UNEVEN NORMAL COVERAGE PERTURBS RESIDUAL ANOMALY METRICS: A SPECTRAL MECHANISM STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual anomaly detectors rely on reference statistics from normal data, making global covariance-based scores vulnerable to unequal coverage of valid normal modes. We present a mechanistic study linking uneven normal-reference coverage to residual-metric changes and target-specific false positive inflation. By isolating the metric pathway, we use a two-mode mixture model to separate within-mode covariance shifts from between-mode mean differences, along with an exact common-basis decomposition to attribute score changes to specific spectral directions. In the primary DINOv3 ViT-B/16 representation, positive net attribution is heavily concentrated in leading spectral directions. Controlled interventions—including sample-matched deletions, reweighting, and spectral adjustments—confirm that these shifts stem from composition and spectral location rather than sample count. As a mechanism-inspired probe, Head-Calibrated Mahalanobis (HCM) mitigates conditional false-positive inflation under reduced coverage, demonstrating that the identified failure mode is actionable. Our work provides a precise mechanistic account of coverage sensitivity in residual anomaly detection.

1 INTRODUCTION

Visual anomaly detection commonly estimates a model of normal data and scores deviations using reconstruction residuals, feature distributions, or memory banks (Bergmann et al., 2019; Defard et al., 2021; Roth et al., 2022; You et al., 2022; Zhang et al., 2025). Residual covariance provides a global quadratic metric for anisotropic normal variation: directions with greater reference variance receive lower precision weights. Its behavior therefore depends on the composition of the normal reference used to estimate that covariance.

Normal reference sets can contain multiple valid modes associated with pose, texture, illumination, device state, or acquisition protocol. Uneven representation can shift the global metric so that a lower-coverage normal mode receives higher scores and more false positives at a threshold calibrated on the full reference. We call this conditional effect *majority-reference bias*; aggregate detection metrics alone may not reveal how errors are distributed across normal modes. Our interventions change reference statistics during fitting and calibration, not the training frequency used to learn the frozen representation.

The analysis asks which parts of the metric change increase scores for a fixed target-normal residual. An eigenvalue or covariance distance describes global movement, but does not by itself give the realized score contribution: the latter depends on how the residual aligns with the perturbation and includes cross-direction terms. The general relation is

$$\begin{aligned}\text{uneven normal coverage} &\longrightarrow \text{global residual-metric change}\\ &\longrightarrow \text{direction-dependent target-normal score change}\\ &\longrightarrow \text{conditional FPR inflation.}\end{aligned}$$

In the primary DINOv3 ViT-B/16 representation (Siméoni et al., 2025), fixed-residual attribution locates most positive net contribution in leading directions. Coverage sensitivity was additionally evaluated across DINOv2 ViT-B/14 and DINOv3 ViT-S/16; the available cross-backbone summaries support score and detection comparisons, while the specific leading-direction concentration finding is reported only for the primary representation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Figure 1 summarizes the tested pathway from coverage change to metric–residual alignment, score change, and conditional-FPR inflation. The visual elements are schematic rather than per-rank measurements.

