WHEN IS TEST-TIME ADAPTATION IDENTIFIABLE FROM UNLABELED EVIDENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) offers many ways to update a deployed model without labels, but choosing the wrong update can make a strong source model worse. Recent methods therefore try to predict which adaptation will work from unlabeled test data. Yet the difficulty is not merely that different deployments can prefer different TTA actions: two deployments can be indistinguishable under the selector's allowed pre-action evidence while requiring opposite actions. We therefore ask a prior question: does the evidence given to the selector contain enough information to determine the best action at all? We show that this is not guaranteed, even with a perfect selector. If an observation channel makes two deployments look the same while their TTA rankings differ, reliable selection is impossible from that channel; richer evidence can restore the decision only when it resolves the relevant ambiguity. TTA selection failure can therefore reflect an information-channel failure, not only a selector-capacity failure. We make this boundary explicit in a finite-batch Gaussian TTA model, where doing nothing beats mean recentering for small shifts, recentering wins beyond a unique critical shift, and the boundary shrinks with batch size n as $1/\sqrt{n}$. Controlled experiments on CIFAR-100-C and DomainNet-126 instantiate the same obstruction with modern TTA methods: across both synthetic corruptions and natural domain shifts, changing only deployment stream structure can reverse the oracle action while global order-blind evidence remains essentially invariant. Graded correlation sweeps show these reversals emerging progressively as temporal structure increases, while order-blind evidence barely changes. This separates two failure modes that are usually mixed together—a weak selector and an information channel that cannot support the desired decision—and suggests a simple question before building a stronger selector: can its inputs support the decision at all?

1 INTRODUCTION

Test-time adaptation (TTA) updates a trained model using unlabeled data after deployment. Methods such as Tent, EATA, SAR, DeYO, and ROID can improve performance under distribution shift (Wang et al., 2021; Niu et al., 2022; 2023; Lee et al., 2024a; Marsden et al., 2024), but the same update can also be harmful: small batches, class imbalance, temporal correlation, and protocol choices can reverse the benefit of adaptation (Gong et al., 2022; Lim et al., 2023; Zhao et al., 2023; Sreeram et al., 2026).

Existing work spans two distinct information budgets: post-adaptation protocols (such as AETTA, TTALine, and Cygert et al.) that evaluate candidate models after adaptation has already run (Lee et al., 2024b; Kim et al., 2024; Cygert et al., 2026), and pre-action selection (such as MORPHEUS) that predicts performance beforehand from frozen-source geometry (Danilowski et al., 2026). Our inquiry is prior to either estimation problem. We therefore ask:

> *Is the unlabeled evidence available before adaptation actually enough to determine which action is best?*

This is not the obvious statement that the best method depends on the candidate set: our menu of TTA actions is fixed. The issue is that a selector sees only a chosen summary of the unlabeled deployment, while relative TTA performance can depend on properties—such as batch composition or arrival

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.

order—that this summary may discard. As illustrated in Figure 1, two deployments can therefore look the same to the selector even though different actions are best; a richer stream-aware channel can resolve the ambiguity when it preserves the relevant deployment structure. When the observation channel discards that structure, a larger regressor, a neural selector, or more selector-training data cannot recover the missing information.

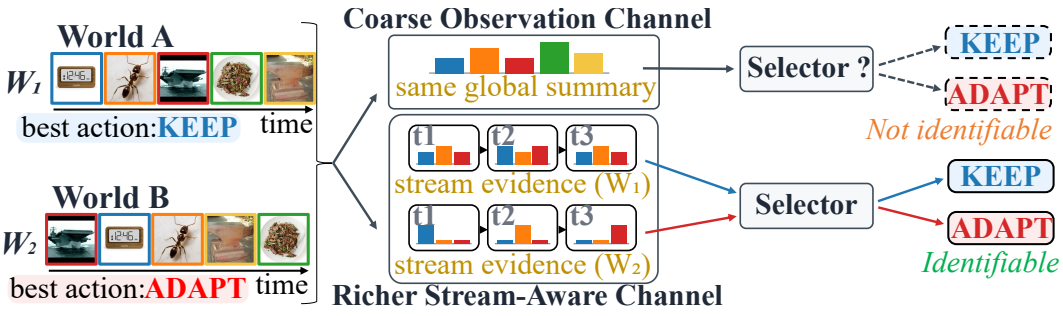

Figure 1: **Observation-channel view of pre-action TTA selection.** Same-sample worlds with different orders can require different best actions: order-blind evidence collapses them, while stream-aware evidence can separate them when order is decision-relevant.

We call the oracle action *identifiable* when the allowed evidence determines at least one action that is optimal for every deployment still compatible with that evidence. The general decision-theoretic idea is established; partial identification and domain-adaptation theory provide important precedents (Kasy, 2016; Pu & Zhang, 2021; Han, 2024; Christensen et al., 2026; Ben-David et al., 2010; Garg et al., 2022; Gulrajani & Hashimoto, 2022; Kong et al., 2022; Hanneke et al., 2023; Dong et al., 2026). Our contribution is to study this question for competing, data-dependent TTA interventions chosen from pre-action unlabeled evidence: deployments requiring different actions can be indistinguishable to the selector, and we characterize how batch size, shift mechanism, stream structure, and evidence richness affect whether the decision is possible.

Our contributions are:

- **TTA action identifiability.** We formalize when a fixed menu of TTA actions is determined by a specified unlabeled observation channel and deployment family.
- **Information limits and a positive criterion.** We show that indistinguishable deployments with different best actions make reliable selection impossible from that evidence, and give the complementary condition under which an optimal action is identifiable.
- **A TTA-specific finite-batch boundary.** In a tractable KEEP-versus-RECENTER model, we prove a unique transition between when adaptation hurts and helps, show that its critical shift scales with adaptation batch size n as $1/\sqrt{n}$, and construct different shift mechanisms with the same coarse evidence but opposite preferred actions.
- **Public-benchmark instantiation.** On CIFAR-100-C and DomainNet-126, changing only stream structure frequently reverses the oracle action while global evidence barely moves; richer stream evidence helps when that missing structure is decision-relevant.

2 Problem Setup

Let f_0 be a fixed pretrained source model and $\mathcal{A} = \{\textsc{Keep}, a_1, \ldots, a_K\}$ a finite action set. KEEP uses the frozen source model; every other action may adapt f_0 using an unlabeled target batch $X_{1:n}$ of size n. A deployment *world* $\theta \in \Theta$ specifies both the target joint distribution $P_\theta(X, Y)$ and the deployment protocol, including batch size, ordering or temporal structure when these affect adaptation. This part is essential because a stateful TTA action can produce a different predictor from the same image multiset when the images arrive differently.

For the population theory, let $(X', Y') \sim P_\theta(X, Y)$ be an independent draw from the target marginal, independent of $X_{1:n}$. For action a, let $f_{a,X_{1:n}}$ be the resulting predictor and define $R_\theta(a) =$

$\mathbb{E}_\theta[\ell(f_{a,X_{1:n}}(X'), Y')]$. The oracle action set is $\mathcal{A}^*(\theta) = \arg\min_{a\in\mathcal{A}} R_\theta(a)$. Target labels are unavailable to the selector; they enter only the oracle definition and retrospective experimental evaluation. In the experiments, the five actions are SOURCE / KEEP, DeYO, Tent, ROID, and BN-test. Section 3.4 uses RECENTER only as a one-dimensional theoretical prototype; it is neither an empirical benchmark method nor the NEO method of Murphy et al. (2026).

Before choosing an action, the selector observes only $Z_n = \phi_n(f_0, X_1, \ldots, X_n)$, where ϕ_n is the specified unlabeled observation channel (for example, entropy, source logits, global moments, or sequential statistics), and $Q_{\theta,n}$ denotes the probability law of Z_n in world θ.

Definition 1 (Oracle-action identifiability)**.** Fix an observation channel and a family of worlds Θ, and let $\mathcal{Q}_n = \{Q_{\theta,n} : \theta \in \Theta\}$ be the induced evidence-law family. For $Q \in \mathcal{Q}_n$, define $\Theta(Q) = \{\theta \in \Theta : Q_{\theta,n} = Q\}$. The oracle action is identifiable at Q if $\bigcap_{\theta\in\Theta(Q)} \mathcal{A}^*(\theta) \neq \varnothing$; it is identifiable over Θ if this holds for every $Q \in \mathcal{Q}_n$. With unique oracle actions, worlds inducing the same evidence law must have the same best action.

This definition asks an information question, not a training question. If we knew the full probability law of the allowed evidence, would it determine a best action? A practical selector sees only finite data, so there is a second statistical problem of learning the decision from limited samples. We keep these questions separate. Experiments use the finite-stream analogue, defining the oracle by retrospective error on each frozen deployment world (Section 4). Target/Mechanism-ID $\Rightarrow$ Risk-ID $\Rightarrow$ Action-ID. Knowing the complete target mechanism is stronger than knowing every action risk, and knowing every action risk is stronger than merely knowing which action wins. The reverse directions need not hold. This hierarchy is a positioning device rather than a claim of a new general decision-theory principle; Section 6 connects it to earlier partial-identification and domain-adaptation work. For selector s, we measure the remaining decision cost by oracle regret: $\mathrm{Reg}_\theta(s) = \mathbb{E}[R_\theta(s(Z_n)) - \min_{a\in\mathcal{A}} R_\theta(a)]$.

