DON'T FIX WHAT ISN'T BROKEN: CALIBRATED REPAIR OF FROZEN FOUNDATION MODELS UNDER PRODUCT SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen foundation models are increasingly deployed in Earth observation, where the scene and its label stay the same but the product changes: satellites are replaced, processing baselines are updated, and resolution, radiometry, or band registration shift. We study how such product shifts break frozen pipelines and what it takes to repair them, across eight encoders, two Sentinel-2 benchmarks, 36 physically grounded degradations, and a real Level-2A-to-Level-1C product change. Only about 4% of the feature displacement reaches the readout, yet this small, concentrated component crosses decision margins and lowers micro-F1 under the real product change by 20 points on average. The information is not lost: a readout refit with labels recovers 92% of the loss. Thirteen test-time adaptation methods recover at most 29% of it, and every one that recovers accuracy also harms clean inputs; batch statistics even mistake class composition for shift, costing up to 37 points on clean streams. We introduce CALIPER, which repairs final embeddings only as far as evidence against a clean null supports, and decides for each image how far it moves. CALIPER recovers 46% of the loss while changing 7 of 15,000 clean predictions, carries over to images outside the benchmark, comes close to the strongest baseline on the real product change without its clean-input loss, and stays within 0.11 points on clean streams where batch-level adaptation fails.

1 INTRODUCTION

Vision Foundation Models (VFMs) are increasingly used as frozen feature extractors: one encoder is reused across tasks, and a lightweight readout is trained on its representations. This works as long as the input product stays compatible with the one used to train the readout. In practice, the scene and its label may remain unchanged while the product changes through spatial resolution, spectral alignment, radiometry, or processing level. The encoder then produces a different representation, while the readout remains fixed. We refer to this setting as *product shift*. It arises whenever the measurement or processing pipeline evolves, from scanner replacement in medical imaging (Castro et al., 2020) to revisions of a satellite processing chain. Earth observation (EO) makes product shift both consequential and unusually measurable: Landsat has maintained a continuous record since 1972, Sentinel-2 has undergone both satellite replacement and processing-baseline changes (Wulder et al., 2022; European Space Agency, 2021; 2022), and changes in resolution, calibration, band registration, and noise can be studied while the scene and label stay fixed.

Benchmarks such as RSPDBench (Faruk et al., 2026) show that such product changes cause substantial, model-dependent losses in frozen EO foundation models. Knowing that performance drops, however, does not explain why it drops or what can be done about it. Tracing paired clean and degraded scenes through the frozen pipeline, our first result is that representation drift alone does not explain failure. Product degradations move frozen features substantially, but only about 4% of that movement lies in the directions the readout uses; errors arise when this small, concentrated component crosses decision margins, across readouts and under a real Sentinel-2 processing-level change. The information is largely not lost: a readout refit with labels recovers 92% of the loss. Deployment, however, provides neither labels nor paired clean images, so recovery must work from unlabeled data alone.

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.

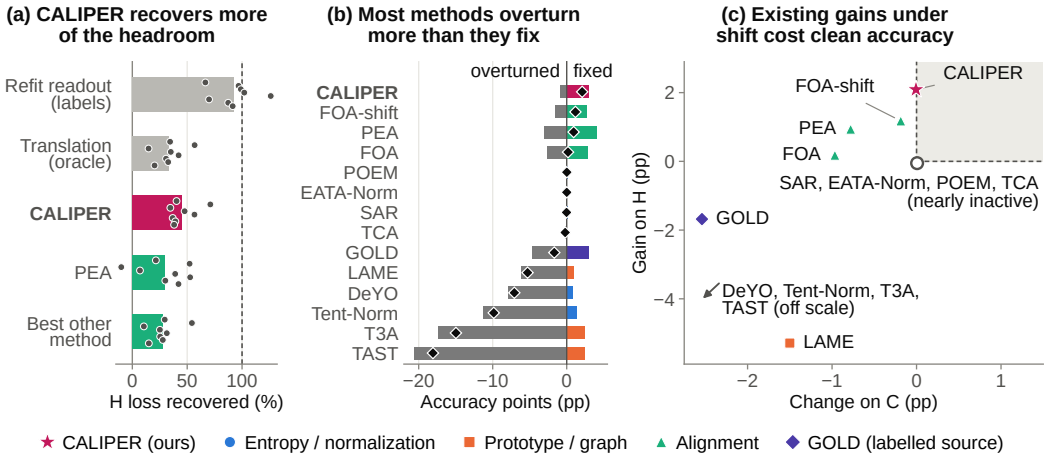


Figure 1: **Existing test-time adaptation leaves most of the recoverable accuracy unused and pays for its gains on clean inputs; CALIPER recovers more at almost no clean cost** (m-EuroSAT). (a) A readout refit with labels recovers 92% of the loss on homogeneous streams (H); existing label-free methods recover under 30%, and CALIPER 46%, more than an oracle translation with paired clean inputs (33%). Dots are encoders. (b) Most existing methods overturn more correct predictions than they fix; CALIPER has the largest net gain (diamonds). (c) Among existing methods, only those that barely act (ring) avoid clean loss; CALIPER gains 2.1 points on H while changing 7 of 15,000 clean predictions. Panels (b) and (c) average over five encoders (Appendix G).

Test-time adaptation (TTA) is designed for this setting, but most methods are developed on benchmark corruptions and adapt whenever target data arrive (Hendrycks & Dietterich, 2019; Wang et al., 2021). We evaluate 13 established methods on homogeneous, heterogeneous, mixed clean–degraded, and clean streams, counting every changed prediction as a corrected error or a new one. Most methods overturn more correct predictions than they fix; those that remain quiet on clean streams barely act under shift; and batch statistics confound product change with class composition, so that even the strongest method, PEA (Xiao et al., 2026), loses up to 37 points on clean streams when labels arrive in correlated runs. A useful method must therefore answer a question that standard adaptation leaves implicit: *is there enough evidence of product shift to justify changing this prediction at all?*

We introduce **CALIPER** (Calibrated Per-image Embedding Repair), which separates *whether* to correct from *which* images to correct and *how far*. From unlabeled clean source embeddings, CALIPER learns how much a stream drifts by chance when nothing has shifted, measures drift on residuals that remove the effect of class composition, and corrects only as far as the evidence against this clean null supports. Each image then moves by its own posterior estimate of shift, so images that look clean stay in place. Unlike batch-level alignment (Lim et al., 2023; Xiao et al., 2026), CALIPER needs only final embeddings and the frozen readout, with no gradients or extra forward passes.

CALIPER recovers 46% of the lost accuracy, more than any existing label-free method (at most 29%) and more than an oracle translation that uses paired clean inputs (33%), while changing 7 of 15,000 clean predictions (Figure 1). Its gains carry over to 1,000 EuroSAT images outside the benchmark; on a real Sentinel-2 change, it comes close to PEA on Level-1C inputs and doubles its gain on mixed streams while leaving clean predictions intact.

Collectively, these results give a decision-level account of product-shift failure, a recoverability analysis of existing test-time adaptation, and a calibrated method that recovers more without harming clean inputs.

2 ANATOMY OF PRODUCT-SHIFT FAILURE

2.1 FROZEN READOUTS UNDER PAIRED PRODUCT SHIFT

Models and tasks. We adopt the classification setup of RSPDBench (Faruk et al., 2026): seven foundation models, namely five EO-specific encoders (Prithvi-EO-2.0 (Szwarcman et al., 2024), DOFA (Xiong et al., 2025), Galileo (Tseng et al., 2025), RAMEN (Houdré et al., 2026), and Panopticon (Waldmann et al., 2025)) and DINOv3 with web-image (DINOv3-Web) and satellite (DINOv3-Sat) pretraining (Siméoni et al., 2026), together with a ViT-B/16 trained from scratch with supervision, on the GEO-Bench versions of single-label m-EuroSAT and multilabel m-BigEarthNet (Lacoste et al., 2023). For model m, f_{θ_m} denotes the frozen encoder with its native feature extraction, producing $\mathbf{z} = f_{\theta_m}(x) \in \mathbb{R}^{p_m}$ (Appendix A). Each model's *original probe* is a readout trained and selected on clean data only (for the supervised ViT, its own classification head) and kept fixed; we write it as an affine map $\boldsymbol{\ell} = W_{0,m}\mathbf{z} + \mathbf{b}_{0,m}$. Predictions are scored by top-1 accuracy on m-EuroSAT and micro-F1 on m-BigEarthNet.

Product degradations. We use the product-degradation suite of RSPDBench (Faruk et al., 2026) without modification. A condition d applies an operator $\mathcal{D}_d$ to the image in dataset-native units, before model-specific preprocessing, and leaves the label unchanged. Four primitives isolate single defects: resolution loss (P1), radiometric gain (P2), band misregistration (P3), and sensor noise (P4). Two chains combine them: acquisition quality (C1) and harmonization (C2), with its resolution-only and gain-only components. With three severities (S1–S3) and both gain directions, this gives 36 conditions for multispectral encoders and 18 for RGB encoders, for which P3 and C2 are undefined (Appendix B). As a real product change, we replace m-BigEarthNet's Level-2A inputs with the Level-1C products of the same acquisitions.

2.2 PRODUCT SHIFTS INDUCE STRUCTURED REPRESENTATION CHANGE

Representation change is structured and largely preserves geometry. For a clean test image x_i with features $\mathbf{z}^0_{i,m}$ and degraded features $\mathbf{z}^d_{i,m} = f_{\theta_m}(\mathcal{D}_d(x_i))$, let $\Delta\mathbf{z}^d_{i,m}$ be their difference; because the scene, label, preprocessing, and weights are fixed, each pair isolates the product change. We measure how far representations move relative to their clean norm (relative displacement, RD) and how much the similarity structure across images changes (geometry disruption, GD $= 1-$CKA; Kornblith et al., 2019; Appendix C). Both increase strictly with severity along all 168 trajectories (Figure 2a), and resolution loss and the acquisition chain move foundation-model features the most (Figure 2b), whereas the supervised ViT is most sensitive to uniform gain. At the most severe level, the median displacement on m-EuroSAT reaches 28% of the clean norm while the median CKA stays at 0.97. Because linear CKA is invariant to a common translation, rotation, and isotropic rescaling, this is consistent with a movement that largely preserves the relative arrangement of images, of the kind a transport map could undo (Section 4).

2.3 FAILURE IS READOUT-RELATIVE, BUT PRODUCT SENSITIVITY PERSISTS

Representation change is necessary but not sufficient for failure. The probe's bias cancels in paired differences, so the logit change is exactly $\Delta\boldsymbol{\ell}^d_{i,m} = W_{0,m}\Delta\mathbf{z}^d_{i,m}$.

Only a small but concentrated component reaches the readout. Let $P_{0,m}$ project onto the k_m-dimensional row space of $W_{0,m}$. Only $P_{0,m}\Delta\mathbf{z}$ changes the logits, and its share of the displacement energy,

$$\nu_{m,d} = \frac{\sum_i \|P_{0,m}\Delta\mathbf{z}^d_{i,m}\|_2^2}{\sum_i \|\Delta\mathbf{z}^d_{i,m}\|_2^2}, \tag{1}$$

averages about 4%, so most of the movement leaves the logits unchanged. This is not merely a consequence of subspace size. A randomly oriented displacement would place a fraction k_m/p_m of its energy in the row space; the observed share exceeds this reference in 87% of cells on m-EuroSAT, by a median factor of 4.5, and in 56% of cells on m-BigEarthNet, by a median factor of 1.3 (Figure 2c). Degradation thus concentrates its movement in the probe's directions, strongly on m-EuroSAT and mildly on m-BigEarthNet.

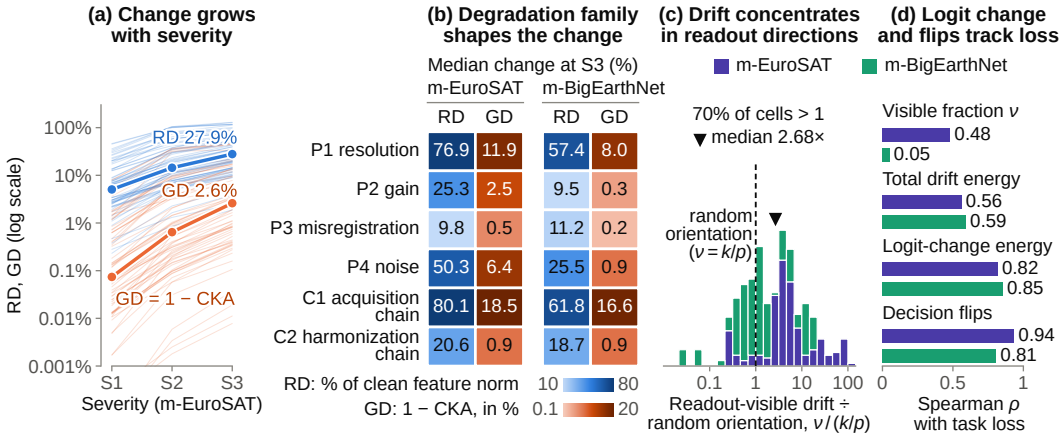


Figure 2: **Product degradations move frozen features far, but only a small, concentrated part of the movement reaches the readout, and that part predicts failure.** Seven foundation models; each test image is compared with its own degraded version. (a) Relative displacement (RD) and geometry disruption (GD $= 1-$CKA) grow with severity on m-EuroSAT. (b) Medians at the most severe level by degradation family. (c) Readout-visible share of the movement relative to a randomly oriented movement; values above 1 (dashed line) mean it concentrates in the readout's directions. (d) Spearman correlation with task loss: logit change and decision flips track loss more closely than total displacement or the visible share.

Logit change and decision margins determine failure. Pooled across models and conditions, logit-change energy tracks task loss (Spearman $\rho = 0.82$ on m-EuroSAT and 0.85 on m-BigEarthNet) far more closely than displacement energy (0.56 and 0.59) or the visible share (0.48 and 0.05; Figure 2d). Within a single model, both rank conditions well (median $\rho = 0.93$–0.96), but only the logit change is measured in units that are comparable across encoders: those of the readout itself. A logit change matters only if it flips a decision, which happens when it exceeds the clean margin to a competing class or, for a multilabel label, crosses zero; small-margin predictions therefore fail first, and decision flips correlate with loss at $\rho = 0.94$ and 0.81.

The sensitivity is not an artifact of one probe. An independently trained linear probe and an MLP, both fit on clean data only, change the size of individual losses but preserve which conditions are most damaging (Spearman $\rho = 0.95$ on m-EuroSAT and 0.90–0.96 on m-BigEarthNet). Readouts trained only on the probe's null space keep 97.5–98.9% of clean accuracy, yet they rank conditions the same way ($\rho = 0.93$–0.98; Appendix D). The sensitivity therefore comes from the representation, not from the particular readout.

