection, and confidence thresholds for pseudo-labels (Bengio et al., 2009; Kumar et al., 2010; Mindermann et al., 2022; Sohn et al., 2020), faces the same feedback between chosen samples and future states, but is usually judged by end performance rather than by comparisons at matched counts. We borrow Shapley val-

ues and interaction indices (Shapley, 1953; Grabisch & Roubens, 1999; Sundararajan et al., 2020) to attribute value to selection stages rather than to individual samples, within one declared family of interventions.

3 Stateful Selection Interventions

Entropy-based TTA such as Tent predicts each incoming test batch with the current model and then updates the model on that batch. Selective methods insert one or more *selection stages* between prediction and update, so that only part of the batch drives the update. This section defines how we measure the value of those choices. Section 3.1 writes selective TTA as a loop, Section 3.2 defines interventions that change only which samples are kept, and Section 3.3 defines the quantities we estimate from them.

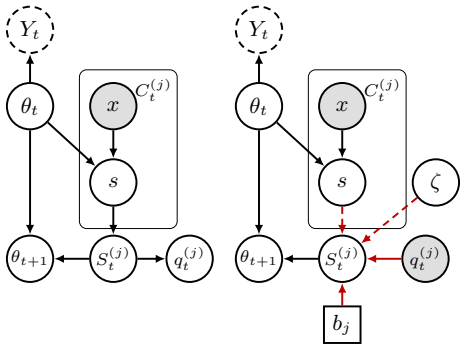
3.1 Selective TTA as a multi-stage selection loop

Test batches $B_1, \dots, B_T$ arrive in sequence. Let θ_t denote the adaptation state before batch t: the model parameters together with everything else the method carries across batches, such as normalization statistics, optimizer momentum, prediction history, and recovery state. The model predicts B_t with θ_t before any update; we write Y_t for its accuracy on B_t, and performance is the resulting online accuracy over the stream.

A selective method then passes B_t through L selection stages. Stage j receives a candidate set $C_t^{(j)}$, scores each candidate x by $s^{(j)}(x; \theta_t)$, and keeps a subset $S_t^{(j)} \subseteq C_t^{(j)}$. The first stage receives the whole batch, $C_t^{(1)} = B_t$, and each later stage receives the survivors of the previous one, $C_t^{(j)} = S_t^{(j-1)}$. The method then updates its state with the kept samples,

$$\theta_{t+1} = U\big(\theta_t,\, S_t^{(1)}, \dots, S_t^{(L)}\big), \tag{1}$$

where U covers everything after selection: loss weighting, the gradient step, history updates, and recovery. EATA, SAR, and DeYO each use $L = 2$ stages (Table 1): a low-entropy reliability stage followed by a method-specific refinement stage, each keeping the candidates whose score passes a fixed threshold. When a method computes between stages, as SAR does with its sharpness-aware perturbation (Foret et al., 2021), the second-stage score uses that intermediate result.



(a) Native (b) Intervened

Figure 1: **One selection stage.** (a) Native method. (b) Intervention (red): fixed count $q_t^{(j)}$, rule chosen by b_j. Shaded: observed or fixed; dashed Y_t: evaluation only.

Figure 1a draws one stage of this loop; with two stages, the second repeats the same pattern on the survivors before the update. The key property is feedback: the samples kept at batch t change θ_{t+1}, and θ_{t+1} determines the scores, and hence the kept samples, at every later batch. The accuracy Y_t depends on θ_t alone, because each batch is predicted before its own update; test labels are used only to compute it, never by the method.

3.2 Identity interventions at native counts

Changing a stage's rule usually changes both *which* samples it keeps and *how many*. To vary which samples are kept while holding how many fixed, we proceed in two steps. First, we run the native method and record how many samples each stage keeps at each batch, $q_t^{(j)} = |S_t^{(j)}|$. Second, we rerun the stream from the same source model, requiring every stage to keep exactly $q = q_t^{(j)}$ of its own candidates, chosen by one of two rules:

$$\underbrace{S_t^{(j)} = \mathrm{Top}_q\big(C_t^{(j)};\, s^{(j)}(\cdot; \theta_t)\big)}_{\text{native ranking}}, \qquad \underbrace{S_t^{(j)} \sim \mathrm{Unif}\big\{S \subseteq C_t^{(j)} : |S| = q\big\}}_{\text{random}}. \tag{2}$$

Native ranking keeps the q highest-scoring candidates, with scores computed from the rerun's own state θ_t; random keeps a uniformly random subset of the same size. The downstream weighting, update, and recovery rules remain unchanged, although the weights, gradients, and later states these rules produce respond to the new samples. The recorded counts are produced by the native run and consumed by the rerun, without forcing the rerun back onto the native states; they are an offline device for defining the comparison, not something a deployed method could use.

Crossing the two rules at the two stages gives four policies P_{ab}, where a and b give the rule at stages 1 and 2 ($1 =$ native ranking, $0 =$ random). P_{11} ranks natively at both stages, P_{00} is random at both, and the *hybrid* policies P_{10} and P_{01} are native at one stage only; with L stages the same construction gives 2^L policies. Figure 1b shows the change: the count becomes a fixed input, and a switch b_j chooses whether the kept set follows the scores ($b_j = 1$) or a random draw ζ ($b_j = 0$). P_{11} reproduces the native run exactly, which we verify by replay, because on native states the top-$q_t^{(j)}$ candidates are exactly those passing the threshold. In the hybrids the state drifts away from the native run, so keeping $q_t^{(j)}$ samples by ranking can include candidates that the native threshold would now reject: 0.27–2.73% of those kept by stage 1 in P_{10} and 10.05–56.54% of those kept by stage 2 in P_{01} (Appendix E).

Two consequences of this design matter for interpreting the results. First, *equal counts do not imply equal loss weights*. DeYO and EATA weight each kept sample by its entropy (DeYO also uses PLPD), so the low-entropy samples that native ranking keeps receive larger weights than random ones, and the native loss puts more total weight on the same number of samples. Section 5.2 shows that matching these weights removes most of the ViT-B/16 value. Second, *the two stages differ in selection capacity*, that is, in how much room they have to change the kept samples. Stage 1 chooses $q_t^{(1)}$ of the $|B_t|$ batch samples, whereas stage 2 chooses $q_t^{(2)}$ of only $q_t^{(1)}$ survivors. A native and a random choice of q out of n candidates differ on $2q(1 - q/n)$ samples in expectation (Appendix E), so when stage 2 keeps most of its candidates, switching its rule changes few samples, whatever its criterion. Section 5.2 finds that larger credit largely goes with larger room.

3.3 Identity value and its decomposition

Let Y_{ab} be the online accuracy of policy P_{ab}, averaged over random seeds (P_{11} is deterministic). The *identity value*

$$\tau_{\text{id}} = Y_{11} - Y_{00} \tag{3}$$

is the accuracy gained by choosing samples with the native scores rather than at random, with the same counts at every stage and batch. It is the total downstream effect of native selection under the method's own weighting and update rules, not a sample-quality effect with weights and states held fixed.

What is being changed? The intervention replaces a stage's rule at every batch. Each kept set shapes later states, so τ_{id} includes all downstream consequences. Freezing later decisions would define a different estimand and can become infeasible, because later candidates and scores are themselves produced by earlier choices. A finite-horizon performance-difference identity (Kakade & Langford, 2002; Schulman et al., 2015) decomposes the policy effect into native-versus-random choices with their future consequences (Appendix B). Thus the observation is a whole trajectory, not an individual update or test sample.

Splitting the value between stages. Two paths lead from P_{00} to P_{11} (Figure 2): switch stage 1 to native first ($P_{00} \to P_{10} \to P_{11}$), or stage 2 first ($P_{00} \to P_{01} \to P_{11}$). The paths order policy substitutions, not computation: every policy runs stage 1 before stage 2, and every node is a separate run from the same source model. Along each path, a stage's contribution is the accuracy change when it is switched on, and this can depend on the other stage. Stage 1 contributes $m_1^{(0)} = Y_{10} - Y_{00}$ while stage 2 is random and $m_1^{(1)} = Y_{11} - Y_{01}$ while stage 2 is native; symmetrically, $m_2^{(0)} = Y_{01} - Y_{00}$ and $m_2^{(1)} = Y_{11} - Y_{10}$. We credit each stage with the average of its two contributions, the two-player Shapley value (Shapley, 1953), and measure their disagreement by the interaction:

$$\phi_j = \tfrac{1}{2}\big(m_j^{(0)} + m_j^{(1)}\big), \qquad I_{12} = m_2^{(1)} - m_2^{(0)} = m_1^{(1)} - m_1^{(0)}. \tag{4}$$

Because both paths start at Y_{00} and end at Y_{11}, the credits add up exactly to the identity value:

$$\tau_{\mathrm{id}} = \phi_1 + \phi_2. \tag{5}$$

Equation (5) splits the value of native selection between the two stages, and I_{12} measures how much that split depends on the order in which the stages are switched on. (ϕ_j is also the single number closest, in the worst case, to both of its contributions, with worst-case error $|I_{12}|/2$; Appendix B.1.)

Reading the interaction. If $I_{12} < 0$, each stage helps less once the other is already native, as when both stages remove unreliable samples and the second finds fewer left to remove. If $I_{12} > 0$, each stage helps more when the other is native. A negative interaction alone does not distinguish diminishing returns from harmful composition: if stage 2 adds 3 points when stage 1 is random but only 1 point when stage 1 is native, then $I_{12} = -2$ although stage 2 helps in both cases. The following null model makes this precise.

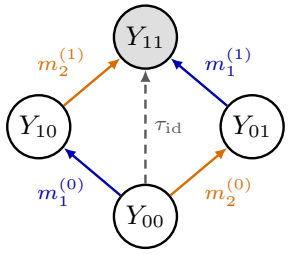


Figure 2: **Policy lattice.** Subscripts: rule at stages 1 and 2 (1 = native, 0 = random). Blue: stage 1 switched on; orange: stage 2.

Proposition 1 (Diminishing-returns null) *Suppose* $Y_{ab} = f(z_0 + a\alpha + b\beta)$ *with* $\alpha, \beta \geq 0$ *and* f *nondecreasing and concave. Then every contribution* $m_j^{(0)}, m_j^{(1)}$ *is nonnegative and* $I_{12} \leq 0$.

Here α and β are how much each stage adds to a latent adaptation quality, and concavity means that accuracy gains saturate as that quality grows. Under this null both stages help in every context and the interaction is nonpositive, so negative interaction alone is unsurprising (proof in Appendix B.2). What the null cannot produce is a stage that *hurts* in some context. A *composition reversal* is the clearest case: each stage helps on its own ($Y_{10} > Y_{00}$ and $Y_{01} > Y_{00}$), yet the two together do worse than random selection at both stages ($Y_{11} < Y_{00}$). This forces stage 2 to hurt once stage 1 is native, $m_2^{(1)} = Y_{11} - Y_{10} < 0$. Section 5.3 finds such reversals in 8 of 12 candidate SAR cells.

4 Experimental Setup

Shifts and data. We use all 15 ImageNet-C corruptions at severity 5 (Hendrycks & Dietterich, 2019). A fixed subset of 10,000 validation images is shared by all corruptions and presented in two fixed random orders, streams A and B, in batches of 64. Each corruption is adapted separately, starting from the source model. ColoredMNIST provides a gate-level boundary case; for the population analysis we additionally adapt on a 4,770-image Waterbirds stream and evaluate on a disjoint 1,024-image probe (Sagawa et al., 2020).

Methods and models. We study DeYO (Lee et al., 2024), EATA (Niu et al., 2022), and SAR (Niu et al., 2023) with the specified native configurations, on ResNet-50 with batch normalization and ViT-B/16 with layer normalization. Table 1 lists the two selection stages of each method. EATA's anti-forgetting regularizer uses a Fisher estimate from 1,400 clean source images (Appendix A); all other settings are in Appendix A.

Protocol. A method–model–corruption–stream combination is a *cell*: $3 \times 2 \times 15 \times 2 = 180$ cells. Each uses one native trajectory and 5 seeds for each of P_{00}, P_{10}, P_{01} at its recorded stage counts (Section 3.2), totaling 2,880 runs; P11 replay is exact (Appendix A). A sign is *resolved* when its 95% bootstrap interval excludes zero; these intervals condition on the fixed streams and models. ImageNet-C measures online accuracy and Waterbirds final probe accuracy, with differences in percentage points (pp). Follow-ups comprise a separate full-50,000-image Gaussian study with 20 seeds, a five-corruption weight-matching study with 10 seeds per policy, and expansion to 20 seeds in 12 SAR cells; the latter two reuse their corresponding native trajectories.

Research questions. **RQ1**: Does it matter which samples are selected once the number selected is fixed? **RQ2**: How is that value divided between the two stages, and is the division stable? **RQ3**: Do the two stages combine additively? **RQ4**: Does native selection help every population?

Table 1: **Selection stages of the studied methods.** Our interventions change which samples each stage keeps; weighting and update rules remain those of the method.

