OBSERVATION AND TRANSPORT LIMITS OF SOURCE-ONLY CONCEPT ERASURE UNDER DOMAIN SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept erasure is often evaluated by whether a chosen probe can recover a nuisance variable from the resulting representation. This criterion does not establish whether a detector trained on source domains will retain its operating point on an unseen domain. Across pathology and natural-image benchmarks, we find that source-fitted linear erasure can remove the observed domain signal while leaving target false-positive rates largely unchanged. We explain this separation through two limitations. First, domain means expose only the content-averaged nuisance response; content-dependent responses can cancel under averaging, and finite-sample grouped means can misleadingly appear to recover them. Second, even perfect knowledge of the source nuisance geometry need not identify the target-optimal erasure subspace without assumptions connecting source and target domains. Target-oracle controls distinguish this transport limitation from limitations of the projector class. Finally, we compare adaptation strategies under two deployment objectives: false-positive-rate calibration and discrimination at matched false-positive rate. Direct threshold adaptation is effective for the first in the studied settings, although small adaptation sets have substantial uncertainty; feature-space adaptation can improve the second by changing the ordering of examples. These results support choosing an adaptation target according to the deployment objective rather than treating source-side erasure as a sufficient criterion for robustness.

1 INTRODUCTION

Consider a detector trained on normal examples from several source domains and deployed on a new hospital or visual style. A common strategy for improving robustness is to remove representation directions associated with domain identity. A probe then measures whether the domain label remains recoverable. Success against a specified probe establishes a form of statistical guardedness, but the deployment question is different: will normal examples from the new domain still receive scores below the detector's threshold?

We observe a separation between these criteria. Source-estimated first-moment erasure can eliminate the fitted domain contrast while leaving the detector's false-positive rate (FPR) on an unseen domain almost unchanged. A nonlinear probe may still recover domain identity, but incomplete statistical erasure is only part of the explanation: the directions removed at the source can also differ from those relevant at the target. Understanding this distinction requires separating what source data reveals from whether that information transfers.

The first limitation concerns *observation*. If a domain change moves different kinds of content in different directions, their responses can cancel when averaged. A method informed by differences between domain means therefore observes only part of the nuisance variation. The second concerns *transport*. Even if source nuisance geometry were known perfectly, it need not determine the geometry at an unobserved domain. We study these limitations separately using controlled nuisance transformations, finite-sample corrections, and target-side oracle projectors.

These limitations concern representation-level erasure, whereas deployment may require a weaker intervention. Controlling the FPR at a single operating point requires estimating a target score quan-

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.

tile, rather than recovering a nuisance subspace. Improving discrimination at that FPR is a different objective: threshold changes preserve score ordering and cannot improve the underlying receiver operating characteristic (ROC). Throughout, *calibration* means operating-point or FPR calibration, not probabilistic calibration of class probabilities.

Our contributions are threefold:

1. **Observation.** We characterize the nuisance information retained by first-moment domain summaries. We explain why noisy grouped means can mimic additional information and use cross-fitting to separate this estimation bias from the benefit of semantic conditioning (Sec. 5).
2. **Transport.** We distinguish source estimation from target subspace selection using target-oracle controls, operator reconstruction, and transport regret. A source-indistinguishable construction establishes non-identifiability without assumptions connecting source and target nuisance operators (Sec. 6).
3. **Deployment.** We compare threshold and feature adaptation under explicit calibration and power objectives. Standard order-statistic and ROC properties explain both the effectiveness and the limits of direct threshold adaptation (Sec. 7).

Scope of the theoretical contribution. Several mathematical ingredients used below are standard, including conditional second-moment identities, spectral perturbation tools, order-statistic calibration, and ROC invariance. Our contribution is not to rederive these facts in isolation, but to use them to separate two distinct failure modes of source-side nuisance removal: which nuisance information the source statistic can observe, and whether the resulting source geometry identifies the target-relevant subspace. The results specific to this setting are the observation-channel formulation and its finite-sample consequence, the transport-regret characterization, and the source-indistinguishability construction, which shows that source information alone cannot provide a uniform extrapolation guarantee without additional assumptions linking source and target domains.

2 Related Work

Concept erasure. LEACE provides closed-form linear guardedness (Belrose et al., 2023); INLP iterates nullspace projections (Ravfogel et al., 2020); RLACE formulates erasure adversarially (Ravfogel et al., 2022); MANCE constrains updates to an estimated manifold (Avitan et al., 2026); and Obliviator targets nonlinear dependence (Akbari et al., 2025). Chowdhury et al. (2025) study information-theoretic limits on perfect concept erasure and the associated utility tradeoff. Their limits concern the erasure objective; our question is whether source-observed nuisance information identifies the geometry needed for a detector on an unseen domain. Elazar et al. (2021) argue that probe recoverability alone cannot establish behavioral consequences and use counterfactual erasure to assess them. Our detector evaluation addresses the related deployment question under domain shift.

Naowarat et al. (2026) establish that concept-subspace criteria need not generalize to unseen data. We take this failure as a starting point rather than as our primary novelty: our aim is to determine why source-side erasure can fail and to separate failures of source observation from failures of source-to-target transport. We identify two distinct reasons for source-side failure: content-dependent nuisance responses can be hidden by the observation statistic, and even complete source geometry can leave target geometry unidentified. We characterize the corresponding information and transport limits, use target-oracle controls to distinguish transport failure from a limitation of the projector class, and evaluate corrections against downstream operating-point and matched-FPR discrimination objectives. The first-moment characterization applies to LEACE's domain cross-covariance, not to all erasure methods.

Invariance and domain generalization. IRM and its critiques study invariant prediction (Arjovsky et al., 2019; Rosenfeld et al., 2021; Gulrajani & Lopez-Paz, 2021), while domain-adaptation analyses identify limits of invariant representations and the importance of support and invertibility (Ben-David et al., 2010; Zhao et al., 2019; Johansson et al., 2019). First- and second-order information can differ substantially in their ability to identify invariant subspaces (Wang et al., 2022a); class-conditional invariance also requires care (Mahajan et al., 2021). We use conditional-expectation

identities to characterize the information discarded specifically by content averaging, and then distinguish this population limitation from finite-sample bias. Work on gradual adaptation motivates coverage of intermediate domains (Kumar et al., 2020; Wang et al., 2022b). Here the transported object is a rank-constrained nuisance subspace, and the target operator may remain outside the observed source family.

Detection and target adaptation. Score-based OOD detection (Hendrycks & Gimpel, 2017; Lee et al., 2018; Sun et al., 2022), benign-versus-semantic shift (Gao et al., 2024), covariance alignment (Sun et al., 2016), and conformal calibration (Angelopoulos & Bates, 2021) provide the deployment context. Geometric accounts of nuisance robustness emphasize matching representations to nuisance variation (Rajput, 2026). We ask when source observations identify the relevant target geometry and when deployment requires estimating it at all. Threshold calibration is an established baseline; our comparison uses it to distinguish operating-point correction from improvements in discrimination.

3 SETUP

Deployment task. A frozen encoder maps an observation to $z \in \mathbb{R}^D$. We fit a one-class residual detector on source-domain normals,

$$s(z) = \|P_\perp(z - \mu)\|^2,$$

where μ is the source mean and $P_\perp$ projects outside a fitted principal subspace. This is the residual component of a probabilistic-PCA model (Tipping & Bishop, 1999). A separate source-normal split sets a threshold τ at nominal FPR $\alpha = 0.05$; larger scores indicate anomalies. At an unseen target domain we evaluate FPR on normals, true-positive rate (TPR) on anomalies when available, and AUROC. Splits separate adaptation from evaluation and group patient or content identities where applicable. The primary observational benchmarks are Camelyon17-WILDS (Bandi et al., 2019; Koh et al., 2021), PACS (Li et al., 2017), and Office-Home (Venkateswara et al., 2017); TCGA and 20 Newsgroups provide additional audits. Appendix B distinguishes their protocols.

Nuisance directions. We write the frozen representation as $\Phi(c, d)$, with content c carrying the semantic information to preserve and domain factor d representing a nuisance. Varying d while holding c fixed traces local directions in representation space. Their span is the *nuisance tangent space*, $V_{c,d} = \operatorname{Im} D_d\Phi(c, d)$. For a differentiable eraser E, its Jacobian J_E determines the transmission of those directions:

$$\eta_E(c, d) = \|J_E(\Phi(c, d))P_{V_{c,d}}\|_{\mathrm{op}}.$$

Here P_V is an orthogonal projector. This worst-direction gain can exceed one; it is not a probability. Controlled transformations supply nuisance derivatives directly. On observational domains, source and held-out domain contrasts serve as proxies, so their measured gains do not establish attenuation of every nuisance direction.

Evaluation quantities. We distinguish *probe recoverability* (linear and MLP domain accuracy), *nuisance transmission* (η_E), *preservation outside the nuisance subspace* (singular values of J_E restricted to $V^\perp$), and *detector behavior* (FPR, TPR, AUROC). Preservation on $V^\perp$ is a geometric proxy, not a guarantee that all semantics survive. Residual probe information, failure to attenuate target nuisance directions, and poor detector behavior are distinct outcomes; none is interchangeable with another.

Even perfect erasure of a source tangent space does not establish target attenuation. Adding and subtracting source terms gives

$$\|J_tP_t\|_{\mathrm{op}} \le \underbrace{\|J_sP_s\|_{\mathrm{op}}}_{\text{source transmission}} + \underbrace{\|(J_t - J_s)P_s\|_{\mathrm{op}}}_{\text{Jacobian change}} + \underbrace{\|J_t\|_{\mathrm{op}}\|P_t - P_s\|_{\mathrm{op}}}_{\text{subspace mismatch}}. \tag{1}$$

For a linear eraser the Jacobian-change term vanishes, but the source and target subspaces may still differ. This upper bound motivates separate measurements; it is not an equality attributing observed FPR to geometry (App. A.1).

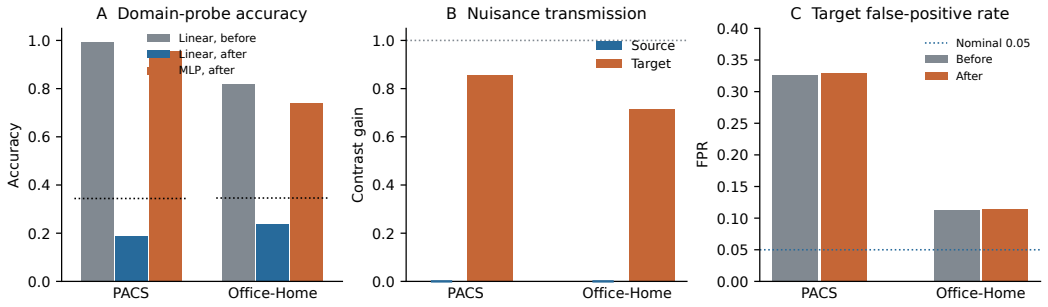


Figure 1: Source first-moment erasure and target detector behavior on natural images. Reported PACS and Office-Home aggregates: (A) linear and MLP domain probes after erasure, with chance baselines; (B) source and target contrast transmission; (C) target FPR before and after erasure, with nominal source FPR. Source transmission is zero to reported precision, but target transmission remains substantial and target FPR changes little. Values are the semantic-partition aggregates reported in App. C.3.1; panels do not mix them with the distinct single-partition PACS audit.

Role of controlled and observational experiments. The controlled transformations test the proposed geometric mechanism, because content can be held fixed and nuisance derivatives can be measured directly. The observational benchmarks instead test the deployment consequence, namely whether source-side erasure improves unseen-domain detector behavior. We therefore do not infer the full target nuisance tangent geometry from observational domain contrasts alone. Agreement between the two experiment types is evidence that the mechanism is plausible, not an identification result for naturally occurring shifts.

4 Source first-moment erasure and unseen-domain detector behavior

Linear guardedness and target attenuation. On Camelyon17, LEACE reduces the linear domain probe below chance and removes the fitted source contrast to numerical precision. An MLP nevertheless retains high domain accuracy, and target transmission remains near its unmodified value. The target FPR changes only from 0.848 to 0.855 (App. C.3.2). These measurements establish removal of the observed linear signal, not general domain invariance. A low accuracy from a particular fitted probe is also not, by itself, a certificate against every linear classifier.

Replication across semantic partitions. The same pattern appears on PACS and Office-Home (Fig. 1). The evaluation includes random normal/anomaly partitions and an exhaustive sweep of the 35 four-class normal sets on PACS. Average target FPR changes by $+0.003$ on PACS and $+0.001$ on Office-Home. The latter exhibits milder shift and more target attenuation, but no appreciable average improvement in its operating point. This supports the scoped conclusion that source first-moment erasure need not improve target FPR, rather than a conclusion about invariance in general.

Preservation and detection must be checked together. LEACE and the evaluated MANCE variants largely preserve anomaly discrimination while leaving FPR unchanged. Some stronger interventions reduce FPR together with substantial loss of TPR and contraction outside the nuisance subspace; FPR alone therefore cannot certify preservation. The nonlinear comparisons in Appendix C.3.2 are descriptive because the Obliviator run did not reach its stopping criterion. The Camelyon17 eraser audit has two independent target hospitals, so its FPR result is also descriptive. Additional controls appear in Apps. C.3.4 and C.3.6. Our strongest mechanistic characterization concerns erasers whose source nuisance information is carried by first-moment domain statistics. The comparisons with more expressive erasers test whether the empirical phenomenon extends beyond this case, but they do not establish a universal impossibility result for all nonlinear or higher-order concept-erasure methods.

5 WHAT FIRST-MOMENT SUPERVISION REVEALS

Suppose a domain change moves two content classes in opposite directions. Averaging within a domain can cancel these responses even when individual examples move substantially. We call the statistic through which a method receives nuisance information its *observation channel*. For LEACE with one-hot domain labels, the domain cross-covariance is determined by differences between domain means.