Real products confirm the failure. Replacing Level-2A inputs with the Level-1C products of the same scenes lowers micro-F1 by 19.8 points on average under the original probes, and by similar amounts under the independently trained probes, while re-extracting the same Level-2A products costs only 1.2 points; the loss therefore reflects the processing level rather than patch extraction (Appendix E).

3 How Much of the Failure Is Recoverable?

Section 2 located the failure in small, readout-visible changes that cross decision margins. We now ask how much of the resulting loss can be recovered, and with what information.

Most of the lost performance remains in the frozen embedding. A linear readout refit with labels on features of all degradation conditions recovers 92% of the loss on homogeneous streams, averaged over the eight encoders, while changing clean accuracy by -0.7 to $+2.5$ points (Figure 1a).

For RAMEN and Prithvi-EO-2.0 it recovers more than 100%, because the refit readout is also more accurate than the original one on clean inputs. The task information therefore survives in the frozen embedding. Translating each image's embedding by the average clean-to-degraded displacement of its degradation recovers 33%. Both references use labels or paired clean inputs, which deployment does not provide. They are reference points rather than bounds: the translation applies one average correction per degradation, and a per-image correction can exceed it (Section 4).

Label-free recovery on realistic streams. We evaluate 13 test-time adaptation methods: seven that use only the checkpoint and unlabeled target images (Tent (Wang et al., 2021), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), T3A (Iwasawa & Matsuo, 2021), TAST (Jang et al., 2023), LAME (Boudiaf et al., 2022), TCA (You et al., 2025)), five that also use unlabeled clean source statistics (EATA (Niu et al., 2022), POEM (Bar et al., 2024), PEA (Xiao et al., 2026), FOA (Niu et al., 2024), and FOA-shift, its activation-shift component run alone), and GOLD (Lai et al., 2026), which uses labelled source data. Each method sees four unlabeled streams of the same 1,000 test scenes in batches of 64: homogeneous (H; one condition and severity), heterogeneous (D; mixed conditions), mixed (M; half clean, half degraded), and clean (C). It resets for every stream, never sees labels or degradation metadata, and keeps its authors' order of updating and predicting; we count the errors it fixes and the correct predictions it overturns (Appendix F).

Most existing methods overturn more correct predictions than they fix. Averaged over the five encoders of Figure 1b, 8 of the 13 methods overturn more predictions than they fix, and only FOA-shift and PEA gain more than negligibly. Near-zero methods are mostly inactive: SAR, EATA, and TCA use thresholds released for 1,000-class ImageNet, which rarely allow an update at ten classes (Appendix F). The harmful ones are active but imprecise: T3A and TAST replace the trained readout rather than correct the shift, and lose as much accuracy on clean streams as on degraded ones.

No existing label-free method gains under shift without harming clean inputs. Every existing method that recovers accuracy under shift also changes clean predictions, and those that leave clean inputs untouched recover essentially nothing (Figure 1c). Batch statistics cannot separate a product shift from a change in class composition: with correlated labels, PEA loses up to 37 points on clean streams, and its clean-stream loss grows about fivefold as batches shrink from 64 to 16 (Appendix G).

Existing label-free methods reach only a fraction of the headroom. PEA and the best other method recover 29% and 27% of the loss, about as much as the translation reference and far less than the refit readout (Figure 1a). Recovering more requires a correction that is estimated for each image rather than the batch, does not mistake class composition for shift, and stays inactive on clean inputs without blocking useful corrections; CALIPER (Section 4) is designed to meet all three.

4 Calibrated Per-image Embedding Repair

CALIPER repairs the final embedding before the unchanged readout and corrects each image only as far as the evidence supports. Like PEA (Xiao et al., 2026), it uses unlabeled clean source embeddings; at test time it needs only the final embeddings of arriving images and the frozen readout, with no labels, gradients, intermediate layers, or extra forward passes.

Setting. For a fixed model, let $\mathbf{z} \in \mathbb{R}^p$ be a final embedding, $\boldsymbol{\ell}(\mathbf{z}) = W\mathbf{z} + \mathbf{b}$ the readout, and $\hat{\mathbf{y}}(\mathbf{z})$ its decision: the one-hot arg max, or $\mathbf{1}[\boldsymbol{\ell}(\mathbf{z}) > 0]$ for multilabel tasks. CALIPER works in whitened coordinates $\mathbf{u} = T(\mathbf{z})$, in which clean source embeddings have zero mean and unit variance in every direction, so departures are measured relative to how much clean data naturally vary. It keeps a memory of the embeddings seen so far in the stream, cleared at the start of every stream, and for each arriving batch re-estimates its corrections from the memory before predicting the batch. Three principles decide whether to correct, which images, and how far; three corrections then apply them (Algorithm 1).

Residual detection keeps class composition from looking like shift. When some classes arrive more often than in the source data, the mean embedding moves even though nothing is degraded.

Algorithm 1 CALIPER on one arriving batch. Steps indexed by i apply to every embedding in the memory; corrections are recomputed from the memory at every batch.

Input: arriving batch; memory $\mathcal{M}$ of n whitened embeddings; readout $\boldsymbol{\ell}$
Precomputed from clean source data: whitening T; decision means B; clean nulls $\mathbb{E}_0[\cdot \mid n]$; source centroids
Output: predictions for the batch

Add the batch to $\mathcal{M}$; $\mathbf{r}_i \leftarrow \mathbf{u}_i - B^\top[1; \hat{\mathbf{y}}_i]$ ▷ residual detection
$g_{\mathrm{tr}} \leftarrow g(\|\bar{\mathbf{r}}\|^2; n)$; $w \leftarrow (1 - \mathrm{df}/\chi^2)_+$ ▷ null shrinkage
$\tilde{\mathbf{u}}_i \leftarrow \mathbf{u}_i - g_{\mathrm{tr}}\big[(1-w)\,\hat{s}_i^{\mathrm{u}}\mathbf{v}^{\mathrm{u}} + w\,\hat{s}_i^{\mathrm{c}}\mathbf{v}^{\mathrm{c}}\big]$ ▷ translation
for eigen-directions $\mathbf{v}_j$ of the translated residuals **do**
  $g_j \leftarrow g(\lambda_j; n)$; **if** $g_j = 0$ **break**
  $\tilde{\mathbf{u}}_i \leftarrow \tilde{\mathbf{u}}_i - (1 - g_{\mathrm{tr}})\, g_j\, \hat{s}_{j,i}\, \mathbf{v}_j$ ▷ mixture directions
end for
$q_i \leftarrow g_{\mathrm{tr}}\, \omega_i$; fit $\hat{f}$ by q-weighted entropic OT; $g_{\mathrm{aff}} \leftarrow g(S_{\mathrm{aff}}; n_{\mathrm{eff}})$
$\tilde{\mathbf{u}}_i \leftarrow \tilde{\mathbf{u}}_i + g_{\mathrm{aff}}\, q_i\big(\hat{f}(\tilde{\mathbf{u}}_i) - \tilde{\mathbf{u}}_i\big)$ ▷ affine transport
return $\boldsymbol{\ell}\big(T^{-1}\tilde{\mathbf{u}}_i\big)$ for the images of the current batch

CALIPER therefore measures departures on residuals against the readout's own decisions, $\mathbf{r}_i = \mathbf{u}_i - B^\top[1; \hat{\mathbf{y}}_i]$, where B holds the average source embedding for each decision. A change in class frequencies alone leaves the mean residual at zero when the readout's decisions are correct (Proposition 1), and no labels are needed.

Null shrinkage decides whether to correct. Even a clean stream differs from the source by chance. CALIPER therefore compares each statistic S of the memory, such as the squared norm of the mean residual, with its average over clean memories of the same size, $\mathbb{E}_0[S \mid n]$, computed once from source images (Appendix I), and scales its correction by

$$g(S; n) = \Big(1 - \frac{\mathbb{E}_0[S \mid n]}{S}\Big)_+. \tag{2}$$

A statistic at its clean level gives $g = 0$ and no correction; twice that level gives $g = 1/2$. Small memories therefore need stronger evidence, and no threshold has to be chosen. This g is a positive-part James–Stein factor, which under a Gaussian model never makes the estimated shift worse on average than applying it in full (Proposition 2; James & Stein, 1961; Baranchik, 1970).

Per-image evidence decides which images move and how far. A stream-level factor cannot separate clean from degraded images in a mixed stream. Along a shift direction $\mathbf{v}$, CALIPER writes each image's projection $a_i = \mathbf{u}_i^\top \mathbf{v}$ as a clean part, distributed like held-out clean source projections, plus a shift s_i, whose distribution in the stream is estimated by nonparametric maximum likelihood (Kiefer & Wolfowitz, 1956; Koenker & Mizera, 2014). Each image moves by its posterior mean shift $\hat{s}_i = \mathbb{E}[s_i \mid a_i]$, with posterior probability $\omega_i = P(|s_i| > h \mid a_i)$ of being shifted. Images that look clean stay in place even when $g = 1$.

Three corrections of increasing flexibility. The *translation* removes a shift shared by the stream. Its direction blends the mean embedding ($\mathbf{v}^{\mathrm{u}}$), which is reliable when images arrive in random order, with the mean residual ($\mathbf{v}^{\mathrm{c}}$), which is reliable under any class mix, using the weight $w = (1 - \mathrm{df}/\chi^2)_+$; w rises when the mix of predicted classes changes across batches. The *mixture directions* remove opposite shifts that cancel in the mean: directions of largest residual variance are tested one at a time against their clean null. The *affine transport* corrects changes in spread and orientation with an entropic optimal-transport map to the clean source (Cuturi, 2013; Perrot et al., 2016; Courty et al., 2017), weighting each image by $q_i = g_{\mathrm{tr}}\, \omega_i$. Corrections are recomputed from the memory at every batch, so they never accumulate; all settings are listed in Appendix J.

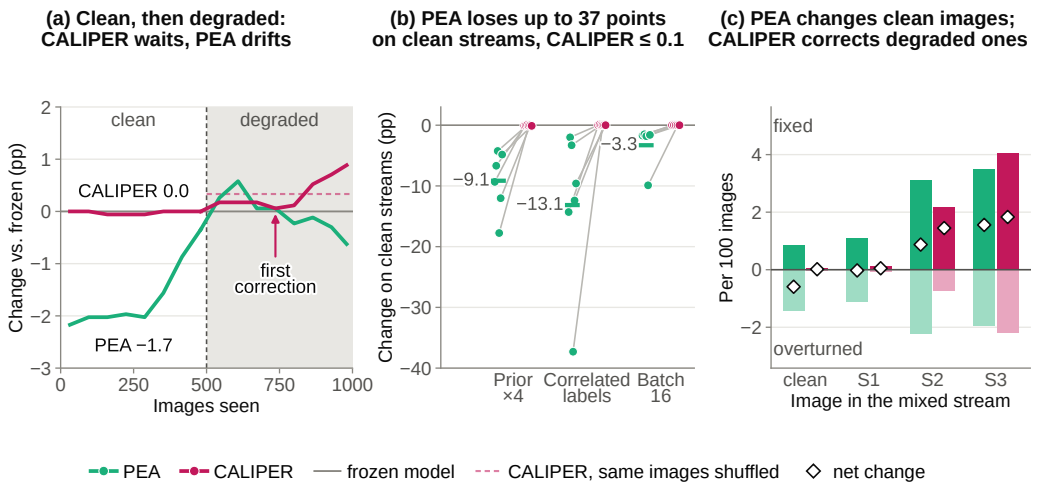


Figure 3: **CALIPER stays reliable where batch-level adaptation breaks down** (m-EuroSAT). (a) With all clean images first, PEA loses 1.7 points before the switch; CALIPER stays unchanged, corrects a few batches after it (arrow), and then gains as much as on shuffled streams (dashed). (b) Under a $\times 4$ class prior, correlated labels, or batches of 16, PEA loses up to 37 points on clean streams and CALIPER at most 0.11 (lines join each encoder's two results). (c) In mixed streams, PEA changes clean images, while CALIPER changes almost none and gains more on degraded ones (diamonds).

5 Experiments

We show that CALIPER recovers more accuracy than existing methods without harming clean inputs, that it stays reliable in exactly the streams where batch-level adaptation fails, and that each of its components is needed.

5.1 Setup

We use the eight encoders of Section 2.1, each with its original clean-trained readout, and the four streams of Section 3. On real m-BigEarthNet, the homogeneous stream holds Level-1C products, the clean stream the benchmark Level-2A inputs, and the mixed stream both. We compare CALIPER with the 13 methods of Section 3 on m-EuroSAT, and with PEA on m-BigEarthNet. On both datasets we also compare with six embedding-level baselines: mean shift, CORAL (Sun & Saenko, 2016), quantile matching, EM re-estimation of the class prior (Saerens et al., 2002), AdaNPC (Zhang et al., 2023), and optimal-transport domain adaptation (Courty et al., 2017). Every method that uses source data receives the same 2,000 unlabeled clean training images, and no method sees target labels or degradation metadata.

Unseen images. We also evaluate on 1,000 EuroSAT images outside the m-EuroSAT benchmark, class-balanced and processed identically, to test whether the results carry over beyond the benchmark split.

5.2 CALIPER Recovers More Without Harming Clean Inputs

CALIPER outperforms PEA and every embedding-level baseline on unseen images. On the unseen images, CALIPER gains more than PEA and every embedding-level baseline on each degraded stream type, and it leaves clean accuracy unchanged where PEA loses 1.4 points (Table 1). The test split confirms the ranking with three readouts per encoder: CALIPER leads PEA on homogeneous, mixed, and clean streams with each of the three readouts (mean over encoders), and the two are tied on heterogeneous streams, where PEA leads only on DINOv3-Sat. Measured against the headroom of Section 3, CALIPER recovers 46% of the homogeneous-stream loss: more than

Table 1: **CALIPER recovers accuracy under product shift while leaving clean predictions unchanged.** Gain over the frozen predictor in points, mean over eight encoders. Unseen: 1,000 EuroSAT images outside the benchmark, evaluated once (one readout per encoder); test: three readouts. Best other: for each column, the best other same-budget method (on unseen images, the embedding-level baselines).

	EuroSAT, unseen				m-EuroSAT, test				m-BigEarthNet, real		
Method	H	D	M	C	H	D	M	C	L1C	M	L2A
PEA	+3.03	+1.35	−0.18	−1.41	+3.36	+1.46	+0.02	−0.59	+3.92	+0.91	−1.44
Best other	+3.21	+0.69	+0.08	−0.04	+3.44	+0.87	+0.00	**+0.00**	**+7.25**	+1.58	**−0.03**
CALIPER	**+4.73**	**+1.86**	**+0.81**	**+0.00**	**+4.48**	**+1.54**	**+0.53**	−0.01	+3.60	**+1.77**	−0.04

PEA (29%), and more than an oracle that translates each image using its degradation and clean counterpart (33%; Figure 1a).