3 When Is the Oracle TTA Action Identifiable?

3.1 Indistinguishable worlds force selection error

Theorem 1 (Exact non-identifiability)**.** *Suppose two worlds θ_0, θ_1 have the same evidence law, $Q_{\theta_0,n} = Q_{\theta_1,n}$, but different unique oracle actions $a_0^* \neq a_1^*$. Under an equal prior over the two worlds, every selector based only on Z_n has average oracle-action error at least $1/2$. If every non-oracle action costs at least $\Delta > 0$ extra risk, average oracle regret is at least $\Delta/2$.*

The selector receives statistically identical evidence in the two worlds, so it cannot reliably know which action is correct. The result can persist with unlimited unlabeled data if the allowed observation remains identical. Appendix A.1 gives the proof and an explicit two-world construction with identical $P(X)$ but reversed action risks.

Corollary 1 (Approximate non-identifiability)**.** *Let the two evidence laws be Q_0, Q_1 and let their unique oracle actions differ. Under an equal prior,*

$$\mathbb{P}(\textit{wrong oracle action}) \geq \frac{1 - \mathrm{TV}(Q_0, Q_1)}{2}.$$

If every wrong action costs at least Δ, then $\mathbb{E}[\mathrm{Reg}] \geq \frac{\Delta}{2}\big(1 - \mathrm{TV}(Q_0, Q_1)\big)$.

Thus near-indistinguishable evidence still imposes a quantitative error floor; see Appendix A.2.

3.2 The positive side

Theorem 2 (Compatibility criterion)**.** *Fix an induced evidence law $Q \in \mathcal{Q}_n$. An action depending only on Q has zero oracle regret for every compatible world if and only if $\bigcap_{\theta\in\Theta(Q)} \mathcal{A}^*(\theta) \neq \varnothing$.*

In simple terms, given an evidence law Q, consider every compatible world. If they share at least one best action, the decision is determined even if the full target world is not. The proof is in Appendix A.3.

3.3 ORDER-BLIND EVIDENCE CAN HIDE THE CORRECT ACTION

Proposition 1 (Permutation/invariance obstruction). *Let g transform a deployment batch $\mathcal{B}$ (for example, by permuting its order), and suppose the deployment family contains a world θ and its transformed world $g\theta$ as distinct deployment processes, as can occur for non-exchangeable or temporally correlated streams. If $\phi_n(g\mathcal{B}) = \phi_n(\mathcal{B})$ for every $\mathcal{B}$ and the two worlds have different unique oracle actions, then the oracle action is not identifiable from ϕ_n over any family containing both worlds.*

This is the bridge to the matched experiments: if the selector discards order but a stateful TTA action reacts to order, the same image multiset can require a different action while looking identical to the selector. This is what our matched CIFAR-100-C and DomainNet-126 experiments are designed to test. The formal proof is in Appendix A.4.

3.4 A TTA-SPECIFIC FINITE-BATCH KEEP-VERSUS-RECENTER BOUNDARY

The previous statements are general. We now give a concrete TTA mechanism where the oracle action changes with batch size and shift magnitude. KEEP incurs no estimation noise, but suffers increasing error as the true distribution shift $|\delta|$ grows; RECENTER eliminates the true shift δ, but incurs a finite-sample noise floor because the shift is estimated from a noisy batch of size n.

Consider balanced binary classification $Y \in \{-1, +1\}$ with source distributions $\mathcal{N}(\pm\mu, \sigma^2)$ ($\mu, \sigma > 0$) split by a source threshold at zero. In the target domain, data undergo a translation shift δ: $X = \delta + \mu Y + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$, and KEEP leaves the threshold at zero. RECENTER receives n unlabeled samples, sets $\hat{\delta} = \bar{X}_n$, and classifies an independent fresh target point using threshold $\hat{\delta}$. KEEP has risk that grows strictly with the true translation $|\delta|$:

$$R_K(\delta) = \frac{1}{2}\left[\Phi\left(\frac{-\mu - \delta}{\sigma}\right) + \Phi\left(\frac{\delta - \mu}{\sigma}\right)\right]. \tag{1}$$

Let $R_R(n)$ denote RECENTER's fresh-point risk. The true δ cancels after recentering, so $R_R(n)$ is independent of δ; Appendix A.5 gives its exact finite-mixture expression.

Theorem 3 (Finite-batch KEEP/RECENTER boundary). *For every finite $n \geq 1$, there is a unique $\delta_c(n) > 0$ satisfying $R_K(\delta_c(n)) = R_R(n)$, with KEEP better for $|\delta| < \delta_c(n)$ and RECENTER better for $|\delta| > \delta_c(n)$. Moreover, as $n \to \infty$,*

$$\delta_c(n) = \frac{\sqrt{\mu^2 + \sigma^2}}{\sqrt{n}} + O(n^{-3/2}) = \Theta(n^{-1/2}).$$

At zero shift, RECENTER only adds sampling jitter; as $|\delta|$ grows, KEEP worsens while RECENTER stays at a finite-batch noise floor. The crossing therefore occurs when the shift reaches the natural $1/\sqrt{n}$ estimation scale. Figure 2 shows the exact finite-n crossings; Appendix A.5 proves uniqueness and the asymptotic rate without a normal approximation.

3.5 THE SAME MEAN CAN IMPLY OPPOSITE ACTIONS

A target mean shift need not be a physical translation. If the class-conditionals remain $\mathcal{N}(\pm\mu, \sigma^2)$ but $P(Y = +1) = \pi \neq 1/2$, the target mean becomes $m = \mu(2\pi - 1)$ even though neither class distribution moved.

Proposition 2 (Prior shift). *Under pure class-prior shift with $\pi \in (0, 1) \setminus \{1/2\}$, population RECENTER has strictly larger classification risk than KEEP.*

Corollary 2 (Population-mean channel: same mean, opposite best actions). *Fix $0 < |m| < \mu$. A balanced translation world with $\delta = m$ and an untranslated prior-shift world with $\pi = (1 + m/\mu)/2$ have the same population target mean m, yet population RECENTER is better in the translation world and KEEP is better in the prior-shift world.*

For finite-batch RECENTER, the same reversal requires $|m| > \delta_c(n)$ and hence holds for all sufficiently large n; the finite-sample mean laws themselves need not coincide. Mean recentering interprets both changes as translations, although the Bayes threshold moves in the opposite direction

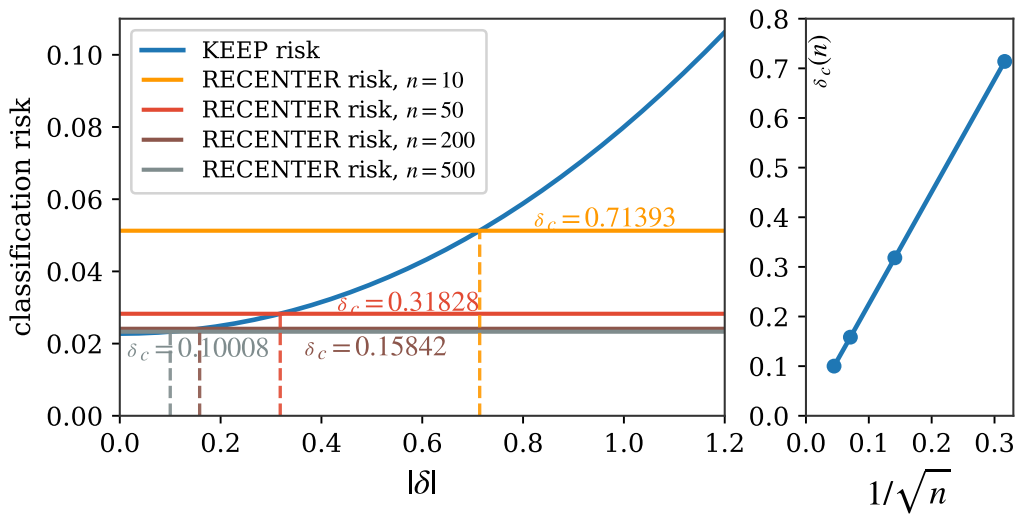

Figure 2: **Finite-batch KEEP–RECENTER boundary.** *Left:* prospective risk for $\mu = 2, \sigma = 1$; each crossing defines $\delta_c(n)$. *Right:* $\delta_c(n)$ scales approximately linearly with $1/\sqrt{n}$ over the plotted batch sizes, consistent with Theorem 3.

under prior shift. The result concerns a *mean-only* channel: the full target distributions, and generally finite-sample mean laws, differ. Richer evidence may therefore resolve the ambiguity. Proofs are in Appendices A.6–A.7. Together, these give the main theoretical message: the best TTA action depends not only on the shift itself, but also on batch size, the shift mechanism, the deployment structure, and what information the selector is allowed to use.

4 EXPERIMENTAL DESIGN

The experiments use the finite-stream analogue above. Here N denotes the full pre-action deployment horizon and B the execution mini-batch size; the theoretical n denotes the unlabeled batch size in the population analysis.