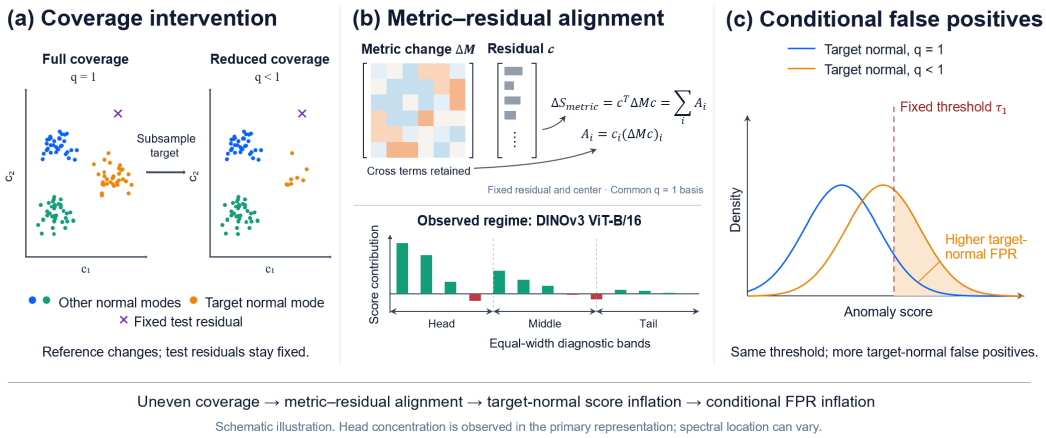


Figure 1: **Mechanism overview (schematic).** Coverage changes the global metric; its interaction with fixed target-normal residuals changes scores and conditional FPR at a fixed threshold. The leading-direction pattern is established only in the primary DINOv3 ViT-B/16 representation. Curves, matrices, and bars are illustrative, not per-rank measurements.

Our contributions are:

1. We characterize how uneven normal-reference coverage changes a global residual metric and can raise false positives for the under-covered normal mode. A finite-mixture covariance decomposition and an exact fixed-residual, common-basis attribution trace this effect from reference composition to the target-normal score while retaining eigenvector rotation and cross-direction contributions.
2. In the primary DINOv3 ViT-B/16 representation, we find that positive score attribution is concentrated in leading directions. Matched controls, additional representations, and cross-dataset tests establish the scope and boundaries of the coverage effect.
3. We use HCM as a mechanism-inspired intervention to reduce target-normal conditional FPR inflation under fixed thresholds, while reporting the accompanying detection cost and the limits of transfer across representations.

2 Related Work

Normal modeling in visual anomaly detection. Visual anomaly detection models normality through several distinct scoring families. Reconstruction and inpainting methods detect deviations from learned normal-image structure (Bergmann et al., 2018; Gong et al., 2019; Akçay et al., 2018; Schlegl et al., 2019; Zavrtanik et al., 2021). Distributional methods fit local feature statistics or density models (Defard et al., 2021; Gudovskiy et al., 2022; Yu et al., 2021), while pretrained-feature and student–teacher approaches compare representations extracted from frozen or adapted networks (Reiss et al., 2021; Deng & Li, 2022; Zhang et al., 2023a; Liu et al., 2023). Memory-based methods instead score samples by their distance to stored normal features or structured normal prototypes (Roth et al., 2022; Hou et al., 2021; Batzner et al., 2024; You et al., 2022). These methods differ in how they represent and score normality. Our study does not propose a new general anomaly-detection architecture. It isolates one narrower question: how the composition of a normal reference set affects a global covariance score applied to a fixed residual representation.

Covariance estimation and spectral scoring. Covariance-based residual monitoring has a long history in multivariate detection. PCA residual statistics provide a classical basis for screening

deviations in correlated variables (Jackson & Mudholkar, 1979). In high dimensions, covariance estimation is sensitive to limited samples and poor conditioning, motivating sparse estimators, shrinkage, and spectral regularization (Bickel & Levina, 2008; Ledoit & Wolf, 2004; Chen et al., 2010; Ledoit & Wolf, 2020; Bun et al., 2017; Marchenko & Pastur, 1967). These works primarily ask how to estimate a covariance matrix more reliably from finite data or how to regularize unstable spectral directions. Our question is different. We keep the representation fixed and deliberately change the relative coverage of valid normal modes. We then ask how this change alters the fitted metric and, for a fixed target-normal residual, how the metric change alters its score. Thus the central object is not covariance estimation error alone, but the conditional score effect induced by a change in reference composition.

Long-tailed anomaly detection and domain variation. Recent anomaly-detection work has also studied imbalanced or heterogeneous normal data. Long-tailed anomaly detection considers imbalance across product classes and has explored class-aware modeling, including learned class names (Ho et al., 2024). TailedCore addresses long-tailed and noisy normal memories through sampling and memory construction (Jung et al., 2025). AeBAD studies anomaly detection under normal-domain changes such as viewpoint and illumination differences between training and testing conditions (Zhang et al., 2023b). Related long-tailed recognition methods address unequal supervised class frequencies through margin adjustment, representation decoupling, logit correction, or calibration (Cao et al., 2019; Kang et al., 2020; Menon et al., 2021; Zhong et al., 2021). These settings establish that imbalance and domain variation can matter, but they study a different source of failure from the one isolated here.