The gains transfer to a real Sentinel-2 product change. On Level-1C inputs, CALIPER comes close to PEA (3.6 versus 3.9 micro-F1 points), doubles its gain on mixed streams, and leaves clean Level-2A predictions unchanged, where PEA loses 1.4 points. Quantile matching gains more on Level-1C inputs but harms clean ones; among methods that leave clean inputs intact, CALIPER recovers the most.

Clean images stay in place. Across the five encoders on which every method ran, CALIPER changes 7 of 15,000 clean predictions (Figure 1c). Within mixed streams, its corrections fall almost entirely on degraded images, whereas PEA changes clean and degraded images alike (Figure 3c).

5.3 CALIPER Stays Reliable Where Batch-Level Adaptation Fails

A product change mid-stream. A processing-baseline update switches a stream from one product to another at a single moment. On streams that switch from clean to degraded half-way, PEA loses up to 3.8 points before the switch, adapting to shifts that are not there. CALIPER leaves the clean segment untouched, detects the switch within two batches on 92% of streams, and then recovers half of the gain it reaches when the whole stream is degraded.

Class frequencies, label order, and batch size. These are exactly the changes that batch statistics mistake for a shift. With a fourfold class prior, correlated labels, or batches of 16, PEA loses up to 37 points on streams of clean images, while CALIPER never changes clean accuracy by more than 0.11 (Figure 3b). The clean null is what protects it: CALIPER still registers the skewed class frequencies, but its correction stays too small to change decisions. The limit of this protection is label-correlated arrival: CALIPER stays safe on clean streams, but it no longer recovers accuracy on heterogeneous and mixed streams.

5.4 What Each Component Contributes

The clean null and the per-image amounts carry the method. Without the clean null, CALIPER acts on chance fluctuations: it loses accuracy on every clean stream, by up to 7 points, and even gains less under shift (Table 2). Without per-image amounts, a single correction for all images halves the gain on homogeneous streams and removes almost all of it on real Level-1C inputs. The mixture and affine steps add about one point beyond translation alone. Class conditioning helps only under label-correlated arrival, by an amount comparable to the variation across readouts; the clean null, not class conditioning, protects against changes in class frequency.

Sensitivity, cost, and scope. Halving or doubling any of CALIPER's main settings changes its mean gains by at most 0.4 points (Appendix L), and CALIPER adds no pass through the encoder, whereas PEA needs six forward passes per batch; on CPU, however, our unoptimized implementation takes 0.6 to 30 seconds per batch of 64 as the embedding dimension grows from 1,024 to 8,192, comparable to PEA's 2 to 19 seconds on a GPU. Because it repairs only the final embedding,

Table 2: **Each component protects the case it was designed for.** Gain in points, mean over eight encoders (prior: six). Each row removes one component (✓ = included); red marks results worse than full CALIPER by more than twice the variation across readouts (0.3 points on shuffled streams, 0.8 under correlated labels). One readout per encoder; full ablation in Appendix K.

Components				m-EuroSAT			Prior	Corr.	Real BEN	
Clean null	Class cond.	Per-image	Mix. & affine	H	D	C	×4 C	D	L1C	L2A
✓	✓	✓		+3.32	+0.76	+0.01	−0.06	−0.63	+2.84	−0.04
✓	✓		✓	+2.42	+0.96	−0.01	−0.05	−0.45	+0.13	−0.09
✓		✓	✓	+4.56	+1.76	−0.01	+0.01	−0.99	+3.96	−0.04
	✓	✓	✓	+4.18	+1.15	−0.67	−7.01	−2.30	+0.85	−4.31
✓	✓	✓	✓	+4.56	+1.71	−0.03	−0.05	−0.16	+3.70	−0.04

CALIPER recovers less when a shift is absorbed by earlier layers: on standard ImageNet-C with ResNet-50, re-estimating BatchNorm statistics throughout the network gains 24–30 points, while CALIPER gains 3.6 (Appendix M).

6 Related Work

Product shift and frozen readouts. Synthetic-corruption robustness (Hendrycks & Dietterich, 2019) need not transfer to natural shifts (Taori et al., 2020; Koh et al., 2021). EO benchmarks span tasks, sensors, geography, and corruptions (Lacoste et al., 2023; Marsocci et al., 2026; Li et al., 2025; Doerksen & Kerner, 2026; Tamazyan et al., 2025); we build on the paired product degradations of RSPDBench (Faruk et al., 2026). Frozen readouts suit shift: fine-tuning distorts features (Kumar et al., 2022), last-layer retraining restores robustness (Kirichenko et al., 2023), and shifts favor different layers (Lee et al., 2023). CKA (Kornblith et al., 2019) measures representation change, not whether it reaches the decision.

Test-time adaptation. TTA uses entropy minimization (Wang et al., 2021; Niu et al., 2022; Lee et al., 2024; Niu et al., 2023), output adjustment (Iwasawa & Matsuo, 2021; Jang et al., 2023; Boudiaf et al., 2022), or feature-statistic alignment (Schneider et al., 2020; Lim et al., 2023; Niu et al., 2024; Xiao et al., 2026), tested under mixed domains (Marsden et al., 2024; Tomar et al., 2024; Fan et al., 2026), correlated labels (Gong et al., 2022; Yuan et al., 2023), no shift (Mummadi et al., 2021; Bar et al., 2024), and negative flips (Yan et al., 2021); we combine these regimes on identical scenes.

Transport and calibrated correction. Domain adaptation applies one map to all inputs, via correlation alignment (Sun & Saenko, 2016) or optimal transport (Courty et al., 2017; Perrot et al., 2016; Cuturi, 2013). Whether to adapt has been cast as shift detection (Rabanser et al., 2019) or entropy betting (Bar et al., 2024); label-shift methods separate class-frequency from covariate change (Lipton et al., 2018; Saerens et al., 2002). CALIPER instead calibrates each correction against a clean null and moves each image by its empirical-Bayes posterior shift (Kiefer & Wolfowitz, 1956; Efron, 2010; Koenker & Mizera, 2014).

7 Conclusion

Product shift breaks frozen foundation-model pipelines through a small, readout-visible part of a large but structured representation change, while most of the lost information survives in the frozen embedding. Existing test-time adaptation recovers little of it and harms clean inputs, because batch-level corrections cannot tell a product shift from a change in class composition, or degraded images from clean ones. CALIPER decides against a clean null whether to correct and decides for each image how far to move it; it recovers more than existing methods on average, holds its advantage on images outside the benchmark, and leaves clean predictions essentially unchanged. Reliable recovery from product shift therefore depends not only on how a model is corrected, but on knowing when a correction is warranted.

AI USE STATEMENT

In this work, we used generative AI tools to implement the method and the evaluation pipelines. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in writing proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, clean or reformat datasets, or interpret results, and translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to polish and edit the text of the paper and to find and summarize related literature. We have reviewed all AI-assisted work: all code was reviewed by the authors, and every reported number, table, and figure was verified by the authors against the underlying result files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use publicly available datasets (GeoBench versions) and pretrained encoders; the degradation suite follows the companion benchmark; all corrections, detectors, predictors, and baselines operate on cached features and logits with fixed seeds (42/43/44). Baseline outcomes were pre-registered in run metadata before execution, and a machine-checked hygiene audit of the data flow (split disjointness; no test labels anywhere) ships with the code. Code, calibration protocols, and full per-cell tables will be released.

ETHICS STATEMENT

This work analyzes and improves the reliability of Earth-observation pipelines under realistic product defects; we foresee no direct negative societal impact beyond general EO dual-use considerations.

REFERENCES

Yarin Bar, Shalev Shaer, and Yaniv Romano. Protected test-time adaptation via online entropy matching: A betting approach. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9b35a0a20d617dc68ae98a7a57df2f51-Abstract-Conference.html.

A. J. Baranchik. A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *The Annals of Mathematical Statistics*, 41(2):642 – 645, 1970. doi: 10.1214/aoms/1177697104. URL https://doi.org/10.1214/aoms/1177697104.

Alvin John Baranchik, United States Office of Naval Research, and Stanford University Department of Statistics. Multiple regression and estimation of the mean of a multivariate normal distribution, 1964. URL http://statistics.stanford.edu/~ckirby/techreports/ONR/CHE%20ONR%2051.pdf.

M. E. Bock. Minimax Estimators of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 3(1):209 – 218, 1975. doi: 10.1214/aos/1176343009. URL https://doi.org/10.1214/aos/1176343009.

Malik Boudiaf, Romain Müller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8334–8343. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00816. URL https://doi.org/10.1109/CVPR52688.2022.00816.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL https://doi.org/10.1038/s41467-020-17478-w.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921. URL https://doi.org/10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Kelsey Doerksen and Hannah Kerner. Earthshift: a benchmark for measuring robustness to real-world distribution shifts in earth observation. *CoRR*, abs/2605.29330, 2026. doi: 10.48550/ARXIV.2605.29330. URL https://doi.org/10.48550/arXiv.2605.29330.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.

European Space Agency. Copernicus Sentinel-2: Major Products Upgrade Upcoming - Sentinel Online — sentinels.copernicus.eu. https://sentinels.copernicus.eu/-/copernicus-sentinel-2-major-products-upgrade-upcoming, 2021. [Accessed 25-09-2026].

European Space Agency. Imminent deployment of Sentinel-2 processing baseline 04.00 on 25 January 2022. Sentinel Online, January 2022. URL https://sentinels.copernicus.eu/-/imminent-deployment-of-sentinel-2-processing-baseline-04.00-on-25-january-2022. Accessed 24 September 2026.

Xiao Fan, Jingyan Jiang, Zhaoru Chen, Fanding Huang, Xiao Chen, Qinting Jiang, Bowen Zhang, Xing Tang, and Zhi Wang. Moetta: Test-time adaptation under mixed distribution shifts with moe-layernorm. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 21011–21019. AAAI Press, 2026. doi: 10.1609/AAAI.V40I25.39243. URL https://doi.org/10.1609/aaai.v40i25.39243.

Tanjim Bin Faruk, Khondaker Masfiq Reza, Shrideep Pallickara, and Sangmi Lee Pallickara. Rspdbench: Benchmarking vision foundation models on earth observation tasks under physically grounded remote-sensing product degradations, 2026. URL https://arxiv.org/abs/2609.23427.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: robust continual test-time adaptation against temporal correlation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ae6c7dbd9429b3a75c41b5fb47e57c9e-Abstract-Conference.html.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Nicolas Houdré, Diego Marcos, Hugo Riffaud de Turckheim, Dino Ienco, Laurent Wendling, Camille Kurtz, and Sylvain Lobry. Ramen: Resolution-adjustable multimodal encoder for earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27838–27848, June 2026.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2427–2440, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1415fe9fea0fa1e45dddcff5682239a0-Abstract.html.

W. James and Charles Stein. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 361–379, Berkeley, Calif., 1961. University of California Press. URL http://projecteuclid.org/euclid.bsmsp/1200512173.

Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=EzLtB4M1SbM.

J. Kiefer and J. Wolfowitz. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics*, 27(4):887 – 906, 1956. doi: 10.1214/aoms/1177728066. URL https://doi.org/10.1214/aoms/1177728066.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014. doi: 10.1080/01621459.2013.869224. URL https://doi.org/10.1080/01621459.2013.869224.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Alexandre Lacoste, Nils Lehmann, Pau Rodríguez, Evan D. Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vázquez, Dava J. Newman, Yoshua Bengio, Stefano Ermon, and Xiaoxiang Zhu. Geo-bench: Toward foundation models for earth monitoring. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a0644215d9cff6646fa334dfa5d29c5a-Abstract-Datasets_and_Benchmarks.html.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. *CoRR*, abs/2603.21928, 2026. doi: 10.48550/ARXIV.2603.21928. URL https://doi.org/10.48550/arXiv.2603.21928.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=APuPRxjHvZ.

EL Lehmann and George Casella. *Theory of point estimation*. Springer, 2 edition, 1998.

Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiaoxiang Zhu, and Tianjin Huang. Reobench: Benchmarking robustness of earth observation foundation models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/58af38e2fba04b5e00c3450018c56406-Abstract-Datasets_and_Benchmarks_Track.html.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=EQfeudmWLQ.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Robert A. Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 2543–2553. IEEE, 2024. doi: 10.1109/WACV57701.2024.00254. URL https://doi.org/10.1109/WACV57701.2024.00254.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, 14(1):245–285, 2026. doi: 10.1109/MGRS.2025.3628194.

Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=yUxUNaj2Sl.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 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 *The Eleventh International*

Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38298–38315. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/niu24a.html.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4197–4205, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/26f5bd4aa64fdadf96152ca6e6408068-Abstract.html.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1394–1406, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Comput.*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/85690f81aadc1749175c187784afc9ee-Abstract.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns. Featured Certification.

Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, pp. 443–450, 2016. doi: 10.1007/978-3-319-49409-8_35. URL https://doi.org/10.1007/978-3-319-49409-8_35.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Thorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arévalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell D. Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabé-Moreno. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *CoRR*, abs/2412.02732, 2024. doi: 10.48550/ARXIV.2412.02732. URL https://doi.org/10.48550/arXiv.2412.02732.

Hakob Tamazyan, Ani Vanyan, Alvard Barseghyan, Anna Khosrovyan, Evan Shelhamer, and Hrant Khachatrian. Geocrossbench: Cross-band generalization for remote sensing. *CoRR*, abs/2511.02831, 2025. doi: 10.48550/ARXIV.2511.02831. URL https://doi.org/10.48550/arXiv.2511.02831.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18583–18599. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Unmixing test-time normalization statistics: Combatting label temporal correlation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 50967–50988, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/de08b3ee7c0043a76ee4a44fe68e90bc-Paper-Conference.pdf.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gqZO3eSZRy.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J. Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 2204–2214. Computer Vision Foundation / IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/EarthVision/html/Waldmann_Panopticon_Advancing_Any-Sensor_Foundation_Models_for_Earth_Observation_CVPRW_2025_paper.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Michael A. Wulder, David P. Roy, Volker C. Radeloff, Thomas R. Loveland, Martha C. Anderson, David M. Johnson, Sean Healey, Zhe Zhu, Theodore A. Scambos, Nima Pahlevan, Matthew Hansen, Noel Gorelick, Christopher J. Crawford, Jeffrey G. Masek, Txomin Hermosilla, Joanne C. White, Alan S. Belward, Crystal Schaaf, Curtis E. Woodcock, Justin L. Huntington, Leo Lymburner, Patrick Hostert, Feng Gao, Alexei Lyapustin, Jean-Francois Pekel, Peter Strobl, and Bruce D. Cook. Fifty years of landsat science and impacts. *Remote Sensing of Environment*, 280: 113195, 2022. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2022.113195. URL https://www.sciencedirect.com/science/article/pii/S0034425722003054.