Method	Stage 1	Stage 2	Loss weighting	Update
DeYO	Low entropy	Large prediction change under patch shuffling (PLPD)	Entropy and PLPD	Weighted entropy minimization
EATA	Low entropy	Prediction differs from recent history	Entropy	Entropy plus Fisher penalty
SAR	Low entropy	Low entropy after a sharpness-aware perturbation	Uniform	SAM step with model recovery

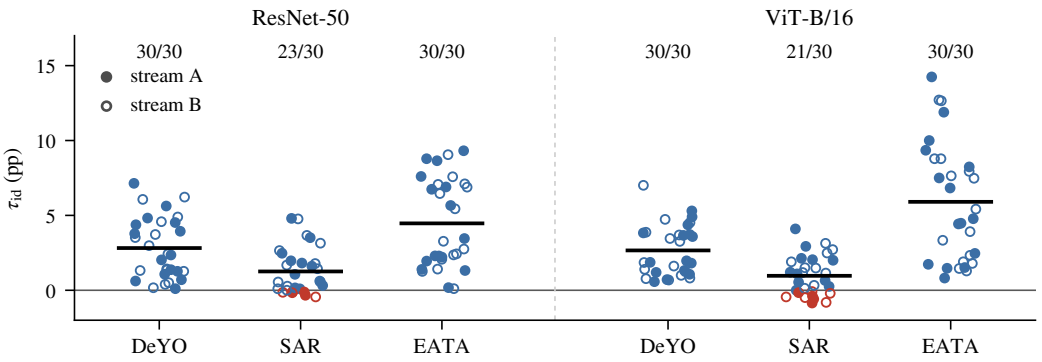


Figure 3: **Identity value in all 180 cells.** Each dot is one corruption and stream; filled and open dots are streams A and B, and red dots are negative. Black bars are means over the 30 cells of each method and model; numbers count positive cells. Random policies use 5 seeds.

5 Results

5.1 RQ1: Which samples are selected matters

Native selection outperforms exact-count random selection in 164/180 cells (Figure 3): all 60 DeYO and 60 EATA cells are positive, while all 16 negatives belong to SAR and 13 are within 0.5 pp of zero. ResNet-50/ViT-B/16 means are 2.8/2.7 pp for DeYO, 1.3/1.0 for SAR, and 4.5/5.9 for EATA. Matching counts at every stage and batch attributes this contrast to *which* samples are chosen, including their resulting loss weights (RQ2).

Shift effects range from near zero to over 12 pp and increase with source error (Appendix I).

5.2 RQ2: The split between stages depends on the comparison

Under the original intervention, stage 1 receives more credit than stage 2 in 151 of 180 cells and in every method–model mean (Table 2). Three checks show that this ordering partly reflects the comparison rather than the criteria.

Criterion or position? Swapping DeYO's stages, so that PLPD filters the whole batch and entropy filters its survivors, moves the larger credit to the new first stage: it wins in 20 of 20 cells and entropy in none, compared with 14 of 20 for entropy in the original order (5 seeds; Appendix J). The preliminary DualTTA-style selector implemented from the paper's description (official code unavailable), whose first decision uses prediction stability, also gives its first decision more credit in 4 of 5 preliminary cells, although its second decision keeps only a handful of samples (Appendix K).

Selection capacity. Stage 2 chooses among fewer candidates than stage 1 and keeps most of them (Table 2, last two columns); ViT-B/16 SAR's second stage keeps 96% of its candidates, so switching

Table 2: **Stage credit under the original intervention.** Means over the 30 cells of each method and model (pp); $\tau_{\mathrm{id}} = \phi_1 + \phi_2$. $Y_{11} < Y_{10}$: cells in which stage 2 lowers accuracy when stage 1 is native. Last two columns: share of the batch kept by stage 1 and share of its survivors kept by stage 2 (ratios pooled over batches within each cell, then averaged over cells); the closer to 100%, the less room a stage has to change which samples are kept.

Model	Method	τ_{id}	ϕ_1	ϕ_2	I_{12}	$Y_{11} < Y_{10}$	$q^{(1)}/\lvert B\rvert$	$q^{(2)}/q^{(1)}$
ResNet-50	DeYO	2.82	1.94	0.88	−0.08	0/30	69%	75%
	SAR	1.26	1.20	0.06	−1.05	26/30	42%	73%
	EATA	4.47	3.48	0.99	−0.19	0/30	56%	74%
ViT-B/16	DeYO	2.66	1.84	0.82	−0.29	0/30	86%	81%
	SAR	0.97	1.17	−0.21	−1.08	30/30	63%	96%
	EATA	5.91	3.95	1.95	−1.40	0/30	75%	75%

Table 3: **Matching loss weights redistributes stage credit.** Original → matched means (10 cells per row; 10 seeds per policy; pp). Last column: cells with $\phi_1 > \phi_2$; parentheses count independent-arm 95% intervals above zero. SAR has uniform weights.

Model	Method	τ_{id}	ϕ_1	ϕ_2	I_{12}	$\phi_1 > \phi_2$
ResNet-50	DeYO	$3.66 \to 3.33$	$2.35 \to 1.99$	$1.31 \to 1.34$	$0.01 \to -0.07$	$9(7) \to 6(4)$
ResNet-50	EATA	$5.53 \to 6.43$	$4.13 \to 4.57$	$1.40 \to 1.86$	$0.02 \to -1.11$	$9(9) \to 9(6)$
ViT-B/16	DeYO	$1.59 \to 0.38$	$1.15 \to 0.05$	$0.43 \to 0.32$	$-0.09 \to 0.04$	$7(5) \to 2(1)$
ViT-B/16	EATA	$3.64 \to 0.32$	$2.56 \to 0.03$	$1.08 \to 0.29$	$-0.89 \to -0.05$	$10(10) \to 2(0)$

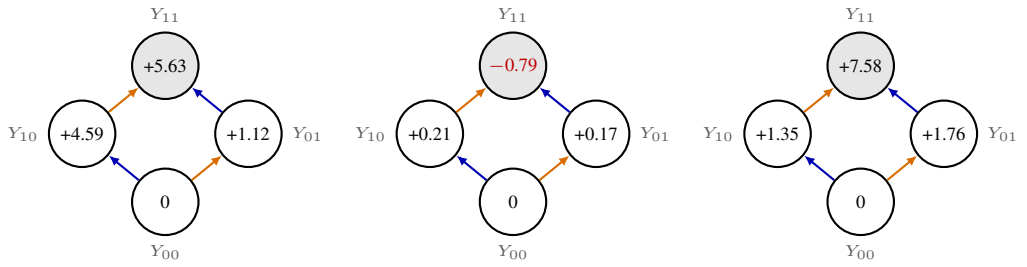
its rule changes few samples. Stage 1 has more room in 143 of 180 cells; among these it receives more credit in 130 (91%), versus 21 of the remaining 37 (57%) (Appendix E).

Loss weights. DeYO and EATA weight kept samples by entropy (and DeYO also by PLPD) and average over the number kept, so native sets carry 1.1–3.2 times the average weight of random sets. Rescaling the random and hybrid policies' weights to match native at every update (10 seeds, 40 cells; Table 3) removes 76% and 91% of the identity value of ViT-B/16 DeYO and EATA, with resolved changes, and reduces first-stage dominance from 35 to 19 of 40 cells, with 8 of 17 reversals resolved (9 with joint resampling; Appendix A). Because $\phi_1 - \phi_2 = Y_{10} - Y_{01}$, which stage receives more credit depends only on the two hybrid policies. On ResNet-50, identity value decreases by 0.33 pp for DeYO and rises for EATA, almost entirely because of one cell (contrast, stream B; excluding it, the mean change is +0.065 pp). Tuning the loss-weight scale of random selection makes the same point more strongly: with the best of three weight scales ($0.5\times$, $1\times$, $2\times$ native, chosen post hoc per cell), random selection outperforms native selection in 25 of 40 cells, and at $2\times$ it does so in all 10 ViT-B/16 DeYO cells (5 seeds; Appendix F). The comparisons with the method's own weights and with matched weights are both well defined but answer different questions, and stage credit depends on which one is asked.

Refinement is method-specific. Given a native first stage, stage 2 helps in all 60 DeYO and all 60 EATA cells but hurts in 56 of 60 SAR cells (Table 2, column $Y_{11} < Y_{10}$). Averages can hide this: ResNet-50 SAR's second stage has almost no credit ($\phi_2 = 0.06$) because it helps when stage 1 is random (+0.585 pp) and hurts when stage 1 is native (-0.467 pp).

5.3 RQ3: Stages that help alone can hurt together

Interaction is negative in 155 of 180 cells (Appendix C), but by Proposition 1 diminishing returns alone can produce this, so a negative interaction does not show that the stages conflict: ViT-B/16 EATA has the most negative mean interaction (-1.40 pp; Table 2), yet its second stage helps in every cell. What the null cannot produce is a stage that hurts, and the clearest case is a composition reversal (Figure 4b). The 5-seed estimates flag 12 SAR cells as candidates. With 20 seeds, 8 of the 12 have all three effects resolved on the predicted side of zero (Figure 5); using only the 15 seeds added after the candidates were chosen, 7 of 12 do. None of these candidate trajectories triggers



(a) Near-additive
ResNet-50 DeYO, Gaussian A
$I_{12} = -0.08$

(b) Composition reversal
ViT-B/16 SAR, contrast A
$I_{12} = -1.17$, 20 seeds

(c) Positive interaction
ResNet-50 EATA, contrast B
$I_{12} = +4.48$
(matched, 10 seeds: $+0.88$)

Figure 4: **Three cells on the policy lattice** (Figure 2). Node values are accuracy relative to random selection at both stages, $Y_{ab} - Y_{00}$ (pp); blue edges switch stage 1 to native ranking, orange edges stage 2. In (b), all three signs are resolved by 95% intervals over 20 seeds. (a) and the lattice nodes in (c) use 5 seeds; the matched interaction annotated in (c) uses 10 seeds. The positive interaction in (c) shrinks under weight matching (Section 5.2).

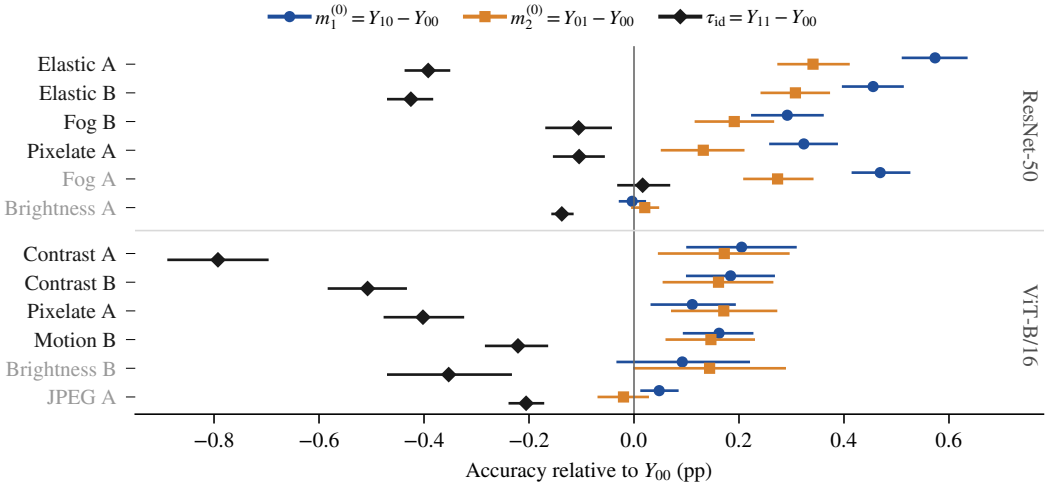


Figure 5: **Composition-reversal check in 12 SAR cells (20 seeds).** Accuracy change when only stage 1 (blue), only stage 2 (orange), or both stages (black) switch from random to native ranking, relative to random selection at both stages, with 95% intervals over seeds. A reversal requires the blue and orange intervals above zero and the black interval below zero; this holds in the 8 cells with dark labels.

SAR recovery (Appendix G). Interaction can also be positive (Figure 4c): for ResNet-50 EATA on contrast, stream B, the original five-seed interaction is $+4.48$ pp, while the matched ten-seed interaction is $+0.88$ pp.

5.4 RQ4: Aggregate gains can hide group harms tied to loss weighting

Waterbirds groups cross bird type and background; waterbirds on land (g_2, 111/1,024 probe images) are worst (one image is 0.9 pp). We analyze DeYO; transferred EATA/SAR thresholds suppress or undo adaptation (Appendix H). Native selection (N), exact-count random selection (R), and random additionally matching native group counts at both stages (G) give $N - R = (G - R) + (N - G)$; group labels serve only this diagnostic. Worst-group gains of 19.5/22.9 pp in streams A/B, equivalent to about 22/25 images, split into allocation 8.3/12.1 pp and within-group 11.2/10.8 pp (Figure 6a). Joint 95% bootstrap intervals over probe images and seeds exclude zero for both parts (Appendix H).