Positioning of this work. Our setting differs from prior long-tail and domain-shift studies in that the imbalance occurs among valid modes within a single normal class, while the test set and frozen representation remain fixed. We study how reference-composition changes enter the fitted residual metric and affect the conditional score of the under-covered mode. This distinguishes our problem from supervised class imbalance, train–test domain shift, and covariance estimation error, while the fixed-residual analysis isolates the metric-mediated pathway.

3 Setup and Coverage Interventions

Let $e_p(x) \in \mathbb{R}^D$ be a frozen feature-reconstruction residual for patch p, $r_p = e_p - \mu$, and

$$\Sigma = U \operatorname{diag}(\lambda_1, \ldots, \lambda_D) U^\top, \qquad \lambda_1 \geq \cdots \geq \lambda_D. \tag{1}$$

Writing $c_{p,i} = u_i^\top r_p$, the quadratic Mahalanobis patch score is

$$S_{\mathrm{Mah}}(p) = r_p^\top \Sigma^{-1} r_p = \sum_i \lambda_i^{-1} c_{p,i}^2. \tag{2}$$

Image scores use the largest 1% of patch scores. We report image AUROC (area under the receiver-operating-characteristic curve), TPR (true-positive rate), pixel AUROC, and AUPRO (area under the per-region-overlap curve) where available. Thresholds are selected from an independent full-coverage normal validation set; test images are never used for calibration in the primary protocol.

For the main MVTec AD study (Bergmann et al., 2019), normal training images are partitioned into three fixed modes using frozen feature embeddings. Each mode is in turn designated as target and retained at $q \in \{1, 0.5, 0.25, 0.1\}$, with five sampling seeds. The encoder, residuals, assignments, test set, and aggregation rule are shared across paired conditions. The $q = 0.1$ toothbrush target-delete split is unavailable because its support is below the preregistered minimum and is never imputed. K-means is an intervention device only; HCM does not require clusters or a known mode count at deployment.

The conditional robustness quantity is

$$\Delta\mathrm{FPR}^{(q)}_{m,t} = \Pr[S^{(q)}_m > \tau^{(1)}_{m,t} \mid g = t] - \Pr[S^{(1)}_m > \tau^{(1)}_{m,t} \mid g = t], \tag{3}$$

where g is the fixed normal-mode assignment. We use fixed-$q = 1$ thresholds for operational stress testing and per-q recalibration only as a diagnostic.

4 THE COVERAGE FAILURE MECHANISM

4.1 WHY COMPOSITION CHANGES A GLOBAL METRIC

For two fixed conditional distributions, target and remainder, let their means and covariances be (μ_t, Σ_t) and (μ_r, Σ_r), and write $d = \mu_t - \mu_r$. At target mixture proportion π, the population covariance is

$$\Sigma(\pi) = \pi\Sigma_t + (1-\pi)\Sigma_r + \pi(1-\pi)dd^\top. \tag{4}$$

If the target mode is retained at fraction q and the remainder is unchanged, its ideal population weight becomes $\pi_q = q\pi/[q\pi + (1-\pi)]$. With $\pi' = \pi_q$, the finite population covariance change is

$$\Delta\Sigma = (\pi' - \pi)(\Sigma_t - \Sigma_r) + [\pi'(1-\pi') - \pi(1-\pi)]dd^\top. \tag{5}$$

The first term reflects within-mode covariance differences and the second reflects the between-mode mean difference. For positive-definite Σ and Σ', the exact fixed-residual quadratic-score change is

$$\Sigma'^{-1} - \Sigma^{-1} = -\Sigma^{-1}(\Sigma' - \Sigma)\Sigma'^{-1}, \qquad \Delta S(r) = -r^\top\Sigma^{-1}\Delta\Sigma\Sigma'^{-1}r. \tag{6}$$

These population identities hold with the two conditional distributions fixed. Finite-sample deletion also introduces sampling variation. The between-mode term in the global covariance remains even when scoring fixes the center at μ_1; holding the center fixed removes center movement from the score comparison, not the between-mode contribution used to fit covariance. When the score uses flooring, ridge, or another stabilization, its effective precision matrices must be used for the score attribution.