Averaging removes content-dependent variation. For a scalar nuisance path, write $U(c,d) = D_d\Phi(c,d)$ and $\bar{U}(d) = \mathbb{E}_c[U(c,d)]$. Assume a fixed content distribution across d, finite second moments, and regularity permitting differentiation under the expectation. Then the domain mean $\mu(d) = \mathbb{E}_c[\Phi(c,d)]$ satisfies $D_d\mu(d) = \bar{U}(d)$. Define the second-moment operators $M_{\tan} = \mathbb{E}_{d,c}[UU^\top]$ and $M_{\text{mean}} = \mathbb{E}_d[\bar{U}\bar{U}^\top]$. The conditional second-moment identity yields

$$\boxed{M_{\tan} = M_{\text{mean}} + \mathbb{E}_d[\mathrm{Cov}_c(U \mid d)].} \tag{2}$$

The first term on the right is the nuisance energy exposed by mean responses; the second is the content-dependent variation lost by averaging. This covariance term is positive semidefinite, hence $M_{\text{mean}} \preceq M_{\tan}$. The *content-cancellation gap* is its trace divided by $\operatorname{tr} M_{\tan}$. It measures missing energy, not necessarily directions orthogonal to the observable subspace. Finite domain means provide samples of the mean trajectory, not direct access to all of its derivatives.

The identity itself is standard. Its use here identifies the population information limitation of the domain first-moment channel. LEACE also uses the marginal feature covariance to choose its transformation; the claim concerns its domain cross-covariance, not every statistic in its implementation. Second-moment or nonlinear domain information is outside this characterization. If content proportions vary with domain, differentiating their distribution adds a term, so the tangent interpretation above requires the controlled-content assumption.

Controlled measurement and semantic conditioning. We apply known nuisance transformations to fixed content, avoiding the matching and class-prior ambiguities of observational domains. The core stain study differentiates frozen features of $1{,}260$ NCT-CRC-HE tiles across 25 nuisance states and nine tissue classes. Domain means omit 68–86% of tangent energy across the reported histology configurations. Additional encoders, RGB gains, blur, and natural-image controls are reported in Apps. C.1.5, C.1.4, and C.2.5.

Conditioning on content can expose responses that domain averaging removes. For information sets $\mathcal{G}_1 \subseteq \mathcal{G}_2$, the operators $M_{\mathcal{G}} = \mathbb{E}[\mathbb{E}(U \mid \mathcal{G})\mathbb{E}(U \mid \mathcal{G})^\top]$ satisfy $M_{\mathcal{G}_1} \preceq M_{\mathcal{G}_2}$ (App. A.1). We assess the selected subspaces on both held-out content and held-out nuisance states. For a projector P, *capture* is $\operatorname{tr}(PM)/\operatorname{tr} M$, the fraction of nuisance energy removed by $I - P$. Rank budgets are matched, but realized dimensions can differ: the domain-mean channel in this comparison supplies at most eight directions.

Finite-sample bias can mimic additional information. To isolate semantic information, we compare class conditioning with random partitions of matched cell number and size. An independent random label adds no population information: $\mathbb{E}[U \mid D, R] = \mathbb{E}[U \mid D]$. Nevertheless, naive grouped means appear to improve capture. If cell g contains m_g independent samples with covariance Σ_g, then, conditional on fixed cell counts and weights,

$$\mathbb{E}[\widehat{M}_{\mathcal{G}}] = M_{\mathcal{G}} + \mathbb{E}_g[\Sigma_g/m_g]. \tag{3}$$

Squaring a noisy group mean converts sampling variance into apparent signal. Dividing a cell into smaller random groups increases this bias without adding population information.

Independent within-cell halves remove the additive bias in expectation:

$$\widetilde{M}_{\mathcal{G}} = \mathbb{E}_g\left[\operatorname{sym}\left(\widehat{\mu}_g^A(\widehat{\mu}_g^B)^\top\right)\right], \qquad \mathbb{E}[\widetilde{M}_{\mathcal{G}}] = M_{\mathcal{G}}.$$

The estimate is symmetric but need not be positive semidefinite; we retain and report negative eigenvalues. Unbiasedness of this operator does not imply unbiasedness of its selected eigenspace. At attainable ranks $r \leq 8$, cross-fitting reduces the random-partition capture gain from $+0.016$ to $+0.001$, while the semantic component increases from $+0.015$ to $+0.033$ (Fig. 2). The remaining high-rank comparison also reflects unequal available dimensions (App. C.1.3).

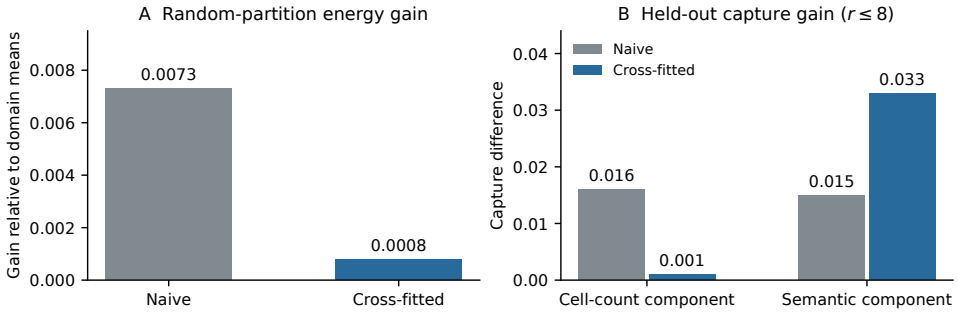


Figure 2: Cross-fitting separates finite-sample bias from semantic conditioning. (A) Extra observable energy from a matched random partition, relative to domain means. (B) Held-out capture gains at rank budgets $r \leq 8$: the cell-count component is random minus domain conditioning; the semantic component is class minus random conditioning. The two metrics are shown on separate axes.

Observation is only one limitation. Improved held-out capture does not necessarily improve the detector. On controlled extrapolation tasks, source-built projectors from domain means, class conditioning, and even raw source tangents leave substantial target FPR across residual, Mahalanobis, and kNN detectors (App. C.2.4). A spectral capacity analysis further separates rank budget, available information, and subspace selection (App. C.1.2). Neither an energy decomposition nor capture alone determines the operating point. We next test whether access to target geometry changes the result within the same projector class.

6 Transport to unseen domains

The *target nuisance operator* $M_t = \mathbb{E}_t[UU^\top]$ describes nuisance directions under the target distribution; M_s is its source counterpart. Their rank-relevant eigenspaces are what we mean by source and target *geometry*. All operator comparisons use a common tangent convention within an experiment. We measure empirical calibration error as

$$CE_\alpha = |\widehat{\mathrm{FPR}} - \alpha|,$$

computed per evaluation setting before aggregation. It therefore differs from the absolute error of an average FPR.

A target oracle separates intervention from transport. A three-way disjoint content split fits the detector and source threshold, estimates M_t, and evaluates the result. A target-built, energy-optimal linear projector improves calibration in each of the five materially miscalibrated configurations, with improvements across tested ranks from 2 to 64 and mean CE_α reduction 0.073 (App. C.2.1). The target tangent information makes this a diagnostic control of projector capacity and transport. Its energy objective need not be FPR-optimal, since FPR also depends on scores near the threshold.

Recombining source operators leaves a residual. We optimize convex and affine combinations of per-state source operators against M_t. Target capture improves little relative to the pooled source operator, remaining below target-oracle capture. Projection onto the source affine span leaves relative Frobenius residual 1.14, versus 0.32 for an omitted state inside the observed range (App. C.2.2). This measures mismatch for the tested operator family. Distinct operators can share leading eigenspaces, so operator residual and rank-constrained capture remain separate quantities.

Transport regret quantifies subspace mismatch. Let $P^\star(M)$ be a rank-r orthogonal projector maximizing $\mathrm{tr}(PM)$ and let $\mathrm{Opt}_r(M) = \sum_{j \leq r} \lambda_j(M)$. The loss in target energy capture from using the source-optimal subspace is

$$\mathcal{T}_r(s \to t) = \mathrm{Opt}_r(M_t) - \mathrm{tr}(P^\star(M_s)M_t).$$

This definition assumes exact knowledge of M_s and equal rank budgets. If $q = \|\sin\Theta(P_s, P_t)\|_F^2$ measures the squared principal-angle discrepancy and $\gamma_t = \lambda_r(M_t) - \lambda_{r+1}(M_t)$, then

$$\gamma_t q \leq \mathcal{T}_r(s \to t) \leq (\lambda_1(M_t) - \lambda_D(M_t))q.$$

Thus the cost of rotation depends on the target spectrum. Appendix A.3 gives the proof and complementary operator-norm, eigengap, and Wasserstein bounds. These bounds control geometry; numerical FPR prediction additionally requires the target score distribution near the threshold.

Non-identifiability without a source–target restriction. Smoothness alone does not identify the target-optimal subspace outside the source range. For $D \geq 2r$, we construct two smooth representations that agree on every source observation and derivative but have target operators P_{V_1}/r and P_{V_2}/r for orthogonal rank-r subspaces. Every source-only rank-r selection rule then has expected transport regret at least $1/2$ in one world (Thm. 3). This is a worst-case statement for an unrestricted smooth family, not a claim that every real target is unrecoverable. Assumptions linking source and target operators can permit transfer. Empirical extrapolation and alternative correction controls appear in Apps. C.2.3, C.2.7, and C.4.6.

7 Matching adaptation to the deployment objective

We do not propose threshold adaptation as a new algorithm. We use this standard low-dimensional baseline to expose an objective mismatch: representation-level nuisance removal is a substantially stronger intervention than is required when the deployment objective is only calibration of a single operating point. We distinguish obtaining an FPR near α from retaining high TPR at that FPR. A *target adaptation batch* is a sample used to fit an intervention, disjoint from evaluation. We compare threshold changes with mean, diagonal-moment, CORAL, and rank-8 feature corrections. Feature corrections act on target examples at inference while the source detector remains fixed. For a fitted map followed by its own threshold, disjoint subsets fit the map and calibrate its scores. The all-data geometry arm is a deliberately larger-budget comparator; it uses all available target adaptation data in every column of Table 1.

What threshold adaptation estimates. For a fixed score with continuous target-normal CDF F_t, let $S_{(j)}$ be the jth order statistic of k independent target-normal scores. A standard identity gives the conditional population FPR of the resulting cutoff:

$$p_j = 1 - F_t(S_{(j)}) \sim \mathrm{Beta}(k+1-j, j), \qquad \mathbb{E}[p_j] = \frac{k+1-j}{k+1}. \tag{4}$$

The expectation is over calibration batches. This is the standard order-statistic law underlying split conformal prediction (Angelopoulos & Bates, 2021), not a new distributional result (App. A.4). The rank index is $j = \lceil (k+1)(1-\alpha) \rceil$. At budgets where $j > k$, a valid deterministic conformal cutoff is $+\infty$. The finite maximum used in the small-budget experiments is instead an empirical cutoff, and we evaluate it as such.

At $k = 19$ and $\alpha = 0.05$, the maximum gives $p \sim \mathrm{Beta}(1, 19)$ and $\mathbb{E}[p] = 0.05$. Individual batches remain variable: the 90th and 95th percentiles of p are approximately 0.114 and 0.146. Indeed, the analytically implied $\mathbb{E}|p - 0.05| = 0.1(0.95)^{20} \approx 0.0358$, despite zero signed marginal bias. Nineteen clean normals therefore support a comparison of sample efficiency, not a guarantee that each deployment batch attains FPR near 0.05. The exact law assumes independent clean normals and a score fixed independently of the calibration sample; contamination and patient dependence require separate consideration.

Calibration and discrimination respond to different interventions. With all available adaptation data, the rank-8 projector reaches $CE_\alpha = 0.201$, whereas the threshold reaches 0.037 with 19 normals and 0.010 with all data. Composing feature maps with independently fitted thresholds produces calibration errors close to threshold adaptation alone. Their potential benefit is improved discrimination: in separate matched-FPR comparisons, CORAL and the rank-8 projector improve TPR by 0.097 and 0.080, respectively (App. C.4.1). These matched-FPR results should not be equated with TPR differences between finitely calibrated methods in Table 1.

The relevant distinction is score ordering. A standard ROC fact is that, for strictly increasing g, $\{g(S) > g(t)\} = \{S > t\}$, hence

$$\mathrm{ROC}(g \circ S) = \mathrm{ROC}(S).$$

Table 1: Operating-point adaptation on real domains. Mean per-setting CE_α, with all-data TPR in parentheses, across evaluated held-out domains and three split seeds. Lower CE_α is better. "All" denotes the full adaptation pool. At $k < 19$ the finite rank threshold clips the order-statistic index and is an empirical cutoff without nominal conformal validity. Geometry-all uses the full pool in every column; its compositions are descriptive all-data controls rather than matched-k comparisons. Additional results appear in Appendix C.4.1.

adaptation	$k = 5$	10	19	100	all
none	0.265	0.266	0.265	0.266	0.266 (0.771)
rank threshold	0.118	0.105	**0.037**	**0.018**	**0.010** (0.587)
mean shift	0.273	0.214	0.193	0.178	0.174 (0.742)
diagonal second moment	0.949	0.763	0.537	0.366	0.330 (0.823)
CORAL	0.883	0.824	0.773	0.508	0.093 (0.831)
geometry-all, rank 8	0.198	0.198	0.202	0.198	0.201 (0.745)
CORAL + threshold	0.163	0.180	0.064	0.026	0.012 (0.694)
geometry-all + threshold	0.014	0.014	0.013	0.013	0.014 (0.657)

Table 2: Matched-budget adaptation on naturally occurring shift. Means across the seven folds with unadapted $\mathrm{FPR} > 2\alpha$ among nine leave-one-domain-out PACS and Camelyon17 folds, over five seeds. Every target-aware arm uses the same total budget of $k = 19$ target normals from an adaptation split disjoint from evaluation; the composed arm divides that budget between map fitting and threshold calibration. CE_α is averaged per setting.

adaptation	CE_α	FPR	TPR	AUROC
no correction	0.460	0.510	0.968	0.958
source first-moment erasure	0.466	0.516	0.968	0.959
target geometry, rank 8	0.401	0.450	0.962	0.967
target threshold	0.035	0.057	0.749	0.958
geometry + threshold	0.035	0.047	0.743	0.967

Threshold changes and strictly increasing score transformations preserve discrimination at matched FPR. Improving it requires changing ordering or tie structure, although such a change is not sufficient to improve it. Empirical quantile maps can introduce ties and therefore are not covered by the strict-increase statement. Representation adaptation can change ordering, but it is neither the only way to do so nor a guarantee of improvement (App. A.4).