MA Xiao, Young D. Kwon, Pan Zhou, and Dong Ma. Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7kLNGaAHaw.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation, 2025. URL https://arxiv.org/abs/2403.15356.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 14299–14308. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01407. URL https://openaccess.thecvf.com/

content/CVPR2021/html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

Linjing You, Jiabao Lu, and Xiayuan Huang. Test-time correlation alignment. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/you25c.html.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 15922–15932. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01528. URL https://doi.org/10.1109/CVPR52729.2023.01528.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41647–41676. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23am.html.

A Models and Native Feature Extraction

All encoders are frozen and used under the clean-selected native protocol of RSPDBench (Faruk et al., 2026): each model receives the bands, input size, and normalization of its own input contract, and its image-level feature vector $\mathbf{z} = f_{\theta_m}(x)$ is produced by the pooling recipe that RSPDBench selected on clean validation data. We change none of these choices. Table 3 lists the recipe for each model. The same recipe, applied to the degraded image, gives $\mathbf{z}^d$, so every measure in the paper compares two outputs of one fixed function. Every input normalization uses fixed statistics: per-band means and standard deviations, constant divisors, or per-band percentiles computed once on the clean training set. None uses per-image statistics. A uniform gain therefore reaches the input of every encoder. For the DINOv3 encoders on m-EuroSAT, whose scaled inputs are clipped to $[0, 1]$, clipping can absorb part of an upward gain.

Readouts. For each foundation model, the original probe is the readout that RSPDBench trained and selected on clean training and validation data. For the supervised ViT, it is the network's own classification head, trained end to end on clean data; this head reads exactly the concatenated feature in Table 3. We treat every readout as an affine map, $\boldsymbol{\ell} = W_{0,m}\mathbf{z} + \mathbf{b}_{0,m}$, with fixed feature standardization and inference-mode BatchNorm folded in. On m-EuroSAT, we take this map directly from the stored probe weights. On m-BigEarthNet, the probes were saved as training checkpoints, so we fit the map to the saved features and logits; it reproduces the saved logits to within 2×10^{-5} and changes no prediction. Each foundation model is evaluated with three readouts, trained with different seeds on the same frozen encoder. The supervised ViT is instead trained three times, end to end.

B Product Degradations

All synthetic degradations are the RSPDBench suite (Faruk et al., 2026), used without modification. We summarize it here for completeness; the physical motivation for each operator and parameter is given in that paper. Operators act on an image $x \in \mathbb{R}^{C\times H\times W}$ in the dataset's native units, band by band where relevant, before any model-specific band selection, normalization, or resizing. They leave the label unchanged.

Primitives.

- **P1, effective spatial-resolution loss:** $\hat{x} = U_N(D_N(G_\sigma * x))$. The image is blurred by a Gaussian point-spread function G_σ, reduced by area averaging over $N \times N$ blocks (D_N), and resampled bilinearly to its original size (U_N).

Table 3: **Native feature-extraction recipes** (RSPDBench protocol; models are cited in Section 2.1). Bands use Sentinel-2 names. Where the two datasets differ, entries read m-EuroSAT / m-BigEarthNet; p_m is the feature dimension.

Model	Backbone (pretraining)	Input bands	Input size and normalization	Feature $\mathbf{z}$	p_m
EO-specific foundation models					
Prithvi-EO-2.0	ViT, 600M (HLS)	B02, B03, B04, B8A, B11, B12	224 px; dataset / Prithvi statistics	Mean of final-layer patch tokens	1,280
DOFA	ViT-L (wavelength-conditioned)	B02–B08, B11, B12	224 px; dataset statistics	Mean of final-layer tokens	1,024
Galileo	Galileo-Base	B02–B08, B8A, B11, B12	Native size (64 / 120 px), patch size 4; Galileo normalization, no clipping	Mean of encoder tokens	768
RAMEN	RAMEN-Base (resolution-adaptive)	All 13 / 12 bands	Native size; input GSD 30 / 40 m, selected on clean validation; PANGAEA statistics	Pooled multi-layer features	3,072
Panopticon	ViT-B/14 (any-sensor)	All 13 / 12 bands	224 px; GEO-Bench band statistics	[CLS] of last 1 / 4 blocks $\|\|$ mean patch token	1,536 / 3,840
DINOv3 foundation models					
DINOv3-Web	ViT-7B/16 (web, LVD-1689M)	B04, B03, B02 (RGB)	256 px; fixed per-band 2–98% training percentiles, clipped to $[0, 1]$ / reflectance $\times 10^{-4}$; model-card statistics	Last-block [CLS] $\|\|$ mean patch token	8,192
DINOv3-Sat	ViT-7B/16 (satellite, SAT-493M)	B04, B03, B02 (RGB)	256 px; reflectance$/3000$, clipped to $[0, 1]$ / reflectance $\times 10^{-4}$; model-card statistics	Last-block [CLS] $\|\|$ mean patch token	8,192
Supervised baseline					
ViT-B/16	ViT-B/16, trained from scratch on each dataset	All 13 / 12 bands	224 px	Spatial means of blocks 3, 5, 7, 11, concatenated	3,072

- **P2, uniform radiometric gain shift:** $\hat{x} = p\,x$, with $p = 1 - \alpha$ (down) or $1 + \alpha$ (up), applied to every band without clipping.
- **P3, band misregistration:** the bands of a group S are translated by a sub-pixel displacement Δ_s (bilinear sampling, border padding); all other bands stay on the product grid. Each image keeps its shift direction at every severity, and only the length of the shift increases.
- **P4, band-dependent sensor noise:** $\hat{x}_c = x_c + k_s \mu_c z_c$. Here μ_c is the clean training mean of band c, and z_c is a standard-normal field determined by the image, the band, and the seed, so the added-noise SNR is $1/k_s$.

Compound chains. Each chain at severity s uses the Ss setting of every component.

- **C1, acquisition-quality chain:** gain scaling as in P2, then the P1 blur and area reduction, then P4 noise added at the reduced resolution, then bilinear resampling to the original size.
- **C2, product-harmonization chain:** the P1 resolution loss and a P3-type shift, both on a geometric band group G, followed by a signed gain on a radiometric band group H. Bands outside G and H are untouched.
- **C2 components:** RSPDBench also evaluates the resolution-only and gain-only parts of C2 separately, as single-component controls.

Band groups and applicability. P3 moves the red-edge, narrow-NIR, and SWIR bands, $S = \{$B05, B06, B07, B8A, B11, B12$\}$, against the fixed visible and broad-NIR bands. The C2 groups follow each dataset's RSPDBench contract:

- m-EuroSAT (foundation models): $G = \{$B05, B06, B07, B8A, B11, B12$\}$ and $H = \{$B08, B8A, B11, B12$\}$;

Table 4: **Severity levels** of the RSPDBench primitives (Faruk et al., 2026). Chains combine the settings of the same severity. P3 shifts are in native pixels.

Severity	P1 (blur σ, factor N)	P2 (gain)	P3 (shift)	P4 (k_s; SNR)
S1	$\sigma = 0.5$ px, $N = 1$	$\pm 3\%$	0.3 px	0.01; 100
S2	$\sigma = 1.0$ px, $N = 2$	$\pm 8\%$	1.0 px	0.03; 33
S3	$\sigma = 2.0$ px, $N = 4$	$\pm 15\%$	2.0 px	0.05; 20

- m-BigEarthNet (foundation models and supervised ViT): $G = H = \{$B08, B8A, B11, B12$\}$;
- supervised ViT on m-EuroSAT: RSPDBench's generic C2, with a single group for G and H.

The supervised ViT's pipeline selects the P3 and C2 bands by the band names that the dataset loader returns. For m-EuroSAT, these names do not match the stored band order, B01–B08, B09, B10, B11, B12, B8A. Because of this band-ordering error, the ViT's m-EuroSAT P3 conditions shift $\{$B05, B06, B07, B09, B12, B8A$\}$, and its C2 conditions act on $\{$B08, B09, B12, B8A$\}$: B09 is perturbed in place of B11.

- **Unaffected conditions.** The ViT's other 18 m-EuroSAT conditions (P1, P2, P4, C1) act on all bands, and their noise is scaled by each band's own clean mean, so they are unaffected.
- **Unaffected dataset.** On m-BigEarthNet the stored order matches the band names, which we verified from each band's native resolution.
- **Consequences.** Every method receives the same degraded inputs, so comparisons between methods on the ViT remain valid. We keep the ViT out of condition-level comparisons across models; the family medians of the representation analysis use the seven foundation models only.

On m-EuroSAT, the supervised ViT's C2 conditions therefore apply a different operator from those of the foundation models, and we do not compare the two on C2. Degradations act on the full product, so a model is affected only through the bands it consumes. P3 and C2 are undefined for the RGB encoders. Multispectral encoders therefore see 36 conditions:

- P1, P3, P4: 3 each;
- P2, C1, C2: 6 each, in both gain directions;
- C2 resolution-only: 3;
- C2 gain-only: 6.

RGB encoders see 18, from P1, P2, P4, and C1.

Random components. The noise field (P4, C1) and the shift direction (P3, C2) are deterministic functions of the image identifier, the band, and an operator seed. The same realization is reused at every severity, so moving along a severity axis changes only the strength of the defect, never its random draw. All records of a model share one operator seed (42 on m-EuroSAT and for the supervised ViT), except for DOFA and Prithvi-EO-2.0 on m-BigEarthNet, for which RSPDBench used each run's seed (Appendix C).

C Representation-Change Measures

Features. All measures use the image-level feature vectors returned by each model's native feature-extraction recipe (Appendix A), on the same $N = 1{,}000$ test images for the clean and degraded inputs. Image identifiers and labels are checked for alignment before any measure is computed.

Relative displacement.

$$\mathrm{RD}_{m,d} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{z}_{i,m}^{d} - \mathbf{z}_{i,m}^{0}\|_2}{\|\mathbf{z}_{i,m}^{0}\|_2}.$$

Normalizing by the clean norm makes displacement comparable across conditions of one encoder, whose feature scale is fixed; it does not make displacement comparable across encoders (Section 2.3).

Linear CKA. Let $Z_{m,0}, Z_{m,d} \in \mathbb{R}^{N\times p_m}$ hold the clean and degraded features of the same images in the same order, and let $\tilde{Z}$ denote column-centered matrices. Then

$$\mathrm{CKA}_{\mathrm{lin}}(Z_{m,0}, Z_{m,d}) = \frac{\|\tilde{Z}_{m,d}^{\top}\tilde{Z}_{m,0}\|_F^2}{\|\tilde{Z}_{m,0}^{\top}\tilde{Z}_{m,0}\|_F\,\|\tilde{Z}_{m,d}^{\top}\tilde{Z}_{m,d}\|_F}, \qquad \mathrm{GD}_{m,d} = 1 - \mathrm{CKA}_{\mathrm{lin}}.$$

Linear CKA is invariant to orthogonal transformations and isotropic scaling of either representation (Kornblith et al., 2019), and centering makes it invariant to a common translation. A change that moves all images by the same vector, rotates them, or rescales them uniformly therefore leaves GD unchanged, which is why a large RD with a small GD indicates movement that largely preserves the relative arrangement of images. For large p_m, CKA is computed from the $N \times N$ centered Gram matrices.

Cells, trajectories, and aggregation. A *cell* is a (dataset, model, condition) triple. A *trajectory* fixes the dataset, the model, and the degradation specification (family, gain direction, and C2 component) while severity varies from S1 to S3.

- **Counts.** On each dataset, the six multispectral models (five foundation models and the supervised ViT) contribute all 12 specifications. The two RGB DINOv3 encoders contribute the six that apply to RGB inputs: P1, P4, and P2 and C1 in both gain directions. This gives 84 trajectories per dataset, 168 in total, over 504 cells. The seven foundation models alone account for 144 trajectories and 432 cells.
- **Records.** Every cell has three run records.
  - For the foundation models, the encoder is shared across records. On m-EuroSAT, and for five of the seven m-BigEarthNet foundation models, the three records share identical features, so RD and GD are identical across them.
  - For DOFA and Prithvi-EO-2.0 on m-BigEarthNet, the random degradation components used each run's seed. The degraded features of the 18 affected conditions (of 36) therefore differ slightly across records, with standard deviations of at most 0.0007 in RD and 0.0046 in CKA.
  - For the supervised ViT, each record is a separately trained network evaluated on one common degradation realization. Its records differ by at most 0.025 in RD and 0.0073 in GD (standard deviation).

  Every reported value is the mean over the three records.
- **Monotonicity.** "Strictly increasing" means S1 $<$ S2 $<$ S3 for both measures on every trajectory, with no tolerance. It holds for all 168 trajectories, for the stored values and for the double-precision recomputation, and for each supervised-ViT record separately.
  - RD rises by at least 0.004 at every step, so the severity ordering does not rest on small differences.
  - The smallest GD step, 5.5×10^{-6}, still exceeds the single–double precision discrepancy bound of 10^{-6}.

  Both minima occur for RAMEN under P4 on m-BigEarthNet.
- **Medians in Figure 2b.** These are taken within each dataset over all S3 cells of a family for the seven foundation models: every encoder the family applies to, both gain directions, and, for C2, the full chain and both single-component controls. Each median pools 14 (C1), 25 (C2), 7 (P1), 14 (P2), 5 (P3), and 7 (P4) cells per dataset.

- **The supervised ViT.** On the operators it shares with the foundation models, its profile differs markedly. The S3 median RD under uniform gain (P2) is 0.55 against 0.25 on m-EuroSAT and 0.38 against 0.10 on m-BigEarthNet. Under sensor noise (P4) it is 0.04 against 0.50 and 0.08 against 0.26 (Table 5).

Table 5: **Representation change at the most severe level, by model and degradation family.** Relative displacement (RD) and geometry disruption (GD = 1−CKA, shown $\times 10^3$) at S3: the median over the model's S3 conditions in each family (both gain directions; for C2, the full chain and its two single-component controls), averaged over the three run records. P3 and C2 are undefined for the RGB DINOv3 encoders. Every value increases strictly from S1 to S3 (Appendix C). The released CSV lists RD and GD for all 504 cells (168 trajectories × 3 severities).