4.2 FIXED-RESIDUAL METRIC-ONLY EVIDENCE

The operational intervention changes which normal reference samples estimate the score metric while holding the test membership fixed. The fixed-center, fixed-residual analysis is a separate mechanism-isolation calculation, not a shared assumption of every experiment: it fixes the full-coverage center μ_1, each test residual $r = e - \mu_1$, and the test sample, while q changes the reference covariance and its effective precision matrix M_q (including the scoring stabilization). Let U_1 be the eigenbasis of the effective $q = 1$ precision matrix M_1, ordered by the corresponding reference spectral ranks. The comparison uses $q = 1$ as the common basis rather than separately rotating coordinates at each q:

$$\Delta S_{\text{metric}}(r; q) = r^\top(M_q - M_1)r. \tag{7}$$

The complete effective-metric difference is expressed in the common $q = 1$ basis, retaining eigenvector rotation and cross-block terms. In DINOv3 ViT-B/16, the selected leading 15% band has positive net attribution and exceeds the selected middle and tail 15% bands at the tested coverages; the remaining 55% of ranks are not included in this three-band summary. The decomposition closes to 2.73×10^{-12}. This exact identity applies to the fixed-residual patch score; because the image score reselects the top 1% of patches separately in each condition, it does not establish exact closure for image-level aggregated scores. The corresponding band-level totals are reported in Table 1.

Table 1: Fixed-residual attribution for three selected 15% rank bands in the common $q = 1$ basis. Values are means over available category–target–seed tuples; the final column gives the fraction of tuples with positive head contribution and head contribution greater than both other bands. The unreported 55% is outside this summary.

q	tuples	head net	middle net	tail net	head + / dominant
.50	225	123.28	0.158	0.011	100% / 100%
.25	225	245.30	0.238	0.012	100% / 100%
.10	210	240.19	0.299	0.018	100% / 100%

For an exact direction-level attribution, let $c = U_1^\top r$ and $\Delta M = U_1^\top(M_q - M_1)U_1$. The symmetric contribution assigned to rank i is

$$A_i = c_i^2\Delta M_{ii} + \sum_{j\neq i} c_i c_j \Delta M_{ij}, \qquad \sum_i A_i = \Delta S_{\text{metric}}. \tag{8}$$

This is a score attribution, not a ranking of eigenvalue shifts. Symmetric cross terms are divided equally between their two participating directions. The identity closes for fixed-residual patch scores;

image-level top-1% aggregation can select different patches across conditions and is not covered by this closure. The DINOv3 ViT-B/16 analysis completed 2,700 jobs with maximum absolute closure error 1.46×10^{-10}. Equal-width diagnostic thirds account for 86%, 8%, and 6% of total attribution; separately, the first-15% HCM head and remaining 85% contribute 5,550.17 and 3,404.82. The selected 15% analysis bands, equal-width thirds, and HCM taper are separate reporting schemes. Leading-direction concentration is established here for the primary representation only. Table 2 gives the associated energy, diagonal, cross-direction, and total contributions.

Table 2: Fixed-residual spectral attribution on DINOv3 ViT-B/16. Values are category–target–seed means. The equal-width thirds are a diagnostic partition; the separately reported HCM head covers the first 15%. Cross terms are included and split symmetrically across directions.

Band	residual energy	diagonal	cross	total
Head (equal-width)	9240.80	6883.72	813.26	7696.98
Middle (equal-width)	12.41	502.95	210.62	713.57
Tail (equal-width)	7.78	474.32	70.12	544.44
HCM head (15%)	–	–	–	5550.17
Non-head	–	–	–	3404.82

4.3 Dose response and controls

For Mahalanobis, target-normal score inflation grows from 11.03% at $q = .5$ to 32.69% at $q = .1$, while target $\Delta\mathrm{FPR}$ grows from .267 to .630. Matched random deletion and fixed-support reweighting test whether this increase can be explained by deletion count alone or by changing mixture mass without changing sample identity. Per-coverage recalibration is examined separately as a diagnostic of threshold transfer rather than as primary mechanism evidence. The main dose-response values are collected in Table 3.