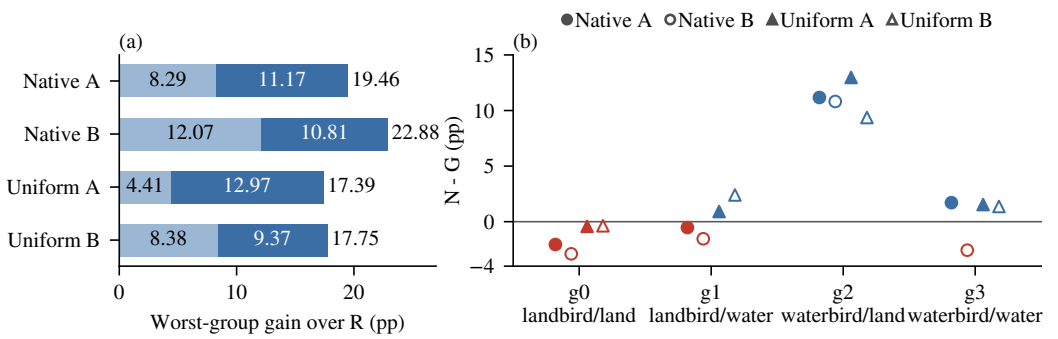


Figure 6: **Where DeYO's worst-group gain on Waterbirds comes from.** (a) Worst-group gain over exact-count random selection (R), split into group allocation ($G - R$) and within-group choice ($N - G$), for native and uniform loss weights and streams A/B. (b) Within-group effect $N - G$; negative estimates largely disappear under uniform weights. Groups (probe images): g_0 landbird/land (380), g_1 landbird/water (416), g_2 waterbird/land (111), g_3 waterbird/water (117). R uses 10 seeds and G uses 5.

Under native weights, within-group choices hurt landbirds on land (g_0, 380 images) by 2.1/2.9 pp, equivalent to about 8/11 images, with smaller negative estimates elsewhere (Figure 6b). With uniform weights, a separate policy family at the same counts with its own random baselines, the g_0 loss shrinks to 0.4/0.4 pp, other groups have nonnegative estimates, and g_2 retains within-group gains of 13.0/9.4 pp. The within-group benefit thus persists without native weights while other-group harms largely attenuate, paralleling the identity–weighting dependence in RQ2.

A boundary case. On ColoredMNIST (Arjovsky et al., 2019), a label-free gate accepts entropy-minimization updates improving perturbation consistency. There is no resolved overall-accuracy difference from count-matched random gates in six conditions, while color-conflicting accuracy is lower than BatchNorm-only in all six (Appendix L). Selection need not add value when adaptation rewards a shortcut; this gate-level control differs from our per-stage interventions.

6 Conclusion

Selective TTA methods choose which test samples drive adaptation, and because every choice reshapes the states in which later samples are scored, the value of a selection stage cannot be read from a single update or from a one-component ablation. Measured with stateful interventions at the native counts, native selection has higher accuracy than random selection in 164 of 180 cells, so which samples are kept does matter. The credit that selection receives, however, is intervention-dependent: it follows a stage's position and capacity more than its criterion, it shrinks or vanishes once random selection receives matched or better-chosen loss weights, stages that help alone can hurt together, and on Waterbirds the harms to non-worst groups largely weaken under uniform loss weights. When the adaptation objective itself rewards a shortcut, as on ColoredMNIST, selection shows no resolved advantage over random gating. For evaluating selective TTA, these results suggest three practices: compare against random selection at the same counts rather than against adaptation on all samples; report each stage's effect both with and without the other stage; and state, and where possible match, the loss weights before attributing value to a selection criterion.

Limitations. Because our interventions reuse the native counts, they measure which samples to keep, not how many; the value of choosing the counts remains open. The three main methods all begin with an entropy stage; the swapped-order DeYO and the DualTTA-style selector, a preliminary five-cell study of our own implementation, are the only exceptions. The two models differ in architecture, normalization, and pretraining at once, so differences between them cannot be attributed to any single factor. Finally, our intervals reflect variation over random seeds for fixed streams and models, not over new shifts or checkpoints.

AI USE STATEMENT

Generative AI tools were used primarily to assist with manuscript editing and software implementation, including code generation, debugging, and refactoring. They were also used to provide feedback on experimental design and to assist with the organization and interpretation of experimental results. All methodological decisions, experiments, quantitative results, and final claims were reviewed and verified by the authors. The authors take full responsibility for the content of this work, including text, code, and analyses produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The appendix provides detailed descriptions of the selection interventions, experimental configurations, Fisher construction, evaluation protocol, and validation procedures, together with proofs of the theoretical results. An anonymized implementation, including experiment configurations and analysis scripts, is available at `https://anonymous.4open.science/r/iclr27-tta-anon-3660/`. The supplementary material additionally contains numerical result tables and CSV files underlying the analyses reported in the paper.

REFERENCES

Shyma Alhuwaider, Yasmeen Alsaedy, Merey Ramazanova, Silvio Giancola, and Bernard Ghanem. GoTTA be diverse: Rethinking memory policies for test-time adaptation, 2026. URL `https://arxiv.org/abs/2605.19890`. arXiv:2605.19890.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL `https://arxiv.org/abs/1907.02893`. arXiv:1907.02893.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *International Conference on Machine Learning, ICML*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380. URL `https://ronan.collobert.com/pub/matos/2009_curriculum_icml.pdf`.

Hyeongheon Cha, Dong Min Kim, Hye Won Chung, Taesik Gong, and Sung-Ju Lee. SNAP: Low-latency test-time adaptation with sparse updates. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/820e42f39773c6cbbd875553db45658f-Abstract-Conference.html`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://arxiv.org/abs/2010.11929`.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL `https://arxiv.org/abs/2010.01412`.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2242–2251, 2019. URL `https://proceedings.mlr.press/v97/ghorbani19c.html`.

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, pp. 27253–27266, 2022. doi: 10.52202/068431-1976. URL `https://papers.neurips.cc/paper_files/paper/2022/hash/ae6c7dbd9429b3a75c41b5fb47e57c9e-Abstract-Conference.html`.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. SoTTA: Robust Test-Time Adaptation on Noisy Data Streams. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14070–14093, 2023. doi: 10.52202/

075280-0620. URL https://papers.neurips.cc/paper_files/paper/2023/hash/2da53cd1abdae59150e35f4693834f32-Abstract-Conference.html.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125. URL https://link.springer.com/article/10.1007/s001820050125.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://arxiv.org/abs/1512.03385.

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.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent Test-time Adaptation in Recurring Testing Scenarios. In *Advances in Neural Information Processing Systems*, volume 37, pp. 123402–123442, 2024. doi: 10.52202/079017-3923. URL https://papers.nips.cc/paper_files/paper/2024/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding Predictions with Data and Data with Predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9525–9587, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Siru Jiang, Yuwei Liang, Jian Liang, Ran He, and Tieniu Tan. To adapt or not to adapt? selective adaptation for vision-language models. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/3715.

Yong-Yeon Jo, Junho Song, and Joon-myoung Kwon. RECAST: Recent & context-aware sampling for test-time adaptation in streaming biosignals, 2026. URL https://arxiv.org/abs/2608.28271. arXiv:2608.28271.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002. URL https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/KakadeLangford-icml2002.pdf.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.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, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-Paced Learning for Latent Variable Models. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

Yongchan Kwon and James Zou. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 8780–8802, 2022. URL https://proceedings.mlr.press/v151/kwon22a.html.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025. doi: 10.1007/s11263-024-02181-w. URL https://link.springer.com/article/10.1007/s11263-024-02181-w.

Yuejiang Liu, Parth Kothari, Bastien 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, pp. 21808–21820, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b618c3210e934362ac261db280128c22-Abstract.html.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized Training on Points that are Learnable, Worth Learning, and not yet Learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15630–15649, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

Saypraseuth Mounsaveng, Florent Chiaroni, Malik Boudiaf, Marco Pedersoli, and Ismail Ben Ayed. Bag of tricks for fully test-time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. URL https://arxiv.org/abs/2310.02416.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating Prediction-Time Batch Normalization for Robustness under Covariate Shift, 2020. URL https://arxiv.org/abs/2006.10963. arXiv: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*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 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://arxiv.org/abs/2302.12400.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 27074–27113, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Nam Nguyen Phuong, Duc Nguyen The Minh, Phi Le Nguyen, Ehsan Abbasnejad, and Minh Hoai. Dual strategies for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 2483–2492, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Phuong_Dual_Strategies_for_Test-Time_Adaptation_CVPRF_2026_paper.html.

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*, volume 36, pp. 39915–39935, 2023. doi: 10.52202/075280-1735. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7d640f377893fc5f22b5610e175ef7c3-Abstract-Conference.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

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*, volume 33, pp. 11539–11551, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/85690f81aadc1749175c187784afc9ee-Abstract.html.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games*, volume 2. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei 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, pp. 9229–9248, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9259–9268, 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

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://openreview.net/forum?id=uXl3bZLkr3c.

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 (CVPR)*, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Zehao Xiao and Cees G. M. Snoek. Beyond Model Adaptation at Test Time: A Survey, 2024. URL https://arxiv.org/abs/2411.03687. arXiv:2411.03687.

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 (CVPR)*, pp. 15922–15932, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yuan_Robust_Test-Time_Adaptation_in_Dynamic_Scenarios_CVPR_2023_paper.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, pp. 38629–38642, 2022. doi: 10.52202/068431-2799. URL https://papers.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

A IMPLEMENTATION DETAILS

This section supports the execution-matched protocol in Section 4 by specifying the models, update rules, and evaluation context.

Models and preprocessing. ResNet-50 (He et al., 2016) uses `resnet50-19c8e357.pth`; ViT-B/16 (Dosovitskiy et al., 2021; Steiner et al., 2022) uses `vit_base_patch16_224.augreg_in21k_ft_in1k`, with 12 blocks and ImageNet-21K to ImageNet-1K pretraining. ImageNet-C images use center crop to 224, RGB tensor conversion, and ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$, without an extra resize; clean Fisher views use Resize256 followed by CenterCrop224. The ViT checkpoint's alternative mean/std and interpolation are not substituted. Waterbirds uses its binary source checkpoint, direct 224×224 resize, and ImageNet normalization. Architecture, normalization, pretraining, and adapted parameters change together between the model families.

Table 4: Native ImageNet-C settings. There are $C = 1000$ classes; H is entropy, H' is entropy after the SAM perturbation, and d is absolute cosine similarity to prediction history. BN/LN counts refer to affine tensors.

Model	Method	Learning rate	Stage 1	Stage 2	Parameters	Weights
ResNet-50	DeYO	2.50×10^{-4}	$H < 0.50 \log C$	PLPD > 0.30	86 BN	w_D
	SAR	2.50×10^{-4}	$H < 0.40 \log C$	$H' < 0.40 \log C$	86 BN	1
	EATA	2.50×10^{-4}	$H < 0.40 \log C$	$d < 0.05$	106 BN	w_E
ViT-B/16	DeYO	1.00×10^{-3}	$H < 0.50 \log C$	PLPD > 0.20	36 LN	w_D
	SAR	1.00×10^{-3}	$H < 0.40 \log C$	$H' < 0.40 \log C$	36 LN	1
	EATA	1.00×10^{-3}	$H < 0.40 \log C$	$d < 0.05$	50 LN	w_E

All methods use SGD momentum 0.90, zero weight decay, and one update opportunity per batch without episodic resets within a corruption. DeYO/SAR exclude ResNet layer 4 and adapt ViT blocks 0–8 (30,592/27,648 scalars); EATA adapts all BN/LN affine parameters (53,120/38,400 scalars). Detached weights are $w_D = \exp(0.40 \log C - H) + \exp(\text{PLPD})$ and $w_E = \exp(0.40 \log C - H)$, divided by the selected count. DeYO shuffles a 4×4 patch grid. SAR uses nonadaptive SAM radius 0.05, a full-batch second forward, entropy EMA 0.90, and recovery threshold 0.20; EATA uses history EMA 0.90 and an all-survive stage 2 before history exists. Waterbirds DeYO uses learning rate 1.25×10^{-3}, binary entropy thresholds, and PLPD threshold 0.50.

Fisher estimate. Each model's EATA regularizer uses 1,400 clean source images spanning 762 synsets, source-model argmax pseudo-labels, and the equal-batch average of squared batch-mean cross-entropy gradients over 22 batches (21 of size 64, one of size 56). The source anchor is unchanged and the regularizer multiplier is 2000. This differs from the official 2,000-image protocol, for which no sensitivity result is claimed; all 1,400 clean-image identities also occur as corrupted views in the 10,000-image target subset, so their identities are not disjoint.

Streams and seeds. All 15 corruptions share 10,000 image identities; streams A/B permute them with seeds 2024/2025 into 157 batches of size 64, including a final batch of 16. Each corruption restarts from source. Every random policy in the primary panel uses 5 seeds, 30000–30004 for ResNet and 40000–40004 for ViT; adaptation randomness is initialized separately with seed 2024. The Gaussian study uses 50,000 images and 782 batches; the 12 SAR candidates extend each random policy to 20 seeds while retaining its first five runs. Waterbirds R/G use 10/5 seeds, 10000–10009 and 20000–20004.

Replay validation. P11 exactly replays native scores, selections, logits, loss weights, parameters, optimizer/normalization/history/recovery states, and adaptation randomness; P00 reproduces the nested exact-quota random policy. Hybrid policies recompute rankings at their own current or intermediate states, and stage 2 receives that policy's stage-1 survivors, with native weighting, SAM, Fisher, history, empty-mask and recovery branches preserved. Checks cover all 2,880 trajectories,

Table 5: Sensitivity of interval classifications to joint identity-root resampling. "Primary" compares new independent and paired diagnostic intervals on the same 180 cells, not previously published primary-cell interval labels; primary point-count claims are unchanged. Full per-cell intervals and classifications are included in the accompanying scalar tables.