Joint objectives. Under the joint requirement $CE_\alpha \leq 0.05$ and TPR at least 80% of unadapted TPR, the reported feasible methods use diagonal or CORAL corrections followed by thresholds, with clean batches of at least 100 examples. At 90%, only CORAL plus a threshold with all available clean data qualifies; none of the evaluated methods meets the 95% requirement. These are empirical feasibility results for the evaluated methods, not consequences of ROC invariance alone. The reference TPR is measured at the unadapted, potentially inflated FPR, so retention is an operational constraint rather than a matched-FPR discrimination measure.

Matched target information on real shifts. Table 2 compares source first-moment erasure with target geometry, threshold calibration, and their composition. On the seven materially miscalibrated folds, source first-moment erasure leaves mean FPR essentially unchanged (0.516 versus 0.510 without correction). With the same total adaptation budget of $k = 19$ target normals, geometry reduces FPR to 0.450, threshold calibration reaches 0.057, and the composition reaches 0.047. Geometry raises AUROC from 0.958 to 0.967, and the composition retains that gain; TPR at the fitted thresholds is lower for the threshold-based arms. Thus calibration and discrimination change differently even when target information is matched. These results do not establish that threshold adaptation dominates feature adaptation in general. They show that the interventions target different quantities in these experiments: a threshold directly adjusts the operating point, whereas geometry-changing interventions can alter score ordering and hence discrimination. The contribution of this comparison is therefore diagnostic rather than algorithmic, identifying when the representation itself must change and when adapting only the decision rule suffices. The nine-fold analysis and adaptation protocol appear in Appendix C.4.2.

Contaminated adaptation batches. When all target-aware methods receive contaminated batches, both threshold and geometry adaptation lose TPR. At 20% contamination, the rank-8 geometry arm's TPR falls from 0.746 to 0.606. A separate unequal-input control holds geometry's inputs clean and therefore does not establish contamination robustness (App. C.4.5). Small absolute calibration error also need not indicate a useful detector: always accepting normals gives $CE_{0.05} = 0.05$ while detecting no anomalies.

8 Discussion and limitations

The results support three claims of different scope. First, for source erasers driven by first-moment domain summaries, the observable source contrast can substantially underrepresent content-dependent nuisance variation. Second, even complete knowledge of source nuisance geometry does not by itself identify the target-optimal subspace without assumptions linking source and target domains. Third, the representation-level objective may be stronger than the deployment objective, since operating-point calibration and matched-FPR discrimination require different interventions. The first claim is specific to the observation channel studied here, and the second is a source-only identifiability statement; neither constitutes an impossibility result for every concept-erasure method under every form of domain shift.

The first-moment analysis does not cover all nonlinear or second-moment erasers, and controlled nuisance paths do not exhaust naturally occurring domain shift. Geometric preservation outside a chosen nuisance subspace is only a proxy for semantic preservation. The target oracle has privileged tangent information, while ordinary adaptation requires representative target samples. Independence, clean-normal access, and stable target distributions are assumptions of the order-statistic analysis, not guaranteed properties of deployment.

The empirical comparisons should be read at their stated information budgets. Geometry-all uses more target data than the column budgets. Camelyon17 FPR and TPR may average different available outcomes because one hospital has no anomaly tiles. The observation identity and source-only non-identifiability result do not depend on these aggregation choices (App. C.4.7). The number of independent naturally occurring deployment domains is modest. Multiple semantic partitions and random seeds provide robustness checks within a domain but do not create additional independent domain shifts. Accordingly, the real-domain experiments should be interpreted as replicated evidence across the available held-out domains rather than as a large-sample estimate over a population of deployment environments.

9 Conclusion

Removing a source-observed domain signal does not ensure a well-calibrated detector on an unseen domain. First-moment summaries lose content-dependent nuisance variation, and even complete source geometry need not identify the target-optimal erasure subspace. Cross-fitting distinguishes additional information from grouped-mean sampling bias; target-oracle controls distinguish source transport limitations from the capability of the intervention class.

Deployment objectives determine which of these quantities must be adapted. In the studied settings, threshold adjustment effectively addresses operating-point calibration, with appreciable uncertainty for small batches. Recovering discrimination at matched FPR requires changing score ordering, which feature adaptation can provide when target information is sufficient. These results support adapting the lowest-dimensional quantity sufficient for the stated objective, while evaluating calibration, discrimination, and representation preservation separately. More expressive source-only erasers may recover nuisance information that first-moment methods miss, and additional structural assumptions may permit successful source-to-target transport. Our results therefore do not argue against concept erasure in general; they identify the information and transport assumptions that must be justified before source-side erasure can be treated as evidence of deployment robustness.

Reproducibility Statement

Complete proofs of the theoretical results are provided in Appendix A. Appendix B specifies the datasets, representations, data splits, evaluation metrics, and reporting conventions used throughout the experiments, while Appendix C provides additional experimental results, controls, and robustness analyses. Target-domain adaptation data are kept disjoint from evaluation data, and repeated experiments report results across multiple random seeds where applicable. The paper also distinguishes controlled nuisance transformations, for which nuisance derivatives can be measured directly, from observational domain shifts, for which domain contrasts are used only as proxies.

Statement on the use of AI assistants

An AI assistant was used for language editing, improving clarity of presentation, identifying potentially relevant literature, and providing feedback on manuscript organization and presentation. All cited literature, theoretical claims and proofs, experimental procedures, numerical results, and scientific conclusions were independently verified by the authors, who take responsibility for the final manuscript.

References

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, S. Biderman. LEACE: Perfect linear concept erasure in closed form. *NeurIPS*, 2023.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *ACL*, 2020.

S. Ravfogel, M. Twiton, Y. Goldberg, R. Cotterell. Linear adversarial concept erasure. *ICML*, 2022.

M. Avitan, Y. Goldberg, Y. Elazar. MANCE: Manifold aware concept erasure. arXiv:2607.03973, 2026.

R. Akbari, M. Afshari, V. N. Boddeti. Obliviator reveals the cost of nonlinear guardedness in concept erasure. *NeurIPS*, 2025.

B. Naowarat, H. Tang, S. Goldwater. A framework for analyzing concept representations in neural models. *CoNLL*, 2026. https://arxiv.org/abs/2605.01381.

S. B. R. Chowdhury, K. A. Dubey, A. Beirami, R. Kidambi, N. Monath, A. Ahmed, S. Chaturvedi. Fundamental limits of perfect concept erasure. *AISTATS*, PMLR 258:901–909, 2025. https://proceedings.mlr.press/v258/chowdhury25a.html.

Y. Elazar, S. Ravfogel, A. Jacovi, Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *TACL*, 9:160–175, 2021. https://aclanthology.org/2021.tacl-1.10/.

B. Sun, J. Feng, K. Saenko. Return of frustratingly easy domain adaptation. *AAAI*, 2016.

W. E. Johnson, C. Li, A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 2007.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17, 2016.

M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz. Invariant risk minimization. arXiv:1907.02893, 2019.

E. Rosenfeld, P. Ravikumar, A. Risteski. The risks of invariant risk minimization. *ICLR*, 2021.

I. Gulrajani, D. Lopez-Paz. In search of lost domain generalization. *ICLR*, 2021.

H. Zhao, R. T. des Combes, K. Zhang, G. Gordon. On learning invariant representations for domain adaptation. *ICML*, 2019.

F. D. Johansson, D. Sontag, R. Ranganath. Support and invertibility in domain-invariant representations. *AISTATS*, 2019.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79, 2010.

H. Wang, H. Si, B. Li, H. Zhao. Provable domain generalization via invariant-feature subspace recovery. *ICML*, 2022.

D. Mahajan, S. Tople, A. Sharma. Domain generalization using causal matching. *ICML*, 2021.

A. Kumar, T. Ma, P. Liang. Understanding self-training for gradual domain adaptation. *ICML*, 2020.

H. Wang, B. Li, H. Zhao. Understanding gradual domain adaptation. *ICML*, 2022.

V. Rajput. The matching principle: a geometric theory of loss functions for nuisance-robust representation learning. arXiv:2605.22800, 2026.

D. Hendrycks, K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

K. Lee, K. Lee, H. Lee, J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 2018.

Y. Sun, Y. Ming, X. Zhu, Y. Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022.

Z. Gao, B. Li, M. Salzmann, X. He. Generalize or detect? Towards robust semantic segmentation under multiple distribution shifts. *NeurIPS*, 2024.

M. E. Tipping, C. M. Bishop. Probabilistic principal component analysis. *JRSS-B*, 61(3), 1999.

K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations. *PNAS*, 35(11), 1949.

C. Davis, W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7(1), 1970.

J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 1961.

P. W. Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. *ICML*, 2021.

P. Bandi et al. From detection of individual metastases to classification of lymph node status (CAMELYON17). *IEEE TMI*, 38(2), 2019.

D. Li, Y. Yang, Y.-Z. Song, T. M. Hospedales. Deeper, broader and artier domain generalization. *ICCV*, 2017.

J. N. Kather, N. Halama, A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo, 2018.

Y. Le, X. Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford CS231N, 2015.

A. Filiot et al. Scaling self-supervised learning for histopathology with masked image modeling. 2023.

M. Oquab et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

A. Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

H. Xu et al. A whole-slide foundation model for digital pathology. *Nature*, 2024.

de Jong et al. Current pathology foundation models are unrobust to medical center differences. arXiv:2501.18055, 2025.

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021. https://arxiv.org/abs/2107.07511.

H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan. Deep hashing network for unsupervised domain adaptation. *CVPR*, 2017. https://openaccess.thecvf.com/content_cvpr_2017/html/Venkateswara_Deep_Hashing_Network_CVPR_2017_paper.html.

A THEORY

A.1 OBSERVATION CHANNEL

Lemma 1 (Differential transfer)**.** For source and target points x_s, x_t with nuisance projectors P_s, P_t,

$$\|J_t P_t\|_{\mathrm{op}} \;\le\; \|J_s P_s\|_{\mathrm{op}} + \|(J_t - J_s)P_s\|_{\mathrm{op}} + \|J_t\|_{\mathrm{op}}\|P_t - P_s\|_{\mathrm{op}}.$$

Proof. $J_t P_t = J_s P_s + (J_t - J_s)P_s + J_t(P_t - P_s)$; take norms and use submultiplicativity. □

Proposition 1 (Observation-channel decomposition)**.** *Assume a fixed distribution of content across domains, finite second moments, and a scalar nuisance path. With $U(c,d) = D_d\Phi(c,d)$, $\bar{U}(d) = \mathbb{E}_c[U]$, $M_{\tan} = \mathbb{E}_{d,c}[UU^\top]$ and $M_{\mathrm{mean}} = \mathbb{E}_d[\bar{U}\bar{U}^\top]$,*

$$M_{\tan} = M_{\mathrm{mean}} + \mathbb{E}_d[\mathrm{Cov}_c(U \mid d)] \succeq M_{\mathrm{mean}}.$$

Proof. Write $U = \bar{U} + (U - \bar{U})$ with $\mathbb{E}_c[U - \bar{U} \mid d] = 0$. Expanding the outer product and taking $\mathbb{E}_c[\cdot \mid d]$, the cross terms vanish, giving $\mathbb{E}_c[UU^\top \mid d] = \bar{U}\bar{U}^\top + \mathrm{Cov}_c(U \mid d)$. Take $\mathbb{E}_d$. The remainder is an average of covariances, hence PSD. □

Under regularity permitting differentiation through the expectation, $D_d\mathbb{E}_c[\Phi] = \mathbb{E}_c[D_d\Phi] = \bar{U}$, so a method whose only view of the nuisance is a per-domain first moment has access to the mean trajectory, whose derivative is $\bar{U}$. LEACE is in this class: for a one-hot domain label Z, column g of $\mathrm{Cov}(X, Z)$ is $\pi_g(\mu_g - \bar{\mu})$. We do not claim it for INLP, CORAL, MANCE or Obliviator.

Proposition 2 (Observability hierarchy)**.** *For σ-fields $\mathcal{G}_1 \subseteq \mathcal{G}_2$ refining the domain label, with $M_{\mathcal{G}} = \mathbb{E}[\mathbb{E}[U \mid \mathcal{G}]\mathbb{E}[U \mid \mathcal{G}]^\top]$, we have $M_{\mathcal{G}_1} \preceq M_{\mathcal{G}_2}$.*

Proof. Fix w and set $Z = w^\top U$; then $w^\top M_{\mathcal{G}} w = \mathbb{E}[(\mathbb{E}[Z \mid \mathcal{G}])^2]$. Conditional expectation is the orthogonal projection of L^2 onto $\mathcal{G}$-measurable functions, and the tower property gives $\mathbb{E}[Z \mid \mathcal{G}_1] = \mathbb{E}[\mathbb{E}[Z \mid \mathcal{G}_2] \mid \mathcal{G}_1]$. A projection is an L^2 contraction, so the smaller σ-field gives the smaller second moment. As w was arbitrary, the Loewner inequality follows. □

A.2 FINITE-SAMPLE GROUPED-MEAN ESTIMATION

Let cell g have fixed weight w_g, m_g independent tangent samples, population mean μ_g, and covariance Σ_g. With $\sum_g w_g = 1$,

$$\widehat{M} = \sum_g w_g \widehat{\mu}_g \widehat{\mu}_g^\top, \qquad \mathbb{E}[\widehat{M}] = \sum_g w_g \mu_g \mu_g^\top + \sum_g w_g \Sigma_g / m_g.$$

This follows by expanding the second moment of each sample mean. The statement conditions on cell counts and weights; estimated or data-dependent partitions require additional care. Equal random refinement preserves the population conditional mean but reduces cell sizes, increasing the additive term.