	RD at S3						GD at S3 ($\times 10^3$)					
Model	P1	P2	P3	P4	C1	C2	P1	P2	P3	P4	C1	C2
m-EuroSAT												
Prithvi-EO-2.0	0.189	0.299	0.024	0.100	0.342	0.134	8.2	29.8	0.1	1.1	35.8	3.3
DOFA	0.431	0.141	0.098	0.193	0.523	0.144	131.0	25.1	7.4	78.0	190.7	18.9
Galileo	0.769	0.103	0.191	0.503	0.801	0.354	97.3	2.5	6.6	47.5	184.2	26.9
RAMEN	0.089	0.578	0.019	0.105	0.588	0.227	2.3	116.7	0.1	2.8	115.0	4.0
Panopticon	0.864	0.355	0.184	0.681	0.981	0.273	118.9	13.7	5.2	64.1	189.6	9.9
DINOv3-Web	1.129	0.256	–	0.639	1.287	–	442.7	44.5	–	156.1	456.1	–
DINOv3-Sat	1.174	0.119	–	0.897	1.168	–	335.3	8.2	–	228.4	450.6	–
ViT-B/16 (sup.)	0.228	0.552	0.115	0.043	0.592	0.318	6.4	21.5	1.3	0.1	27.9	9.6
m-BigEarthNet												
Prithvi-EO-2.0	0.181	0.109	0.044	0.112	0.219	0.125	2.8	1.1	0.5	1.6	6.9	3.2
DOFA	0.417	0.080	0.119	0.055	0.458	0.103	208.6	9.6	16.9	8.8	272.0	12.1
Galileo	0.629	0.064	0.112	0.255	0.626	0.364	79.8	1.6	4.0	30.6	165.6	51.2
RAMEN	0.079	0.259	0.022	0.012	0.276	0.285	1.5	27.8	0.1	0.0	30.4	5.9
Panopticon	0.574	0.157	0.130	0.309	0.623	0.252	23.9	3.2	1.6	5.1	33.5	6.3
DINOv3-Web	1.152	0.060	–	0.287	1.173	–	611.4	2.6	–	55.7	607.6	–
DINOv3-Sat	1.330	0.095	–	0.565	1.361	–	535.2	3.0	–	111.6	560.1	–
ViT-B/16 (sup.)	0.492	0.377	0.359	0.076	0.620	0.700	8.0	6.0	3.5	0.2	13.0	24.3

D Task-Subspace Controls and Reader Dependence

These controls use the seven foundation models on m-BigEarthNet and six on m-EuroSAT. RAMEN's m-EuroSAT feature cache for this analysis fails the readout reconstruction check, so it is excluded there. Values are first averaged over seeds 42–44 within a model, then weighted equally across models; correlations are taken over model–condition cells.

D.1 Clean Decodability of Canonical Row and Null Components

Using the canonical projection P_0, we define

$$\mathbf{z}^{\text{row}} = P_0\mathbf{z}, \qquad \mathbf{z}^{\text{null}} = (I - P_0)\mathbf{z}. \tag{3}$$

Matched linear and MLP readers are trained on each projected representation using clean training data only. The same fixed projection is used for clean and degraded evaluation.

Table 6: **Clean decodability and degradation ordering in canonical row and null components.** Clean retention is the projected reader's clean score divided by that of its matched full-feature reader, computed per model and seed and then averaged. ρ_S is the Spearman correlation between the row- and null-reader relative losses across model–condition cells.

Dataset	Reader	Row retention	Null retention	ρ_S
m-EuroSAT	Linear	98.99%	98.82%	0.931
m-EuroSAT	MLP	99.88%	98.86%	0.931
m-BigEarthNet	Linear	97.44%	97.51%	0.969
m-BigEarthNet	MLP	99.16%	98.77%	0.982

The high null retention shows that substantial clean-decodable task signal exists outside the canonical row space. Because the null space has dimension $p - r$, whereas the row space has dimension r, this result establishes redundant decodability rather than equal information density.

D.2 Rank-Matched PCA and Random Controls

To test whether the canonical row space performs well merely because it retains r dimensions, we compare it against two rank-matched controls:

- the top r principal components of the clean representation;
- random r-dimensional subspaces.

Clean scores of linear readers (accuracy on m-EuroSAT, micro-F1 on m-BigEarthNet):

Dataset	**Canonical row**	**Top-r PCA**	**Random rank-r**
m-EuroSAT	95.18%	86.41%	65.15%
m-BigEarthNet	61.85%	55.30%	48.55%

The canonical row representation outperforms both rank-matched controls in all 39 model–seed comparisons: 18 on m-EuroSAT and 21 on m-BigEarthNet. The random control is compared through its mean over draws. The row space is therefore a task-aligned low-dimensional subspace rather than an arbitrary r-dimensional projection.

D.3 The Null Component Is Not Systematically More Robust

Despite strong clean decodability, readers trained on the canonical null component inherit nearly the same degradation ordering as readers trained on the row component. Row–null relative-loss correlations range from $\rho_S = 0.931$ to 0.982. Under S3 resolution loss (P1) and the acquisition chain (C1), the null readers' mean relative drops remain between 48.07% and 61.81% across datasets and reader types.

Thus, the representation contains multiple clean-decodable routes to the task, but the alternative route exposed by the canonical null space is not a hidden robust subspace.

D.4 Different Clean Readers Select Different Subspaces

Let $\widetilde{W}$ denote the independently trained clean linear reader. Its row space differs substantially from that of the canonical W_0. Averaged over independently trained heads, the mean squared-cosine row-space overlap is 0.494 on m-EuroSAT and 0.309 on m-BigEarthNet. The corresponding mean principal angles are 45.8° and 58.7°.

The energy decomposition changes accordingly. Pooled over the 396 model–condition cells and weighted by displacement energy, the visible fractions are

$$Q_{W_0} = 4.01\%, \qquad Q_{\widetilde{W}} = 1.35\%. \tag{4}$$

The unweighted cell means are 3.93% and 1.70%, respectively.

Pooled across models, the alternative reader's visible fraction is essentially uncorrelated with that reader's relative degradation loss:

$$\rho_S \left(q_{\widetilde{W}}, L_{\widetilde{W}} \right) = 0.014. \tag{5}$$

This pooled value is descriptive. Within a model, the association ranges from -0.53 to 1.00 (median 0.47): it is positive for all six m-EuroSAT models and mixed in sign on m-BigEarthNet. The visible fraction is therefore not comparable across encoders as a predictor of loss.

These results show that there is no single universal reader-visible task subspace. Different clean-trained linear solutions use different feature directions while preserving a highly similar degradation-failure ordering.

D.5 Projected-Hybrid Diagnostics

The exact identity

$$W_0 \Delta \mathbf{z}^{d,\perp} = \mathbf{0} \tag{6}$$

applies only to the canonical reader W_0. A displacement that is null to W_0 can be visible to $\widetilde{W}$ or to a nonlinear MLP.

We additionally evaluate row-only and null-only projected hybrid representations for alternative readers as a diagnostic. These hybrids should not be interpreted as additive causal partitions of task loss: projection can produce feature vectors that do not lie on the natural encoder representation manifold, and row and null contributions can cancel or mask one another when recombined. We therefore use these experiments only as supporting reader-specific diagnostics rather than as estimates of mediated performance fractions.

E Real-Product Cohorts and Results

Cohort construction. We use all 1,000 patches of the GeoBench m-BigEarthNet test split; no patch is selected or excluded. For each patch, we download from the public Sentinel-2 archive the L1C product from which BigEarthNet derived the patch (the product identifier recorded in its metadata; 102 products at processing baselines 02.05 and 02.06, with sensing times agreeing to within 5 s). We crop the patch's footprint in the product's own UTM projection and resample each band from its native 10, 20, or 60 m resolution onto the patch's 120×120, 10 m grid by bilinear interpolation. We convert L1C digital numbers to top-of-atmosphere reflectance and L2A values to surface reflectance using the quantification value in the product metadata, adding the per-band radiometric offset introduced in Processing Baseline 04.00 where present (European Space Agency, 2022), and store reflectance $\times 10^4$, the benchmark's scale, without clipping. None of the L1C products carries an offset. Both products then pass through each model's identical preprocessing and feature extraction, and the original labels are kept.

Same-level control. We extract archive L2A patches for the same acquisitions and footprints in the same way. Of the 1,000 patches, 927 have a usable archive L2A crop: the archive holds no L2A product for the acquisitions of 61 patches (five granules), and 12 further crops have fewer than 99.9% valid pixels.

- **Processing versions.** The archive L2A products come from ESA processing at baselines 02.05–02.08 (224 patches) or from the Collection-1 reprocessing at baseline 05.00 (703 patches). The benchmark patches were produced by the BigEarthNet authors with Sen2Cor 2.5.5, so the control also contains processor-version differences.
- **Alignment.** Phase correlation at pixel precision finds no offset between the benchmark patch and the L1C crop for any of the 1,000 patches (median per-band correlation 1.000). For the control, 676 of the 927 L2A crops have no offset. The other 251, all baseline-05.00 products, are offset by 1–2.2 pixels, consistent with the geolocation refinement in that reprocessing. We neither correct nor exclude these offsets, so the control also includes a small misregistration.
- **Clouds and invalid pixels.** No cloud masking is applied, since each crop comes from the same acquisition as its benchmark patch. Pixels with digital number 0 (no data) or 65,535 (saturated) are invalid. Every crop used has at least 99.9% valid pixels in all 12 bands, and its few invalid pixels are set to 0.

Representation change. Table 7 reports the relative displacement and geometry disruption of each encoder for two comparisons: the L2A-to-L1C change on all 1,000 scenes, and the same-level control on its 927 scenes. As a reference, it also gives the median over the encoder's synthetic S3 conditions on m-BigEarthNet (Appendix C).

- The L1C product moves every representation 2.1–8.8 times further than re-extracting the same L2A product does.
- For five of the eight encoders, it moves the representation further than the encoder's median synthetic S3 condition.
- It does not always disrupt geometry more than the control. For Prithvi-EO-2.0 the control's GD is slightly larger, reflecting the misregistration of 251 control crops.

- Displacement alone does not determine the task loss. DINOv3-Web moves by 0.26 of its clean feature norm and loses 1.3 points; Galileo moves by 0.27 and loses 21.3 (Table 8).

Physical severity is not matched between the synthetic and real comparisons.

Table 7: **Real-product representation change on m-BigEarthNet.** Relative displacement (RD) and geometry disruption (GD $= 1-$CKA) of each encoder's features between benchmark L2A and L1C products of the same 1,000 scenes. Ctrl: benchmark L2A against archive L2A of the same 927 scenes. S3 median: median over the encoder's synthetic S3 conditions. Mean over seeds 42–44.

	L2A $\rightarrow$ L1C		Ctrl		Synthetic S3 median	
Encoder	RD	GD	RD	GD	RD	GD
DINOv3-Web	0.263	0.026	0.124	0.012	0.718	0.328
DINOv3-Sat	0.753	0.159	0.222	0.025	0.947	0.323
DOFA	0.233	0.063	0.080	0.013	0.111	0.015
Galileo	0.268	0.023	0.045	0.004	0.310	0.041
Panopticon	0.821	0.062	0.109	0.009	0.281	0.006
Prithvi 2.0	0.291	0.006	0.033	0.007	0.122	0.003
RAMEN	1.077	0.189	0.125	0.034	0.262	0.009
ViT-B/16 (supervised)	2.205	0.148	0.303	0.016	0.552	0.009

Task performance. Table 8 reports micro-F1 for every encoder and readout.

- **Size of the drop.** Under the original probes, replacing benchmark L2A with L1C lowers micro-F1 by 19.7 points averaged over the eight encoders. On the seven foundation models, the drop is 16.7 points under both linear probes and 22.4 under the MLP.
- **927-scene subset.** The L1C drop is essentially unchanged on the 927 scenes available for the control: 16.7, 16.8, and 22.4 points on the foundation models, against 16.7, 16.7, and 22.4 on all 1,000, and 41.1 against 41.0 for the supervised ViT.
- **Archive L2A as reference.** Using the archive L2A patches instead of the benchmark patches as the reference gives drops of 15.6, 15.9, and 21.3 points for the three readouts on the foundation models, and 39.9 for the supervised ViT. The loss is therefore not an artifact of the benchmark extraction.

F Test-Time Adaptation Baselines: Protocol and Fidelity

This appendix specifies the streams and the accounting behind Section 3, how each method was ported to frozen EO encoders, and the checks of those ports.

Test streams. Every method sees the same 1,000 m-EuroSAT test scenes, degraded with the RSPDBench operators of Appendix B: 36 conditions for the multispectral encoders and the supervised ViT, and the 18 RGB-applicable conditions for the two DINOv3 encoders. For each encoder and seed we build 43 streams (25 for DINOv3):

- **Homogeneous (H):** one stream per degraded condition, holding all 1,000 scenes under that condition (36 streams; 18 for DINOv3).
- **Clean (C):** one stream of the 1,000 clean scenes.
- **Mixed (M):** a fixed random half of the scenes is degraded and the other half stays clean. The degraded half is assigned the conditions in rotation, so each condition covers 13–14 scenes (27–28 for DINOv3). Three streams.
- **Heterogeneous (D):** every scene is degraded, with the conditions again assigned in rotation, so each condition covers 27–28 scenes (55–56 for DINOv3). Three streams.

H and C streams present the scenes in one fixed random order. The three M streams and the three D streams each use a different random order, but the scene-to-condition assignment is the same in

Table 8: **Replacing L2A inputs with L1C products lowers m-BigEarthNet micro-F1 under all clean-trained readouts, while replacing them with archive L2A products of the same acquisitions barely does.** Micro-F1 (%, threshold 0.5) on benchmark L2A inputs and on L1C products of the same 1,000 scenes. Drop is L2A minus L1C in points; Ctrl is the drop when benchmark L2A is replaced by archive L2A of the same scenes (927 available). All readouts are trained on clean benchmark features and stay fixed. Values average seeds 42–44 (Prithvi 2.0: 42–43). For the supervised ViT each seed is a separately trained network; its independent linear and MLP probes are trained on its clean features with the same recipe and clean-validation selection as for the foundation models. Mean rows weight encoders equally; the 927-scene rows restrict the L1C comparison to the control's scenes. Ctrl is always computed on those 927 scenes, so it is the same in the all-scene and 927-scene rows. Drops are computed before rounding.