Panel	Contrast or summary	Changed classifications
10-seed weight panel	$Y_{10} - Y_{01}$ (80 program/cells)	2
	Mean cross-program $\Delta\tau_{\mathrm{id}}$ (4 groups)	1
Primary 180 cells	$m_1^{(0)}$, $m_2^{(0)}$, I_{12}	3, 7, 12
	ϕ_1, ϕ_2, $Y_{10} - Y_{01}$	2, 12, 5
	τ_{id}, $m_2^{(1)}$	0, 0
Swapped-order / preliminary panels	Reported cell and pooled intervals	0 / 0

preserve incoming reduction order after sampling without replacement, and use evaluation labels only for measurement.

Uncertainty. The original ImageNet intervals resample whole random-policy trajectories independently within policies with native fixed. The shared identity roots do couple policies; the joint-resampling sensitivity below preserves that dependence. These intervals condition on source, stream, quotas and numerical context, without estimating variation over new shifts or checkpoints. Waterbirds uses 4,000 joint bootstrap draws over within-group probe identities and R/G trajectories, with common probe weights across policies and the worst-group metric recomputed in each draw.

Identity roots and coupling sensitivity. Each stochastic trajectory initializes a dedicated CPU generator with its identity seed. It advances only when a random stage calls `randperm`; native-ranked stages leave it unchanged. Thus equal roots couple the streams even though P00, P10 and P01 consume different numbers and shapes of draws and need not select equal identities. The initial five roots are shared across these policies, across the original and weight-matched programs, and across cells within a model family (30000–30004 for ResNet; 40000–40004 for ViT). The five added roots in the 10-seed panel are disjoint across policies and programs: for ResNet, P00/P10/P01 use 710000–710004, 711000–711004, 712000–712004 in the original program and 713000–713004, 714000–714004, 715000–715004 in the matched program; ViT adds 10000 to each. Roots remain shared across cells within each such list. Swapped DeYO uses shared roots 610000–610004 (ResNet) and 620000–620004 (ViT); the preliminary selector uses 810000–810004 across its policies and cells. The P00 scale sweep shares the original five roots across scales; it reports point comparisons without bootstrap intervals. Selection randomness is isolated from the adaptation generator.

The reported independent-arm intervals are retained as the original analysis; their seed labels do not justify treating the simulations as independent. We additionally resample actual shared-root indices jointly across arms, programs and fixed cells, keeping native endpoints fixed (50,000 percentile draws). In the 10-seed panel, this preserves the five-old/five-new composition, couples the old roots, and resamples the new disjoint roots independently. A separate sensitivity pairs all ten ordinal indices; this imposes an arbitrary coupling on the new roots. Both sensitivities resolve 9/17 forward reversals, versus 8/17 under the original independent analysis; the reverse reversal remains resolved, and original first-stage dominance remains resolved in 31/35 cells. They change no point estimates. Stratifying the independent analysis into the same two cohorts already resolves the ninth reversal, so this difference is not attributed solely to coupling.

For the 10-seed panel, matched ViT/DeYO defocus A changes from a resolved positive stage difference, $[0.007, 0.218]$, to unresolved, $[-0.006, 0.224]$ pp; matched ViT/EATA contrast B changes from unresolved, $[-0.242, 0.014]$, to resolved negative, $[-0.233, -0.012]$. The ResNet/DeYO mean matched-minus-original identity-value change remains -0.3268 pp, but its interval changes from $[-0.6533, 0.0125]$ to $[-0.5242, -0.1002]$. Across the primary panel, 41 cell–contrast diagnostic classifications change in 36 cells; these include $m_1^{(0)}$, $m_2^{(0)}$, I_{12} and both credits, rather than only the stage-order contrast. The swapped-order and preliminary-selector panels retain every reported resolved/unresolved classification under joint resampling. These sensitivities remain conditional on the evaluated trajectories and fixed cells, without image- or shift-population coverage.

B PROOFS

This section supports the policy-level interpretation, stage-credit summary, and diminishing-returns null in Section 3. Write $\theta_t^{(j)}$ for the state immediately before stage j of batch t, with $\theta_t^{(1)} = \theta_t$ and $\theta_t^{(L+1)} = \theta_{t+1}$. For this proof, the state includes the model, normalization, optimizer, history and recovery variables, the current candidate set $C_t^{(j)}$, within-batch intermediates such as SAR's perturbation, and the reference quota $q_t^{(j)}$. The action is the kept set $S_t^{(j)} \subseteq C_t^{(j)}$ of that size. Condition on the initial state, stream, numerical context, and native quota trace, which is endogenous to the native trajectory. A policy π specifies the selection kernel $\pi_{t,j}(\cdot \mid \theta_t^{(j)})$; all compared policies share the transition program $\mathsf{T}_{t,j}$, including weighting and computation between stages. Let π^R denote the reference policy.

Performance difference. Let $N_{\text{img}} = \sum_{t=1}^{T} |B_t|$. For ImageNet-C, assign reward $r_{t,1}(\theta_t) = |B_t| Y_t / N_{\text{img}}$ before selection at stage 1, and $r_{t,j} = 0$ for $j > 1$. This gives image-weighted online accuracy. To include terminal evaluation, define

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{j=1}^{L} r_{t,j}(\theta_t^{(j)}) + g(\theta_{T+1}) \right].$$

ImageNet-C uses $g \equiv 0$; Waterbirds uses $r_{t,j} \equiv 0$ and g equal to the final probe metric. Define the reference continuation value by the terminal condition $V^R_{T+1,1} = g$ and the batch-boundary convention $V^R_{t,L+1} = V^R_{t+1,1}$. Working backward, for a state θ and legal kept set S, set

$$\begin{aligned} Q^R_{t,j}(\theta, S) &= r_{t,j}(\theta) + \mathbb{E}_{\mathsf{T}_{t,j}}\left[V^R_{t,j+1}(\theta') \mid \theta, S \right], \\ V^R_{t,j}(\theta) &= \mathbb{E}_{S \sim \pi^R_{t,j}(\cdot|\theta)} Q^R_{t,j}(\theta, S), \\ \mathcal{A}^R_{t,j}(\theta, S) &= Q^R_{t,j}(\theta, S) - V^R_{t,j}(\theta). \end{aligned}$$

Here θ' is the next stage state (or the next batch state when $j = L$). For integrable rewards, common initialization μ_0, shared transition/reward rules, and legal policies at the reached states, the tower rule gives

$$\begin{aligned} &\sum_{t=1}^{T} \sum_{j=1}^{L} \mathbb{E}_\pi \mathcal{A}^R_{t,j}(\theta_t^{(j)}, S_t^{(j)}) \\ &\quad = \mathbb{E}_\pi \sum_{t=1}^{T} \sum_{j=1}^{L} \left[r_{t,j}(\theta_t^{(j)}) + V^R_{t,j+1}(\theta_t^{(j+1)}) - V^R_{t,j}(\theta_t^{(j)}) \right] \\ &\quad = J(\pi) - \mathbb{E}_{\theta_1 \sim \mu_0} V^R_{1,1}(\theta_1) = J(\pi) - J(\pi^R). \end{aligned}$$

The second equality telescopes within and across batches, retaining the terminal term g. This finite-horizon performance-difference identity (Kakade & Langford, 2002; Schulman et al., 2015) uses states reached under π and continuation under π^R; replacing those states by the reference policy's states requires an approximation. With $\pi = P_{11}$ and $\pi^R = P_{00}$, it decomposes τ_{id} into stage advantages along native trajectories with random-policy continuation. The effect includes downstream state changes. Final-batch adaptation cannot change that batch's cached online prediction; it can change a terminal probe metric. Neither independent images nor differentiable accuracy is required; local gradient utility is a different loss surrogate.

B.1 THE MIDPOINT SUMMARY

For the contextual marginals $u = m_j^{(0)}$ and $v = m_j^{(1)}$, the triangle inequality gives $|u - v| \leq 2\max\{|u - a|, |v - a|\}$ for any scalar a. The midpoint $a = (u + v)/2 = \phi_j$ attains the bound, so $\min_a \max\{|u-a|, |v-a|\} = |I_{12}|/2$; uniqueness follows because the closed intervals of that radius around u, v meet only at the midpoint (also when $u = v$). This is the two-stage Shapley average (Shapley, 1953); the algebra holds for true values and estimates but does not resolve uncertain signs.

B.2 Diminishing-returns null

For Proposition 1, let $s = \alpha + \beta > 0$. Concavity gives

$$f(z_0 + \alpha) \geq (\beta/s)f(z_0) + (\alpha/s)f(z_0 + s),$$
$$f(z_0 + \beta) \geq (\alpha/s)f(z_0) + (\beta/s)f(z_0 + s).$$

Adding yields $I_{12} \leq 0$; when $s = 0$, all endpoints coincide. Monotonicity makes all contextual marginals nonnegative. A reliably negative marginal or resolved composition reversal contradicts the combined assumptions, without identifying which assumption fails.

C Per-cell results

This section supports RQ1–RQ3 by retaining every signed identity value, interaction, and stage profile across the 180 cells.

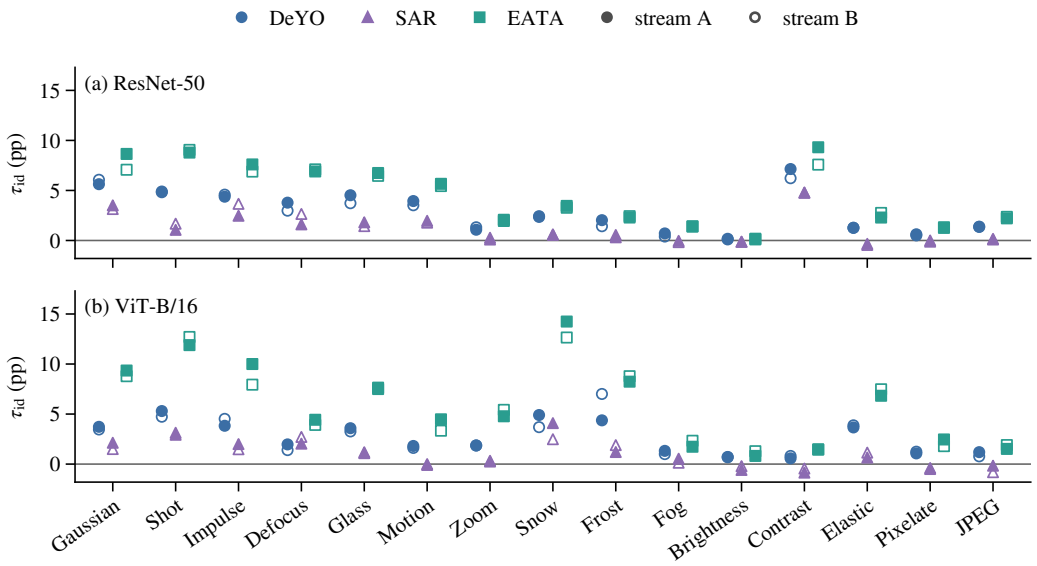


Figure C1: Identity value by corruption for ResNet-50 (top) and ViT-B/16 (bottom), in pp; 5 seeds per random policy. Colors and marker shapes identify methods; filled/open points are streams A/B. Small and negative estimates remain visible.

D Robustness

This section supports RQ1 by checking the signed identity-value pattern across stream length, order, numerical context, and model/configuration family.

Stream length. The separate full-50,000-image Gaussian-noise severity-5 study uses ResNet-50 and 20 seeds per random policy. Native beats all 20 controls in each method/stream, and these longer trajectories are not pooled with the 10,000-image factorial.

Stream order. Streams A/B have the same sign in 88/90 identity-value pairs and 71/90 interaction pairs. The largest interaction discrepancy is ResNet-50 EATA contrast, -2.64 versus $+4.48$ pp; stream permutations also change their native quota traces, so this is a comparison of the resulting conditional policies.

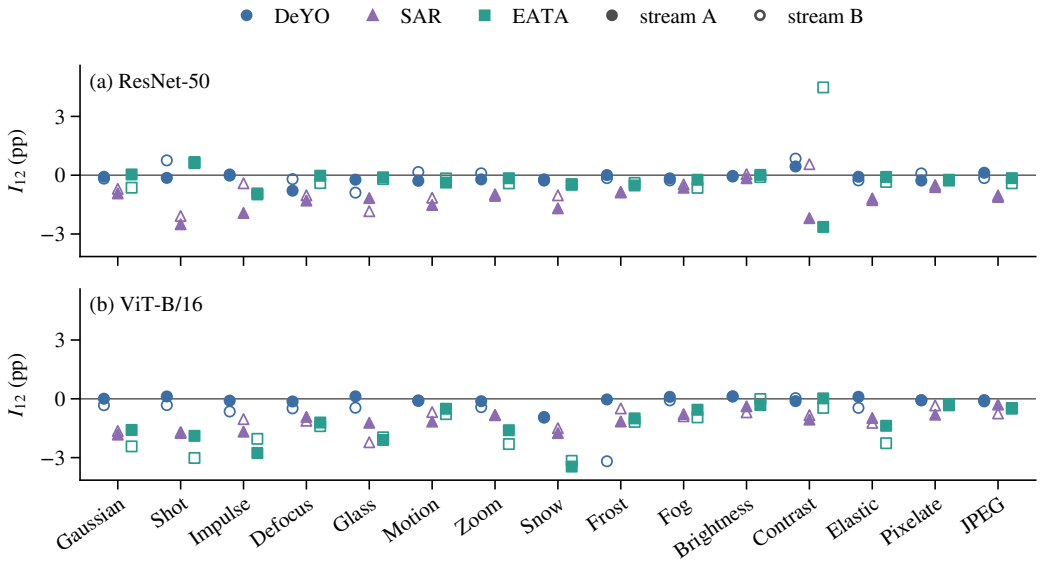


Figure C2: Interaction I_{12} by corruption, with the same methods, streams, and 5 seeds per random policy. Negative interaction is compatible with beneficial conditional refinement.