For independent halves A, B within each cell, define $\widetilde{M} = \sum_g w_g \,\mathrm{sym}(\widehat{\mu}_g^A(\widehat{\mu}_g^B)^\top)$, where $\mathrm{sym}(A) = (A + A^\top)/2$. Independence gives $\mathbb{E}[\widehat{\mu}_g^A(\widehat{\mu}_g^B)^\top] = \mu_g\mu_g^\top$, proving unbiasedness. Finite-sample negative eigenvalues and noisy eigenspaces remain possible. Accordingly, selected subspaces are evaluated on disjoint content and nuisance states; operator unbiasedness is not interpreted as an unbiased capture estimator after data-dependent eigenspace selection.

A.3 TRANSPORT

Let $P^\star(M) \in \arg\max_{\mathrm{rank}\,P \le r} \mathrm{tr}(PM)$ over orthogonal projectors, and $\mathrm{Opt}_r(M) = \mathrm{tr}(P^\star(M)M) = \sum_{j \le r} \lambda_j(M)$ by Ky Fan. Define the transport regret $\mathcal{T}_r(s \to t) = \mathrm{Opt}_r(M_t) - \mathrm{tr}(P^\star(M_s)M_t)$.

Lemma 2. For an orthogonal projector P of rank r and symmetric A, $|\,\mathrm{tr}(PA)| \le r\|A\|_{\mathrm{op}}$.

Proof. $\mathrm{tr}(PA) = \sum_{i \le r} u_i^\top A u_i$ for an orthonormal basis $\{u_i\}$ of $\mathrm{ran}\,P$, and $|u^\top A u| \le \|A\|_{\mathrm{op}}$ for unit u. □

Theorem 1 (Transport-regret bound)**.** $0 \le \mathcal{T}_r(s \to t) \le 2r\|M_t - M_s\|_{\mathrm{op}}$.

Proof. Nonnegativity is the optimality of $P^\star(M_t)$ for M_t. Writing $P_t = P^\star(M_t)$, $P_s = P^\star(M_s)$,

$$\mathcal{T}_r = \operatorname{tr}\big(P_t(M_t - M_s)\big) + \underbrace{\big[\operatorname{tr}(P_t M_s) - \operatorname{tr}(P_s M_s)\big]}_{\le 0} + \operatorname{tr}\big(P_s(M_s - M_t)\big),$$

the bracket being nonpositive by optimality of P_s for M_s. Lemma 2 bounds each remaining term by $r\|M_t - M_s\|_{\mathrm{op}}$. □

Corollary 1 (Observation and transport separate additively)**.** *If the source can construct only $M_{\mathcal{G},s}$, then $\|M_t - M_{\mathcal{G},s}\|_{\mathrm{op}} \le \|M_s - M_{\mathcal{G},s}\|_{\mathrm{op}} + \|M_t - M_s\|_{\mathrm{op}}$, and the regret of $P^\star(M_{\mathcal{G},s})$ against M_t is at most $2r$ times that sum.*

Proposition 3 (Grassmann form)**.** *Let $M_t = \sum_i \lambda_i u_i u_i^\top$ with $\lambda_1 \ge \cdots \ge \lambda_D \ge 0$, let $P_t = P^\star(M_t)$ project onto $\operatorname{span}(u_1, \ldots, u_r)$, let $P_s = P^\star(M_s)$, and put $q = \|\sin\Theta(P_s, P_t)\|_F^2$ and $\gamma_t = \lambda_r - \lambda_{r+1}$. Then*

$$\gamma_t\, q \;\le\; \mathcal{T}_r(s \to t) \;\le\; (\lambda_1 - \lambda_D)\, q.$$

Proof. Set $a_i = u_i^\top P_s u_i \in [0, 1]$. Since P_s is an orthogonal projector of rank r and $\{u_i\}$ is an orthonormal basis, $\sum_i a_i = \operatorname{tr}(P_s) = r$. For two rank-$r$ projectors $q = r - \operatorname{tr}(P_t P_s)$, and $\operatorname{tr}(P_t P_s) = \sum_{i \le r} a_i$, hence

$$q = r - \sum_{i \le r} a_i = \sum_{i > r} a_i,$$

the second equality using $\sum_i a_i = r$. Now

$$\mathcal{T}_r = \operatorname{tr}(P_t M_t) - \operatorname{tr}(P_s M_t) = \sum_{i \le r} \lambda_i - \sum_i \lambda_i a_i = \sum_{i \le r} \lambda_i (1 - a_i) - \sum_{i > r} \lambda_i a_i.$$

Both $1 - a_i \ge 0$ and $a_i \ge 0$. For the lower bound, $\lambda_i \ge \lambda_r$ for $i \le r$ gives $\sum_{i \le r} \lambda_i (1 - a_i) \ge \lambda_r q$, and $\lambda_i \le \lambda_{r+1}$ for $i > r$ gives $\sum_{i > r} \lambda_i a_i \le \lambda_{r+1} q$; subtracting yields $\mathcal{T}_r \ge (\lambda_r - \lambda_{r+1}) q = \gamma_t q$. For the upper bound, $\sum_{i \le r} \lambda_i (1 - a_i) \le \lambda_1 q$ and $\sum_{i > r} \lambda_i a_i \ge \lambda_D q$, giving $\mathcal{T}_r \le (\lambda_1 - \lambda_D) q$. □

Theorem 2 (Gap-aware transport bound)**.** *Let $\delta = \|M_t - M_s\|_{\mathrm{op}}$ and $\gamma = \lambda_r(M_s) - \lambda_{r+1}(M_s) > 0$. If $\delta < \gamma/2$ then*

$$\mathcal{T}_r(s \to t) \;\le\; \|P_t - P_s\|_*\, \delta \;\le\; \frac{4r\,\delta^2}{\gamma},$$

so transport regret is quadratic *in the operator mismatch and inversely proportional to the source eigengap.*

Proof. Adding and subtracting $\operatorname{tr}(P_t M_s)$ and $\operatorname{tr}(P_s M_s)$,

$$\mathcal{T}_r = \operatorname{tr}\big((P_t - P_s)(M_t - M_s)\big) + \underbrace{\big[\operatorname{tr}(P_t M_s) - \operatorname{tr}(P_s M_s)\big]}_{\le 0},$$

the bracket being nonpositive by optimality of P_s for M_s. By trace–operator norm duality $\operatorname{tr}\big((P_t - P_s)\Delta\big) \le \|P_t - P_s\|_* \|\Delta\|_{\mathrm{op}}$. The nonzero singular values of a difference of two rank-r projectors are the $\sin\theta_i$, each occurring twice, so $\|P_t - P_s\|_* = 2\sum_i \sin\theta_i \le 2\sqrt{r}\|\sin\Theta\|_F \le 2r\|\sin\Theta\|_{\mathrm{op}}$. Davis–Kahan under $\delta < \gamma/2$ gives $\|\sin\Theta\|_{\mathrm{op}} \le 2\delta/\gamma$, hence $\|P_t - P_s\|_* \le 4r\delta/\gamma$. □

Remark 1 (Which bound is informative). Theorem 1 is worst-case and, on our data, empirically vacuous: its median tightness is 0.016. Theorem 2 gives a sufficient perturbation regime when the source eigengap is large relative to the operator mismatch. Appendix C.2.3 reports the empirical norm and eigengap diagnostics, which do not become useful predictors on these configurations.

Corollary 2 (Distributional transport)**.** *If $\|M(d) - M(d')\|_{\mathrm{op}} \le L_M \rho(d, d')$ and $M_\nu = \mathbb{E}_{d \sim \nu} M(d)$, then $\|M_{\nu_t} - M_{\nu_s}\|_{\mathrm{op}} \le L_M W_1(\nu_s, \nu_t)$ and hence $\mathcal{T}_r(\nu_s \to \nu_t) \le 2r L_M W_1(\nu_s, \nu_t)$.*

Proof. For any coupling π of (ν_s, ν_t), $\|M_{\nu_t} - M_{\nu_s}\|_{\mathrm{op}} = \|\mathbb{E}_\pi[M(d') - M(d)]\|_{\mathrm{op}} \le \mathbb{E}_\pi\|M(d') - M(d)\|_{\mathrm{op}} \le L_M \mathbb{E}_\pi \rho(d, d')$ by Jensen and the Lipschitz property. Take the infimum over couplings. □

This is the form that matches our experiments, in which the source projector is built from a *set* of source states: the governing quantity is a Wasserstein distance between the source and deployment state distributions, not the distance to the nearest source state.

Theorem 3 (Quantitative source-only non-identifiability)**.** *Let $D \geq 2r$ and let $V_1 \perp V_2$ be orthogonal r-dimensional subspaces. There exist two C^∞ nuisance fields agreeing on the entire source range to all orders, hence producing identical source statistics of every kind, whose deployment operators are $M_t^{(1)} = \frac{1}{r} P_{V_1}$ and $M_t^{(2)} = \frac{1}{r} P_{V_2}$. Then every source-only algorithm A returning a rank-r projector satisfies*

$$\inf_{A \text{ source-only}} \max_{j \in \{1,2\}} \left[\mathrm{Opt}_r\big(M_t^{(j)}\big) - \mathbb{E}_A[\mathrm{tr}\,\big(P_A M_t^{(j)}\big)] \right] \;\geq\; \tfrac{1}{2}.$$

Proof. The construction is explicit. Let the content variable be $C \sim \mathrm{Uniform}\{1, \ldots, r\}$ and fix orthonormal bases $\{v_{1,i}\}_{i=1}^r$ of V_1 and $\{v_{2,i}\}_{i=1}^r$ of V_2. Let b be a smooth bump supported strictly beyond the source range with $b'(d_\star) = 1$, and choose a common $\Phi_0(c, d) = \phi_0(c)$ independent of d and set $\Phi_j(c, d) = \Phi_0(c, d) + b(d)\, v_{j,c}$. All derivatives of b vanish on the source range, so Φ_1 and Φ_2 agree there to all orders and no source-only algorithm can distinguish them; it must return the same $P = P_A$ in both worlds. At $d_\star$ the nuisance tangent is $U_j(c, d_\star) = v_{j,c}$, so $M_t^{(j)} = \mathbb{E}_C[U_j U_j^\top] = \frac{1}{r} \sum_i v_{j,i} v_{j,i}^\top = \frac{1}{r} P_{V_j}$ as claimed. Both worlds have $\mathrm{Opt}_r(M_t^{(j)}) = 1$, since the top-$r$ eigenspace of $\frac{1}{r} P_{V_j}$ is V_j and captures all of its trace. Writing $c_j = \mathrm{tr}(P M_t^{(j)}) = \frac{1}{r} \mathrm{tr}(P P_{V_j})$ and using $P_{V_1} + P_{V_2} = P_{V_1 \oplus V_2} \preceq I$,

$$c_1 + c_2 = \tfrac{1}{r} \mathrm{tr}\big(P(P_{V_1} + P_{V_2})\big) \leq \tfrac{1}{r} \mathrm{tr}(P) = 1,$$

so $\min_j c_j \leq \frac{1}{2}$ and $\max_j (1 - c_j) \geq \frac{1}{2}$. For a randomized algorithm, take expectations over its internal randomness before selecting the world: $\mathbb{E}_A[c_1] + \mathbb{E}_A[c_2] \leq 1$ gives the same minimax lower bound in expected regret. □

This is the quantitative form of the impossibility below: not merely that source information fails to *guarantee* extrapolation, but that it must forfeit at least half of the attainable deployment capture in one of two source-indistinguishable worlds. Verified numerically on 400 random (D, r, P) with zero violations of $c_1 + c_2 \leq 1$.

Proposition 4 (No source-only guarantee without global regularity)**.** *Let the source occupy $d \in [a, b]$ and $d_\star > b$. For a nonzero residual-score projector and a positive threshold, there exist $\Phi_1, \Phi_2 \in C^\infty$ agreeing on $[a, b]$ to all orders with $s(\Phi_1(c, d_\star)) < \tau < s(\Phi_2(c, d_\star))$.*

Proof. Take $\Phi_2 = \Phi_1 + \psi$ with ψ a smooth bump supported in (b, ∞), scaled so the score crosses τ at $d_\star$; all derivatives of ψ vanish on $[a, b]$. □

This is a standard smooth-bump construction. Its role is to bound the positive results: some restriction connecting the observed and unobserved domains is needed for a uniform extrapolation guarantee. The Lipschitz condition in Corollary 2 is one sufficient restriction, not a necessary one.

A.4 Operating-point adaptation: standard facts

Definition 1 (Adaptation classes)**.** A threshold adaptation changes a cutoff for a fixed score S. A strictly increasing score transformation replaces S by $g \circ S$. An ordering-changing intervention alters relative scores or ties. A representation intervention changes features before scoring and may or may not change ordering. Representation-level invariance is an objective on nuisance dependence, not a fourth nested statistical class.

Fact 1 (ROC invariance under strictly increasing transformations)**.** *For strictly increasing g,* $\mathrm{ROC}(g \circ S) = \mathrm{ROC}(S)$ *and the AUROC and attainable TPR at matched FPR are unchanged.*

Proof. For every threshold t, $\{g(S) > g(t)\} = \{S > t\}$ under both the normal and anomalous laws. The attainable ROC points coincide. When scores have ties, use the same threshold randomization convention in both representations. □

Remark 2 (Necessary versus sufficient changes). Changing ordering or ties is necessary for improving the ROC but not sufficient. Reorderings within a class, or changes whose effects cancel, can leave AUROC unchanged. Likewise, ROC invariance does not assert that threshold adaptation never satisfies a joint calibration-and-power objective: this depends on the baseline ROC and the chosen power requirement.