	Original probe				Independent linear probe				MLP probe			
Encoder	L2A	L1C	Drop	Ctrl	L2A	L1C	Drop	Ctrl	L2A	L1C	Drop	Ctrl
DINOv3-Web	64.0	62.8	1.3	0.7	66.2	64.3	1.8	0.4	68.0	65.8	2.2	0.4
DINOv3-Sat	63.5	54.9	8.6	1.1	67.4	53.8	13.6	1.1	68.3	59.0	9.3	0.8
DOFA	58.0	43.7	14.3	0.6	60.4	49.5	10.9	0.2	62.3	46.6	15.7	0.8
Galileo	62.8	41.5	21.3	1.0	62.8	46.6	16.2	0.4	65.7	37.2	28.5	0.7
Panopticon	66.5	39.8	26.7	0.9	67.8	35.8	32.0	1.1	68.3	37.8	30.5	0.7
Prithvi 2.0	56.0	43.0	13.0	0.2	62.9	50.4	12.5	0.5	65.3	31.6	33.7	1.4
RAMEN	52.9	21.2	31.7	3.7	56.7	26.7	30.0	2.1	61.0	24.3	36.8	3.2
ViT-B/16 (supervised)	64.9	23.9	41.0	1.2	65.0	26.9	38.1	1.0	65.9	21.9	43.9	1.4
Mean, foundation models	60.5	43.8	16.7	1.2	63.5	46.7	16.7	0.8	65.6	43.2	22.4	1.1
Mean, all eight	61.1	41.3	19.7	1.2	63.6	44.2	19.4	0.9	65.6	40.5	25.1	1.2
Mean, foundation models, 927 scenes	60.4	43.7	16.7	1.2	63.4	46.6	16.8	0.8	65.5	43.1	22.4	1.1
Mean, all eight, 927 scenes	61.0	41.2	19.8	1.2	63.6	44.2	19.5	0.9	65.6	40.5	25.1	1.2

all three and for every encoder and seed. A stream arrives in consecutive batches of 64 images, 15 full batches followed by one of 40. Seeds 42–44 index three linear readouts trained independently on the same frozen features; for the supervised ViT, each seed is a separately trained network. The degraded inputs and the stream orders are identical across seeds.

Adaptation protocol. Every method adapts a frozen encoder with its original clean-trained linear readout. It receives the batches of a stream in order and keeps its state across them: parameters, optimizer, memories, and running statistics. At the start of every stream it returns to the unadapted model, and to its source preparation where it has one. It never receives labels, the condition, its severity, or the stream type. Each method keeps the prediction timing of its authors' code:

- **Predict, then update:** Tent-Norm, SAR, EATA-Norm, DeYO, POEM and GOLD return the output of the forward pass that computes their update, so an update first affects the next batch. FOA-shift applies a shift estimated from earlier batches.
- **Adapt to the batch, then predict:** PEA and FOA estimate their correction on the current batch, T3A and TAST add it to their memory, and LAME and TCA operate on the batch alone.

Every method executes its authors' update functions, extracted from the official repositories at pinned commits; only the interfaces to the model and the data are replaced. The exceptions are TCA's alignment fit and LAME's affinity, which we re-implemented and verified against the official code. Hyperparameters come from the released configurations, and none was tuned on EO data or on target labels.

Ports to frozen EO encoders. All eight encoders are Vision Transformers with LayerNorm and no two-dimensional BatchNorm. The released code of Tent and EATA adapts BatchNorm; Tent-Norm and EATA-Norm adapt the affine parameters of every normalization layer instead. SAR, DeYO and POEM use SAR's released parameter rule, which excludes the LayerNorms of blocks 9–11 by name. On 12-block encoders these are the last three blocks; on DOFA (24 blocks), Prithvi-EO-2.0 (32) and DINOv3 (40) they are middle blocks. Table 9 summarizes each port.

Table 9: **Ports of the adaptation methods.** "Norm." denotes the affine parameters of normalization layers; "SAR rule" excludes the LayerNorms of blocks 9–11. $C = 10$ classes. Source data are the probe's own clean splits; no degraded or paired images enter any preparation.

Method	Adapts	Source data	Main settings
Tent-Norm	all norm.	none	Adam, lr 10^{-3}; batch entropy; 1 step
SAR	norm., SAR rule	none	SAM on SGD, lr 10^{-3}; $E_0 = 0.4 \ln C$; weight reset when the entropy average falls below 0.2
DeYO	norm., SAR rule	none	SGD, lr 2.5×10^{-4}; margins $0.5 \ln C$, $0.4 \ln C$; PLPD threshold 0.3; 4×4 patch shuffle
EATA-Norm	all norm.	2,000 clean train images (Fisher)	SGD, lr 2.5×10^{-4}; $E_0 = 0.4 \ln C$; redundancy threshold 0.05; Fisher weight 2,000
POEM	norm., SAR rule	1,000 clean validation images (entropy distribution)	SGD, lr 10^{-3}; $E_0 = 0.4 \ln C$; protector $\gamma = 1/(8\sqrt{3})$, clip 1.8; first batch not adapted
FOA	3 input prompts; output shift	2,000 clean train images (per-block mean, std)	CMA-ES, population $27 + 1$; $\lambda = 0.4$; shift average 0.9
FOA-shift	output shift	2,000 clean train features (mean)	shift average 0.9
PEA	none (per-block token alignment)	2,000 clean train images (per-block moments)	released ImageNet-C settings; 3 views
T3A	class prototypes	none	$K = 20$
TAST	projection ensemble	none	$K = 20$; 10 members; $k = 4$; $\tau = 10$; Adam, lr 10^{-3}
LAME	output probabilities	none	RBF graph, 5 nearest neighbours; 100 iterations
TCA	linear feature map	none	$K = 20$; uniform class prior; 20 Adam steps, lr 10^{-3}
GOLD	all norm.; low-rank adapter; projector	2,000 labelled clean train images (class prototypes; 781-step warm-up without labels)	released configuration; Adam, lr 0.01; teacher average 0.999

Deviations from the released code.

- **Class-count thresholds.** Entropy thresholds that the authors define relative to $\ln 1000$ are rescaled to $\ln 10$ (SAR, DeYO, EATA, POEM). Absolute thresholds are kept as released: SAR's weight-recovery threshold (0.2), EATA's redundancy threshold (0.05), and TCA's $K = 20$. At ten classes, these leave three methods nearly inactive:
  - SAR restores its initial weights after 89–100% of batches (58–73% on DINOv3).
  - EATA's redundancy filter admits updates almost only on the first batch of a stream (5–6% of images).
  - TCA's pseudo-source is the whole batch in 77–100% of batches, and there its alignment is the identity.
- **Learning rates.** DeYO and EATA-Norm use the released ResNet-50 learning rate (2.5×10^{-4}); the authors' ViT setting is 10^{-3}.
- **TCA.** It aligns each batch rather than the whole target set, with a uniform class prior where the released runner counts classes from target labels. The fit runs in double precision.
- **POEM.** It uses temperature 1.0 rather than the 1.025 calibrated for ImageNet ViTs. Its entropy distribution comes from the clean validation split, which gives it 1,000 clean images beyond the 2,000 used by the other source-prepared methods.
- **FOA.** Prompts follow the positional embedding of each encoder's token sequence, and its per-block statistics use the tokens that the encoder's readout pools.
- **PEA.** It runs through hooks on every encoder block and aligns all tokens. Its source statistics use 2,000 images rather than up to 10,000.
- **GOLD.** Its augmentation is a band-shared rotation or flip rather than CoTTA's RGB transforms. Its warm-up changes the predictor before any test data arrive, and its reported gains include this change.
- **Stochastic layers.** The released configuration functions set the model to training mode; we return dropout and stochastic depth to evaluation mode. Only Galileo has active stochastic depth.

Encoder-specific exceptions.

- **Galileo.** Its readout contains a BatchNorm layer. Tent-Norm, SAR, EATA-Norm, DeYO, POEM and GOLD run it with the statistics of the current batch, as their released code does with BatchNorm, and this layer accounts for most of their changes on Galileo. PEA, FOA and the feature-level methods keep its stored statistics.
- **DINOv3.** These encoders see the 18 RGB conditions, so their mixtures are balanced over 18 conditions. Image-level adaptation of the 7B-parameter models runs on 80 GB GPUs with the same batch of 64.
- **Supervised ViT.** Its readout was trained jointly with the encoder, and its positional embedding is zero in every checkpoint. Its unadapted predictions are validated against the stored features rather than by an independent replay, and its P3 and C2 conditions perturb the bands given in Appendix B.

Coverage and exclusions. Eleven of the thirteen methods cover all eight encoders with three seeds each. FOA covers six: its DINOv3 runs were incomplete when the analysis was frozen, and DINOv3-Web finished only afterwards. GOLD's DINOv3 runs are excluded from every statistic: the runner reused a cached optimizer state that one stream could modify for the next, so these runs are flagged for review. Nine further methods from our initial list were not run: CoTTA, TEA, MEMO, TTT, AdaContrast, SoTTA and RMT for compute, and NOTE and RoTTA because they rely on BatchNorm layers that these encoders lack.

Fidelity checks.

- **Unadapted predictor.** Before adapting a stream, an image-level run recomputes the unadapted network's logits on the exact batches and requires agreement with an independent native predictor (relative and absolute tolerance 10^{-4}, identical decisions). Every stream so checked passes. Reused seed-42 runs for DOFA and Panopticon and all supervised-ViT runs are instead checked against the stored features. For the feature-level methods, the unadapted logits are the stored ones by construction.
- **Inputs.** Regenerated degraded images match the SHA-256 checksums of the stored inputs (not recorded for the supervised ViT).
- **Code equivalence.** Unit tests confirm that SAR's update step, POEM, FOA's modified components, FOA-shift, LAME's affinity and TCA's support selection reproduce the official code exactly, and that EATA's Fisher information equals the accumulated squared gradients.
- **Scoring.** Recomputing all 11,862 per-stream gains from the stored per-image predictions reproduces the reported gains (largest difference 7×10^{-15} points).

Decision accounting. For every image of a stream, we compare the method's prediction $\hat{y}$ with the prediction $\hat{y}_0$ of the unadapted model on the same input. A *fix* corrects an error of the unadapted model ($\hat{y}_0 \neq y = \hat{y}$), an *overturn* breaks a correct prediction ($\hat{y}_0 = y \neq \hat{y}$), and a *change* is any $\hat{y} \neq \hat{y}_0$, including changes from one wrong class to another. The accuracy change of a stream equals the number of fixes minus the number of overturns, divided by the number of images. Figure 1b pools fixes and overturns over the H streams and seeds of each encoder, expresses them per 100 predictions, and averages encoders with equal weight. The share of the loss recovered (Figure 1a) is a method's mean accuracy change on H divided by the encoder's mean loss on H, the clean minus the unadapted degraded accuracy. The loss and the two references use the seed-42 readout; method gains average the three seeds.

Reproduction on the original benchmarks. We also ran each method's official code on a benchmark of its original paper and compared the result with the published value (Table 10). All reproductions use a single seed and a current PyTorch runtime. We changed the official code only for compatibility (device placement, imports, disabled downloads), with three exceptions:

- T3A, TAST and TCA use fixed hyperparameters ($K = 20$) instead of the papers' searches.

- TCA's official runner estimates class counts from target labels; we replaced this with a uniform prior.
- TCA has no released checkpoint, so we retrained its source ResNet-18 with the official training script.

For the ten methods with a published result, our adapted accuracy is within 1.0 point of it. Two unadapted models differ more:

- SAR's ViT-B/16 is 1.7 points more accurate than the published one.
- TCA's retrained ResNet-18 is 1.9 points more accurate, so TCA's gain is 1.9 points smaller than published.

LAME's paper reports only averages over several scenarios, so we checked only that the official code runs as specified. POEM and FOA-shift were not run on their original benchmarks. For both, unit tests confirm that our adapters reproduce the official implementations bit for bit on identical inputs. For POEM, this covers outputs and parameters over four batches.

Table 10: **Reproduction of each method on a benchmark of its original paper.** Top-1 accuracy (%) of the unadapted model → the adapted model, averaged over the benchmark's corruptions or domains (severity 5 for the corruption benchmarks); one seed. Published values are taken from the cited table. Where the paper does not print an average, we compute it from its per-corruption values.

Method	Benchmark and model	Published	Ours	Published value from
Tent	CIFAR-10-C, WRN-28-10, reset per corruption	56.5 → 81.4	56.5 → 81.4	authors' code; CoTTA, Tab. 2
SAR	ImageNet-C, mixture of 15 corruptions, ViT-B/16	29.9 → 57.1	31.6 → 57.7	SAR, Tab. 3
DeYO	ImageNet-C, ResNet-50, mild scenario	18.0 → 48.6	18.2 → 48.7	DeYO, Tab. 1
EATA	ImageNet-C, ResNet-50	18.0 → 47.8	18.2 → 47.9	EATA, Tab. 2 (mean)
T3A	PACS, ERM ResNet-50, 4 target domains	83.9 → 84.5	84.3 → 84.9	T3A, Tab. 2
TAST	PACS, ERM ResNet-50, 4 target domains	83.2 → 84.1	83.7 → 85.0	TAST, Tab. 1
LAME	ImageNet-V2, class-grouped episodes, ResNet-50	–	63.1 → 80.8	no single-setting value
TCA	CIFAR-10-C, ResNet-18 (retrained), LinearTCA	50.8 → 61.0	52.7 → 61.0	TCA, Tab. 1
FOA	ImageNet-C, ViT-B/16	55.5 → 66.3	55.5 → 66.0	FOA, Tab. 2
PEA	CIFAR-10-C, continual, ResNet-50, with augmentation	62.5 → 83.4	62.5 → 83.4	PEA, Tab. 2
GOLD	CIFAR-10-C, continual, WRN-28-10	56.5 → 85.9	56.5 → 84.9	GOLD, Tab. 1; source: CoTTA, Tab. 2
POEM	not run on its benchmark; adapter identical to the official code (unit test)			
FOA-shift	not run on its benchmark; identical to the official activation-shift code (unit test)			

G TTA DIAGNOSTICS

Encoders. Figure 1b,c and the tables below use the five encoders on which every method ran: DOFA, Panopticon, Prithvi-EO-2.0, RAMEN, and the supervised ViT, with three seeds each. Not every method ran on the two DINOv3 encoders. We also exclude Galileo: its readout contains a BatchNorm layer, and the gains of Tent-Norm, SAR, EATA-Norm, DeYO, and POEM on it come from that layer normalizing with batch statistics, not from their updates.

Decision accounting. A change in accuracy hides how many decisions a method changes. We count every prediction that differs from the frozen model's as a *fix* (a frozen error becomes correct) or an *overturn* (a correct frozen prediction becomes wrong); predictions that change from one error to another are neither. Table 11 reports both per 1,000 images on the clean stream and on the homogeneous streams. The existing methods that gain under shift also change clean predictions: PEA fixes 10.2 and overturns 18.0 clean predictions per 1,000, and FOA-shift 5.5 and 7.4. The methods that leave clean predictions in place (POEM, EATA-Norm, SAR, TCA) barely act under shift either: they fix or overturn at most 2.9 of 1,000 decisions on homogeneous streams, for a net of at most $+0.01$ points. Most existing methods overturn more correct predictions than they fix. CALIPER fixes 2 and overturns 3 of 15,000 clean predictions and has the largest net gain on homogeneous streams ($+2.09$), with 29.8 fixes against 8.9 overturns per 1,000.