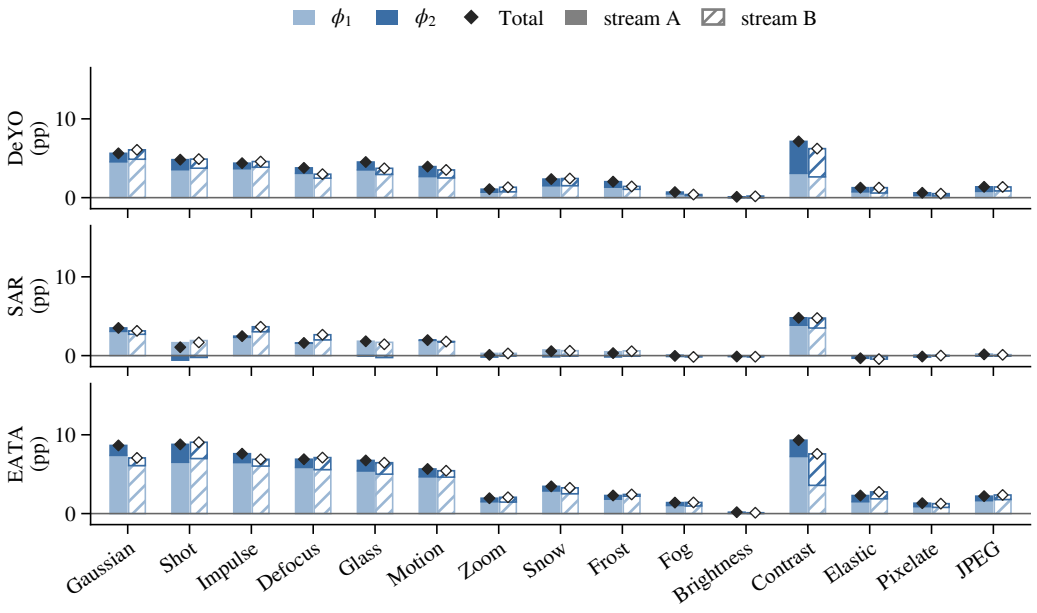


Figure C3: ResNet-50 stage profiles (pp; 5 seeds per random policy). Bars show ϕ_1, ϕ_2, stacked separately by sign; diamonds show $\tau_{\text{id}} = \phi_1 + \phi_2$. Solid/hatched bars and filled/open diamonds identify streams A/B.

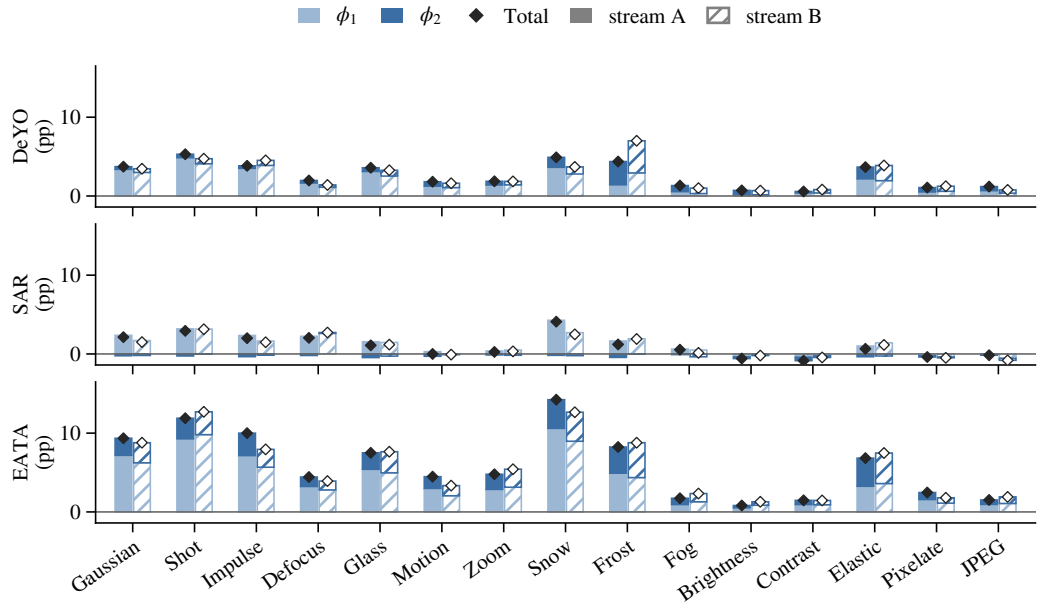


Figure C4: ViT-B/16 stage profiles, using the same encoding. Component insertion order is distinct from stream order: each stream has its own quota trace and factorial.

Table 6: Separate 50,000-image Gaussian-noise severity-5 ResNet-50 study, with 20 seeds per random policy.

Method	Stream	N (%)	R mean±SD	τ_{id} (pp)	95% CI	Beats
DeYO	A	36.66	34.18±0.15	2.48	[2.42, 2.54]	20/20
DeYO	B	36.32	34.03±0.22	2.29	[2.19, 2.38]	20/20
SAR	A	31.62	28.76±0.15	2.86	[2.80, 2.93]	20/20
SAR	B	31.59	28.48±0.18	3.11	[3.03, 3.19]	20/20
EATA	A	35.85	30.91±0.31	4.94	[4.81, 5.08]	20/20
EATA	B	35.80	30.72±0.70	5.08	[4.83, 5.42]	20/20

Model family. ResNet-50 and ViT-B/16 have the same identity-value sign in 80/90 corresponding cells. The comparison changes architecture, normalization, checkpoint/pretraining, learning rate, adapted tensors and some thresholds together, and therefore does not isolate an architecture effect.

E SELECTION CAPACITY

This section supports RQ2's qualification that larger stage credit can accompany greater capacity to change selected identities. For a fixed q-of-n set and a uniform draw of the same size, expected overlap is q^2/n and expected symmetric difference is $2q(1 - q/n)$; empty candidate sets have zero difference. Here n is candidate count and q is kept count. Within a batch, abbreviate B_t as B and $q_t^{(j)}$ as q_j. Stage 1 uses $(n, q) = (|B|, q_1)$ and stage 2 uses (q_1, q_2), conditional on its own survivors. Write Δ_j for the symmetric difference between the native-ranked and random kept sets at stage j. Nested-uniform P00 is instead marginally uniform q_2-of-$|B|$, so its final expected overlap with native is $q_2^2/|B|$, not q_2^2/q_1.

Table 7: Selection capacity across 30 cells per row. Retention ratios pool batches within a cell, then average cells; expected symmetric differences sum over batches, then average cells.

Model	Method	$q_1/\lvert B\rvert$ (%)	q_2/q_1 (%)	$\mathbb{E}\lvert\Delta_1\rvert$	$\mathbb{E}\lvert\Delta_2\rvert$
ResNet-50	DeYO	68.62	74.51	3660.80	2553.47
ResNet-50	SAR	41.53	73.18	4004.31	1242.90
ResNet-50	EATA	56.31	74.43	4256.09	2020.23
ViT-B/16	DeYO	86.23	81.07	2023.21	2519.75
ViT-B/16	SAR	62.51	95.73	4093.44	467.16
ViT-B/16	EATA	74.59	74.85	3287.98	2701.78

Table 8: Cells with $\phi_1 > \phi_2$ / cells in each capacity stratum. Empty strata are shown as dashes.

Model	Method	$\mathbb{E}\lvert\Delta_1\rvert > \mathbb{E}\lvert\Delta_2\rvert$	$\mathbb{E}\lvert\Delta_1\rvert \le \mathbb{E}\lvert\Delta_2\rvert$
All	All	130/143	21/37
ResNet-50	DeYO	22/25	1/5
ResNet-50	SAR	29/30	–
ResNet-50	EATA	27/28	0/2
ViT-B/16	DeYO	10/10	10/20
ViT-B/16	SAR	25/30	–
ViT-B/16	EATA	17/20	10/10

The larger-stage-1 stratum has first-stage dominance in 130/143 cells, versus 21/37 otherwise; these strata also differ in methods and corruptions. Capacity is therefore an association, not a per-changed-identity causal normalization or an intrinsic criterion ranking.

At a hybrid policy's own state, choosing the native-ranked top quota can include samples failing the native numerical threshold. Across method–model means, this share ranges from 0.267 to 2.726% for P10 stage 1 and from 10.046 to 56.542% for P01 stage 2. Each trajectory pools failing and selected counts over batches; means then give equal weight to 30 cells and 5 seeds per cell. This is part of the quota-constrained intervention. All native P11 threshold sets and hybrid native-score rankings agree with their specified policies, including empty masks and EATA's no-history branch.

F WEIGHT MATCHING

This section supports RQ2's finding that stage credit changes when the comparison also controls mean selected loss weight.

For each nonempty selected objective, let q_t be its size, $\widetilde{w}_{t,i}$ its own-state native loss weights, and m_t^N the native reference mean; weight matching applies the detached scalar

$$c_t = \frac{m_t^N}{q_t^{-1} \sum_i \widetilde{w}_{t,i}}, \qquad w_{t,i}^M = \mathrm{stopgrad}(c_t)\widetilde{w}_{t,i}. \tag{6}$$

Table 9: Kept samples failing the native threshold at the hybrid's native-ranked stage (%; 30 cells and 5 seeds per cell).

Model	Method	P10 stage 1	P01 stage 2
ResNet-50	DeYO	2.726	31.465
	EATA	2.717	45.163
	SAR	1.662	56.542
ViT-B/16	DeYO	0.879	10.046
	EATA	1.871	26.726
	SAR	0.267	33.483

Table 10: Historical 5-seed loss-weight audit: mean selected-weight ratios across the ten five-corruption/stream cells per method and model. Each cell compares trajectory-mean loss weights after averaging the five controls; $\bar{w}_{ab}$ is the trajectory-mean selected loss weight under P_{ab}; these are not gradient-norm ratios.

Model	Method	$\bar{w}_{11}/\bar{w}_{00}$	Cell range	$\bar{w}_{10}/\bar{w}_{00}$	$\bar{w}_{01}/\bar{w}_{00}$
ResNet-50	DeYO	1.80	[1.31, 2.36]	1.57	1.21
ResNet-50	SAR	1.00	[1.00, 1.00]	1.00	1.00
ResNet-50	EATA	3.24	[1.56, 5.38]	3.01	1.45
ViT-B/16	DeYO	1.14	[1.03, 1.35]	1.15	1.02
ViT-B/16	SAR	1.00	[1.00, 1.00]	1.00	1.00
ViT-B/16	EATA	1.60	[1.29, 2.36]	1.47	1.28

The gap in average weight arises because DeYO and EATA average weighted losses over the number of kept samples; a self-normalized weighting, which divides by the sum of the weights, would remove it by construction. Within-set relative weights and the selected-count denominator remain unchanged, EATA's Fisher penalty is outside the scalar, and empty-mask, optimizer, history and recovery rules are preserved. Each policy follows its own state trajectory; P11 is reused with scale one. Gaussian noise, defocus blur, fog, contrast and JPEG compression give ten cells per method/model, with 10 seeds per random policy. SAR has uniform weights, making the operation a no-op.

The contrast-B outlier. P00 accuracy changes from 26.05% to 17.60%, with 5 of 10 matched seeds below 15%, while $Y_{11} = 34.09\%$ is reused. Identity value changes from 8.04 to 16.50 pp. The mean P00 scale is 2.44 across 1570 nonempty updates. Figure 4(c) retains its explicitly labeled original five-seed example.

Table 11: ResNet-50 EATA contrast B P00 trajectories (10 per program). The first five roots are shared across programs; the five added roots are disjoint. Appendix A reports joint-resampling sensitivity. Native accuracy is 34.09%.

Original seed	Accuracy (%)	Matched seed	Accuracy (%)
30000	27.99	30000	12.42
30001	24.86	30001	11.90
30002	27.15	30002	14.25
30003	25.72	30003	21.86
30004	26.83	30004	20.46
710000	26.55	713000	8.64
710001	25.30	713001	11.33
710002	23.84	713002	27.11
710003	26.58	713003	28.46
710004	25.70	713004	19.52
Mean	26.05	Mean	17.60

Weight matching defines a different transition program: it matches neither gradient norm/direction nor parameter displacement, momentum, normalization influence or the Fisher/data balance. The sensitivity can attenuate or amplify identity value and does not identify a mediated percentage or a uniquely correct effect.