Remark 3 (Quantiles and monotone maps). Quantile cutoffs and empirical score-CDF maps use the same target score sample. Strictly increasing maps preserve the full ROC; empirical maps may create ties and need not do so. Neither statistical validity nor calibration of all groups follows merely from monotonicity. Source-only non-identifiability of a rank-constrained nuisance projector is a separate statement from these target-informed score adaptations.

Fact 2 (Standard order-statistic law for a rank-based operating point)**.** *Let $S_1, \dots, S_k$ be i.i.d. target-normal scores with continuous CDF F_t and order statistics $S_{(1)} \le \cdots \le S_{(k)}$, let $\widehat{\tau} = S_{(j)}$, and let S^* be an independent fresh target normal. Then $p_j = \Pr[S^* > \widehat{\tau} \mid \widehat{\tau}] = 1 - F_t(S_{(j)})$ satisfies*

$$p_j \sim \mathrm{Beta}(k+1-j,\, j), \qquad \mathbb{E}[p_j] = \frac{k+1-j}{k+1}, \qquad \mathrm{Var}[p_j] = \frac{j(k+1-j)}{(k+1)^2(k+2)},$$

depending on k and j alone, independently of F_t, D, and any rank budget.

Proof. Continuity makes $F_t(S_i)$ i.i.d. uniform and F_t is non-decreasing, so $F_t(S_{(j)})$ is the j-th uniform order statistic, which is $\mathrm{Beta}(j, k+1-j)$; hence $p_j \sim \mathrm{Beta}(k+1-j, j)$. Equivalently and without computation: $S^*, S_1, \dots, S_k$ are exchangeable, so the rank of S^* among the $k+1$ values is uniform and $\Pr[S^* > S_{(j)}] = (k+1-j)/(k+1)$. □

Corollary 3 (Attainable grid, calibration floor, validity threshold)**.** *Among deterministic thresholds restricted to one of the k observed order statistics, the attainable expected FPRs from k target normals are exactly $\{i/(k+1)\}_{i=1}^{k}$, so $\min_j |\mathbb{E}[p_j] - \alpha| = \min_i |i/(k+1) - \alpha|$ and no such threshold can do better in absolute signed marginal error. This is not a bound on $\mathbb{E}|p_j - \alpha|$. Validity $\mathbb{E}[p_j] \le \alpha$ needs $j \ge (k+1)(1-\alpha)$, compatible with $j \le k$ only when $k \ge 1/\alpha - 1$. At $\alpha = 0.05$: $k = 19$ attains α exactly, $k = 10$ has floor $|1/11 - 0.05| = 0.0409$, and $k \le 18$ over-rejects in expectation even at the sample maximum. Allowing $+\infty$ yields a valid zero-FPR cutoff at any k, and randomized thresholds can interpolate marginal levels; thus this is a restriction on finite sample-order thresholds, not a universal impossibility.*

Remark 4 (Relation to split conformal prediction, and to the default quantile). Fact 2 is the standard exchangeability argument behind split conformal prediction; here it specifies what our experiments measure and is far sharper here than the concentration bound below. Note that the interpolated sample quantile used by default in most numerical libraries is not rank-based: it mixes two order statistics, so its exceedance probability depends on the shape of F_t between them and none of the above applies.

Fact 3 (Concentration with an empirical-quantile grid term)**.** *Let F_t be continuous, $\widehat{\tau} = S_{(j)}$, and $j = \lceil k(1-\alpha) \rceil$. The Dvoretzky–Kiefer–Wolfowitz inequality gives, with probability at least $1-\delta$,*

$$|1 - F_t(\widehat{\tau}) - \alpha| \le \sqrt{\frac{\log(2/\delta)}{2k}} + \frac{1}{k}.$$

Proof. On the uniform-CDF concentration event, $|F_t(S_{(j)}) - j/k| \le \epsilon$ and $|j/k - (1-\alpha)| < 1/k$. Apply the triangle inequality. For another order index, replace $1/k$ by $|j/k - (1-\alpha)|$. □

Remark 5 (The asymmetry). Estimating the target nuisance *operator* well enough to choose its top-r eigenspace requires $\|\widehat{M}_t - M_t\|_{\mathrm{op}}$ small relative to the eigengap (Theorem 2), and matrix concentration makes that sample cost grow with the effective rank of the tangent distribution. Estimating the operating point requires one scalar quantile, at a cost independent of D and r. The measurements match: at $k = 10$ target normals the subspace error is still 18.5 and geometry-based calibration error 0.145, while threshold recalibration has already reached 0.040. *The invariance problem is high-dimensional; the deployment calibration problem, at one operating point, is one-dimensional.*

A.5 Decision channel

Proposition 5 (Score rate)**.** *For $s(z) = \|P_\perp(z-\mu)\|^2$ and $r = P_\perp(z-\mu)$, along a nuisance path $\dot{s} = 2\langle r, P_\perp U_E \rangle$; for $s = (z-\mu)^\top \Sigma^{-1}(z-\mu)$, $\dot{s} = 2(z-\mu)^\top \Sigma^{-1} U_E$. Both are alignments, not energies.*

Theorem 4 (Operating-point drift)**.** *If $(c,d) \mapsto s(z(c,d))$ is C^1, $S_d = s(z(C,d))$ has a density continuous at τ, and $\dot{s}$ is uniformly integrable near the level set, then $F'(d) = f_{S_d}(\tau)\, \mathbb{E}[\nabla s^\top U_E \mid S_d = \tau]$.*

Theorem 5 (No-crossing certificate)**.** *With $s_0 < \tau$, $m = \tau - s_0$ and $L = \int_0^1 |\dot{s}|\, dt$, we have $|s_t - s_0| \le L$ and hence $L < m$ implies no crossing. Population form:* $\mathrm{FPR}(d_1) \le \mathrm{FPR}(d_0) + \Pr[s_0 \le \tau,\ L \ge m]$.

Proof. $s_t - s_0 = \int_0^t \dot{s}$, so $|s_t - s_0| \le \int_0^1 |\dot{s}| = L$. A sample counted at d_1 but not d_0 must have started below τ and crossed, which requires $L \ge m$. □

Computing L requires the score derivative *along the path to the target*, so this is a path-conditioned certificate, not a source-only deployment guarantee.

Proposition 6 (Flux operator)**.** *With $a = \nabla s$ and threshold weight w_τ, $H_\tau = \mathbb{E}[w_\tau \,\mathrm{sym}(U a^\top)]$ satisfies: the first-order boundary flux removed by erasing* $\mathrm{ran}\, P$ *is* $\mathrm{tr}(PH_\tau)$*, and* $\max_{\mathrm{rank}\, P \le r} \mathrm{tr}(PH_\tau) = \sum_{j \le r} [\lambda_j(H_\tau)]_+$.

Proof. $a^\top PU = \mathrm{tr}(P\,\mathrm{sym}(Ua^\top))$ for symmetric P; take expectations. For the maximisation, $P = QQ^\top$ with $Q^\top Q = I_k$, $k \le r$, gives $\mathrm{tr}(PH) = \mathrm{tr}(Q^\top HQ) \le$ sum of the k largest eigenvalues by Ky Fan; since $k \le r$ is free and H is indefinite, including a negative eigendirection only decreases the objective. □

A.6 Numerical verification

Every result above is checked against an independently computed quantity: gradients against central differences (4×10^{-9}); Theorem 4 against a family with closed-form $F'(d)$ (agreement 0.003, kernel bias $0.009 \to 0.0001$); Proposition 6 against random search over rank-r projectors on indefinite matrices; Theorem 1 and Proposition 3 on 500 and 400 random operator pairs at three ranks with zero violations; Proposition 1 to 10^{-15} with the remainder confirmed PSD. Additional checks cover observation identities, decision properties, and spectral bounds.

B Experimental protocols and reporting conventions

B.1 Datasets, representations, and splits

The observational eraser audit uses Camelyon17-WILDS with Phikon, TCGA tissue-source sites with Prov-GigaPath, and 20 Newsgroups with MiniLM. Natural-image replications use PACS and Office-Home; controlled nuisance studies use NCT-CRC-HE and Tiny-ImageNet with the encoders specified in the corresponding result sections. The Camelyon17 WILDS audit with two target hospitals, the five-hospital leave-one-centre-out study, and the semantic-partition sweeps are distinct protocols. A hospital with no tumour tiles contributes FPR but not TPR. The matched-budget real-domain comparison holds out four PACS styles and five Camelyon17 hospitals in turn; the broader adaptation table aggregates the evaluated domain settings.

B.2 Detector, nuisance measurements, and erasure

The primary score is a squared residual outside a source-fitted principal subspace, with held-out source normals setting the nominal 5% threshold. The controlled detector comparison retains 95% of source variance and also evaluates Mahalanobis and kNN scores. Controlled tangents hold image content fixed along specified nuisance paths. Observational domain contrasts approximate nuisance directions and do not identify all local derivatives. Source erasers use source information only; target-oracle operators use a separate target content split. The Obliviator values are from an incomplete run, as marked in Table 3.

B.3 Target adaptation

Target-normal adaptation examples are disjoint from evaluation. Feature maps act only at target inference; the source detector is held fixed. Compositions separate fitting and calibration examples. The finite rank threshold uses index $\min\{k, \lceil (k+1)(1-\alpha) \rceil\}$; clipping at small budgets produces an empirical cutoff without nominal conformal validity. Mean correction aligns target and source means; diagonal and CORAL corrections additionally align moments. Geometry-all uses the full target adaptation pool in every budget column. Its compositions are all-data controls and do not provide matched-k comparisons with the rank threshold.

B.4 Aggregation and uncertainty

CE_α is computed per evaluation setting and then averaged. A mean FPR near α can coexist with nonzero mean absolute calibration error. Two one-sided tests (TOST) in the eraser audit use an FPR margin of 0.02 and target domains as units. The analytic Beta law concerns random calibration batches and independent future scores, while observed FPR adds finite evaluation-sample uncertainty. Random seeds and semantic partitions quantify within-domain variability and are not interpreted as independent domain-shift replicates; claims about breadth across naturally occurring shifts are therefore based on the number of distinct held-out domains rather than the raw number of evaluation cells.

C Additional results and controls

C.1 Observation and finite-sample estimation

C.1.1 The nuisance tangent spectrum, and why we do not report a rank

Tangent-energy spectrum. Let $M_V = \mathbb{E}[vv^\top]$ over unit-norm tangents. For orthogonal projectors of rank r, $\min_{\mathrm{rank}(P)=r} \mathbb{E}\|(I-P)v\|^2 = \sum_{j>r} \lambda_j$ (write $\mathbb{E}\|(I-P)v\|^2 = \mathrm{tr}(M_V) - \mathrm{tr}(PM_V)$ and apply

Ky Fan), so the best energy any rank-r global eraser can remove is $\mathrm{Cap}^*(r) = \sum_{j \le r} \lambda_j$. Since a mean-based eraser fitted on G source domains has rank $\le G-1$, its residual nuisance energy is at least $1 - \mathrm{Cap}^*(G-1)$: a *spectral capacity law*.

Reporting a single "rank" here would be misleading, and we avoid it. The numerical rank of M_V at tolerance 10^{-10} is 767 (i.e. full), the participation ratio is 29.7 and the stable rank 8.1; the naive energy ranks are $r_{50} = 13$, $r_{90} = 160$, $r_{95} = 278$. The naive energy ranks are noise-inflated. A cross-covariance spectrum between two *disjoint halves* of the tiles contains no additive-noise variance term, and gives the defensible figures $r_{90} = 45$, $r_{95} = 95$ (a random-tangent null gives $r_{90} = 667$). Stencil width matters and is reported: r_{90}^{cross} moves $45 \to 38 \to 30$ as the finite-difference half-width goes $1 \to 2 \to 3$.

C.1.2 The three-gap decomposition

Capacity, observation, and selection. The decisive question is whether a source eraser fails because *no* rank-r subspace could do better, or because it picks a badly oriented one. Let S be the subspace the source data actually reveals, and $\mathrm{Cap}^*_{r|S}$ the best rank-r capture achievable *inside* S (the top-r eigenspace of the compression $B^\top M_V B$). Then for any eraser C,

$$\underbrace{1 - \mathrm{Cap}_{\mathrm{alg}}}_{\text{residual}} = \underbrace{[1 - \mathrm{Cap}^*_r]}_{\text{capacity}} + \underbrace{[\mathrm{Cap}^*_r - \mathrm{Cap}^*_{r|S}]}_{\text{observability}} + \underbrace{[\mathrm{Cap}^*_{r|S} - \mathrm{Cap}_{\mathrm{alg}}]}_{\text{selection}}, \tag{5}$$

an algebraic identity holding for every C. The third term is nonnegative *if* $\mathrm{ran}(C) \subseteq S$ and $\mathrm{rank}(C) \le r$, meaning C is feasible for the problem defining $\mathrm{Cap}^*_{\le r|S}$. This is sufficient but *not* necessary: with $S = \mathrm{span}(e_1)$, $M_V = \mathrm{diag}(1, 0)$ and $C = \mathrm{span}(e_2)$ the term is $+1$ although $C \not\subseteq S$, because an infeasible subspace can be arbitrarily bad as easily as arbitrarily good. Only a *negative* value is diagnostic: it proves the algorithm used information outside S. We return the containment flag and the leakage $\|(I - P_S)C\|$ as data rather than assuming them. The flag detects when an algorithm uses information outside the stated observed span.