Table 11: **Decision accounting on m-EuroSAT.** Predictions fixed and overturned per 1,000 images relative to the frozen model, and net accuracy change in points, mean over the five encoders of Figure 1b,c and three seeds. Methods are ordered by their net gain on homogeneous streams. †: uses labelled source data.

	Clean stream			Homogeneous streams		
Method	Fixed	Overturned	Net	Fixed	Overturned	Net
FOA-shift	5.5	7.4	−0.19	26.7	14.9	+1.18
PEA	10.2	18.0	−0.78	40.2	30.8	+0.94
FOA	8.3	17.9	−0.97	28.1	26.3	+0.18
POEM	0.3	0.1	+0.01	1.3	1.2	+0.01
EATA-Norm	0.3	0.1	+0.01	0.4	0.3	+0.01
SAR	0.0	0.0	+0.00	0.8	1.0	−0.02
TCA	0.0	0.0	+0.00	0.4	2.5	−0.21
GOLD†	16.4	41.8	−2.54	29.6	46.4	−1.68
LAME	6.3	21.3	−1.50	9.1	62.0	−5.29
DeYO	5.1	65.2	−6.01	8.1	78.8	−7.08
Tent-Norm	8.8	108.1	−9.93	14.1	112.6	−9.86
T3A	12.7	181.5	−16.88	24.4	174.1	−14.97
TAST	13.0	214.3	−20.13	24.7	205.4	−18.07
CALIPER	0.1	0.2	−0.01	29.8	8.9	+2.09

Per encoder. Table 12 gives the net gain of every method on each of the five encoders. CALIPER gains more on homogeneous streams than every existing label-free method on DOFA, Panopticon, Prithvi-EO-2.0, and RAMEN; on the supervised ViT, PEA gains more (+1.30 versus +0.91). On clean streams CALIPER changes accuracy by at most 0.03 points on every encoder.

Table 12: **Net gain per encoder on m-EuroSAT.** Accuracy change in points relative to the frozen model on homogeneous (H) and clean (C) streams, mean over three seeds. †: uses labelled source data.

	DOFA		Panopticon		Prithvi		RAMEN		ViT	
Method	H	C	H	C	H	C	H	C	H	C
FOA-shift	+1.50	+0.07	+0.33	+0.07	+1.69	−0.40	+1.74	−0.47	+0.63	−0.20
PEA	+1.09	−0.50	−0.33	−0.53	+0.47	−1.97	+2.16	−0.47	+1.30	−0.43
FOA	+1.17	−0.17	+0.24	−0.13	−2.57	−3.33	+1.65	−0.83	+0.40	−0.37
POEM	+0.01	+0.07	+0.00	+0.00	−0.05	+0.03	+0.13	−0.03	+0.00	+0.00
EATA-Norm	+0.01	+0.03	+0.00	+0.00	−0.01	+0.07	+0.03	−0.03	+0.00	+0.00
SAR	+0.04	+0.00	+0.00	+0.00	−0.14	+0.00	+0.00	+0.00	+0.00	+0.00
TCA	−0.25	+0.00	−0.06	+0.00	−0.70	+0.00	−0.03	+0.00	+0.00	+0.00
GOLD†	+1.64	+0.40	+1.22	+0.63	−6.31	−7.20	−5.12	−6.17	+0.17	−0.37
LAME	−4.64	−0.67	−11.02	−1.93	−5.81	−2.03	−3.59	−2.03	−1.38	−0.83
DeYO	−0.03	−0.03	+0.00	+0.00	−1.18	−0.23	−34.18	−29.77	+0.00	+0.00
Tent-Norm	−1.36	−0.63	+0.35	+0.13	−22.63	−20.70	−25.68	−28.40	+0.03	−0.07
T3A	−5.66	−7.07	−3.54	−4.53	−22.19	−25.50	−33.61	−37.27	−9.84	−10.03
TAST	−8.35	−9.17	−6.36	−7.30	−24.05	−28.60	−40.69	−44.90	−10.92	−10.70
CALIPER	+2.05	+0.00	+1.15	+0.00	+3.21	+0.00	+3.13	+0.00	+0.91	−0.03

Batch statistics. Methods that adapt to batch statistics read any change in the batch as a shift, including changes of class composition on clean images. On clean streams whose labels arrive in correlated runs (Dirichlet $\alpha = 0.1$, five to six classes per batch), PEA loses 2.0 to 37.3 points (six encoders); CALIPER changes accuracy by at most 0.1 points (eight encoders). With a fourfold prior on a random third of the classes, PEA loses 4.2 to 17.8 points and CALIPER at most 0.11 (six encoders). Smaller batches make batch statistics noisier: on the five encoders evaluated at batch size 16 (seed 42), PEA's clean-stream loss grows from 0.64 to 3.30 points on average, from 0.2 to 1.8 (Panopticon) and from 1.7 to 9.9 (Prithvi-EO-2.0), whereas CALIPER's mean gains change by at most 0.1 points between batch sizes of 16 and 64, because it estimates its corrections from the stream memory rather than the current batch.

H CALIPER: PROPOSITIONS AND PROOFS

Proposition 1 (Class-frequency invariance of the residual mean)**.** *Let $\hat{\mathbf{y}}$ be the readout's decision for a whitened embedding $\mathbf{u}$, and $\mathbf{r} = \mathbf{u} - B^\top[1; \hat{\mathbf{y}}]$. Suppose the target differs from the source only in how often each decision occurs, $p_t(\mathbf{u} \mid \hat{\mathbf{y}}) = p_s(\mathbf{u} \mid \hat{\mathbf{y}})$ for every $\hat{\mathbf{y}}$, and suppose $B^\top[1; \hat{\mathbf{y}}] = \mathbb{E}_s[\mathbf{u} \mid \hat{\mathbf{y}}]$. Then $\mathbb{E}_t[\mathbf{r}] = \mathbf{0}$ for every decision distribution $p_t(\hat{\mathbf{y}})$.*

Proof. By the law of total expectation and the first condition, $\mathbb{E}_t[\mathbf{r}] = \sum_{\hat{\mathbf{y}}} p_t(\hat{\mathbf{y}})\, \mathbb{E}_s[\mathbf{r} \mid \hat{\mathbf{y}}]$. By the second condition, $\mathbb{E}_s[\mathbf{r} \mid \hat{\mathbf{y}}] = \mathbb{E}_s[\mathbf{u} \mid \hat{\mathbf{y}}] - B^\top[1; \hat{\mathbf{y}}] = \mathbf{0}$ for every $\hat{\mathbf{y}}$, so every term vanishes whatever $p_t(\hat{\mathbf{y}})$ is. □

When the conditions hold. Under a change of class frequencies alone ($p_t(\mathbf{x} \mid y) = p_s(\mathbf{x} \mid y)$), the first condition holds exactly when the readout's decisions are correct, because $\hat{\mathbf{y}}$ then determines y, and approximately when $p(y \mid \hat{\mathbf{y}})$ changes little. For single-label readouts $\hat{\mathbf{y}}$ is one-hot, so the least-squares regression of $\mathbf{u}$ on $[1; \hat{\mathbf{y}}]$ has one free mean per decision and $B^\top[1; \hat{\mathbf{y}}]$ is the source mean of $\mathbf{u}$ among images with that decision; the second condition always holds. For multi-label readouts the regression is additive in the labels, and the second condition holds exactly only if the decision-conditional means are additive. The result concerns the mean only: the residual covariance, which the mixture directions use, is a re-weighted mixture of per-decision covariances and changes with the class frequencies; Section 5.3 measures the effect.

Finite samples. CALIPER estimates B on part of the source set, so out-of-sample clean residuals have a small nonzero mean $\boldsymbol{\mu}_0$. Cross-fitting carries it into the null expectation $\mathbb{E}_0[S \mid n] = \|\boldsymbol{\mu}_0\|^2 + \operatorname{tr}\Sigma_0/n$ for $S = \|\bar{\mathbf{r}}\|^2$, where Σ_0 is the covariance of the clean residuals (Appendix I).

Proposition 2 (James–Stein-type shrinkage of the translation)**.** *Let $\bar{\mathbf{r}} \sim \mathcal{N}(\boldsymbol{\theta}, \tau^2 I_p)$ with $\tau^2 = \sigma^2/n$ and $p \geq 3$, and let $c = \|\boldsymbol{\mu}_0\|^2 + p\tau^2$ be the clean-null expectation of $S = \|\bar{\mathbf{r}}\|^2$. If $\|\boldsymbol{\mu}_0\|^2 \leq (p-4)\tau^2$, the estimate $\hat{\boldsymbol{\theta}} = g(S; n)\, \bar{\mathbf{r}} = (1 - c/S)_+\, \bar{\mathbf{r}}$ satisfies $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \leq \mathbb{E}\|\bar{\mathbf{r}} - \boldsymbol{\theta}\|^2 = p\tau^2$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$, with strict inequality when $\|\boldsymbol{\mu}_0\|^2 < (p-4)\tau^2$.*

Proof. Write $\mathbf{x} = \bar{\mathbf{r}}$. *Without truncation.* For $\boldsymbol{\delta}(\mathbf{x}) = (1 - c/\|\mathbf{x}\|^2)\, \mathbf{x}$, Stein's identity $\mathbb{E}[(\mathbf{x} - \boldsymbol{\theta})^\top \mathbf{h}(\mathbf{x})] = \tau^2\, \mathbb{E}[\nabla \cdot \mathbf{h}(\mathbf{x})]$ with $\mathbf{h}(\mathbf{x}) = c\,\mathbf{x}/\|\mathbf{x}\|^2$ and $\nabla \cdot \mathbf{h} = c\,(p-2)/\|\mathbf{x}\|^2$, which is integrable for $p \geq 3$, gives

$$\mathbb{E}\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = p\tau^2 - c\,\big(2(p-2)\tau^2 - c\big)\,\mathbb{E}\big[\|\mathbf{x}\|^{-2}\big].$$

The second term is non-negative when $0 < c \leq 2(p-2)\tau^2$, that is, when $\|\boldsymbol{\mu}_0\|^2 \leq (p-4)\tau^2$, and positive when the inequality is strict (James & Stein, 1961; Baranchik, 1970). *Positive part.* Let $a(\mathbf{x}) = 1 - c/\|\mathbf{x}\|^2$. Where $a \geq 0$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\delta}$. Where $a < 0$, $\hat{\boldsymbol{\theta}} = \mathbf{0}$ and $\|\boldsymbol{\theta}\|^2 - \|a\mathbf{x} - \boldsymbol{\theta}\|^2 = -a^2\|\mathbf{x}\|^2 + 2a\,\boldsymbol{\theta}^\top\mathbf{x}$. Conditionally on $\|\mathbf{x}\|^2 = t < c$, a is a negative constant, and $\mathbb{E}[\boldsymbol{\theta}^\top\mathbf{x} \mid \|\mathbf{x}\|^2 = t] \geq 0$: on that sphere the density of $\mathbf{x}$ is proportional to $\exp(\boldsymbol{\theta}^\top\mathbf{x}/\tau^2)$, and pairing $\mathbf{x}$ with $-\mathbf{x}$ gives $\boldsymbol{\theta}^\top\mathbf{x}\,(e^{\boldsymbol{\theta}^\top\mathbf{x}/\tau^2} - e^{-\boldsymbol{\theta}^\top\mathbf{x}/\tau^2}) \geq 0$. The difference therefore has non-positive expectation, and $\mathbb{E}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \leq \mathbb{E}\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2$ (Baranchik et al., 1964; Lehmann & Casella, 1998). □

The constant. James and Stein's constant is $(p-2)\tau^2$. CALIPER uses the clean-null expectation of S instead, $p\tau^2$ plus the clean offset, so that a statistic at its clean level gives $g = 0$. This constant lies in the dominance range $(0, 2(p-2)\tau^2]$ under the stated condition.

Non-isotropic residuals. For $\bar{\mathbf{r}} \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma_0/n)$, the same Stein argument bounds the risk of the untruncated estimate by that of $\bar{\mathbf{r}}$ whenever $0 < c \leq 2\big(\operatorname{tr}\Sigma_0 - 2\lambda_{\max}(\Sigma_0)\big)/n$ (Bock, 1975). With CALIPER's constant $c = \|\boldsymbol{\mu}_0\|^2 + \operatorname{tr}\Sigma_0/n$, this reads $\operatorname{tr}\Sigma_0 \geq 4\lambda_{\max}(\Sigma_0) + n\|\boldsymbol{\mu}_0\|^2$: the effective dimension $\operatorname{tr}\Sigma_0/\lambda_{\max}(\Sigma_0)$ must exceed four plus the scaled clean offset. For the cross-fitted clean residuals of our encoders the effective dimension is 64 to 342, and the condition holds up to a memory of 1,350 images on every encoder and dataset, above the largest memory CALIPER uses (1,000). The positive-part step of the proof uses spherical symmetry, so for non-isotropic residuals we claim the risk bound for the untruncated factor only.

Scope. Both results concern the stream-level estimate of the mean shift. CALIPER distributes that estimate over images by empirical Bayes, and neither proposition says anything about classification accuracy.

I CALIPER: CLEAN NULLS

Source splits. All nulls use only the 2,000 unlabeled clean source images. The whitening and the decision means B are fitted on an 80% reference split. The remaining 20% (400 images) provide the clean distributions of the per-image posteriors: projections of whitened embeddings for the unconditional translation, and projections of residuals for the class-conditional translation and the mixture directions. The residual nulls use 5-fold cross-fitting over all 2,000 images: the whitening and the regression are refitted on four folds and residuals are taken on the fifth, which gives 2,000 out-of-sample clean residuals. The affine map uses a separate half split: one half defines its principal components and the transport target, and the other half calibrates its clean bias and null.

Translation. The null is $\mathbb{E}_0[\|\bar{\mathbf{r}}\|^2 \mid n] = \|\boldsymbol{\mu}_0\|^2 + \operatorname{tr}\Sigma_0/n$, the exact expectation for n independent clean residuals, computed from the cross-fitted residuals. It is tabulated at $n \in \{16, 32, 64, 128, 256, 512, 1000\}$ and interpolated in $\log n$.