Actions, worlds, and matched deployments. Both benchmarks use the same five candidate actions: SOURCE / KEEP, DeYO (Lee et al., 2024a), Tent (Wang et al., 2021), ROID (Marsden et al., 2024), and BN-test, which recomputes BatchNorm statistics from the current test batch (Schneider et al., 2020). All run in single-pass online mode, are reset to the frozen source state at the start of each world, and use published configurations without post-hoc tuning. For frozen stream $w = (X_{1:N}, Y_{1:N})$, let $\widehat{R}_w(a) = N^{-1} \sum_{t=1}^{N} \ell(f^a_{w,t-1}(X_t), Y_t)$ be action a's realized cumulative stream error and $\widehat{A}^*(w) = \arg\min_a \widehat{R}_w(a)$ its empirical oracle; labels are accessed only retrospectively.

CIFAR-100-C. We use CIFAR-100-C (Hendrycks & Dietterich, 2019) with the RobustBench pretrained ResNeXt-29 (Croce et al., 2021). The 675 worlds span 15 corruptions, three severities $\{1, 3, 5\}$, three seeds, and five regimes: independent and identically distributed (IID) balanced (uniform class quota, randomly permuted), IID prior-60, IID prior-80, correlated balanced, and correlated prior-80. The prior-shift regimes assign 60% or 80% of the samples to a frozen 25-class major set; matched IID/correlated worlds (270 pairs) preserve the selected image multiset and class quota and differ only in arrival order. Each controlled world uses $N = 2{,}000$ images at execution batch size $B = 200$. A *strong flip* changes oracle action with margin ≥ 0.5 percentage points (pp) in both worlds (used to exclude near-ties; overall flip counts do not depend on this threshold). Target labels are used solely to construct orderings and score actions retrospectively; neither TTA methods nor selectors observe them.

DomainNet-126. We use DomainNet-126 (Peng et al., 2019) with a pretrained ResNet-50. Combining six directional transfers among real, clipart, painting, and sketch with three seeds and five regimes

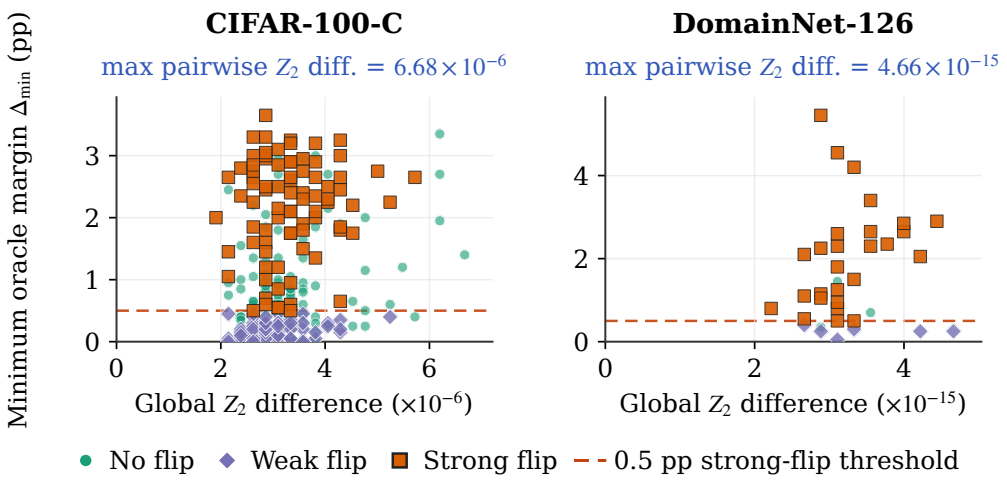

Figure 3: **Matched-pair order-invariance diagnostic.** CIFAR-100-C (left) and DomainNet-126 (right) matched pairs contain the same selected images and class quota but differ in stream order. Despite near-invariant Z_2, the oracle action flips in 157/270 and 33/36 pairs: 82 strong and 75 weak flips on CIFAR-100-C, and 27 strong and 6 weak flips on DomainNet-126. Z_2 differences are numerical diagnostics, not TV estimates.

gives 90 worlds; each uses $N = 2{,}000$ target images at batch size $B = 32$, following the frozen benchmark-specific protocol. The 36 matched IID/correlated pairs again share the same selected images and class quota and differ only in order. Full construction details are in Appendix B.

4.1 Evidence and selector evaluation

All selector inputs are computed from the frozen SOURCE model over the full unlabeled deployment sequence before adaptation; selection is therefore transductive pre-action. Z_1 contains global uncertainty averages; Z_2 adds order-blind quantiles, predicted-class histograms, and global logit moments; Z_3 adds local-window moments and 11 inter-window drift statistics; and Z_4 adds a fixed sparse projection of the ordered logit sequence. Feature extraction follows the same protocol across benchmarks, yielding $\dim(Z_3) = 92 + 23C$ for C classes across 10 evaluation windows: exactly 2,392-D on CIFAR-100-C ($C = 100$) and 2,990-D on DomainNet-126 ($C = 126$). Thus Z_3 is an order-aware observation channel, not a new TTA method. ORDER11 denotes the same 11 predeclared stream statistics already contained in Z_3. On DomainNet-126, we additionally evaluate window sizes $W \in \{32, 64, 128, 200\}$; $W = 32$ matches the TTA execution batch size, while $W = 200$ is retained only as a window-scale robustness comparison. Unless otherwise stated, Z_1–Z_4 and ORDER11 use one action-conditioned random-forest regressor $g(z, a)$: each training world is paired with every candidate action, the target is post-action accuracy, and selection uses $\arg\max_a g(z, a)$. Appendix B reports a Z_3-MultiOutput control matching MORPHEUS's multi-output regression form.

SOURCE / KEEP provides the no-selection baseline. For MORPHEUS (Danilowski et al., 2026), we report its entropy ablation and Neural-Collapse (NC) baseline; the author-clarified feature and regression implementation is documented in Appendix B.6. Cygert et al. (2026) use selection statistics available only after candidate adaptations have run, so we treat them as a neighboring protocol with a stronger information budget rather than a same-budget pre-action baseline. Selectors are evaluated on held-out shifts using leave-one-corruption-out (LOCO) on CIFAR-100-C and leave-one-transfer-out (LOTO) on DomainNet-126. Primary metrics are oracle regret and exact action accuracy. Paired method comparisons use 5,000-replicate hierarchical bootstrap confidence intervals, with the resampling hierarchy and hypothesis tests detailed in Appendix B.

5 Results

5.1 Same images, different order, different oracle action

The matched experiments provide the finite-stream analogue of Proposition 1. On CIFAR-100-C, 157/270 matched IID/correlated pairs change empirical oracle action even though each pair contains exactly the same selected images; 82 are strong flips. On DomainNet-126, 33/36 pairs change oracle action, including 27 strong flips. In both datasets, global order-blind evidence remains essentially invariant within each pair; Figure 3 visualizes this matched-pair separation.

5.2 CIFAR-100-C: RICHER EVIDENCE SHARPLY REDUCES ORACLE REGRET

Table 1: CIFAR-100-C controlled LOCO selection. Lower regret and higher exact action accuracy are better.

Selector / evidence	Regret (pp) ↓	Action acc. (%) ↑
SOURCE / KEEP	4.980	50.2
MORPHEUS-entropy	4.135	53.0
MORPHEUS-NC	1.383	66.1
Z_1 (global)	1.705	63.7
Z_2 (richer global)	1.247	65.0
Z_3 **(stream-aware)**	**0.312**	**76.6**
Z_4 (stream + projection)	0.332	75.7
Oracle	0.000	100.0

Table 1 shows the key separation. On the 82 strong matched flips, Z_3 changes action in 92.7% of pairs with 0.093 pp mean pair regret, whereas the MORPHEUS-NC selector chooses the same action for both members of every strong matched pair (0/82 pairwise action changes) and has 2.782 pp mean pair regret. The NC evidence used by this baseline is global and order-blind, so reordering the same image multiset leaves the available evidence essentially unchanged. Across all CIFAR-100-C LOCO worlds, the paired regret difference (MORPHEUS-NC minus Z_3) is $+1.071$ pp (hierarchical-bootstrap 95% CI $[0.611, 1.625]$; one-sided Wilcoxon signed-rank test across 15 held-out corruptions, $p = 3.05 \times 10^{-5}$). Z_4 does not improve on Z_3, which does not support a simple feature-count explanation. Cygert et al. (2026) is not included in Table 1 because they evaluate unsupervised selection after candidate adaptations have already been executed, so their selector has access to post-adaptation statistics that are unavailable in our pre-action setting; Appendix B reports our official-code sanity check and explains the information-budget difference.

5.3 DOMAINNET-126: THE SIGNAL IS LOW-DIMENSIONAL AND EXECUTION-SCALE

Table 2: DomainNet-126 LOTO selection over 90 worlds. ORDER11 uses the 11 predeclared stream-summary coordinates in Z_3; Z_3 and ORDER11 are separately fitted and make identical predictions on all 90 worlds. Z_4 and Z_3-MultiOutput are in Appendix B.