Selection within the observed mean span. The G centred domain means span at most $G-1$ dimensions, so a rank-$(G-1)$ budget accommodates all of them and the estimator's subspace *is* S_{mean}, the best rank-r subspace inside S_{mean}. Hence the within-span selection term vanishes identically when the whole observed span is removed (measured: 0.000 at every G). The theorem class is "erasers whose nuisance information enters only through per-domain first moments". LEACE is in it exactly: for a one-hot domain label Z, column g of $\mathrm{Cov}(X, Z)$ is $\pi_g(\mu_g - \bar{\mu})$, so its erasure subspace is fixed by the centred domain means. We do *not* claim it for INLP, whose iterated linear probes coincide with a first-moment contrast only under a Gaussian equal-covariance model, nor for CORAL, MANCE or Obliviator, which use second-order or nonlinear statistics.

		S = domain means		S = source tangents	
G (rank $G-1$)	capacity	OBSERV.	ESTIM.	OBSERV.	ESTIM.
5 (4)	0.701	0.099	0.000	0.000	0.099
9 (8)	0.579	0.125	0.000	0.000	0.125
17 (16)	0.454	0.201	0.000	0.000	0.201
25 (24)	0.379	0.244	0.000	0.000	0.244

The same number appears as selection under one reading of S and as observability under the other. Both are correct, and together they say the mean eraser works with less information.

Here a zero "estimation" term is algebraic relative to the chosen observed span S. It does not say empirical domain means have zero sampling error relative to their population values. LEACE is generally an oblique map in the unwhitened feature metric; identifying its behavior with orthogonal capture requires specifying the whitening or metric used. The first-moment cross-covariance characterization remains valid without that identification.

C.1.3 Matched-cardinality controls for the conditioning gain

But a cardinality control shows the gain is only partly semantic. "Nine tissue classes reveal more than one domain mean because there are nine times as many group summaries" is a complete alternative explanation, so we ran matched controls: partitions with *exactly* the same number and sizes of cells, assigned at random or by permuted class labels. In population these add nothing, since $\mathbb{E}[U \mid D, R] = \mathbb{E}[U \mid D]$ for R independent of the tangent. Out of sample they add a great deal:

component	Δcapture	95% CI	wins
total (class − domain means)	$+0.152$	$[+0.113, +0.187]$	$15/15$
cell-count part (random − domain means)	$+0.090$	$[+0.067, +0.113]$	$15/15$
content part (class − random)	$+0.061$	$[+0.044, +0.078]$	$14/15$

The naive estimates attribute much of the high-rank capture gain to cell count. This cannot be interpreted as population information supplied by a random label. Empirical means fluctuate in the within-cell tangent covariance, and selecting their eigenspaces can exploit that sampling variation. Cross-fitting removes the additive operator bias in expectation. At comparable attainable ranks $r \leq 8$, the cell-count component decreases from $+0.016$ to $+0.001$ (reported $p = 0.98$), while the semantic component increases from $+0.015$ to $+0.033$ ($p = 0.005$). The reported 96% reduction uses unrounded aggregates. A residual random-partition gain beyond rank 16 also reflects the domain channel's limited available dimension. It is not evidence that independent labels add population nuisance information.

C.1.4 The content-cancellation gap, per configuration

Per-configuration content-cancellation gap. ρ_{obs} is the fraction of nuisance tangent energy visible to domain means, so "hidden by averaging" is $1 - \rho_{\text{obs}}$; the last column contrasts the 90%-energy rank of the mean field with that of the tangent field it is meant to represent. The identity $M_{\text{tan}} = M_{\text{mean}} + \mathbb{E}_d[\text{Cov}_c(U|d)]$ was verified numerically to 10^{-15} and the remainder confirmed PSD in every row.

encoder	nuisance	ρ_{obs}	hidden by averaging	r_{90} mean $\to$ tangent
Phikon	stain	0.198	**80.2%**	$3 \to 115$
Phikon	hue	0.323	67.7%	$5 \to 159$
Phikon	blur	0.233	76.7%	$2 \to 156$
DINOv2	stain	0.137	**86.3%**	$2 \to 216$
DINOv2	hue	0.213	78.7%	$3 \to 224$

C.1.5 Encoder and nuisance breadth, environment selection, and coverage

Additional encoders and nuisance paths. We repeat the measurement for a second encoder (DINOv2, no histology supervision) and two nuisances that never touch the H&E decomposition: an RGB channel-gain path and a Gaussian-blur texture path.

encoder	nuisance	mean-field r_{90}	top-2 energy	tangent r_{90}
Phikon	Macenko stain	2	0.915	43
Phikon	RGB hue-gain	3	0.852	22
Phikon	Gaussian blur	1	0.995	59
DINOv2	Macenko stain	2	0.952	112
DINOv2	RGB hue-gain	2	0.948	55

The mean field never exceeds $r_{90} = 3$ and always carries $\geq 85\%$ of its energy in two directions; the ratio to the tangent rank runs from $7.3\times$ to $59\times$. It is *largest* for blur, which is not a colour transform at all.

Selection of source environments. Only under a tight per-anchor budget. With 900 tiles in $D = 768$, a *single* source state's tangent cloud already spans a median of 570 directions, which is why the observability gap under $S =$ tangents is 0.000 everywhere. Restricting each state to its top-m tangent directions makes the question well posed: the spread in observable capture across selection strategies (greedy spectral / equispaced / extremes / random) is 0.073 at $m = 1$ but only 0.009 at $m = 8$. Greedy spectral coverage wins everywhere, but the margin is not a design principle once each environment contributes more than a few resolvable directions.

Global capacity and local coverage. A global eraser pays the spectral price above. A local eraser pays a *covering* price instead: if $z \mapsto P_{V(z)}$ is L_V-Lipschitz and K anchors have covering radius h_K, then using the nearest anchor's tangent space gives $\sup_z \eta_{\text{local}}(z) = \sup_z \|P_{V(z_k)} - P_{V(z)}\|_{\text{op}} \leq L_V h_K$, which we verify to be tight in the controlled world (e.g. $\sup \eta = 0.246$ against $L_V h_K = 0.250$). Capacity and coverage are different currencies, and which one binds depends on the eraser class.

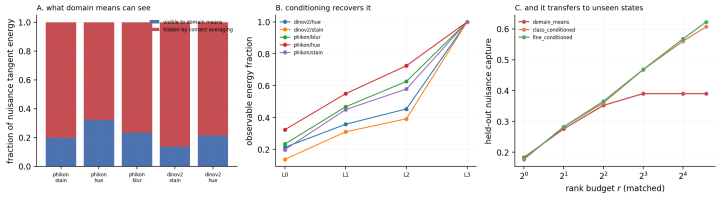


Figure 3: The source observation channel. (A) Domain means reveal only the content-averaged part of the nuisance tangent energy; 68–86% is structurally invisible across two encoders and three nuisance families, and the split is exact (verified to 10^{-15}). (B) Conditioning on content recovers it monotonically, as the Loewner ordering requires. (C) At matched rank budget, a content-conditioned source eraser captures more nuisance energy at *held-out* states; the domain-mean curve is flat beyond $r{=}8$ because its span has only 8 dimensions.

C.1.6 The source observation channel, in three panels

C.1.7 Capacity and nuisance transmission conventions

For a distribution of unit tangents v, $M_V = \mathbb{E}[vv^\top]$ has trace one. The best average capture of a rank-r orthogonal projector is $\sum_{j\le r}\lambda_j(M_V)$, by Ky Fan. Average residual energy is $1-\mathrm{Cap}(P;M_V)$. This average identity must not be confused with worst-direction transmission: for $E=I-P$, $\eta_E^2 = 1-\lambda_{\min}(B^\top PB)$ on a nuisance space with orthonormal basis B. Equality with $1-\mathrm{Cap}$ only holds for a compatible one-direction or worst-direction capture definition. These formulas do not directly apply to an arbitrary oblique eraser.

C.2 Transport and detector consequences

C.2.1 The target-dose curve

Subsampling k target normals and rebuilding the oracle projector, against threshold recalibration given the same budget, on the controlled trajectory. The threshold arm here uses the interpolated percentile; App. C.4.3 reports the same sweep with the rank-based estimator and against the exact prediction.

k target normals	2	5	10	20	50	100
subspace error $\|\sin\Theta\|_F^2$	22.1	20.6	18.5	14.6	8.2	3.7
CE_α, target-geometry projector	0.159	0.150	0.145	0.134	0.127	0.200
CE_α, threshold recalibration	0.149	0.134	**0.040**	0.055	**0.021**	**0.016**

C.2.2 Convex and affine hulls of the source operators

Gate B, testing whether the deployment operator is a re-weighting of the observed source operators. Weights are optimised over the per-state source operators $\{M_1,\dots,M_G\}$ against M_t, on the simplex (convex) and with weights merely summing to one (affine, so the family may extrapolate a trend across states). "% of gap closed" is measured against the pooled-source-to-target capture gap.

subspace built from	capture of M_t	$\|\sin\Theta\|_F^2$	CE_α	% of gap closed
pooled source M_s	0.328	17.92	0.149	n/a
nearest single source state	0.311	18.19	0.153	−1.7%
best *convex* mixture	0.350	17.37	0.142	7.2%
best *affine* combination	0.349	17.40	0.144	7.0%
target operator M_t	**0.752**	0.00	**0.124**	100%

C.2.3 Transport diagnostics and eigengap normalization

Subspace and operator diagnostics. Sweeping source blocks against target states and correlating within configuration×rank:

link	correlation	
coverage distance $h \to \|\sin\Theta\|_F^2$	$+\mathbf{0.70}$	positive association
$\|\sin\Theta\|_F^2 \to \mathcal{T}_r$	$+\mathbf{0.81}$	consistent with the subspace bound
$h \to \mathcal{T}_r$	$+0.56$	
$\|M_t - M_s\|_{\mathrm{op}} \to \|\sin\Theta\|_F^2$	-0.37	negative association in this sweep
δ/γ_r (eigengap-normalised) $\to \|\sin\Theta\|_F^2$	-0.24	normalization does not improve association

Subspace rotation and transport regret are strongly associated in this sweep, with correlation $+0.99$ in a second measurement holding the target fixed. This is consistent with the target-spectrum-weighted bounds, which do not themselves require a particular empirical correlation. Coverage distance is positively associated with rotation, whereas operator-norm mismatch and its eigengap-normalized version are not useful monotone predictors in these data. The operator-norm theorem remains a worst-case upper bound; these measurements do not establish a general causal relation between distance, rotation, and FPR.

C.2.4 Detector behavior after source geometry recovery

Detector behavior after source geometry recovery. We construct a one-class task on the controlled trajectory, using five tissue classes as normals and the remaining classes as semantic anomalies. The detector is fitted on erased source-state normals, its threshold on a held-out source slice, and evaluation uses held-out content and nuisance states. Source states form a contiguous block, so target states require extrapolation. Source FPR is 0.050 in every arm. The "oracle tangents" row below receives raw source tangents; it is not the target-operator oracle in Sec. 6.

	oracle tangents	permuted cond	fine cond	class cond	domain means	domain cov
FPR at unseen states	0.170	0.178	0.179	0.179	0.180	0.187
TPR	0.579	0.582	0.585	0.583	0.585	0.507

All three evaluated detector families (residual, Mahalanobis, and kNN) retain substantial target miscalibration after source geometry recovery. For a residual detector with retained projector P_{ret}, the projected tangent operator is $M_{\mathrm{det}} = (I - P_{\mathrm{ret}})M_{\tan}(I - P_{\mathrm{ret}})$. It contains 30–46% of the reported tangent energy. A rank-24 eraser with capture 0.375 on $M_{\tan}$ captures only 0.070 on M_{det}. Selecting directions using detector-weighted energy increases capture of that operator but does not improve calibration (reported FPR change $+0.0005$, $11/25$ comparisons). Capture has correlations -0.41 and -0.07 with FPR for the two operators.

These measurements show that captured tangent energy alone does not determine the detector's operating point. In particular, residual score changes also depend on alignment with each example's residual and proximity to the threshold. The successful target-oracle control further shows that source information can be insufficient even within a projector class that improves calibration when informed by the target.

C.2.5 Controlled natural-image replication

Replication on natural images. All controlled results above come from one underlying dataset. We rebuilt the identical protocol on Tiny-ImageNet natural photographs with two non-histology encoders (DINOv2, CLIP) and two nuisances, and tested each link separately, since they are logically independent and a mixed outcome is the informative one.

link	natural-image result	
domain means hide most nuisance geometry	50.5–94.5% hidden	✓
and the identity holds numerically	max residual 3×10^{-15}	✓
conditioning is Loewner-monotone	$4/4$ configurations	✓
conditioning improves held-out capture	$4/4$, class $>$ random $>$ domain	✓
energy tracks ΔFPR given distance	partial -0.10	✗
flux adds little beyond distance	partial -0.04	✓
path length beats tangent norm at ranking crossings	0.859 vs 0.561	✓
with zero certificate violations	0 violations	✓
first-order model breaks outside the source range	max rel. err. 1.32	✓

Seven of eight links replicate. The observation bottleneck is at least as large on these natural images as on histology, and the pathwise diagnostic transfers. The exception concerns prediction of FPR change: on natural images neither geometric quantity improves on extrapolation distance.

C.2.6 SCORE SENSITIVITY, BOUNDARY FLUX, AND PREDICTIVE CONTROLS

This section relates nuisance motion to changes in a residual detector's score and FPR. The identities identify relevant quantities but do not by themselves predict aggregate FPR under large extrapolation.

The score rate is an alignment. For a residual detector $s(z) = \|P_\perp(z-\mu)\|^2$ with residual $r = P_\perp(z-\mu)$, the chain rule gives

$$\dot{s} \;=\; 2\,\langle P_\perp(z-\mu),\, P_\perp U\rangle \;=\; 2\,r^\top P_\perp U, \tag{6}$$

an *alignment* between the nuisance tangent and the current residual, rather than $\|P_\perp U\|^2$. A tangent with large residual-space energy orthogonal to r moves the score by nothing; a tiny tangent aligned with r moves it a lot. The relevant length of a nuisance direction for local score change is its length in the decision-induced metric $G_s = \nabla s \nabla s^\top$, rather than its Euclidean norm alone.