Mixture directions. $\mathbb{E}_0[\lambda_j \mid n]$ is the mean of the j-th eigenvalue of the residual covariance, restricted to the top 256 principal components, over 40 i.i.d. draws of n cross-fitted residuals. It is tabulated at every memory size a stream reaches, all multiples of 16 up to 1,000, without interpolation, because eigenvalues near the rank edge do not vary smoothly with n. A draw of n images from a finite pool repeats some of them, so its covariance has rank below $n-1$ and $\mathbb{E}_0[\lambda_j \mid n]$ is zero beyond that rank. Directions are tested in order of eigenvalue, each with factor $(1 - \mathbb{E}_0[\lambda_j \mid n]/\lambda_j)_+$; the step-down stops at the first zero factor and never tests a direction beyond the smallest rank among the draws.

Affine map. For 48 i.i.d. draws of n held-out clean embeddings, CALIPER fits the same map as at test time. The mean of these maps, $\bar{W}(n)$, is the clean bias. The statistic is the source-variance-weighted Frobenius norm of $\hat{W} - \bar{W}(n)$ over the 64 principal components, and its null expectation is the mean of that deviation over the draws. At test time, n is the effective size $n_{\text{eff}} = (\sum_i q_i)^2 / \sum_i q_i^2$ of the weighted memory, the map is fitted only when $n_{\text{eff}} \geq 16$, and the map applied is the debiased deviation $\hat{W} - \bar{W}(n_{\text{eff}})$, scaled by its factor.

Exchangeability. Under exchangeable arrival, Pearson's χ^2 for a table with fixed margins has expectation $\mathrm{df} \cdot N/(N-1) \approx \mathrm{df}$, where N is the number of entries, so the exchangeability weight $(1 - \mathrm{df}/\chi^2)_+$ has the same form as g.

Why i.i.d. draws. The null must describe a stream of independent clean images. Drawing n images without replacement from the pool of 2,000 understates the sampling noise by the finite-population factor $(N-n)/(N-1)$, about one half at $n = 1{,}000$, and would make clean streams look shifted. Drawing i.i.d. from the pool matches independent sampling, up to the rank effect handled above.

J CALIPER: SETTINGS

All settings are fixed across encoders, datasets, and streams; Appendix L varies the four with the largest possible effect.

K FULL ABLATION

Table 14 removes or disables one component of CALIPER's final design at a time; no variant adds a step. Every variant runs on the same streams, seed, and cached features as CALIPER.

- **No clean null**: every gate is forced open.

Table 13: **Every numerical setting of CALIPER.**

Setting	Value	Role
Whitening shrinkage	0.1	Eigenvalues $\lambda \mapsto 0.9\lambda + 0.1\bar{\lambda}$
Principal components, affine map	64	Subspace in which the map is fitted
Principal components, mixture directions	256	Subspace searched for directions
Identity-prior strength	$n_0 = 1{,}024$	Ridge penalty toward the identity map
Transport target	256 k-means centroids	Summary of the clean source half, with cluster masses
Entropic regularization	$0.02\times$ median cost	Sinkhorn transport; two alternations of transport and map fitting
Empirical Bayes	80-point grid plus 0	NPMLE of the shift distribution; Gaussian kernel with Silverman bandwidth h for the clean density
Source splits	80/20; 5 folds; 50/50	Whitening and decision means vs. clean priors; cross-fitted residual nulls; affine map
Null draws	40 (eigenvalues); 48 (map)	i.i.d. draws per memory size
Null table sizes	16 to 1,000	Every multiple of 16 for eigenvalues; interpolated in $\log n$ elsewhere
Smallest memory	16	Below it, a step is skipped
Memory limit	1,000 embeddings	Binds only on the standard ImageNet-C streams

- **No class conditioning**: the gate is computed from the raw stream mean, and the translation uses the unconditional estimate (no class-conditional residual and no exchangeability test).
- **Global, not per-image**: one translation amount for all images instead of the per-image empirical-Bayes posteriors.
- **Translation only**: no affine step and no mixture directions.
- **Always unconditional** and **Always conditional**: the exchangeability test is removed, and the translation always uses the unconditional or the class-conditional estimate.
- **No translation**, **No mixture dirs.**, and **No affine step**: that step is removed.
- **Mixture dirs. full weight**: the mixture directions act at full strength instead of only on what the translation left unexplained (weight $1 - g_1$).

In the released code, the no-clean-null path contained a bug that left the class-conditional residual undefined; this variant was run from a copy with that bug fixed, which does not change any other variant.

Table 14: **Full ablation.** Gain in points over the frozen predictor, seed 42, mean over the eight encoders ($\times 4$ class prior: six encoders). Correlated labels: Dirichlet $\alpha = 0.1$ arrival order on the m-EuroSAT test streams. Real m-BigEarthNet: Level-1C inputs (L1C) and clean benchmark Level-2A inputs (L2A).

	m-EuroSAT, shuffled				Prior	Correlated labels			Real m-BigEarthNet	
Variant	H	D	M	C	$\times 4$ C	D	M	C	L1C	L2A
CALIPER	+4.56	+1.71	+0.56	−0.03	−0.05	−0.16	−0.57	+0.00	+3.70	−0.04
No clean null	+4.18	+1.15	−0.28	−0.67	−7.01	−2.30	−3.45	−1.93	+0.85	−4.31
No class conditioning	+4.56	+1.76	+0.53	−0.01	+0.01	−0.99	−1.00	−0.05	+3.96	−0.04
Global, not per-image	+2.42	+0.96	+0.41	−0.01	−0.05	−0.45	−0.82	+0.00	+0.13	−0.09
Translation only	+3.32	+0.76	+0.27	+0.01	−0.06	−0.63	−0.80	−0.03	+2.84	−0.04
Always unconditional	+4.56	+1.71	+0.57	−0.03	−0.06	−1.56	−1.42	−0.04	+3.99	−0.04
Always conditional	+2.50	+0.78	+0.15	−0.03	+0.00	+0.43	+0.05	+0.00	+0.93	−0.06
No translation	−0.16	+0.74	+0.18	−0.01	+0.01	+0.65	+0.08	+0.01	−1.90	−0.04
No mixture dirs.	+4.56	+1.58	+0.47	+0.00	−0.08	−0.34	−0.74	−0.03	+3.70	−0.03
No affine step	+3.32	+0.89	+0.40	−0.01	−0.03	−0.41	−0.66	+0.01	+2.84	−0.04
Mixture dirs. full weight	+2.88	+1.64	+0.56	−0.03	−0.04	+0.38	−0.42	+0.00	+1.07	−0.05

- **Clean null.** Without it, CALIPER loses 0.67 points on clean test streams, 7.01 on clean streams with a fourfold class prior (4.0 to 13.4 per encoder), 1.93 on clean streams with

correlated labels, and 4.31 on clean Level-2A inputs, and its real Level-1C gain falls from 3.70 to 0.85.

- **Translation.** Without it, the homogeneous-stream gain disappears (-0.16) and the real Level-1C gain becomes negative (-1.90). Under correlated labels, however, the variant without translation loses less than CALIPER ($+0.65$ and $+0.08$ on heterogeneous and mixed streams, versus -0.16 and -0.57): the translation is where correlated arrival costs accuracy.
- **Per-image amounts.** A single correction for all images halves the homogeneous-stream gain (2.42 versus 4.56) and removes almost all of the real Level-1C gain (0.13 versus 3.70).
- **Class conditioning.** Under a fourfold class prior, clean accuracy stays within 0.06 points with or without it, so it does not matter for class-frequency changes on clean streams; the clean null does. Without it, CALIPER loses 0.83 points more on correlated heterogeneous streams and gains 0.26 points more on real Level-1C inputs. Always using the unconditional estimate matches CALIPER on exchangeable streams but loses 1.40 and 0.85 points more on correlated heterogeneous and mixed streams. Always using the class-conditional estimate protects correlated streams but costs 2.06 points on exchangeable homogeneous streams. The exchangeability test chooses between the two.
- **Affine step and mixture directions.** Removing the affine step lowers the gain on homogeneous streams from 4.56 to 3.32 points and on real Level-1C inputs from 3.70 to 2.84. Removing the mixture directions leaves homogeneous streams and real inputs unchanged and lowers the heterogeneous- and mixed-stream gains by 0.13 and 0.09 points, within seed variation. Letting them act at full strength, rather than only on what the translation left unexplained, lowers the homogeneous-stream gain to 2.88 and the real Level-1C gain to 1.07.

Across seeds 42–44, CALIPER's mean gains vary by at most 0.31 points on shuffled streams, 0.76 under correlated labels, and 0.14 on real Level-1C inputs; smaller differences in Table 14 lie within seed variation.

L SENSITIVITY

Table 15 changes one setting of CALIPER at a time. The rank of the affine map ($k = 64$), the dimension of the subspace searched for mixture directions ($d_k = 256$), and the shrinkage of the whitening covariance (0.1) are halved and doubled. The identity-prior strength of the affine map ($n_0 = 1024$) is divided and multiplied by four. No change moves a mean gain by more than 0.40 points, and none changes clean accuracy by more than 0.03 points.

Table 15: **Sensitivity to CALIPER's settings.** Gain in points over the frozen predictor on the m-EuroSAT test streams, seed 42, mean over the eight encoders; default values in parentheses.

Setting (default)	Value	H	D	M	C
CALIPER		$+4.56$	$+1.71$	$+0.56$	-0.03
Affine-map rank k (64)	32	$+4.26$	$+1.41$	$+0.49$	-0.01
	128	$+4.69$	$+1.59$	$+0.51$	-0.03
Mixture subspace d_k (256)	128	$+4.55$	$+1.71$	$+0.56$	-0.01
	512	$+4.56$	$+1.70$	$+0.56$	-0.01
Whitening shrinkage (0.1)	0.05	$+4.56$	$+1.72$	$+0.54$	-0.03
	0.2	$+4.57$	$+1.61$	$+0.57$	-0.03
Identity prior n_0 (1024)	256	$+4.79$	$+1.62$	$+0.55$	-0.03
	4096	$+4.16$	$+1.60$	$+0.51$	-0.01

M IMAGENET-C AND CIFAR-10-C

Protocols. We run the frozen CALIPER, without retuning, under two protocols. The standard ImageNet-C protocol (Hendrycks & Dietterich, 2019) follows Tent: a torchvision ResNet-50, corruption severity 5, and one stream per corruption that holds the 48,000 validation images outside the

2,000-image source set in a fixed random order, adapted in batches of 64 and reset between corruptions; the clean 48,000 images form a stream without shift. On these long streams CALIPER keeps the last 1,000 embeddings in memory. Our stream protocol uses the streams of Section 3 at severity 5: 10 disjoint homogeneous streams of 1,000 images per corruption (150 streams), 10 heterogeneous streams in which each image carries a random corruption, 10 mixed streams that are half clean, and 10 clean streams, with the memory reset for every stream. It uses a ViT-B/16 on ImageNet-C and a ResNet-50 on CIFAR-10-C. Methods that use source data receive the same 2,000 unlabeled clean images.

Standard ImageNet-C. Methods that re-estimate normalization statistics throughout the network gain 24.5 (Tent) to 30.5 points (DeYO) over the 15 corruptions (Table 16). CALIPER, which sees only final embeddings, gains 3.6 points, from 18.2% to 21.8% accuracy. The normalization-based methods pay for their gains on clean images: on the clean stream they lose 1.0 to 1.4 points and change 9.4% to 11.7% of the predictions, whereas CALIPER changes 0.01% and leaves accuracy unchanged.

Table 16: **Standard ImageNet-C protocol** (ResNet-50, severity 5). Accuracy and gain over the frozen model in points, mean over the 15 corruptions (48,000 images each); clean stream: the 48,000 clean images.

	15 corruptions		Clean stream	
Method	Accuracy	Gain	Gain	Changed (%)
Frozen	18.2	–	–	–
Tent	42.6	+24.46	−1.02	9.57
SAR	43.6	+25.41	−0.99	9.37
EATA	47.9	+29.70	−1.10	10.43
DeYO	48.6	+30.47	−1.41	11.68
POEM	42.7	+24.57	−0.99	9.81
CALIPER	21.8	+3.59	+0.00	0.01

Stream protocol. On ImageNet-C, PEA recovers the most on homogeneous and heterogeneous streams (+6.73 and +1.84 points) but loses 0.61 points on mixed and 2.34 on clean streams (Table 17). CALIPER gains +3.79 and +1.45, more than every other method, has the largest gain on mixed streams (+0.44), and leaves clean accuracy unchanged. On CIFAR-10-C, PEA gains +20.93 on homogeneous streams and CALIPER +6.10; on heterogeneous and mixed streams CALIPER loses 0.41 and 0.08 points, where PEA gains +7.18 and +3.07 (Table 18).

Table 17: **ImageNet-C, stream protocol** (ViT-B/16, severity 5). Gain over the frozen model in points on homogeneous (H), heterogeneous (D), mixed (M), and clean (C) streams of 1,000 images, batches of 64. Every method is evaluated on the same 157 of the 180 streams (131 H, 8 D, 9 M, 9 C); CALIPER's remaining 23 streams were still running.

Method	H	D	M	C
PEA	+6.73	+1.84	−0.61	−2.34
FOA-shift	+3.42	+1.09	+0.20	+0.03
FOA	+3.07	+1.17	−0.02	−0.11
Tent	+0.86	+0.44	−0.09	−0.08
EATA	+0.56	+0.52	−0.01	−0.14
SAR	+0.37	+0.39	−0.17	−0.12
POEM	+0.36	+0.24	−0.06	−0.02
DeYO	+0.03	+0.72	+0.23	−0.24
CALIPER	+3.79	+1.45	+0.44	+0.00

How much the final embedding can recover. Two oracles bound what repairing the final embedding can achieve on these corruptions. Both use the clean version of every corrupted image, and neither uses labels. The best per-corruption translation of the final embedding gains +3.64 points on

Table 18: **CIFAR-10-C, stream protocol** (ResNet-50, severity 5). Gain over the frozen model in points on all 180 streams (150 H, 10 D, 10 M, 10 C) of 1,000 images, batches of 64.

Method	H	D	M	C
PEA	$+20.93$	$+7.18$	$+3.07$	$+0.09$
CALIPER	$+6.10$	-0.41	-0.08	$+0.01$

ImageNet-C and $+4.74$ on CIFAR-10-C over the homogeneous data; an affine map fitted by ridge regression on half of the pairs and applied to the other half gains $+6.15$ and $+15.24$. Without pairs, CALIPER gains more than the translation oracle on both datasets but stays below the affine one. PEA exceeds even the affine oracle ($+6.73$ and $+20.93$), so part of its gain comes from repairing layers before the final embedding, which no method that sees only the final embedding can reproduce. On m-EuroSAT the order is reversed on average: even the paired translation recovers more of the homogeneous-stream loss than PEA (33% versus 29%; Section 3 and Figure 1a).