Selector / evidence	Regret (pp) ↓	Action acc. (%) ↑
SOURCE / KEEP	1.933	45.6
MORPHEUS-entropy	4.886	46.7
MORPHEUS-NC	1.933	45.6
Z_2	1.933	45.6
Z_3	**0.034**	**93.3**
ORDER11 RF, $W = 32$	**0.034**	**93.3**
Ridge ORDER11, $W = 32$	0.061	92.2

DomainNet-126 shows that the information-boundary phenomenon also appears on natural domain-transfer benchmarks. Across all 90 deployment worlds (Table 2), SOURCE/KEEP, MORPHEUS-NC, and Z_2 all obtain 1.933 pp regret and 45.6% accuracy because the two learned order-blind selectors predict SOURCE in every LOTO test world; Z_3 and its 11-D ORDER11 summary ($W = 32$) instead obtain 0.034 pp regret and 93.3% accuracy. The matched-world analysis isolates why: on the 27 strong same-image flips, order-blind Z_2 and MORPHEUS-NC never switch actions ($0/27$), suffering

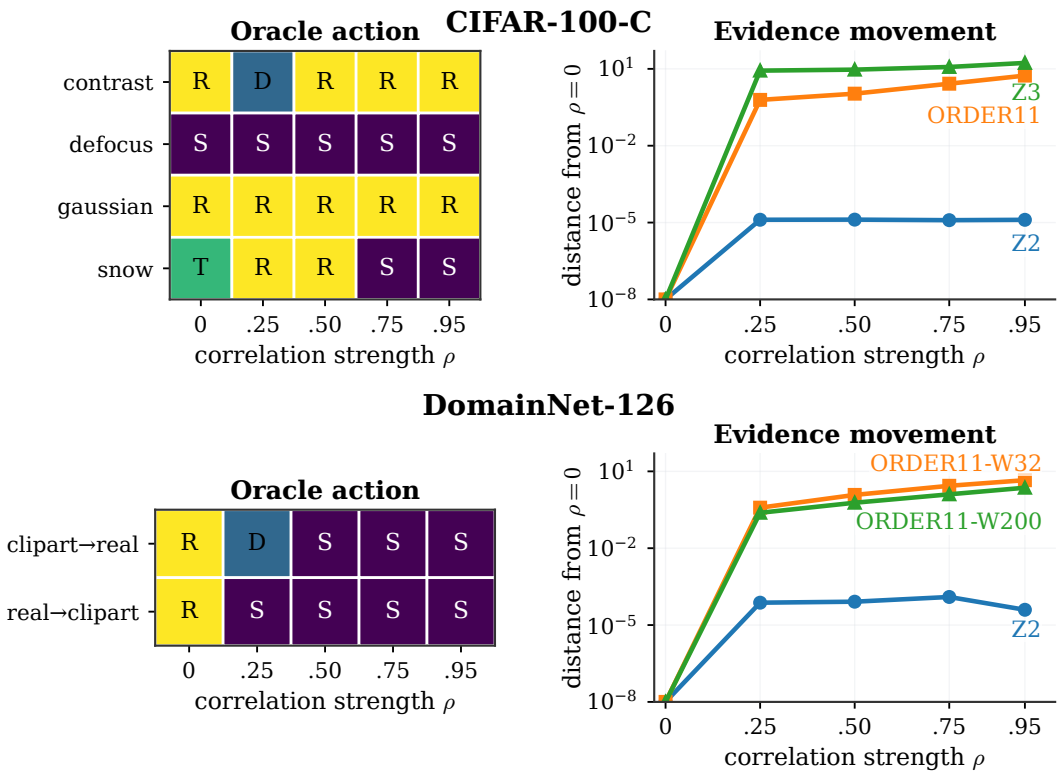

Figure 4: **Correlation-strength sweeps on fixed image sets.** *Left:* oracle action maps for fixed seed 0 as the local class-correlation strength ρ increases. Cell labels denote the oracle action: S = SOURCE, R = ROID, D = DeYO, and T = Tent. *Right:* movement of summary evidence away from its $\rho = 0$ baseline (log scale). In both datasets, the global order-blind evidence Z_2 remains nearly flat, while stream-aware summaries move as temporal structure is strengthened. ORDER11-$W200$ is shown only as a window-scale robustness comparison. Raw distances are unstandardized and interpreted only within each representation; their vertical magnitudes are not comparable across channels of different dimensionality.

1.942 pp mean pair regret, whereas Z_3 and ORDER11 switch actions in all 27 pairs with zero pair regret. The identical Z_3 and ORDER11 metrics arise from separately fitting ORDER11 on its 11-D summary; nearby controls differ (ORDER11-W64: 0.106 pp; Ridge: 0.061 pp), confirming that the decision-relevant stream signal is low-dimensional. Across the six DomainNet-126 held-out transfers, the paired MORPHEUS-NC minus Z_3 regret difference is +1.899 pp (95% CI [1.024, 2.815]; one-sided Wilcoxon signed-rank test across six transfers with the pre-specified alternative that MORPHEUS-NC regret exceeds Z_3 regret, $p = 0.0156$). Full window sweeps and robustness checks are in Appendix B.

5.4 Deployment family matters

To show that the matched IID-versus-correlated contrast is not merely a binary endpoint effect, we ran frozen correlation-strength sweeps while holding the selected image set fixed and increasing the local class-correlation parameter from $\rho = 0$ to $\rho = 0.95$. Figure 4 shows the same mechanism in a graded form. In both CIFAR-100-C and DomainNet-126, the oracle action changes as ρ increases, while the global order-blind evidence Z_2 stays nearly unchanged. By contrast, the stream-aware summaries move steadily away from their $\rho = 0$ baselines as more temporal structure is injected. Thus, the effect is not a one-off shuffled-versus-blocked artifact: the preferred action can change progressively as stronger local ordering is introduced. Detailed selector-by-ρ diagnostics are deferred to Appendix B.5. **A complementary counter-result is equally important:** under broader stochastic CIFAR-100-C deployments, the large controlled advantage disappears, yielding stochastic-LOCO

regrets of 0.380, 0.338, and 0.366 pp for Z_2, Z_3, and Z_4, respectively (paired $Z_2 - Z_3$ difference: $+0.043$ pp, 95% CI $[-0.091, 0.160]$, Holm-adjusted $p = 0.957$). Thus, stream-aware evidence helps when temporal structure is decision-relevant, rather than being universally superior.

6 Related Work and Positioning

TTA reliability, estimation, and selection. While methods such as Tent, EATA, SAR, DeYO, and ROID improve adaptation reliability (Wang et al., 2021; Niu et al., 2022; 2023; Lee et al., 2024a; Marsden et al., 2024), NOTE (Gong et al., 2022) and TTN (Lim et al., 2023) show that temporal correlation, small batches, and non-IID streams can induce severe adaptation failure. Class imbalance, protocol choices, and time-varying dynamics further shift method rankings (Zhao et al., 2023; Sreeram et al., 2026). OATTA (Kim et al., 2026) actively exploits temporal transition structure as an adaptation signal; our focus is complementary: whether a specified pre-action channel preserves sufficient evidence to distinguish deployments requiring different actions. For performance estimation, AETTA (Lee et al., 2024b) and TTALine (Kim et al., 2024) evaluate models without labels or via agreement relationships, whereas Cygert et al. (2026) perform model selection post-adaptation. Our closest pre-action neighbor is MORPHEUS (Danilowski et al., 2026), which predicts TTA performance from entropy and Neural-Collapse geometry and notes order-dependent DeYO failures, but does not characterize observation-channel sufficiency. Thus, order sensitivity itself is not our novelty; we instead characterize when a specified pre-action observation channel is sufficient to identify the oracle TTA action. **Limits of unlabeled adaptation and transfer.** Recent TTA theory studies recovery complexity and underspecification (Zhou et al., 2026; Shamsi et al., 2026), while domain-adaptation theory bounds identifiability and hardness (Ben-David et al., 2010; Gulrajani & Hashimoto, 2022; Kong et al., 2022; Dong et al., 2026). Unlabeled OOD estimation and transfer model selection likewise require structural assumptions (Garg et al., 2022; Hanneke et al., 2023; Ng et al., 2025); as permutation-invariant metrics, ATC (Garg et al., 2022) and scalar surrogates fall under Proposition 1 whenever stream order is decision-relevant. Unlike works targeting predictor synthesis or transfer rates, our downstream object is strictly the index of the best TTA intervention from pre-action evidence. Moreover, whereas NEO (Murphy et al., 2026) uses latent recentering as an empirical algorithm, our RECENTER action is strictly a theoretical construct exposing the finite-batch boundary. **Partial identification.** Optimal decision-making without full payoff identification is established in econometrics (Kasy, 2016; Pu & Zhang, 2021; Han, 2024; Christensen et al., 2026). We do not claim Action-ID as general decision theory; our contribution is instantiating these information limits specifically for TTA via the finite-batch boundary, mechanism reversals, permutation obstruction, and matched-world benchmarks.