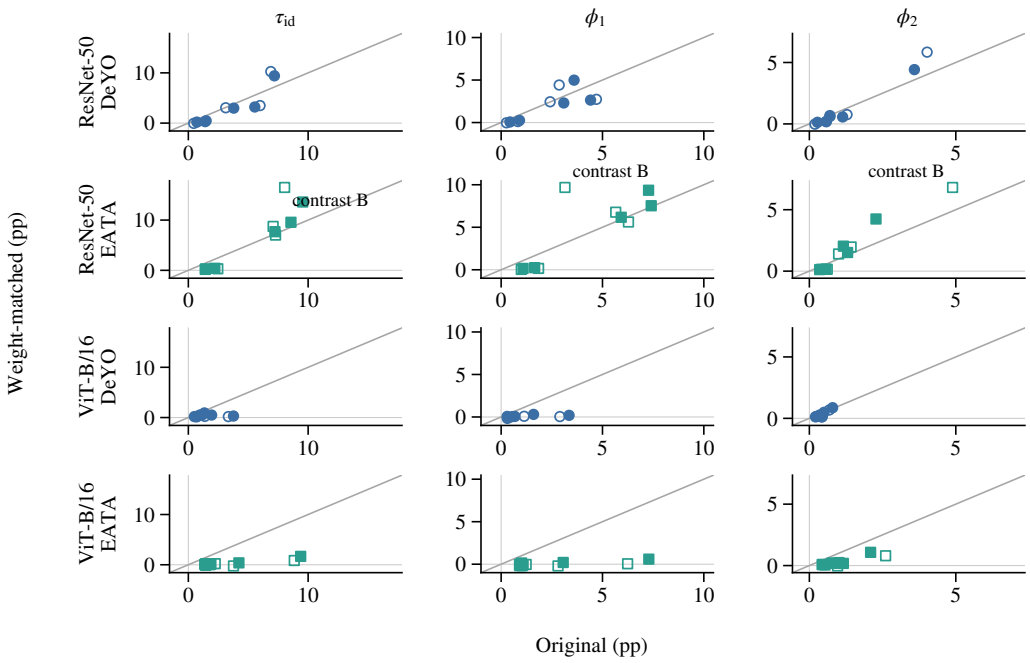


Figure F1: Original versus weight-matched identity value and stage credit (pp). Each row contains ten cells; filled/open markers are streams A/B, with method colors as in Appendix C. The diagonal is equality, and columns share limits across rows. ResNet-50 EATA contrast B is labeled.

Seed variation in stage ordering. With 10 seeds per stochastic policy, stage-1 dominance is resolved in 31 of its 35 cells under the original comparison. Weight matching leaves stage 1 dominant in 19 of 40 cells, with 17 positive-to-negative flips and 8 resolved reversals, plus 1 reverse flip. We independently resample whole trajectories within each P00/P10/P01 arm and program (10 seeds; 50,000 draws, 2.50/97.50 percentiles), reusing native P11. For stage ordering, $\phi_1 - \phi_2 = Y_{10} - Y_{01}$. A resolved positive-to-negative reversal requires an original interval strictly above zero and a matched interval strictly below zero. For each method/model, $\Delta\tau_{\mathrm{id}}$ averages original minus matched over ten fixed cells; native cancels, leaving matched P00 minus original P00. These conditional intervals do not resample corruptions or images or adjust for multiple comparisons. The analysis uses seed 20260924 and input index $2i + p$ for cell i and program $p = 0/1$, drawing P00, P10 and P01 in order. New identity seeds are disjoint from earlier ranges. Supplementary CSVs retain every per-cell interval and both flip directions.

Table 12: Historical independent-arm trajectory bootstrap (pp), with 10 seeds per stochastic policy. Stage-1 counts are out of ten cells; parentheses count intervals strictly above zero. Contrast intervals average ten fixed cells; $\Delta\tau_{\mathrm{id}}$ is original minus matched.

Model	Method	Program	Stage 1 > 2	$Y_{10} - Y_{01}$ [95% CI]	$\Delta\tau_{\mathrm{id}}$ [95% CI]
ResNet-50	DeYO	Original	9 (7)	1.05 [0.86, 1.22]	–
ResNet-50	DeYO	Matched	6 (4)	0.66 [0.41, 0.91]	0.33 [-0.01, 0.65]
ResNet-50	EATA	Original	9 (9)	2.73 [2.55, 2.90]	–
ResNet-50	EATA	Matched	9 (6)	2.72 [2.32, 3.12]	-0.90 [-1.51, -0.29]
ViT-B/16	DeYO	Original	7 (5)	0.72 [0.68, 0.76]	–
ViT-B/16	DeYO	Matched	2 (1)	-0.27 [-0.31, -0.23]	1.21 [1.14, 1.28]
ViT-B/16	EATA	Original	10 (10)	1.47 [1.42, 1.53]	–
ViT-B/16	EATA	Matched	2 (0)	-0.26 [-0.31, -0.21]	3.32 [2.97, 3.55]

Outlier sensitivity and medians. With 10 seeds per arm and program, the mean change in τ_{id} (matched minus original) for ResNet-50 EATA is $+0.9041$ pp over all ten cells and $+0.064889$

pp over the nine excluding contrast B. Median changes over the ten cells are -0.6615, $+0.0995$, -0.5865, and -1.9080 pp for ResNet-50 DeYO, ResNet-50 EATA, ViT-B/16 DeYO, and ViT-B/16 EATA, respectively. These signed changes use the opposite convention to $\Delta\tau_{\mathrm{id}}$ in Table 12.

Scaling P00 weights. With 5 seeds on 40 cells, we scale P00 weights to $0.5\times$ or $2\times$ the native mean per update, reusing the existing $1\times$ controls. At $2\times$, mean P00 accuracy exceeds native in 10/10 ViT/DeYO cells (mean $\tau_{\mathrm{id}} = -1.82$ pp). The ResNet-50 EATA increase at $1\times$ reverses under the best per-cell scale: 3.34 versus 5.40 pp under the original five-seed comparison. Best-scale selection is post hoc and makes identity value negative in 22 additional cells. With hybrids fixed, $Y_{10} - Y_{01}$ and stage ordering are unchanged; this is not a complete factorial at each scale.

Table 13: P00 weight scales (5 seeds): mean τ_{id} over ten cells, in pp. Best maximizes mean P00 accuracy per cell; the last column counts winning scales (ties retained). The $1\times$ controls reuse the five-seed weight-matched cohort, separate from the expanded panel.

Model	Method	0.5×	1×	2×	Best	Wins 0.5/1/2
ResNet-50	DeYO	3.43	3.51	7.69	2.37	3/3/4
ResNet-50	EATA	4.27	6.57	14.24	3.34	6/0/4
ViT-B/16	DeYO	3.23	0.37	-1.82	-1.82	0/0/10
ViT-B/16	EATA	3.63	0.50	-0.07	-1.18	0/3/7

G COMPOSITION-REVERSAL CHECK

This section separates the signs of isolated and joint stage effects in the twelve SAR candidates. Each stochastic policy retains its initial 5 seeds and adds 15, for 20 seeds, with the native endpoint fixed.

Held-out seed check. The twelve candidates were selected using the first five trajectories per policy. Repeating the same analysis using only the fifteen subsequently added seeds gives 7/12 supported reversals, compared with 8/12 when all twenty are retained. We use 50,000 whole-trajectory bootstrap draws, analysis seed 20260924, and 2.50/97.50 percentile intervals. The historical twenty-seed analysis resamples policy arms independently; the fifteen-seed analysis uses disjoint random streams across arms. Each draw reuses its P00 mean in $m_1^{(0)} = Y_{10} - Y_{00}$, $m_2^{(0)} = Y_{01} - Y_{00}$ and $\tau_{\mathrm{id}} = Y_{11} - Y_{00}$. Support requires both isolated intervals strictly above zero and the joint interval strictly below zero. Table 14 retains the original independent-arm twenty-seed intervals for comparison, while Table 15 reports the held-out effects. The coupling sensitivity below addresses the shared roots in the original five seeds.

Table 14: SAR composition-reversal check (pp). Intervals use the historical independent-arm 20-seed bootstrap; Held-out uses only the 15 added seeds. RN: ResNet-50; ViT: ViT-B/16. Support uses unrounded interval signs.

Cell	$m_1^{(0)}$ [95% CI]	$m_2^{(0)}$ [95% CI]	τ_{id} [95% CI]	20 seeds	Held-out
RN Brightness A	-0.00 [-0.03, 0.02]	0.02 [-0.01, 0.05]	-0.14 [-0.16, -0.12]	No	No
RN Elastic A	0.57 [0.51, 0.64]	0.34 [0.27, 0.41]	-0.39 [-0.44, -0.35]	Yes	Yes
RN Elastic B	0.46 [0.40, 0.51]	0.31 [0.24, 0.37]	-0.42 [-0.47, -0.38]	Yes	Yes
RN Fog A	0.47 [0.41, 0.53]	0.27 [0.21, 0.34]	0.02 [-0.03, 0.07]	No	No
RN Fog B	0.29 [0.22, 0.36]	0.19 [0.12, 0.27]	-0.11 [-0.17, -0.04]	Yes	Yes
RN Pixelate A	0.32 [0.26, 0.39]	0.13 [0.05, 0.21]	-0.10 [-0.15, -0.06]	Yes	Yes
ViT Brightness B	0.09 [-0.03, 0.22]	0.14 [-0.00, 0.29]	-0.35 [-0.47, -0.23]	No	No
ViT Contrast A	0.21 [0.10, 0.31]	0.17 [0.05, 0.30]	-0.79 [-0.89, -0.70]	Yes	Yes
ViT Contrast B	0.18 [0.10, 0.27]	0.16 [0.05, 0.27]	-0.51 [-0.58, -0.43]	Yes	Yes
ViT JPEG A	0.05 [0.01, 0.09]	-0.02 [-0.07, 0.03]	-0.21 [-0.24, -0.17]	No	No
ViT Motion B	0.16 [0.09, 0.23]	0.15 [0.06, 0.23]	-0.22 [-0.28, -0.16]	Yes	Yes
ViT Pixelate A	0.11 [0.03, 0.19]	0.17 [0.07, 0.27]	-0.40 [-0.48, -0.32]	Yes	No

Table 15: Held-out SAR effects using only the 15 seeds added after candidate selection, with independent-arm 95% intervals (pp).

Cell	$m_1^{(0)}$ [95% CI]	$m_2^{(0)}$ [95% CI]	τ_{id} [95% CI]
RN Brightness A	-0.01 [-0.04, 0.02]	0.02 [-0.01, 0.05]	-0.14 [-0.17, -0.12]
RN Elastic A	0.56 [0.49, 0.64]	0.34 [0.26, 0.42]	-0.41 [-0.47, -0.36]
RN Elastic B	0.46 [0.39, 0.52]	0.31 [0.23, 0.39]	-0.42 [-0.47, -0.37]
RN Fog A	0.49 [0.43, 0.56]	0.29 [0.22, 0.38]	0.03 [-0.03, 0.10]
RN Fog B	0.33 [0.26, 0.41]	0.21 [0.13, 0.30]	-0.09 [-0.15, -0.02]
RN Pixelate A	0.31 [0.24, 0.38]	0.13 [0.04, 0.22]	-0.10 [-0.15, -0.05]
ViT Brightness B	0.04 [-0.11, 0.19]	0.12 [-0.04, 0.28]	-0.40 [-0.53, -0.26]
ViT Contrast A	0.24 [0.13, 0.34]	0.20 [0.06, 0.33]	-0.77 [-0.87, -0.68]
ViT Contrast B	0.19 [0.09, 0.28]	0.14 [0.03, 0.26]	-0.53 [-0.61, -0.45]
ViT JPEG A	0.03 [-0.01, 0.07]	-0.04 [-0.10, 0.01]	-0.22 [-0.26, -0.19]
ViT Motion B	0.10 [0.03, 0.17]	0.11 [0.01, 0.21]	-0.27 [-0.33, -0.21]
ViT Pixelate A	0.09 [-0.01, 0.19]	0.14 [0.02, 0.27]	-0.41 [-0.50, -0.31]

Multiplicity and joint signs. Bonferroni-adjusted percentile bounds over the 36 isolated/joint contrasts use tail probability $0.05/(2 \times 36)$, giving 4/12 supported cells with the historical twenty-seed bootstrap and 2/12 with the held-out fifteen seeds. Table 16 reports the fraction P^* of bootstrap draws satisfying $m_1^{(0)} > 0$, $m_2^{(0)} > 0$ and $\tau_{\mathrm{id}} < 0$ jointly, retaining their shared P00 reference. This is a bootstrap fraction, not a posterior probability or a separate significance test; values of one mean all 50,000 draws satisfied the inequalities. These are conditional seed intervals for the fixed cells. Multiplicity adjustment does not remove selection bias from the twenty-seed analysis, which includes the original candidate-selection trajectories.

Table 16: Joint bootstrap fraction $P^* = P^*(m_1^{(0)} > 0, m_2^{(0)} > 0, \tau_{\mathrm{id}} < 0)$ and support using Bonferroni-adjusted intervals over 36 contrasts. The 20-seed columns use the historical independent-arm bootstrap; 15-seed columns use held-out seeds.