The operating point drifts by boundary flux. With $S_d = s(z(C,d))$ and $F(d) = \Pr[S_d > \tau]$, under smoothness and a density continuous at τ,

$$F'(d) \;=\; f_{S_d}(\tau)\,\mathbb{E}\Big[\nabla s^\top U \;\Big|\; S_d = \tau\Big]. \tag{7}$$

Only samples *at* the operating point contribute, through the *signed* projection of nuisance motion onto the score gradient. Neither $M_{\tan}$ nor $M_{\det}$ contains the sign or the distance to threshold, so neither can determine (7). We verify the estimator against a family where $F'(d)$ is closed-form (agreement to 0.003; kernel bias $0.009 \to 0.0001$ as the bandwidth narrows).

Predictive quantities and extrapolation distance. Correlating each quantity with the realised FPR change at held-out states gives raw correlations of -0.48 for nuisance energy, -0.40 for residual energy, and $+0.46$ for boundary flux. Held-out states sit at different distances beyond the source block, and both FPR change and the geometric quantities vary with distance. Controlling for distance reverses the ordering:

quantity	raw corr. with ΔFPR	**partial corr. given distance**
$\mathbb{E}\|U\|^2$ (raw nuisance energy)	-0.48	$\mathbf{+0.80}$
$\mathbb{E}\|P_\perp U\|^2$ (residual energy)	-0.40	$\mathbf{+0.68}$
B_τ (boundary flux)	$+0.46$	-0.11

Once extrapolation distance is held fixed, nuisance energy is *strongly and positively* associated with the FPR change, while boundary flux adds little. In a nested regression pooled over configurations, distance alone gives $R^2 = 0.20$, distance plus residual energy 0.61, and distance plus boundary flux only 0.25. Equation (6) and Theorem 4 describe score sensitivity and boundary flux, but these data do not establish boundary flux as a better empirical predictor than residual energy.

On natural images the picture is weaker still: after controlling for distance, *neither* energy (-0.10) nor flux (-0.04) predicts the FPR change. Across both datasets, extrapolation distance is the most reliable predictor of FPR change. The target-oracle experiment also shows the importance of information from beyond the observed source range.

A pathwise certificate that works per sample. For content c with source score $s_0 < \tau$ and margin $m = \tau - s_0$, the decision path length $L = \int |\dot{s}|$ satisfies $|s_t - s_0| \le L$, so

$$L < m \quad\Longrightarrow\quad \text{this normal never crosses the threshold,} \tag{8}$$

a one-sided guarantee. There are **zero violations** across all configurations, as there must be if the implementation is correct. The fraction of source-accepted normals the certificate can protect varies enormously with the size of the shift, from 0.3% on Phikon/stain, where almost everything crosses, to 88.4% on Phikon/blur. We report the range rather than a mean that would hide it. Median tightness is 0.48, i.e. the bound is roughly $2\times$ loose. Ranking which normals actually become false positives at the extrapolated states:

predictor	AUROC
decision path length L	**0.960**
risk ratio $R = L/m$	0.930
residual tangent norm $\|P_\perp U\|$	0.893
tangent norm $\|U\|$	0.813
margin alone	0.670

A path-conditioned diagnostic, not a source-only predictor. One scope point matters for how (8) may be used. Computing L exactly requires the score derivative *along the path to the target*, which a deployed system does not have at source time. So this is a certificate given the nuisance path, as in our controlled audit, not a source-only deployment guarantee. Whether L admits a useful source-side upper bound is open; the impossibility argument below suggests why it may not.

A detector-weighted operator. Since $a^\top PU = \mathrm{tr}(P\,\mathrm{sym}(Ua^\top))$ for symmetric P, the first-order boundary flux removed by erasing a rank-r subspace is $\mathrm{tr}(PH_\tau)$ with $H_\tau = \mathbb{E}[w_\tau\,\mathrm{sym}(U\nabla s^\top)]$, and $\max_{\mathrm{rank}\le r}\mathrm{tr}(PH_\tau) = \sum_{j\le r}[\lambda_j(H_\tau)]_+$. This has the same shape as our capacity law with $M_{\tan} = \mathbb{E}[UU^\top]$ replaced by an *indefinite* H_τ and a different answer, since the optimum takes only positive eigenvalues and its rank may be less than r.

We tested it, and it does not repair the detector. The flux-optimal subspace removes far more first-order flux than any observation channel (by an order of magnitude) while capturing 0.063 rather than 0.234 of $M_{\tan}$, and its FPR is not better ($3/10$ folds, with no consistent improvement; outcomes are worse far from the source block). Nor is refitting to blame: the same holds with the detector held fixed. The obstruction is the shift size: the measured linearisation error of s already exceeds 100% one step beyond the observed nuisance range, so a first-order correction cannot govern a shift of this size.

C.2.7 Local Taylor models and extrapolation

Measured extrapolation error. If the obstruction is extrapolation, the possible extension is a higher-order local model. For a residual detector the second derivative is exact,

$$\ddot{s} \;=\; \underbrace{2\|P_\perp U\|^2}_{\text{decision curvature}} \;+\; \underbrace{2\langle P_\perp(z-\mu),\,P_\perp A\rangle}_{\text{path acceleration}}, \qquad A = \tfrac{\partial^2 z}{\partial d^2}, \tag{9}$$

which identifies residual nuisance energy $\|P_\perp U\|^2$ as one of two exact second-order terms. Empirically the two terms are large and of opposite sign ($+1746$ against -7295 on Phikon/stain), and adding the quadratic term makes prediction *worse* in 100% of state-configurations. Neither the first- nor the second-order model ever reaches a median relative error below 0.25 at any held-out state we tested ($0/30$). A low-order jet at the source simply does not describe the path.

Why this should be expected. Let the source occupy $d \in [a,b]$ and let $d_\star > b$. Two C^∞ representation maps can agree on $[a,b]$ to all orders by using a bump function supported on (b,∞), while at $d_\star$ one puts a given normal below τ and the other above it. So no source-only method, however much local geometry it uses, can guarantee finite extrapolation without further global regularity. The construction is a standard smooth-extension example; its role is to mark the boundary of our positive results. Local tangent and flux information diagnoses behaviour *near* the source range; extrapolation requires coverage of the deployment path, a global regularity assumption, or target information. The oracle above shows that target information suffices.

C.3 Observational audits and preprocessing

C.3.1 Natural-image semantic-partition aggregates

These are the semantic-partition aggregates used in Figure 1. The evaluation includes 20 random partitions per benchmark and a separate exhaustive PACS sweep. The table summarizes 204 PACS and 240 Office-Home evaluation cells, with bracketed variability summaries where available.

	PACS (4 styles, 204 cells)	Office-Home (4 domains, 240 cells)
linear domain probe	$0.995 \to \mathbf{0.189}$	$0.819 \to \mathbf{0.236}$
MLP domain probe	$0.989 \to 0.957$	$0.859 \to 0.738$
chance	0.344	0.346
η_{src}	**0.0000**	**0.0000**
η_{tgt}	$0.855\ [0.748, 0.943]$	$0.714\ [0.499, 0.959]$
FPR, source $\to$ target	$0.050 \to 0.326$	$0.050 \to 0.113$
FPR change from erasure	$+\mathbf{0.003}\ [-0.017, +0.026]$	$+\mathbf{0.001}\ [-0.005, +0.011]$
$\lvert\text{change}\rvert > 0.02$	17/204	1/240
AUROC	$0.912 \to 0.911$	$0.827 \to 0.827$

Table 3: Erasure audit. **Linear erasure removes the fitted source contrast ($\eta_{\mathrm{src}} = 0.0000$), while target transmission and FPR change little.** MANCE++ provides a nonlinear comparison without strong guardedness. Obliviator* is an incomplete run and is shown descriptively. η_{tgt} measures the observed target nuisance direction; FPR is on unseen-domain normals. The 20 Newsgroups measurements appear in Appendix C.3.3.

		probe		transmission		
bench	method	lin.	MLP	η_{src}	η_{tgt}	FPR
Cam17	none	0.990	0.982	1.000	1.000	0.848
	LEACE	0.163	0.971	**0.0000**	**1.007**	0.855
	MANCE++	0.383	0.690	1.434	41.0	0.855
	Obliviator*	0.314	0.794	0.010	n/a	0.245
TCGA	none	0.837	0.734	1.000	1.000	0.036
	LEACE	0.064	0.103	**0.0000**	0.846	0.033
	MANCE++	0.115	0.101	0.369	13.34	0.040

C.3.2 The erasure audit, in full

C.3.3 Probe, geometry, and detection measurements

The 20 Newsgroups block provides an auxiliary text-domain comparison. The linear probe falls, the fitted source contrast is removed, and target transmission and FPR change little. Because newsgroup identity also carries topic information, this block is not used as sole evidence for presentation-style robustness.

	probe		transmission		
method	lin.	MLP	η_{src}	η_{tgt}	FPR
none	0.796	0.784	1.000	1.000	0.130
LEACE	0.098	0.406	**0.0000**	0.895	0.129
MANCE++	0.253	0.328	0.527	3.673	0.126
Obliviator*	0.114	0.237	0.367	n/a	0.081

Residual nonlinear domain information. On Camelyon17 LEACE takes the linear probe from 0.990 to 0.163, below the $1/3$ chance rate, while the **MLP probe stays at** 0.971. Reporting only the linear probe, as is standard, hides that the representation is not guarded at all against a mildly nonlinear adversary.

Source-contrast removal and target transmission. LEACE attains $\eta_{\mathrm{src}} = 0.0000$: it annihilates the observed nuisance subspace to machine precision. Yet $\eta_{\mathrm{tgt}} = 1.007$ on Camelyon17: nuisance motion at the unseen hospital passes through untouched. By Lemma 1 with $\epsilon = 0$ and a constant Jacobian, only the geometry-mismatch term remains in the upper bound. This does not make it an equality or an FPR attribution. The observed FPR changes little ($0.848 \to 0.855$). Formal equivalence to "no benefit" (two one-sided tests, margin 0.02, unit = target domain) is established for 6 of 14 eraser×benchmark comparisons; the remainder are inconclusive for reasons given in the limitations in Sec. 8, and never favour erasure.

Preservation and detector behavior. LEACE, RLACE and the MANCE family largely preserve the measured complement geometry: $\sigma_H \approx 1$, anomaly AUROC preserved (0.95–0.98), FPR unchanged. A domain-adversarial autoencoder and Obliviator are associated with stronger complement contraction: they lower FPR (Obliviator, Camelyon17: $0.848 \to 0.245$) but collapse the horizontal geometry and the detector's power with it ($\sigma_H^{\min} \approx 0.008$, TPR $0.937 \to 0.262$, AUROC $0.978 \to 0.745$).

C.3.4 DANN sweep

Domain-adversarial results. DANN is the cleanest illustration and also the most easily over-read, so we state what it does and does not show. Across a fixed sweep $\lambda \in \{0, 0.1, 0.3, 1, 3, 10, 30\}$, DANN **never achieves invariance**: the domain probe moves only from 0.989 to 0.978. It therefore cannot test the "surgical" arm of the prediction because there is no regime here in which erasure succeeds without damage. What it does show is that anomaly power collapses at the *smallest* nonzero adversarial weight (AUROC $0.920 \to 0.358$–0.585) while $\sigma_H^{\min}$ falls ($0.034 \to 0.003$) and $\sigma_H^{\max}$ explodes ($159 \to 3686$). The horizontal conditioning $\sigma_H^{\min}/\sigma_H^{\max}$ tracks retained power: Spearman $\rho = 0.544$ (Camelyon17, Holm $p = 0.011$) and $\rho = 0.500$ (TCGA, Holm $p = 0.0013$); $\rho = 0.474$ on 20NG is not significant after correction. The domain probe does not track retained power and on 20NG is *anti*-correlated with power ($\rho = -0.648$, $p = 0.012$): the better the probe collapse, the worse the retained detection. In this sweep, the domain probe does not indicate successful

erasure. The Jacobian diagnostics provide complementary evidence of representational distortion, but do not establish a universal replacement for probe evaluation.

C.3.5 PACS, LEAVE-ONE-STYLE-OUT

GATE D, leave-one-style-out on PACS, averaged over the four folds. Three source styles supply the normals and the fourth is the unseen domain; the detector and its threshold are fitted on source-style normals only, with a held-out slice for the threshold. Semantic anomalies are drawn from the *target* style so that domain and semantics are not confounded in the anomaly set.

	no erasure	first-moment style erasure
domain probe (linear; chance 0.343)	0.986	**0.204**
domain probe (MLP)	0.981	0.944
nuisance transmission η_{src}	1.000	**0.0000**
nuisance transmission η_{tgt} (unseen style)	1.000	**0.823**
FPR, held-out *source*-style normals	0.052	0.052
FPR, unseen-style normals	**0.217**	**0.218**
TPR, semantic anomalies	0.723	0.724

C.3.6 STAIN-NORMALIZATION CONTROL

Our headline histology failure is a hospital shift whose most visible component is stain, so the natural objection is that a nuisance-specific *input* correction, not a latent one, is the answer. Macenko normalisation is the standard such correction. We ran it on raw Camelyon17 tiles, leave-one-centre-out, normalising every tile onto the canonical reference basis and re-extracting Phikon features, so the only difference between conditions is the input.