7 Discussion and Limitations

TTA selection fails if the selector is weak or its evidence erases action-ranking differences. Our matched experiments isolate the latter by fixing the images and varying only stream order. DomainNet-126 further shows that the useful signal need not be high-dimensional: the 11 predeclared ORDER11 statistics at the execution scale match the frozen Z_3 decisions, and a linear Ridge model remains near oracle. Stream-aware evidence is not universally superior: the stochastic counter-result shows that its large controlled advantage disappears under broader stochastic deployment variation. The correlation sweeps make the complementary point: as decision-relevant temporal structure is introduced gradually, stream-aware evidence moves while order-blind evidence stays nearly fixed. Identifiability therefore belongs to the pair *(deployment family, observation channel)*, not to a selector architecture alone. **Several limitations define the scope of this work.** Theoretically, our closed-form boundary assumes a finite action menu and a 1D homoscedastic Gaussian mechanism, which does not directly model anisotropic or heteroscedastic shifts. Empirically, evaluations use a single backbone per benchmark, synthetically ordered static images, a mean-only shift construction, and a simpler action boundary on DomainNet-126. Operationally, our transductive selector requires a full initial pass; while ORDER11 ($W = 32$) demonstrates execution-scale utility, it does not establish prefix-only selection. Finally, Z_3 scales linearly with class count C, leaving class-agnostic summaries at scale open. Natural continuous video streams, broader architectures, and open action menus remain important next directions. The foundational design question is therefore not only which selector to train, but what information it must observe for reliable adaptation to be possible at all.

Reproducibility Statement

The experiment notebooks freeze the target worlds, actions, random seeds, evidence definitions, selectors, and statistical tests. Appendix A contains complete proofs, and Appendix B records the experimental audit and robustness checks. The anonymous supplementary material includes the code and frozen result artifacts needed to reproduce the reported tables and figures.

AI USE STATEMENT

Generative AI tools were used to refine hypotheses and research framing; assist with theoretical formulation, mathematical claims, and the checking of mathematical arguments and proofs; provide feedback on research methodology and experimental design; assist with code debugging and editing; support literature search; interpret intermediate results; and edit the manuscript for clarity and presentation. Generative AI tools were not used to generate synthetic datasets, translate research materials, clean or reformat datasets, or perform qualitative or thematic data analysis. The remaining required-disclosure activities were not applicable to this work. All AI-assisted theoretical suggestions were independently checked against the authors' derivations, and AI-assisted code changes were verified against the implementation, tests, and frozen result artifacts. Literature-related outputs were checked against the cited sources. The authors reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims, proofs, code, experiments, and reported results.

References

Shai Ben-David, Tyler Lu, Teresa Luu, and David Pal. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136. PMLR, 2010.

Timothy Christensen, Hyungsik Roger Moon, and Frank Schorfheide. Optimal decision rules when payoffs are partially identified. *The Review of Economic Studies*, 2026. doi: 10.1093/restud/rdag017.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Sebastian Cygert, Damian Sójka, Tomasz Trzciński, and Bartłomiej Twardowski. Realistic evaluation of test-time adaptation: Unsupervised model selection. In *Proceedings of the 21st International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 75–86. SciTePress, 2026. doi: 10.5220/0014320100004084.

Michal Danilowski, Alexander Murphy, Young D. Kwon, and Abhirup Ghosh. Morpheus: Meta test-time adaptation via neural collapse geometry. In *ICLR 2026 Workshop on Test-Time Updates*, 2026. URL https://openreview.net/forum?id=YfyApJH1Rj.

Zhiyi Dong, Zixuan Liu, and Yongyi Mao. On the hardness of unsupervised domain adaptation: Optimal learners and information-theoretic perspective. In *Proceedings of the 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 89–111. PMLR, 2026.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ishaan Gulrajani and Tatsunori Hashimoto. Identifiability conditions for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7982–7997. PMLR, 2022.

Sukjin Han. Optimal dynamic treatment regimes and partial welfare ordering. *Journal of the American Statistical Association*, 119(547):2000–2010, 2024. doi: 10.1080/01621459.2023.2238941.

Steve Hanneke, Samory Kpotufe, and Yasaman Mahdaviyeh. Limits of model selection under transfer learning. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5781–5812. PMLR, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Maximilian Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *The Review of Economics and Statistics*, 98(1):111–131, 2016. doi: 10.1162/REST_a_00528.

Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, and J. Zico Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3820.

Young Kyung Kim, Oded Schlesinger, Qiangqiang Wu, J. Matías Di Martino, and Guillermo Sapiro. Order-aware test-time adaptation: Leveraging temporal dynamics for robust streaming inference. *arXiv preprint arXiv:2601.21012*, 2026.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial identifiability for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11455–11472. PMLR, 2022.

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*, 2024a. URL https://openreview.net/forum?id=9w3iw8wDuE.

Taeckyung Lee, Sorn Chottananurak, Taesik Gong, and Sung-Ju Lee. Aetta: Label-free accuracy estimation for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28643–28652, 2024b.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations*, 2023.

Robert A. Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2555–2565, 2024.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo — no-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations*, 2026.

Ignavier Ng, Yan Li, Zijian Li, Yujia Zheng, Guangyi Chen, and Kun Zhang. A general representation-based approach to multi-source domain adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45911–45933. PMLR, 2025.

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.

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.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Hongming Pu and Bo Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):318–345, 2021. doi: 10.1111/rssb.12413.

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, 2020.

Afshar Shamsi, Xiao-Yu Guo, Hamid Alinejad-Rokny, Arash Mohammadi, Damien Teney, and Ehsan Abbasnejad. Multi-hypothesis test-time adaptation to mitigate underspecification. *arXiv preprint arXiv:2607.00259*, 2026.

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

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.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023.

Zhi Zhou, Ming Yang, Shi-Yu Tian, Kun-Yang Yu, Lan-Zhe Guo, and Yu-Feng Li. On the learnability of test-time adaptation: A recovery complexity perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A Appendix

This appendix records theorem proofs, additional experimental details, audits, and robustness checks supporting the concise main-text results. Before each theorem proof we first state the idea in plain language.

A.1 Proof of Theorem 1

Idea. The selector receives statistically identical evidence in both worlds. Therefore it must use the same action probabilities in both. Since the two correct actions are different, the probability mass assigned to "correct in world 0" and "correct in world 1" cannot add to more than one.

Let $p_j(z)$ be the selector's conditional probability of outputting a_j^* after observing z, including any internal randomization. Because the two evidence laws are the same, call the common law Q. Then

$$\mathbb{P}_{\theta_0}(s(Z_n) = a_0^*) + \mathbb{P}_{\theta_1}(s(Z_n) = a_1^*) = \int \big[p_0(z) + p_1(z)\big]\, dQ(z) \leq 1,$$

because $a_0^* \neq a_1^*$ implies $p_0(z) + p_1(z) \leq 1$ for every z. Under an equal prior, average correctness is therefore at most $1/2$, so average action error is at least $1/2$. If every wrong action has excess risk at least Δ, regret is at least Δ whenever the selector is wrong, giving average regret at least $\Delta/2$. □

Explicit Discrete Two-World Construction. To instantiate Theorem 1 with an explicit discrete example, let $\mathcal{X} = \{x_1, x_2, x_3\}$ with the same target marginal distribution $P(X) = (0.45, 0.25, 0.30)$ in both worlds. Consider two candidate classifiers:

$$f_{\text{KEEP}}(x_1) = 1, \quad f_{\text{KEEP}}(x_2) = 0, \quad f_{\text{KEEP}}(x_3) = 0,$$
$$f_{\text{ADAPT}}(x_1) = 0, \quad f_{\text{ADAPT}}(x_2) = 1, \quad f_{\text{ADAPT}}(x_3) = 0.$$

Let $\eta_\theta(x) = \mathbb{P}_\theta(Y = 1 \mid X = x)$ denote the target class-posterior probability:

$$\begin{aligned} \text{World A:} \quad & \eta_A = (0.250, 0.675, 0.1875), \\ \text{World B:} \quad & \eta_B = (0.750, 0.315, 0.0625). \end{aligned}$$

Under 0-1 loss $R_\theta(f) = \sum_{i=1}^{3} P(X = x_i)\mathbb{P}_\theta(f(x_i) \neq Y \mid X = x_i)$, the exact prospective risks are:

$$\begin{aligned} R_A(\text{KEEP}) &= 0.45(1 - 0.250) + 0.25(0.675) + 0.30(0.1875) = 0.5625, \\ R_A(\text{ADAPT}) &= 0.45(0.250) + 0.25(1 - 0.675) + 0.30(0.1875) = 0.2500, \\ R_B(\text{KEEP}) &= 0.45(1 - 0.750) + 0.25(0.315) + 0.30(0.0625) = 0.2100, \\ R_B(\text{ADAPT}) &= 0.45(0.750) + 0.25(1 - 0.315) + 0.30(0.0625) = 0.5275. \end{aligned}$$

Thus, $\mathcal{A}^*(\text{World A}) = \{\text{ADAPT}\}$ while $\mathcal{A}^*(\text{World B}) = \{\text{KEEP}\}$. Since the evidence law $Z_n \sim P(X)^{\otimes n}$ is identical in both worlds, any pre-adaptation selector incurs an average action error of at least $1/2$ and an expected regret of at least $\Delta/2 = 0.15625$.

A.2 Proof of Corollary 1

Idea. When the two evidence laws are not identical, the selector can partly tell them apart. The best possible two-way discrimination advantage is controlled by total variation. Whatever ambiguity remains becomes an unavoidable action-selection error.