Table 3: Coverage dose response for Mahalanobis under the fixed full-coverage threshold. Values are category-block means over available category–target–seed tuples.

q	target score inflation (%)	target $\Delta\mathrm{FPR}$	score specificity
.50	11.03	.267	1.68
.25	20.85	.481	3.87
.10	32.69	.630	6.86

Targeted deletion has a larger target-$\Delta\mathrm{FPR}$ effect than matched random deletion at every tested coverage, and it also exceeds fixed-support reweighting (Table 4). These controls show that the observed effect is not reproduced by removing the same number of arbitrary samples. Reweighting changes the same mixture-mass direction without changing support, but its effect is smaller than actual target deletion.

Table 4: Matched composition controls under the fixed full-coverage threshold. Entries are Targeted $-$ Control target $\Delta\mathrm{FPR}$; positive values mean that targeted deletion has the larger effect.

q	Targeted $-$ Random	Targeted $-$ Reweight
.50	.088	.251
.25	.107	.423
.10	.134	.475

A separate equal-budget intervention changes the same total amount of quadratic weight at different spectral locations. Table 5 shows the absolute target $\Delta\mathrm{FPR}$ values. The tapered leading-rank intervention is the only tested location and shape that consistently lowers target $\Delta\mathrm{FPR}$ relative to Mahalanobis at all three coverage levels; hard leading-rank, middle, tail, and feasible random perturbations remain close to the baseline. This result supports a localized spectral effect without requiring the same individual rank to dominate every condition.

The metric-to-score-to-FPR associations are visualized in Fig. 2.

Table 5: Equal-budget spectral interventions under the fixed $q = 1$ threshold. All variants use the same quadratic-weight L_1 budget; values are category-block mean target ΔFPR.

Method	$q = .50$	$q = .25$	$q = .10$
Mahalanobis	.2659	.4790	.6286
Tapered leading ranks	**.2540**	**.4591**	**.5990**
Hard leading ranks	.2675	.4791	.6294
Tapered middle	.2658	.4789	.6286
Tapered tail	.2659	.4790	.6286
Feasible random	.2723	.4841	.6286

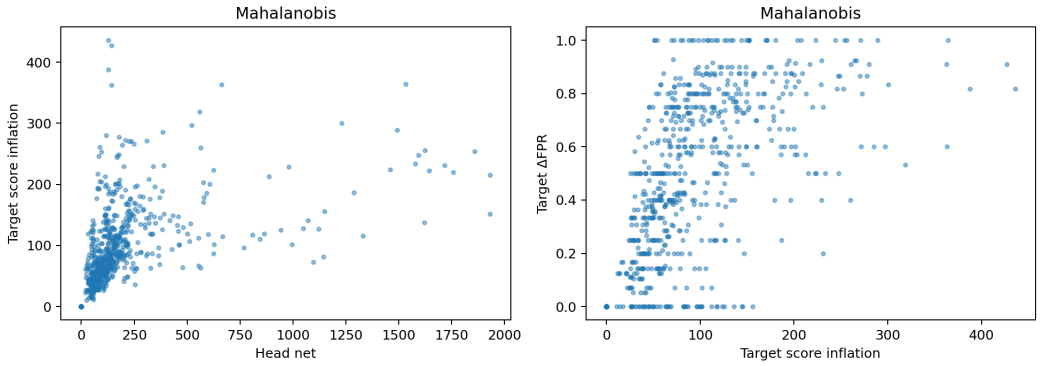


Figure 2: Metric-to-score-to-FPR associations for Mahalanobis under the controlled coverage intervention.

5 Representation Dependence

Coverage sensitivity was additionally evaluated across DINOv2 ViT-B/14 (Oquab et al., 2024) and DINOv3 ViT-S/16. These available cross-backbone summaries support score and detection comparisons; they do not establish transfer of a specific spectral location. The rank-wise concentration result reported here is specific to the primary DINOv3 ViT-B/16 representation.

The paired backbone summary is shown in Table 6.

Table 6: Cross-backbone coverage results. Values are full-coverage image AUROC and target ΔFPR at the indicated coverage. These summaries do not measure spectral-direction attribution.