Cell	P^* (20)	P^* (15)	Bonf. (20)	Bonf. (15)
RN Brightness A	0.37776	0.20350	No	No
RN Elastic A	1.00000	1.00000	Yes	Yes
RN Elastic B	1.00000	1.00000	Yes	Yes
RN Fog A	0.26126	0.14920	No	No
RN Fog B	0.99936	0.99260	Yes	No
RN Pixelate A	0.99916	0.99724	No	No
ViT Brightness B	0.91180	0.67094	No	No
ViT Contrast A	0.99588	0.99778	No	No
ViT Contrast B	0.99826	0.99428	No	No
ViT JPEG A	0.20798	0.06370	No	No
ViT Motion B	0.99962	0.98144	Yes	No
ViT Pixelate A	0.99712	0.95986	No	No

Random-number coupling. The original five identity seeds are shared across P00/P10/P01 (30000–30004 for ResNet-50; 40000–40004 for ViT). Each policy initializes its own CPU generator from that root seed; random stages advance it and ranked stages do not. Thus the roots couple policies despite their different draw consumption. The added fifteen seeds are disjoint across arms: the corresponding roots are 30005–30019, 31005–31019 and 32005–32019 for ResNet-50, with 10000 added for ViT. A design-aware sensitivity resamples the original five seed indices jointly across arms and the fifteen added seeds independently within each arm, preserving the two cohort sizes and reusing P00 in every contrast. It retains 8/12 supported reversals (6/12 after Bonferroni adjustment). ViT brightness B changes from an unresolved $m_2^{(0)}$ interval $[-0.0015, 0.2895]$ to a positive interval $[0.0220, 0.2690]$ pp; its $m_1^{(0)}$ remains unresolved, so the cell remains unsupported. The held-out result is 7/12. As an additional sensitivity, jointly resampling all twenty row indices gives 8/12 (4/12 after adjustment); this artificially pairs the disjoint new seeds and is not the randomization-design analysis. It also resolves brightness-B $m_2^{(0)}$, at $[0.0080, 0.2860]$ pp. Stratifying the independent

bootstrap into the old and new cohorts also resolves this contrast, so the change is not attributable solely to coupling. All $Y_{10} - Y_{01}$ and I_{12} sign classifications are unchanged under these checks. Supplementary readouts retain every interval and joint probability for each resampling scheme.

Unresolved cases and recovery. Under the historical twenty-seed analysis, ResNet fog A has an unresolved joint effect, while ResNet brightness A, ViT brightness B and ViT JPEG A have unresolved isolated effects. The held-out analysis additionally leaves ViT pixelate A unresolved: its $m_1^{(0)}$ interval is $[-0.0067, 0.1947]$ pp. Recovery counts are zero for every policy in every candidate cell: 0 for each of the 12 native trajectories and for each of the 240 P00, 240 P10 and 240 P01 trajectories (732 total, 157 batches each). Thus none of these candidate trajectories triggers SAR recovery.

H WATERBIRDS

This section supports RQ4 by retaining the population-specific allocation and within-group effects, together with the execution reason for omitting EATA/SAR.

Waterbirds (Sagawa et al., 2020) groups are $g = 2y + b$, where y indicates waterbird and b water background: g_0 landbird/land, g_1 landbird/water, g_2 waterbird/land and g_3 waterbird/water. The 4,770-image adaptation stream has group counts 1875/1839/531/525; the disjoint 1,024-image probe has 380/416/111/117 and fixed 64-image observer contexts. R has 10 seeds and G has 5, with G matching native group counts at both stages in every batch, including the final 34-image batch. G uses current-batch group metadata, which contains class information, only for this supervised diagnostic; it uses no future outcomes or probe predictions. Each trajectory's worst-group accuracy is computed before averaging; the minimum is not exchanged with expectation.

Table 17: Waterbirds DeYO with native weights (pp). Total = N − R; allocation = G − R; within = N − G, using means over 10 R and 5 G seeds.

	Stream A			Stream B		
Metric	Total	Allocation	Within	Total	Allocation	Within
Overall	1.33	0.90	0.43	0.38	1.20	-0.82
Macro	4.76	2.18	2.57	4.19	3.23	0.95
Worst group	19.46	8.29	11.17	22.88	12.07	10.81
g_0	-2.68	-0.63	-2.05	-3.45	-0.55	-2.89
g_1	-0.14	0.38	-0.53	-1.75	-0.22	-1.54
g_2	19.46	8.29	11.17	22.88	12.07	10.81
g_3	2.39	0.68	1.71	-0.94	1.62	-2.56

Under uniform weights, worst-group total gains are 17.39/17.75 pp, split into allocation 4.41/8.38 and within-group 12.97/9.37 pp in streams A/B. Uniform N is the score-ranked policy at the native-weight reference's total quotas, and G matches that uniform policy's own group composition.

For native weights, joint 95% intervals over probe images and seeds are $[3.96, 13.15]/[7.12, 17.66]$ pp for worst-group allocation and $[5.11, 17.66]/[4.50, 16.76]$ pp for its within-group term in streams A/B. Group 0 within-group intervals are $[-3.58, -0.79]/[-4.74, -1.32]$ pp, excluding zero in both streams; under uniform weights, the estimates shrink to $-0.42/-0.37$ pp with intervals $[-1.16, 0.00]/[-1.05, 0.00]$. All other uniform-weight group estimates are nonnegative, and the group 2 benefit remains. This contrast concerns independently evolving weighting-policy families, including their scale and selection changes; it does not isolate a weight-only causal pathway.

The algebra $N - R = (G - R) + (N - G)$ is exact for these policy means, but its allocation/within-group interpretation depends on the G intervention: group counts do not fix total group-specific loss weight, gradients or normalization influence. Group 2 benefits coexist with native-weight group 0 within-group harms of $-2.05/-2.89$ pp and stream-B group 3 harm of -2.56 pp; the overall stream-B within-group effect is -0.82 pp with interval $[-2.23, 0.61]$, leaving its sign unresolved.

Why EATA/SAR are omitted. Their binary-task reliability cutoff is $0.40 \log 2$, but EATA retains history-cosine threshold 0.05 and SAR retains recovery threshold 0.20. For binary history m and

probability vector p, $\cos(m, p) \geq \min(m_0, m_1)/\|m\|_2$; the initialized histories imply lower bounds 0.24/0.23, already exceeding 0.05. Because an empty selection leaves the prediction history unchanged, the same bound holds at every later batch by induction, so the exclusion persists: EATA performs only its initial update and rejects stage-2 candidates in all remaining 74 batches. Its separate 2,000-source-training-image Fisher has zero penalty at initialization and subsequent empty masks skip updates, so it does not cause this exclusion. SAR executes 75 nonempty SAM cycles and immediately recovers after all 75 because its entropy EMA stays below 0.20, leaving zero immediately retained parameter updates; momentum/EMA may survive. These transferred configurations do not measure calibrated binary-task method capability, whereas DeYO's adaptation persists across the stream.

I Additional analyses

This section supports the limits of difficulty and local-update explanations of RQ1–RQ3.

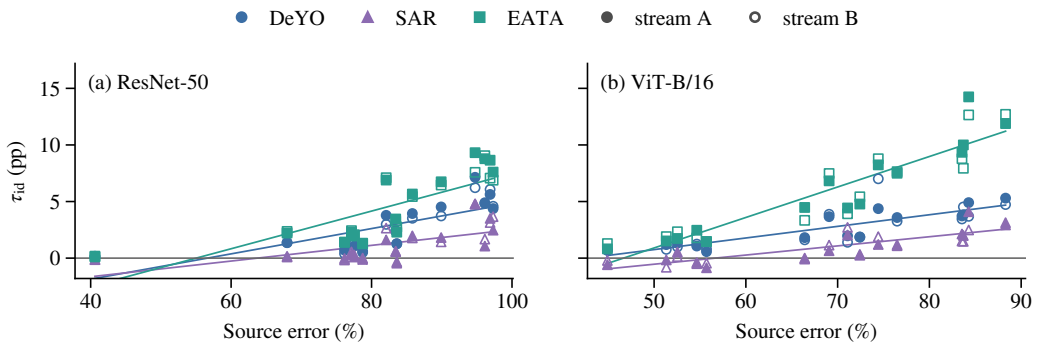


Figure I1: Identity value versus source error. Methods and streams use the encoding in Appendix C. Thin lines fit each method separately, unlike the common-slope regression below.

A method-offset plus common source-error slope explains 51.64%/54.09% of within-method squared variation for ResNet and 64.47%/59.37% for ViT in streams A/B (about 52–65%), using eval-mode source predictions on the same identities. This reduction relative to method offsets alone is not a causal or population variance fraction, and residuals do not establish a shared difficulty-independent profile.

Fixed-state update utility. At five native states per trajectory, offline probe labels define loss L_P and gradient $g_P^* = \nabla L_P$; g_i is the TTA gradient for selected set i, and η denotes step size. We compare native against five random draws at the same state using dot product and cosine, motivated by $L_P(\theta - \eta g_i) - L_P(\theta) \approx -\eta\langle g_P^*, g_i\rangle$; the five-corruption ImageNet probes contain 512 held-out identities and never alter adaptation. DeYO/EATA use total TTA gradients, including Fisher for EATA; SAR uses its second SAM gradient at the perturbed state, leaving unavailable objectives missing. Five-point correlations are descriptive, with EATA counterexamples; on Waterbirds, positive mean utility toward group 0 and stream-B group 3 coexists with harmful N−G accuracy and cross-entropy. Normalization, optimizer state and evolving trajectories limit local explanations.

Finite-difference checks of the probe gradient pass only with TF32 disabled, so we read this diagnostic qualitatively.

Table 18: N − R utility advantage versus endpoint identity value: Pearson / Spearman over five corruptions. SAR uses the second SAM gradient; DeYO/EATA use total gradients.

Method	Stream	Gradient	Dot: r/ρ	Cosine: r/ρ
DeYO	A	Total	0.74 / 0.70	0.72 / 0.60
DeYO	B	Total	0.57 / 0.50	0.40 / 0.30
SAR	A	SAM second	0.83 / 0.90	0.25 / 0.30
SAR	B	SAM second	0.91 / 1.00	0.79 / 0.60
EATA	A	Total	0.31 / 0.40	-0.63 / -0.60
EATA	B	Total	-0.53 / -0.70	-0.57 / -0.50

J CRITERION OR POSITION?

We reverse DeYO on the five-corruption panel: PLPD selects from the whole batch, then the entropy stage selects from its survivors. The model-specific PLPD thresholds (0.3 for ResNet-50, 0.2 for ViT-B/16), entropy cutoff ($0.5 \log 1000$), detached weights, optimizer and other settings remain unchanged. The PLPD transformed forward processes every stage-1 candidate. Each cell has its own native swapped trajectory and recorded quotas; the original-order comparison reuses the original five-seed trajectories.

All 20 cells pass exact P11 and P00 replay and independent checks of quotas, recorded identities, own-state score rankings, loss weights, losses and saved-logit endpoints. The 300 stochastic swapped trajectories use 5 seeds per policy. For $Y_{10} - Y_{01}$, we independently resample whole trajectories within each arm (50,000 draws), with the native endpoint fixed. Model intervals average ten fixed cells, rather than resampling shifts. The swapped and original programs have their own counts and transformation populations; this comparison does not isolate a criterion at fixed capacity.

Table 19: DeYO stage-order comparison (pp). O: entropy then PLPD; S: PLPD then entropy. The difference is first-stage minus second-stage credit, $Y_{10} - Y_{01}$ [95% interval]. F/E counts cells with larger first-stage/entropy credit; model rows are means over ten cells. All stochastic policies use 5 seeds.