Let A_0 be the event that the selector outputs a_0^*. Since $a_1^* \neq a_0^*$,

$$\mathbb{P}_0(\text{correct}) + \mathbb{P}_1(\text{correct}) \leq Q_0(A_0) + Q_1(A_0^c).$$

The right side is

$$1 + Q_0(A_0) - Q_1(A_0) \leq 1 + \text{TV}(Q_0, Q_1).$$

Average correctness is therefore at most $(1 + \text{TV})/2$, so average error is at least $(1 - \text{TV})/2$. Multiplying by the minimum wrong-action cost Δ gives the regret bound. □

A.3 Proof of Theorem 2

Idea. If all worlds compatible with the evidence share one optimal action, choose it. If they do not share any optimal action, no single decision based only on that same evidence law can be optimal in every compatible world.

If $a \in \bigcap_{\theta \in \Theta(Q)} \mathcal{A}^*(\theta)$, then the decision rule $d(Q) = a$ has zero oracle regret in every compatible world. Conversely, if the intersection is empty, every action fails to be optimal in at least one compatible world. Hence no action-valued function of Q can have zero oracle regret simultaneously across the whole compatible set. □

A.4 Proof of Proposition 1

Idea. An invariant observation maps a batch and its transformed version to exactly the same value. Therefore the transformed worlds have the same evidence law. If their best actions differ, Theorem 1 applies.

Let the raw batch in world θ be $\mathcal{B}$, and let the transformed world use the pushforward distribution of $g\mathcal{B}$. Since $\phi_n(g\mathcal{B}) = \phi_n(\mathcal{B})$ pointwise, both worlds induce the same evidence law; for a non-exchangeable deployment process, the transformed world can nevertheless have different action risks. Different unique oracle actions therefore imply non-identifiability by Theorem 1. □

A.5 Proof of Theorem 3

Step 1: write the KEEP risk as one function. Define

$$g(t) = \frac{1}{2}\left[\Phi\left(\frac{-\mu - t}{\sigma}\right) + \Phi\left(\frac{t - \mu}{\sigma}\right)\right].$$

This is exactly the error of the fixed source threshold when the target translation is t, so $R_K(\delta) = g(\delta)$.

The function is even: a left shift and an equally large right shift hurt the symmetric source classifier by the same amount. For $t > 0$,

$$g'(t) = \frac{1}{2\sigma}\left[\varphi\left(\frac{t-\mu}{\sigma}\right) - \varphi\left(\frac{t+\mu}{\sigma}\right)\right] > 0.$$

The inequality holds because $|t-\mu| < t+\mu$ and the standard normal density decreases as $|x|$ moves away from zero. Therefore g strictly increases with $|t|$. Also

$$g(0) = \Phi(-\mu/\sigma), \qquad \lim_{t\to\infty} g(t) = 1/2.$$

Step 2: understand what RECENTER leaves behind. The estimate is

$$\widehat{\delta} = \bar{X}_n = \delta + \frac{1}{n}\sum_{i=1}^{n}(\mu Y_i + \varepsilon_i).$$

Define the estimation error

$$E_n = \widehat{\delta} - \delta.$$

After moving the threshold by $\widehat{\delta}$, the residual translation is $\delta - \widehat{\delta} = -E_n$. Since g is even and the evaluation point is independent of the adaptation batch,

$$R_R(n) = \mathbb{E}[g(-E_n)] = \mathbb{E}[g(E_n)].$$

For every finite n, E_n has a continuous non-degenerate distribution, so $\Pr(E_n = 0) = 0$. Since $g(t) > g(0)$ whenever $t \neq 0$,

$$R_R(n) > g(0).$$

On the other hand $g(t) < 1/2$ for every finite t, hence $R_R(n) < 1/2$.

Because g is continuous and strictly increasing on $t > 0$, there is exactly one positive value $\delta_c(n)$ with

$$g(\delta_c(n)) = R_R(n).$$

Below that value KEEP has lower risk; above it RECENTER has lower risk. This proves the existence, uniqueness, and ordering parts of the theorem.

Step 3: derive the exact finite-n RECENTER risk. Condition on

$$S_n = \sum_{i=1}^{n} Y_i = s.$$

Then

$$E_n \mid S_n = s \sim \mathcal{N}\left(\frac{\mu s}{n}, \frac{\sigma^2}{n}\right).$$

For a Gaussian random variable $Z \sim \mathcal{N}(m, \tau^2)$, the identity

$$\mathbb{E}\left[\Phi\left(\frac{Z-a}{\sigma}\right)\right] = \Phi\left(\frac{m-a}{\sqrt{\sigma^2+\tau^2}}\right)$$

follows by introducing an independent standard normal variable and combining two Gaussian noises. Averaging over the binomial values $S_n = 2k - n$ gives

$$R_R(n) = 2^{-n}\sum_{k=0}^{n}\binom{n}{k}\Phi\left(\frac{\mu(2k-n)/n - \mu}{\sigma\sqrt{1+1/n}}\right). \tag{2}$$

Symmetry of E_n makes the two class-error terms equal after expectation.

Step 4: find the large-n scale of the crossing. Near zero, g is smooth and even, so the first nonzero term is quadratic:

$$g(t) = g(0) + c_2 t^2 + O(t^4), \qquad c_2 = \frac{\mu\,\varphi(\mu/\sigma)}{2\sigma^3} > 0.$$

The summand $\mu Y + \epsilon$ has variance

$$v = \mu^2 + \sigma^2.$$

Here μ^2/n is estimation noise from random class composition in the unlabeled batch, whereas σ^2/n is within-class observation noise; the class-composition term persists independently of the within-class noise level. Therefore

$$\mathbb{E}[E_n^2] = \frac{v}{n}, \qquad \mathbb{E}[E_n^4] = O(n^{-2}).$$

Because $g^{(4)}$ is bounded on $\mathbb{R}$, Taylor's theorem gives $|g(t) - g(0) - c_2 t^2| \le Ct^4$ for some $C < \infty$, and therefore

$$R_R(n) = g(0) + c_2 \frac{v}{n} + O(n^{-2}).$$

Moreover, $E_n \to 0$ almost surely by the strong law; since g is bounded and continuous, $R_R(n) = \mathbb{E}[g(E_n)] \to g(0)$, and strict monotonicity of g on $[0, \infty)$ implies $\delta_c(n) \to 0$. At the crossing, $g(\delta_c(n)) = R_R(n)$, so

$$c_2 \delta_c(n)^2 + O(\delta_c(n)^4) = c_2 \frac{v}{n} + O(n^{-2}),$$

so

$$\delta_c(n)^2 = \frac{v}{n} + O(n^{-2}).$$

Taking the positive square root yields

$$\delta_c(n) = \frac{\sqrt{\mu^2 + \sigma^2}}{\sqrt{n}} + O(n^{-3/2}).$$

This proves the theorem. □

A.6 Proof of Proposition 2

Idea. A change in class proportions moves the mixture mean and the Bayes decision boundary in opposite directions. Mean recentering follows the mixture mean, so it moves away from the Bayes correction.

Under pure prior shift, $\mathbb{P}(Y = +1) = \pi$. For threshold t,

$$R_\pi(t) = \pi \Phi\left(\frac{t - \mu}{\sigma}\right) + (1 - \pi)\Phi\left(\frac{-t - \mu}{\sigma}\right).$$

The Bayes-optimal threshold is

$$t_B = \frac{\sigma^2}{2\mu} \log \frac{1 - \pi}{\pi}.$$

If $\pi > 1/2$, then $t_B < 0$, while the population mean

$$m = \mu(2\pi - 1) > 0.$$

The derivative of $R_\pi(t)$ changes sign only at t_B and is positive for $t > t_B$. Therefore $R_\pi(t)$ is strictly increasing from 0 to m, giving

$$R_\pi(m) > R_\pi(0).$$

Thus moving the threshold to the target mean is worse than keeping it at zero. The case $\pi < 1/2$ follows by symmetry. □

A.7 Proof of Corollary 2

Idea. We build two worlds with the same population mean m. In one world the mean moved because the whole distribution translated, so recentering fixes the problem. In the other world the mean moved only because the class proportions changed, so recentering moves the threshold the wrong way.

In World C, balanced translation by m gives population mean m. Population RECENTER uses the true mean and returns the source risk $g(0)$, while KEEP has $g(m) > g(0)$.

In World P, choose

$$\pi = \frac{1 + m/\mu}{2},$$

which is valid for $|m| < \mu$. Then the target mean is also

$$\mu(2\pi - 1) = m.$$

But Proposition 2 gives

$$R_\pi(0) < R_\pi(m),$$

so KEEP is better than population RECENTER. Thus the same population mean corresponds to opposite best actions.

For the finite-batch RECENTER action, the opposite-action conclusion also holds for all sufficiently large n. In World C, Theorem 3 gives $\delta_c(n) \to 0$, so a fixed nonzero m eventually lies above the boundary. In World P, the sample mean converges almost surely to m, and the bounded fresh-point classification loss converges to the strictly worse population-RECENTER risk. □

B Additional Experimental Details

This appendix section records the frozen reviewer checks behind the concise main-text tables. No result below is chosen by selecting the most favorable random seed or window after seeing the outcome.