One implementation note. The Macenko decomposition is ill-conditioned on these patches: they are bright, few pixels clear the optical-density floor, and the two estimated stain vectors come out nearly collinear, so our normaliser falls back to the canonical reference basis when the per-tile estimate is ill-conditioned, which fires on 4.3% of tiles. Normalising onto the canonical reference reduces the between-centre mean-colour distance from 112.05 to 53.09 (centres 0 and 3, 60 tiles each). We also ran the unguarded variant to check what the guard is worth on this statistic, and the corresponding value is 53.05. The guard matters when the reference basis is itself estimated by pooling per-tile stain matrices, where the ill-conditioning compounds; with a fixed canonical reference it does not. Thus the fallback is a numerical safeguard, not an explanation of the normalization effect.

	FPR source	FPR target	CE_α	TPR	AUROC
raw	0.050	0.484	0.434	0.981	0.963
raw + first-moment erasure	0.050	0.481	0.431	0.980	0.965
raw + threshold (k=19)	0.012	**0.056**	**0.038**	0.801	0.963
Macenko-normalised	0.050	0.379	0.329	0.966	0.940
Macenko-normalised + erasure	0.050	0.372	0.322	0.965	0.939
Macenko-normalised + threshold	0.016	0.064	0.050	0.653	0.940

Stain normalisation *helps, partly*: target FPR falls from 0.484 to 0.379 in this experiment. It does not repair the operating point, since 0.379 is still more than seven times nominal. It costs discriminative power (AUROC $0.963 \to 0.940$), and the same fold that goes $1.000 \to 0.710$ under normalisation goes $1.000 \to 0.027$ under a recalibrated threshold, which reaches 0.056 overall. Composing the two is worse than the threshold alone on both axes (CE_α 0.050 versus 0.038, TPR 0.653 versus 0.801), consistent with the decrease in AUROC after normalization. Thus stain normalization partly reduces this shift, but direct threshold adaptation attains lower FPR error in the reported comparison.

C.4 DEPLOYMENT ADAPTATION AND SCOPE

C.4.1 ADAPTATION, SCORE ORDERING, AND ADDITIONAL REPORTED ARMS

The following matched-FPR comparisons separate discrimination from the finite-batch thresholds in Table 1. Empirical monotone transport is weakly increasing and introduces ties, explaining why its reported AUROC and Kendall correlation are not exactly those of the raw score. Feature interventions can alter ordering, but no proportionality between Kendall correlation and power is asserted.

Table 4: Matched-budget adaptation on naturally occurring shift. Nine leave-one-domain-out folds (four PACS visual styles, five Camelyon17 hospitals), five seeds, restricted to the seven folds that are materially miscalibrated ($\mathrm{FPR} > 2\alpha$). Every target-aware arm sees the *same* $k = 19$ unlabeled target normals, from an adaptation split disjoint from evaluation. The table averages only the seven materially miscalibrated folds; the all-nine-fold paired comparison is reported separately below.

on the 7 materially miscalibrated real domains	CE_α	FPR	TPR	AUROC
no correction	0.460	0.510	0.968	0.958
source first-moment erasure	0.466	0.516	0.968	0.959
target geometry, rank 8 from k=19 normals	0.401	0.450	0.962	0.967
threshold, from the same k=19 normals	**0.035**	0.057	0.749	0.958
both	0.035	0.047	0.743	0.967

adaptation	ΔAUROC	Kendall τ vs. raw	ΔTPR at matched FPR 0.05
none	0.0000	1.0000	+0.0000
conformal threshold	**0.0000**	**1.0000**	**+0.0000**
robust threshold	**0.0000**	**1.0000**	**+0.0000**
monotone transport	0.0008	0.9692	+0.0013 ($p = 0.59$)
mean shift	0.0135	0.9431	+0.046 ($p = 0.018$)
diagonal 2nd moment	0.0157	0.9285	+0.059 ($p = 0.002$)
geometry, rank 8	0.0226	0.9200	+0.080 ($p < 10^{-4}$)
CORAL	0.0312	0.8278	**+0.097** ($p < 10^{-3}$)

Two auxiliary arms provide sensitivity comparisons. The "robust threshold" arm has clean-batch CE_α values $0.146, 0.099, 0.068, 0.024, 0.020$ at $k = 5, 10, 19, 100$, and all, with all-data TPR 0.517. The three-parameter "open-set TTA" arm has corresponding values $0.118, 0.313, 0.439, 0.280, 0.251$ and TPR 0.858. These numerical controls are kept separate from the primary comparison of specified interventions.

At the all-data budget, composing mean, diagonal, CORAL, and geometry corrections with thresholds gives CE_α of $0.011, 0.012, 0.012, 0.014$, versus 0.010 for the threshold. TPR gains for CORAL and geometry compositions are $+0.108$ and $+0.071$; these rounded aggregates arise at their fitted thresholds and are distinct from matched-FPR gains above. Fold-level comparisons favor the threshold over geometry on $12/12$ evaluated folds ($p = 5 \times 10^{-4}$).

C.4.2 Real domains at matched target budget

We compare interventions on nine real leave-one-domain-out folds: four PACS styles and five Camelyon17 hospitals. Every target-aware arm uses the same k unlabeled target normals from an adaptation split disjoint from evaluation. Seven folds have unadapted FPR above 2α (Table 4):

A threshold fitted on nineteen target normals gives average FPR 0.057, versus 0.510 without correction, where the same nineteen spent on a rank-8 feature-space correction reach only 0.450 and source erasure reaches 0.516. Across all nine folds, at k=19 the paired difference is $\Delta CE_\alpha = +0.284$, threshold better on $39/45$ domain×seed cells, with the direction of improvement holding at every budget from 5 to 100, visible separately on each benchmark (Camelyon17 $0.570 \to 0.062$, PACS $0.358 \to 0.044$), and harmless on folds that were already calibrated ($0.043 \to 0.050$). The power cost is the one the contamination sweep exposed: TPR $0.968 \to 0.749$.

Across the domain×seed evaluations, the direction of the effect is highly consistent. These cells are not treated as independent realizations of domain shift: repeated seeds quantify split and algorithmic variability within a domain, while the target domain remains the higher-level replication unit. We therefore report the domain×seed count descriptively rather than as an independence-based significance test.

The 0.284 paired improvement averages all nine folds, whereas the table restricts to seven; subtracting the table's calibration errors therefore gives a different number. Small-batch mean FPR does not guarantee conditional calibration for individual adaptation batches (Sec. 7).

C.4.3 The operating-point law across tail levels

The sweep behind Fact 2 and Corollary 3: $\alpha \in \{0.01, 0.05, 0.10\}$ against target budget k, on nine settings (five controlled encoder×nuisance configurations and four PACS leave-one-style-out folds), 300 resamples of the k adaptation normals per cell, averaged over 8 independent adaptation/evaluation partitions. Averaging over

partitions matters: the law is exact for a *fresh* draw from F_t, so conditioning on one fixed evaluation half leaves an $O(1/\sqrt{n_{\text{eval}}})$ deviation that does not average out within a single split.

	$\mathbb{E}[\text{FPR}]$, exact			measured (histology / PACS)		
k	α=.01	α=.05	α=.10	α=.01	α=.05	α=.10
5	0.167	0.167	0.167	0.166 / 0.164	0.164 / 0.164	0.165 / 0.162
10	0.091	0.091	0.091	0.091 / 0.088	0.092 / 0.088	0.091 / 0.089
19	0.050	0.050	0.100	0.053 / 0.046	0.052 / 0.047	0.101 / 0.097
20	0.048	0.048	0.095	0.050 / 0.044	0.049 / 0.045	0.097 / 0.092
50	0.020	0.039	0.098	0.021 / 0.018	0.042 / 0.035	0.099 / 0.095
100	0.010	0.050	0.099	0.011 / 0.008	0.053 / 0.045	0.101 / 0.095

The non-monotonicity in k is the grid effect of Corollary 3, not noise: $\mathbb{E}[\text{FPR}]$ at $\alpha = 0.05$ is 0.0392 at $k = 50$ (conservatively valid grid point $2/51$) and 0.0495 at $k = 100$ ($5/101$), so signed marginal calibration error need not decrease monotonically with k. This does not imply that larger batches help only at particular grid points: the batch-to-batch variance also decreases. The smallest budgets admitting a finite sample-order cutoff with nominal marginal validity are $k \geq 99$, $k \geq 19$ and $k \geq 9$ at $\alpha = 0.01, 0.05, 0.10$.

Comparing estimators on the same scores, the order statistic beats the interpolated percentile on $45/45$ (dataset, α, k) cells, mean $\Delta CE_\alpha = +0.0088$. The gap is largest at small k and small α, because that is where interpolation between two order statistics moves the threshold furthest relative to the spacing of the grid.

C.4.4 Per-content-class calibration, against a per-content oracle

Content-dependent nuisance responses motivate evaluating class-specific FPR. A global threshold can achieve a small aggregate error while leaving heterogeneous class errors. The per-content oracle below is a content-aware empirical comparator, not a universal lower bound. Its worst-class error matches that of the global threshold in this reported experiment; the global threshold does not provide a general conditional-calibration guarantee.

	CE overall	CE mean-class	CE **worst-class**	TPR
no correction	0.162	0.183	0.283	0.642
target geometry	0.124	0.155	0.268	0.611
threshold recalibration	**0.016**	0.056	**0.098**	0.577
per-content *oracle* thresholds	n/a	0.047	0.098	n/a

C.4.5 Contaminated adaptation batches

Unequal-input control. The threshold uses a contaminated batch while the geometry arm receives clean target nuisance information. This control isolates the effect of contaminating threshold data; its constant geometry TPR does not measure robustness to contamination.

contamination ε	0.00	0.02	0.05	0.10	0.20
CE_α, empirical threshold	0.016	0.020	0.018	0.030	0.037
CE_α, robust score fit	0.030	0.029	0.031	0.030	0.032
CE_α, target geometry	0.124	0.124	0.124	0.124	0.124
TPR, empirical threshold	0.577	0.550	0.526	0.380	0.284
TPR, target geometry	0.611	0.611	0.611	0.611	0.611

Comparison with contaminated inputs to every arm. In the shared-pool real-domain comparison, every target-aware arm receives data from the contaminated pool. At nominal $k = 19$ and 20% contamination, the threshold has $CE_\alpha = 0.047$, TPR 0.136; the geometry-all arm has $CE_\alpha = 0.247$, TPR 0.606. Geometry TPR decreases from its clean value 0.746. Geometry-all retains its larger sample budget. The auxiliary robust-threshold and three-parameter TTA arms yield $(CE_\alpha, \text{TPR}) = (0.053, 0.378)$ and $(0.168, 0.741)$, respectively.

C.4.6 Additional negative controls and discarded hypotheses

We evaluated a Gaussian-model FPR estimator and a nuisance-subspace correction. Additional score-based baselines and independent threshold calibration removed their apparent advantages. The controls below are retained to delimit the positive claims.

Gaussian-model FPR estimation. The residual score under a local Gaussian model is a generalised noncentral χ^2, giving a closed-form $\widehat{\mathrm{FPR}}$ from unlabeled target normals. Against a raw exceedance count it appears to win at very small adaptation batches and under contamination. It does not survive score-based alternatives: Jeffreys shrinkage, a lognormal fit to the detector *scores*, and its median/MAD robust variant. None of these score-based alternatives changes feature geometry. At k=5 the certificate is not significantly better than the lognormal fit ($p = 0.151$), and under contamination $\varepsilon \geq 0.05$ the robust fit *beats* it ($+0.008$ MAE $[+0.002, +0.014]$). Local Gaussianity is in any case violated on all three benchmarks. The construction agrees to 10^{-14} with the exact-χ^2 case but does not improve on these score-based alternatives.

Nuisance-subspace correction compared with CORAL. GATE (*Gated Anisotropic Transport of Erasure*) is our own corrector: having estimated a rank-q nuisance subspace, it rescales the target features *only* along those coordinates to match the source second moment, leaving the orthogonal complement untouched. The design corrects only the estimated nuisance coordinates. Its label-free rank rule selects the smallest q that calibrates on a held-out half of the unlabeled batch, with $p_{\mathrm{Holm}} = 0.006$ in the oracle-equivalence test. When each feature map is evaluated with a threshold fitted after its own transform, GATE and CORAL are statistically equivalent (TOST $p = 4 \times 10^{-9}$ on CE_α and 8×10^{-8} on TPR; paired differences $+0.0002$ and $+0.0004$). Under matched contaminated inputs their TPR difference is $+0.001$ ($p = 0.60$). Threshold recalibration remains better calibrated than GATE alone.

Calibration and power after composition. Threshold recalibration is the best-calibrated correction available and gives up 20 points of TPR doing it; *either* feature-space correction recovers that power at equivalent calibration ($+0.205$ TPR, $p = 6 \times 10^{-9}$, for both GATE and CORAL). The defensible statement is about the class of correction, not about a method: shifting the operating point restores calibration but sacrifices power, correcting the feature geometry recovers it, and the two compose.

C.4.7 Scope of the empirical and geometric results

The Camelyon17 eraser audit has two target hospitals, so its measured FPR change is descriptive. The separate threshold comparison holds out each of five hospitals in turn. A hospital without anomaly tiles contributes FPR but not TPR.

Controlled nuisance paths hold content fixed and permit direct tangent measurements. Observational domain contrasts measure only selected directions and cannot certify attenuation of every local nuisance direction. The spectral measurements span two encoders and three applied nuisance families on NCT-CRC-HE; they do not establish the same numerical gap for every naturally occurring shift. Likewise, preservation of the complement $V^\perp$ is a geometric quantity, while preservation of semantic detection requires separate score and TPR measurements.

For a scalar nuisance, $D_d\Phi(c, d)$ is a single vector, so the pointwise nuisance dimension is one wherever the derivative is nonzero. The tangent-population second moment can nevertheless have participation ratio near eight because this direction rotates with content. This distinction between pointwise rank and population span is central to the observation result. The target-oracle projector uses tangent information unavailable to ordinary adaptation, and the order-statistic calibration law assumes clean, independent target normals and a score fixed before calibration.