Cell	Order	τ_{id}	ϕ_{PLPD}	ϕ_{entropy}	I_{12}	Difference [95% CI]	F/E
ResNet-50							
Gaussian A	O	5.63	1.08	4.55	-0.08	3.47 [3.14, 3.79]	1/1
Gaussian A	S	5.44	4.63	0.81	-0.49	3.81 [3.32, 4.31]	1/0
Gaussian B	O	6.07	1.19	4.88	-0.17	3.68 [3.35, 3.99]	1/1
Gaussian B	S	5.03	4.60	0.43	-0.02	4.17 [3.78, 4.53]	1/0
Defocus A	O	3.78	0.70	3.08	-0.79	2.39 [2.06, 2.77]	1/1
Defocus A	S	4.57	3.91	0.66	0.00	3.25 [2.73, 3.82]	1/0
Defocus B	O	2.98	0.51	2.47	-0.20	1.96 [1.36, 2.48]	1/1
Defocus B	S	3.02	2.42	0.60	-0.16	1.83 [1.23, 2.48]	1/0
Fog A	O	0.70	0.23	0.48	-0.16	0.25 [0.13, 0.37]	1/1
Fog A	S	0.81	0.71	0.11	-0.12	0.60 [0.45, 0.74]	1/0
Fog B	O	0.39	0.14	0.25	-0.26	0.11 [-0.05, 0.27]	1/1
Fog B	S	0.84	0.83	0.01	0.06	0.82 [0.72, 0.92]	1/0
Contrast A	O	7.14	4.06	3.08	0.45	-0.97 [-3.32, 1.12]	0/0
Contrast A	S	9.01	8.23	0.78	-0.70	7.45 [6.09, 8.89]	1/0
Contrast B	O	6.22	3.58	2.64	0.84	-0.94 [-2.07, 0.21]	0/0
Contrast B	S	6.98	6.90	0.08	1.04	6.82 [6.03, 7.51]	1/0
JPEG A	O	1.37	0.57	0.80	0.12	0.23 [-0.03, 0.47]	1/1
JPEG A	S	1.62	1.42	0.21	-0.03	1.21 [1.08, 1.34]	1/0
JPEG B	O	1.36	0.51	0.86	-0.15	0.35 [0.21, 0.52]	1/1
JPEG B	S	1.55	1.47	0.08	-0.12	1.39 [1.20, 1.59]	1/0
Model mean	O	3.57	1.26	2.31	-0.04	1.05 [0.78, 1.31]	8/8
Model mean	S	3.89	3.51	0.38	-0.05	3.14 [2.95, 3.33]	10/0
ViT-B/16							
Gaussian A	O	3.72	0.34	3.38	0.01	3.04 [2.83, 3.24]	1/1
Gaussian A	S	3.79	2.82	0.98	-0.40	1.84 [1.56, 2.11]	1/0
Gaussian B	O	3.46	0.47	2.99	-0.33	2.52 [2.43, 2.63]	1/1
Gaussian B	S	3.68	2.67	1.01	-0.31	1.66 [1.49, 1.87]	1/0
Defocus A	O	1.97	0.32	1.65	-0.14	1.32 [1.15, 1.49]	1/1
Defocus A	S	1.84	1.37	0.47	-0.20	0.90 [0.62, 1.13]	1/0
Defocus B	O	1.40	0.29	1.11	-0.49	0.82 [0.62, 1.06]	1/1
Defocus B	S	1.44	1.13	0.31	-0.55	0.83 [0.67, 1.00]	1/0
Fog A	O	1.32	0.77	0.55	0.10	-0.22 [-0.30, -0.14]	0/0
Fog A	S	1.27	1.01	0.26	0.05	0.75 [0.54, 0.98]	1/0
Fog B	O	1.00	0.69	0.31	-0.08	-0.38 [-0.50, -0.26]	0/0
Fog B	S	0.72	0.73	-0.01	-0.19	0.73 [0.59, 0.90]	1/0
Contrast A	O	0.57	0.18	0.39	-0.12	0.21 [0.07, 0.35]	1/1
Contrast A	S	0.59	0.48	0.12	-0.25	0.36 [0.24, 0.47]	1/0
Contrast B	O	0.81	0.43	0.39	0.03	-0.04 [-0.21, 0.13]	0/0

Cell	Order	τ_{id}	ϕ_{PLPD}	ϕ_{entropy}	I_{12}	Difference [95% CI]	F/E
Contrast B	S	0.57	0.45	0.12	-0.10	0.34 [0.14, 0.54]	1/0
JPEG A	O	1.19	0.52	0.68	-0.06	0.16 [0.07, 0.26]	1/1
JPEG A	S	1.01	0.88	0.14	-0.02	0.74 [0.65, 0.84]	1/0
JPEG B	O	0.77	0.42	0.36	-0.13	-0.06 [-0.10, -0.01]	0/0
JPEG B	S	0.65	0.58	0.08	-0.16	0.50 [0.40, 0.60]	1/0
Model mean	O	1.62	0.44	1.18	-0.12	0.74 [0.69, 0.79]	6/6
Model mean	S	1.56	1.21	0.35	-0.21	0.87 [0.81, 0.92]	10/0

K PRELIMINARY STUDY: A NON-ENTROPY SELECTOR

We evaluate a DualTTA-style selector implemented from the paper's description (official code unavailable) (Phuong et al., 2026). This preliminary study uses our fixed 10,000-image subset, preprocessing and stream A on five severity-5 corruptions with ViT-B/16; it is not an official reproduction. Source predictions match our frozen baseline exactly. The paper's reported accuracies use its own protocol and unspecified exact checkpoint, so Table 20 reports descriptive gaps without a pass/fail threshold.

Decision 1 assigns likely-correct samples from the batch to entropy minimization; decision 2 assigns likely-incorrect samples from the complement to entropy maximization. These are disjoint roles, not nested survivor stages. We use probability-drop cutoffs 0.4/0.7, a 4×4 patch shuffle and feature-statistics perturbation at block 7. Underspecified choices are fixed: all LayerNorm affine tensors, SGD at 10^{-4} with momentum 0.9, summed objectives, detached weights, entropy reference $0.4 \log 1000$, negative-objective multiplier 0.5, population variances, a 10^{-6} denominator clamp and an unchanged class token. The supplement's `additional_d_impl_notes.md` lists all choices.

All five native trajectories, ten exact P11/P00 replays and 75 stochastic trajectories pass quota, disjoint-role, recorded-weight and endpoint checks. Loss weights pass exact same-expression consistency and a float64 reference at relative/absolute tolerances $10^{-5}/10^{-7}$. Intervals use 50,000 independent whole-trajectory bootstrap draws per five-seed arm, conditional on native and these fixed cells.

Table 20: Preliminary paper-described selector. Accuracies are percentages; effects and gaps are pp. Paper values are from its Table 2. q_1/q_2 are total native role assignments; $Y_{11} < Y_{10}$ compares native with the mean hybrid.

Corruption	Source	Native	Paper source	Paper native	Gap	q_1/q_2
Gaussian	16.45	50.05	35.09	54.45	-4.40	6260/3
Defocus	28.92	54.02	31.42	55.74	-1.72	6334/2
Fog	47.46	70.90	54.74	71.29	-0.39	4984/3
Contrast	44.29	62.02	48.98	67.55	-5.53	6104/6
JPEG	48.67	68.99	56.45	68.05	0.94	5527/5

Corruption	τ_{id}	ϕ_1	ϕ_2	I_{12}	$Y_{10} - Y_{01}$ [95% CI]	$Y_{11} < Y_{10}$
Gaussian	7.85	4.95	2.90	-5.54	2.04 [1.77, 2.33]	no
Defocus	0.69	1.00	-0.31	0.66	1.30 [0.86, 1.88]	no
Fog	0.92	0.80	0.12	-0.25	0.68 [0.43, 0.92]	yes
Contrast	0.65	0.24	0.41	0.68	-0.17 [-1.34, 1.00]	no
JPEG	0.63	0.54	0.09	0.07	0.45 [0.34, 0.57]	no
Mean	2.15	1.50	0.64	-0.88	0.86 [0.60, 1.13]	1/5

Native accuracy exceeds source in all five cells.

Decision 2 assigns only 2–6 of 10,000 samples per native stream, sharply restricting its identity choice. These results describe this implementation and five cells; they do not establish a general ordering of non-entropy criteria.

Contrast's stage-credit interval spans zero; the fog cell's $Y_{11} < Y_{10}$ is a 0.006 pp point difference. Mean ϕ_2 and I_{12} intervals also span zero: $[-0.07, 1.78]$ and $[-3.13, 0.54]$ pp.

L COLOREDMNIST BOUNDARY CASE

Data and shifts. ColoredMNIST (Arjovsky et al., 2019) groups digits 0–4 versus 5–9. The first 40,000 MNIST training images form two source environments; labels are flipped with probability 0.25 before red/green color assignment, then colors are flipped with probabilities 0.2 and 0.1. Target conditions use the remaining 20,000 digit identities and clean binary evaluation labels. In flip-0.9, the final color-flip probability is 0.9, giving expected 30% color alignment with the clean label.

Noise-0.2 and noise-0.4 instead retain color-flip probability 0.1 (70% expected alignment) and add Gaussian noise with standard deviation 0.2 or 0.4 to the active color channel in $[0, 1]$ units, clipping before normalization; inactive channels remain zero. Thus the noise conditions preserve color assignment. A/B use split/order seeds 2024/2025 with a common source checkpoint and target caches. Each uses 10,000 adaptation images and a disjoint 1,024-image probe.

Model and adaptation. We use the existing binary ResNet-18 checkpoint with batch normalization. Inputs are 28×28 RGB tensors, normalized by channel means $(0.1307, 0.1307, 0)$ and standard deviation 0.3081. Continual Tent takes one entropy-minimization step per batch with SGD, learning rate 0.00125 and momentum 0.9, adapting 30 normalization-affine tensors. There are 157 batches of size 64, with a final batch of 16; no episodic reset occurs. Source uses checkpoint evaluation statistics; BatchNorm-only uses current-batch statistics without gradient updates, and Tent accepts every update. Probes always use the same 1,024-image normalization context.

Gate and random controls. The label-free consistency gate proposes a Tent update, using the lowest-entropy quarter of the batch (ceiling count, positional tie-breaking) as anchors. It evaluates one-pixel right/down zero-fill translations and brightness scaled by 0.9. Acceptance requires a strict decrease in mean Jensen–Shannon divergence from the fixed pre-update anchor distribution. Rejection restores parameters, optimizer/momentum, gradients, flags, and random-generator state. Neither class/color labels nor probe predictions enter the gate. Ten random controls independently permute its recorded Boolean acceptance mask, preserving the accepted-update count. Because the final batch is smaller, this need not preserve the number of samples used for accepted updates. This gate-level intervention differs from our per-stage identity interventions.

Results and uncertainty. All six overall-accuracy intervals for gate minus random include zero; this does not establish equivalence. Color-conflicting accuracy is below BatchNorm-only in all six conditions (Figure L1). The 95% percentile intervals use 2,000 bootstrap draws, resampling random schedules and unique probe identities with shared identity weights across methods. Fixed-method contrasts resample probe identities alone. These marginal, unadjusted intervals condition on the realized adaptation trajectories and normalization contexts. Cross-entropy and true-class probability remain in the accompanying CSVs.

Table 21: ColoredMNIST final probe accuracy (%). Random reports mean ± SD over ten acceptance masks; each other method has one trajectory per condition.

Shift	Stream	Population	Gate	Random	BN-only	Tent
flip-0.9	A	Overall	41.60	42.04 ± 0.35	46.68	40.53
flip-0.9	A	Conflicting	17.40	18.18 ± 0.45	24.86	16.02
noise-0.2	A	Overall	76.46	76.24 ± 0.20	76.17	75.39
noise-0.2	A	Conflicting	27.63	27.11 ± 0.42	31.25	23.03
noise-0.4	A	Overall	71.97	72.04 ± 0.23	70.90	71.97
noise-0.4	A	Conflicting	20.07	20.53 ± 0.68	27.63	16.78
flip-0.9	B	Overall	40.04	40.04 ± 0.20	43.07	37.50
flip-0.9	B	Conflicting	16.71	16.55 ± 0.31	20.82	13.01
noise-0.2	B	Overall	73.93	74.11 ± 0.30	74.22	73.24
noise-0.2	B	Conflicting	24.36	24.39 ± 0.72	30.13	20.51
noise-0.4	B	Overall	71.09	71.03 ± 0.26	69.73	71.29
noise-0.4	B	Conflicting	17.95	17.05 ± 0.93	23.40	16.03

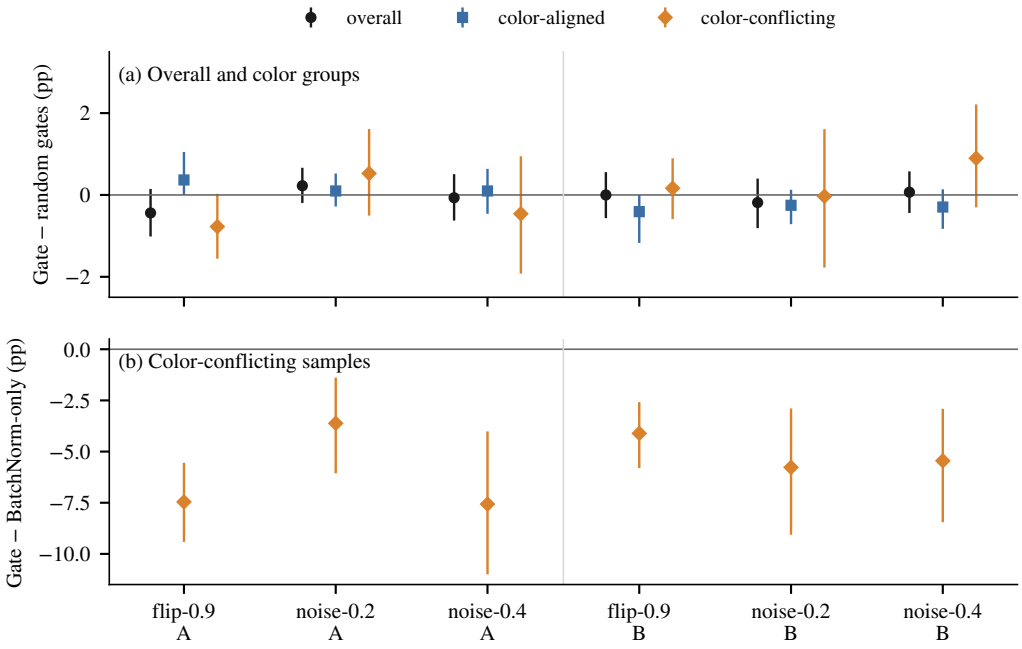


Figure L1: **ColoredMNIST boundary case.** (a) Consistency gate minus the mean of 10 accepted-update-count-matched random gates, by population. All six overall-accuracy intervals include zero. (b) Gate minus BatchNorm-only on color-conflicting samples. Bars are 95% bootstrap intervals; A/B identify split/order seeds.