Observation-channel specification. For source logits $\ell_i \in \mathbb{R}^C$, let $p_i = \mathrm{softmax}(\ell_i)$, $e_i = -\sum_c p_{ic} \log(p_{ic} + 10^{-12})$, $c_i = \max_c p_{ic}$, $m_i = p_{i,(1)} - p_{i,(2)}$, and $\hat{y}_i = \arg\max_c p_{ic}$. Z_1 contains the mean and standard deviation of e_i, c_i, m_i (6 dimensions). Z_2 adds the $0.10, 0.25, 0.50, 0.75, 0.90$ quantiles of each scalar, the normalized predicted-class histogram, and the per-class global mean and standard deviation of the logits, giving $\dim(Z_2) = 21 + 3C$. For the primary Z_3, the ordered sequence is partitioned into $J = 10$ consecutive non-overlapping windows of size $W = 200$. For every window we append the mean and standard deviation of entropy, confidence, and margin, its normalized predicted-class histogram, and its mean-logit vector; we additionally append the 11 ORDER11 statistics defined in Appendix B.4. Hence $\dim(Z_3) = 21 + 3C + 11 + 6J + 2JC = 92 + 23C$ (2,392 on CIFAR-100-C and 2,990 on DomainNet-126). Z_4 concatenates Z_3 with a fixed 512-dimensional sparse random projection of the flattened ordered $N \times C$ source-logit matrix using `SparseRandomProjection(n_components=512, density='auto', random_state=20260830)`. Here `density='auto'` gives density $1/\sqrt{NC}$; each projection entry is zero with probability $1 - d$ and takes values $\pm 1/\sqrt{512d}$ with probability $d/2$ each. Thus $\dim(Z_4) = \dim(Z_3) + 512$ (2,904 on CIFAR-100-C and 3,502 on DomainNet-126). No target labels enter any evidence channel.

B.1 CIFAR-100-C: complete controlled study

The main LOCO results and strong-flip comparison are reported in Section 5.2 and Table 1. Here we record additional audit and robustness checks not shown in the main text.

Primary matched ordering. For each controlled CIFAR corruption/severity/seed/family, the matched IID and correlated worlds use exactly the same selected images and class quota. The IID world uses a random permutation of that fixed selection, whereas the correlated regime uses deterministic class-block ordering. Thus each matched pair differs only in arrival order.

The frozen audit includes exact/tie-aware oracle metrics; five independent RF seeds; feature-dimension control; paired bootstrap confidence intervals; paired significance tests; conventional untouched CIFAR-100-C sanity checks; and a batch-size frontier. Compressing Z_3 to five dimensions performs worse than MORPHEUS-NC, while increasing retained information improves performance. The 16-dimensional projection already has lower point regret than MORPHEUS-NC, with 32–64 dimensions improving further. This is useful evidence that the result depends on preserving decision-relevant information rather than simply naming one representation "better."

The batch-size frontier also confirms that the oracle action itself can move as execution scale changes. Relative to the $B = 200$ reference, action-switch rates are 50.0% at $B = 20$, 31.25% at $B = 50$, 16.67% at $B = 100$, and 29.17% at $B = 500$ in the audited subset.

Proposed selector implementation. The primary selectors use scikit-learn `RandomForestRegressor` with 100 trees, squared-error criterion, no depth limit, `max_features`=1.0, bootstrap sampling, `random_state`=20260831, and `n_jobs`=-1; remaining settings use the defaults of the released environment. For each training world and candidate action, the input concatenates the evidence vector with a one-hot action code and the target is realized post-action accuracy; all five actions are queried at test time and the largest predicted accuracy is selected.

B.2 CIFAR-100-C: STOCHASTIC COUNTER-RESULT

The stochastic suite contains $15 \times 3 \times 3 \times 3 = 405$ worlds: 15 corruptions, severities $\{1, 3, 5\}$, three seeds, and three independently generated deployments per condition. Each world uses $N = 2{,}000$ images at execution batch size $B = 200$. Using RNG seed $7012319+100000s+1000c+100v+d$, we draw $\log_{10} \alpha \sim \mathrm{Unif}(\log_{10} 0.2, \log_{10} 20)$, $\rho \sim \mathrm{Unif}(0, 0.95)$, and a class prior from $\mathrm{Dirichlet}(\alpha \mathbf{1})$. We then sample images without replacement with probabilities induced by that class prior, shuffle examples within class, and generate the stream sequentially by repeating the previous class with probability ρ when available and otherwise sampling the next class proportional to its remaining count. Under stochastic LOCO, mean regret is 0.380 pp for Z_2, 0.338 pp for Z_3, and 0.366 pp for Z_4. The paired $Z_2 - Z_3$ difference is $+0.043$ pp with 95% CI $[-0.091, 0.160]$ pp and Holm-adjusted $p = 0.957$. This counter-result counters a universal-superiority interpretation of stream-aware evidence and supports the observation-channel/deployment-family framing.

B.3 CYGERT ET AL.: STRONGER INFORMATION BUDGET

Our sanity check using the native source/model/data setup of Cygert et al. (2026) closely reproduces their reported CIFAR-100-C SOURCE behavior. Their released evaluation pipeline does not expose every ENT/CON/SND statistic cleanly in our reproducible run, and some released runner paths are inconsistent for candidate methods. We therefore do not alter their source code to force a direct same-table result. More importantly, their model-selection information is available after candidate adaptations have been executed, whereas our main setting chooses from pre-action evidence. We report this as a neighboring protocol with a stronger information budget rather than a direct same-budget baseline.

B.4 DOMAINNET-126: FROZEN NATURAL-DOMAIN STUDY

The study uses six directional transfers: real→clipart, real→painting, real→sketch, clipart→sketch, clipart→real, and clipart→painting. With three seeds and five regimes this gives 90 worlds. Each world uses 2,000 images without replacement. All five TTA actions run at execution batch size 32.

The oracle counts are SOURCE: 41, ROID: 48, DeYO: 1, Tent: 0, and BN-test: 0. All 36 correlated worlds prefer SOURCE, while the 54 IID worlds are mostly ROID. This makes DomainNet-126 a deliberately clean mechanism confirmation rather than the most difficult possible selector benchmark.

Matched IID/correlated worlds share the same selected images and class quota. The oracle changes in 33/36 pairs and 27 are strong flips. Final numerical checks align samples by original dataset index and verify SOURCE-prediction and SOURCE-error consistency. Canonical permutation invariance is checked on the same saved tensor rather than by demanding bitwise equality between separate GPU forwards.

Major-class selection. For DomainNet-126 seed s, we generate a deterministic permutation of the 126 class IDs using `RandomState`(910001+1009s) and take its first 25 classes as the provisional major set. If that set cannot supply the frozen 80% major-class mass without replacement, we deterministically replace its lowest-capacity major classes with the highest-capacity minor classes until the required capacity is available. Prior-60 and prior-80 then assign exactly 60% and 80% of the 2,000 selected images to the major set, respectively, using a max-min/equal-as-possible integer allocation within the major and minor groups subject to per-class capacity.

Primary matched ordering. For each target-domain/seed/family, examples are first sampled without replacement under the frozen class quotas. The IID world is a uniform random permutation of this selected set. The correlated world uses exactly the same selected indices, shuffles examples within each class, and then concatenates class blocks in increasing class-ID order. Sampling and ordering use `RandomState` seed $7712319 + 100000s + 1000i_t + \kappa$, where i_t indexes $(\text{clipart}, \text{painting}, \text{sketch}, \text{real})$ and $\kappa \in \{11, 22, 33\}$ for balanced, prior60, and prior80 families. Thus each matched IID/correlated pair differs only in arrival order.

Implementation compatibility. We use the upstream TTA implementations and frozen configurations without modifying third-party source files. DomainNet-126 requires only two runtime compatibility measures documented in the released audit: BN-test receives the frozen batch size before construction to avoid an upstream initialization-order issue, and rare singleton BatchNorm1d subsets inside DeYO/ROID use the model's existing running statistics for that forward only.

Main and compact evidence. The aggregate LOTO results and paired MORPHEUS-NC comparison are reported in Section 5.3 and Table 2. Here we give the additional window-scale and implementation audits.

ORDER11 features. Let j index consecutive non-overlapping SOURCE-logit windows of size W (retaining a final shorter window), let $\bar{e}_j, \bar{c}_j, \bar{m}_j$ denote the window means of predictive entropy, confidence, and top-1–top-2 probability margin, let h_j and h denote the normalized predicted-class histograms in window j and globally, and let ℓ_j and ℓ denote the corresponding window and global mean-logit vectors. Then $\text{ORDER11} = (\text{sd}_j\, \bar{e}_j, \text{sd}_j\, \bar{c}_j, \text{sd}_j\, \bar{m}_j, \text{mean}_j \|h_j - h\|_1, \max_j \|h_j - h\|_1, \text{mean}_j \|h_j - h\|_2, \max_j \|h_j - h\|_2, \text{mean}_j \|h_{j+1} - h_j\|_1, \max_j \|h_{j+1} - h_j\|_1, \text{mean}_j \|\ell_j - \ell\|_2, \max_j \|\ell_j - \ell\|_2)$. No target labels enter these features.

A CPU-only audit reconstructs the original $W = 200$ Z2/Z3 evidence bitwise exactly, then evaluates the already-predeclared 11-number ORDER11 summary at windows 32, 64, 128, and 200. ORDER11-RF gives 0.034 pp/93.3% at $W = 32$, 0.106 pp/91.1% at $W = 64$, 0.034 pp/93.3% at $W = 128$, and 0.034 pp/93.3% at $W = 200$. Thus the execution-scale $W = 32$ result matches the original $W = 200$ point metrics.