Backbone	Method	AUROC $q = 1$	ΔFPR(.5)	ΔFPR(.25)	ΔFPR(.1)
DINOv2 ViT-B/14	Mahalanobis	.971	.022	.047	.088
	HCM	.542	.005	.007	.012
DINOv3 ViT-S/16	Mahalanobis	.950	.060	.111	.167
	HCM	.937	.047	.116	.187

Table 6 places the full-coverage detection summaries alongside coverage-conditioned target ΔFPR for each tested backbone. On DINOv3 ViT-S/16, the fixed configuration has higher target ΔFPR than Mahalanobis at $q = .25$ and $q = .1$.

Controlled cross-dataset target-cluster score inflation is summarized in Table 7. These means show coverage sensitivity across the listed residual domains; detection outcomes are reported separately in Appendix H.

The cross-dataset score shifts are descriptive boundary tests rather than evidence for a transferable HCM operating point. Full-coverage detection also varies by domain: for example, image AUROC is .671/.542 for Mahalanobis/HCM on Camelyon16 and .554/.476 on MURA (Appendix H). Thus the recurring result is coverage sensitivity, not a universal detection gain from one fixed intervention.

Table 7: Mean target-cluster score inflation (%) across three target clusters under controlled coverage reduction. Values are descriptive means across target clusters.

Dataset	Method	$q = 0.50$	$q = 0.25$	$q = 0.10$
Camelyon16	Mahalanobis	48.18	105.78	200.99
	HCM	31.53	66.01	104.29
MURA	Mahalanobis	57.97	121.86	209.93
	HCM	14.82	27.95	38.91
PlantVillage	Mahalanobis	37.07	60.48	87.56
	HCM	28.93	49.60	72.58
SDNET2018	Mahalanobis	43.85	101.36	190.69
	HCM	19.06	35.59	50.26

Residual retrieval also shows target-specific FPR inflation. Table 8 gives the minimal seed-0 protocol for 1-NN residual retrieval and a 10% PatchCore residual coreset.

Table 8: Residual-retrieval results under fixed-$q = 1$ thresholds. Values are seed-0 summaries; six toothbrush/$q = 0.1$ combinations are unavailable.

Method	q	Target FPR	Target ΔFPR	TPR	Image AUROC	Pixel AUROC	AUPRO
1-NN Euclidean	1	0.140	0	0.705	0.873	0.746	0.222
PatchCore residual	1	0.140	0	0.706	0.873	0.745	0.222
1-NN Euclidean	0.1	0.394	0.244	0.764	0.849	0.734	0.227
PatchCore residual	0.1	0.393	0.244	0.763	0.849	0.733	0.226

At $q = .1$, both retrieval scores show target $\Delta\mathrm{FPR} = .244$ while their image AUROC changes only modestly. This extends the coverage-sensitivity observation beyond the covariance score, but the spectral attribution developed above remains specific to covariance-based quadratic metrics.

6 Mechanism-Aligned Intervention: HCM

HCM is a mechanism-inspired intervention rather than a replacement anomaly detector. The preceding analysis shows that, in the primary representation, coverage reduction changes the effective metric in a way that gives positive score attribution to leading directions. HCM asks whether moderating this part of the spectrum can reduce the resulting conditional false-positive inflation.

It applies a globally stabilized spectral base and an additional exponent taper to the leading ranks. Let $\bar{\lambda} = D^{-1} \sum_i \lambda_i$ and

$$\tilde{\lambda}_i = (1 - \alpha)\lambda_i + (\alpha + \epsilon)\bar{\lambda}, \quad H = \min\{D, \max\{1, \lceil hD \rceil\}\}. \tag{9}$$

For $1 \le i \le H$ with $H > 1$, let $t_i = [1 - (i-1)/(H-1)]^p$; when $H = 1$, set $t_1 = 1$. The exponent and quadratic score are

$$\begin{aligned} \gamma_i &= \begin{cases} \gamma(1 - \beta t_i), & i \le H, \\ \gamma, & i > H, \end{cases} \\ w_i^{\mathrm{HCM}} &= \tilde{\lambda}_i^{-2\gamma_i}, \qquad S_{\mathrm{HCM}}(r) = \sum_i w_i^{\mathrm{HCM}} c_i^2. \end{aligned} \tag{10}$$

The global parameters γ and ϵ affect all ranks, whereas β, h, and p control the additional leading-rank taper. Non-head weights therefore retain the stabilized base and need not equal raw Mahalanobis weights. Relative to this base, $w_i^{\mathrm{HCM}}/w_i^{\mathrm{base}} = \tilde{\lambda}_i^{2\gamma\beta t_i}$ inside the taper region. Standard Mahalanobis is recovered from the unmodified spectrum with exponent one ($\gamma = .5$); setting $\beta = 0$ while keeping $\gamma = .7$ and $\epsilon = .01$ instead gives the stabilized spectral base, not Mahalanobis. We use the fixed probe configuration $(\gamma, \beta, \epsilon, \alpha, h, p) = (.7, .75, .01, 0, .15, 1)$, which gives $H = 116$ for $D = 768$.

Importantly, HCM modifies both the global spectral base and the leading ranks. Its performance alone therefore cannot identify a head-specific cause. HCM serves a different purpose here—to test whether a mechanism-informed spectral modification can reduce the operational failure at a fixed deployment threshold.

Fixed-threshold robustness. Table 9 reports HCM under thresholds calibrated once at full coverage and then kept fixed as target coverage decreases.

Table 9: Fixed-threshold robustness across coverage levels. Gain is paired Mahalanobis minus HCM target $\Delta\mathrm{FPR}$, averaged before rounding.

q	Mahalanobis $\Delta\mathrm{FPR}$	HCM $\Delta\mathrm{FPR}$	Gain
.50	.267	.131	.136
.25	.481	.258	.223
.10	.630	.320	.311

The reduction in target $\Delta\mathrm{FPR}$ becomes larger as the target reference mode becomes more under-covered: the paired differences grow from .136 at $q = .5$ to .311 at $q = .1$. This pattern is consistent with the mechanism study. When coverage is only mildly reduced, the composition-induced metric change is smaller and there is less failure for the intervention to remove; as the coverage perturbation becomes stronger, the gap between Mahalanobis and HCM increases. HCM therefore changes sensitivity to the coverage perturbation rather than simply applying a constant offset to the operating point.

The same table also shows why this result should not be read as a general improvement in anomaly detection. At $q = .5$, HCM lowers target $\Delta\mathrm{FPR}$ from .267 to .131, but TPR decreases from .970 to .805 and image AUROC from .951 to .906. The gain is therefore a robustness–discrimination trade-off: HCM makes the fixed operating point less sensitive to which normal modes are well represented, but it also weakens separation between normal and anomalous samples. This distinction is central to the role of HCM in this paper. The intervention demonstrates that the identified metric pathway is actionable; it does not establish a uniformly better anomaly score.

Where does the trade-off come from? The component study separates the globally stabilized base from the additional taper. Figure 3 and Table 10 summarize the $q = .5$ results.

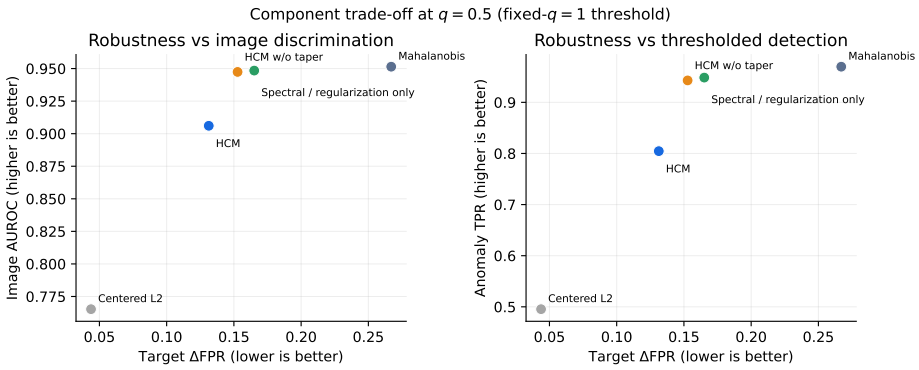


Figure 3: Mechanism-aligned component results at $q = .5$, showing target $\Delta\mathrm{FPR}$ and detection metrics for the tested spectral variants.