At $W = 32$, ORDER11 and the frozen Z_3 selector choose exactly the same action in all 90 worlds. This equality is an audited result rather than reused predictions: ORDER11 is refit from its own 11-D matrix after the $W = 200$ extractor passes the bitwise fidelity gate. Nearby controls differ, including ORDER11-$W64$ and Ridge, so the pipeline is not hard-wired to reproduce Z_3.

Adding global Z2 to ORDER11 does not improve the $W = 32$, 128, or 200 point metrics, which supports the interpretation that stream structure is a key decision-relevant signal in this construction. On the 27 strong matched flips, ORDER11 changes action in all 27 with zero mean pair regret at every tested window.

The low-dimensional sufficiency result is specific to this DomainNet-126 construction; CIFAR-100-C retains the higher-dimensional Z_3 channel as its primary stream-aware representation, and we do not claim that ORDER11 alone is sufficient there.

Training-objective control. The default Z_3 and ORDER11 selectors use the action-conditioned regressor $g(z, a)$ described in Section 4.1: each training world is paired with every candidate action, the action is represented by a one-hot code, and the target is realized post-action accuracy. Z_3-MultiOutput instead predicts the five candidate accuracies jointly and selects their argmax, matching the MORPHEUS-NC multi-output target formulation; it obtains 0.331 pp regret and 83.3% action accuracy on DomainNet-126, compared with 1.933 pp for MORPHEUS-NC. Z_4 uses the default

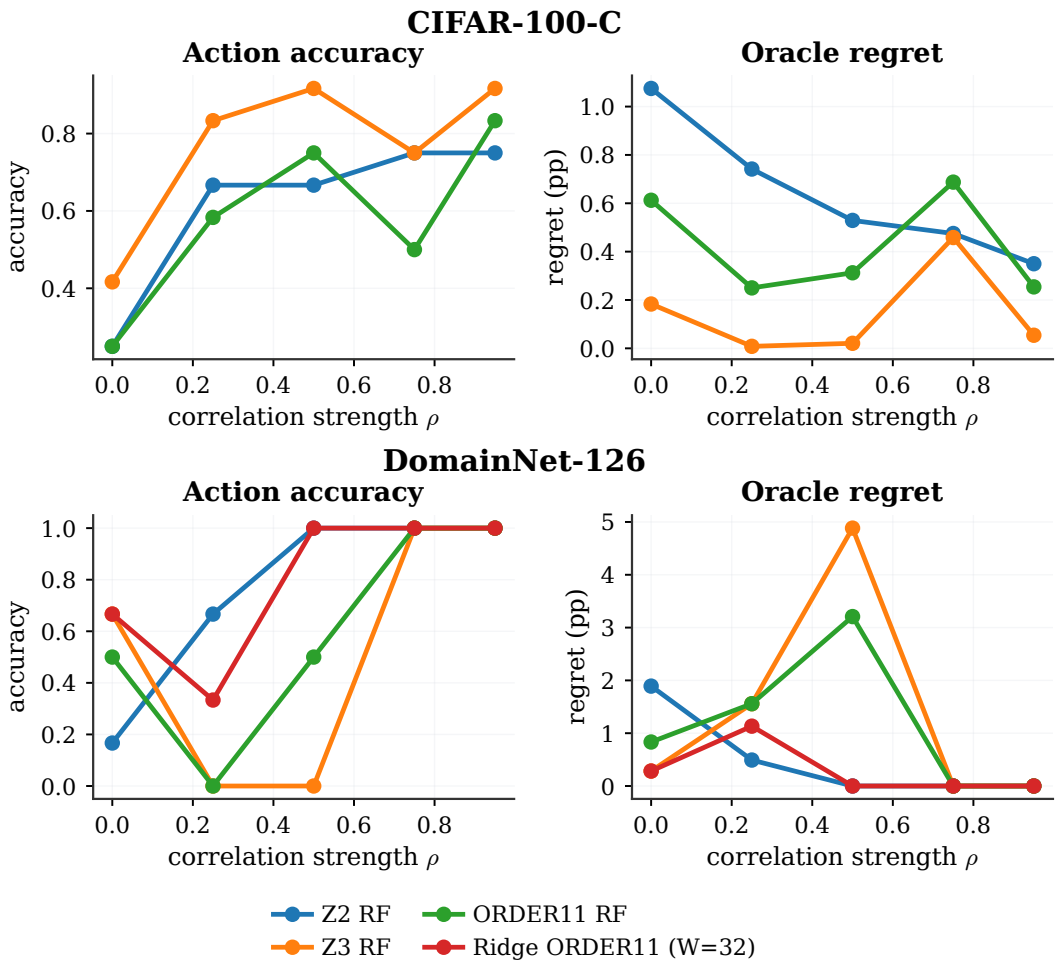

Figure 5: **Selector diagnostics for the correlation-strength sweeps.** *Top:* CIFAR-100-C; *bottom:* DomainNet-126. *Left:* exact action accuracy; *right:* mean oracle regret. DomainNet-126 additionally includes the linear Ridge-ORDER11 ($W = 32$) check. These diagnostics complement the evidence-separation result in Figure 4.

action-conditioned formulation and gives 0.034 pp regret and 93.3% action accuracy, so it is omitted from the compact main table.

Linear-model check. Nested Ridge regression chooses regularization using only the training transfers. With ORDER11 at $W = 32$, Ridge has 0.061 pp mean regret, 92.2% exact action accuracy, 96.7% within 0.5 pp of oracle, and no cases with regret ≥ 5 pp. On the 27 strong matched flips it changes action in 26/27 pairs and has 0.044 pp mean pair regret. The same Ridge architecture on global Z2 remains at 1.933 pp regret, 45.6% action accuracy, and 0/27 action changes.

Robustness and generalization checks. Five RF seeds give identical main Z3 metrics. Fold-only PCA at 5, 16, 32, and 64 dimensions gives the same 0.034 pp/93.3% point result on DomainNet-126, showing that the mechanism is low-dimensional there. Leave-one-target-domain-out regret is 0.180 pp for clipart, 0 for painting, 0.027 pp for real, and 0 for sketch. Leave-one-source-domain-out is harder for the clipart-source family (1.171 pp regret) but still improves substantially over SOURCE (2.522 pp). The conventional native sequence for the real-source setting also behaves normally: ROID improves over SOURCE on the standard domain sequence, confirming that the severe correlated-world adaptation failures are not simply a broken implementation.

B.5 CORRELATION-STRENGTH SWEEP DIAGNOSTICS

The binary matched-world construction is the cleanest main experiment, but we also ran targeted five-level sweeps to show the mechanism in a graded way. For CIFAR-100-C, we used four representative corruptions (contrast, defocus blur, Gaussian noise, and snow), each with three seeds. For DomainNet-126, we used the two held-in transfers clipart→real and real→clipart, again with three seeds. At every ρ, the selected image set was held fixed and only the local class-correlation pattern was changed. Figure 4 gives the compact visual summary, and Figure 5 reports the selector traces. The intended takeaway is modest: as correlation increases, the oracle and stream-aware evidence move, whereas the global order-blind summary changes very little. For readability, the oracle maps in Figure 4 show one fixed seed 0 rather than an aggregate across seeds. We therefore do not resolve three-way seed disagreements inside the displayed map; the figure is intended only to visualize the gradual action transition as ρ increases.

Correlation-sweep construction. We use $\rho \in \{0, 0.25, 0.50, 0.75, 0.95\}$. For each corruption/transfer and seed, the selected image set is fixed across ρ. Within each class, selected examples are shuffled; the stream is then generated sequentially so that, when the previous class still has unused examples, it is repeated with probability ρ, and otherwise the next class is sampled with probability proportional to its remaining count. Thus $\rho = 0$ removes the forced-persistence step, although same-class adjacencies can still occur through the proportional draw. For representation Z, the evidence-movement statistic is the unstandardized Euclidean distance $D_Z(\rho) = \|Z_\rho - Z_0\|_2$ to the corresponding same-condition $\rho = 0$ world. Because these are raw feature-space norms, their absolute magnitudes are interpreted within each representation rather than compared directly across representations. Figure 4 averages these distances over all swept corruption/transfer–seed worlds at each ρ; the 10^{-8} offset is added only for log-scale plotting.

B.6 CLAIM SCOPE AND MORPHEUS IMPLEMENTATION

The studies do not claim that any stream-aware evidence representation is universally best. They also do not claim that every TTA action reversal is caused by order. Their purpose is narrower: to show concrete public-benchmark deployment families in which a restricted observation channel collapses worlds with different oracle actions, and to show that adding the missing kind of information can recover the decision in those families.

MORPHEUS implementation details. Following direct author clarification, the NC baseline extracts the last representation before the classifier, standardizes the NC features, and fits a 600-tree random-forest regressor jointly over the five candidate post-adaptation accuracies using a multi-output wrapper with fixed random seed 0; the action with highest predicted accuracy is selected. We use one canonical Papyan/repository-guided NC1–NC5 vector. Because the original private NC feature-extraction code was not released, we describe this implementation as author-aligned rather than byte-identical. The entropy row is retained as a frozen, simple global-uncertainty ablation; the primary MORPHEUS comparison and significance tests use the NC baseline.