Table 10: Selected component results at $q = .5$ under the fixed full-coverage threshold. Lower target $\Delta\mathrm{FPR}$ is shown together with TPR and image AUROC to expose the robustness–detection trade-off.

Method	Target $\Delta\mathrm{FPR}$	TPR	Image AUROC
Mahalanobis	.2670	.9697	.9514
HCM	.1313	.8045	.9061
HCM without taper	.1527	.9429	.9474
Spectral / regularization only	.1651	.9484	.9484
Taper only	.1481	.8735	.9251
Centered ℓ_2	.0438	.4954	.7653

Several points are important. First, the stabilized spectral base already accounts for a substantial reduction in target $\Delta\mathrm{FPR}$: HCM without the taper gives .1527 compared with .2670 for Mahalanobis

while largely preserving TPR and AUROC. The additional taper further reduces $\Delta\mathrm{FPR}$ to .1313, but its incremental effect is smaller than the full Mahalanobis–HCM gap. Thus the full HCM result should not be interpreted as evidence that the leading-rank taper alone explains all robustness gains.

Second, the taper still changes the trade-off in the direction predicted by the mechanism analysis. The taper-only variant reaches target $\Delta\mathrm{FPR} = .1481$, lower than the stabilized spectral-only variant, while the equal-budget experiment shows that moving the same intervention budget to middle or tail ranks produces almost no improvement. Taken together, these results separate two roles: global stabilization provides broad score regularization, while the localized leading-rank intervention tests the spectral location suggested by the fixed-residual attribution.

Third, the centered ℓ_2 result provides a useful failure case for interpreting robustness metrics. Its target $\Delta\mathrm{FPR}$ is only .0438, much lower than HCM, but TPR falls to .4954 and image AUROC to .7653. A score can therefore appear highly "robust" to reference composition simply because it has lost much of its ability to distinguish anomalies. For this reason, target $\Delta\mathrm{FPR}$ must always be interpreted together with anomaly-detection performance.

What changes after threshold recalibration? Table 11 compares covariance estimators when each method is allowed to recalibrate its threshold at the reduced-coverage condition.

Table 11: Alternative covariance models at $q = 0.5$ with recalibrated thresholds. Values are category-block means.

Method	target FPR	FPR gap	TPR	image AUROC	AUPRO
Mahalanobis	.193	.048	.893	.951	.325
HCM	.194	.065	.801	.925	.295
Cluster-shared	.128	.010	.567	.806	.212
Gaussian-mixture OAS	.075	−.049	.560	.804	.212
Diagonal covariance	.147	.063	.532	.794	.213
Ledoit–Wolf	.148	.060	.504	.767	.208
Ridge ($\alpha = .01$)	.143	.056	.499	.768	.209

With per-coverage recalibration, Mahalanobis and HCM have nearly identical target FPR at $q = .5$ (.193 versus .194), while Mahalanobis retains higher TPR and image AUROC. Other covariance estimators lower target FPR only with larger losses in anomaly discrimination. Thus, HCM's main benefit lies in the fixed-threshold setting, where a previously calibrated operating point must be reused after reference composition changes.

Recalibration reduces the resulting FPR inflation but does not remove the underlying metric and score shift. HCM therefore serves as a limited mechanism-aligned intervention rather than a generally superior anomaly score.

7 Conclusion

Uneven normal coverage can change a global residual metric in a way that selectively increases the score of an under-covered normal mode. Our fixed-residual analysis shows that this effect depends on the alignment between the composition-induced metric change and target-normal residual directions. In the primary DINOv3 ViT-B/16 representation, the positive score attribution is concentrated in leading spectral directions.

HCM provides a mechanism-inspired intervention for this failure mode. It reduces conditional FPR inflation under the reported fixed-threshold settings, but also incurs a detection cost and is not a universal replacement for Mahalanobis.

The broader implication is that aggregate anomaly-detection performance can hide conditional failures caused by reference composition. Evaluating how valid normal modes are represented, and tracing their effect through the fitted score, can reveal deployment risks that aggregate metrics alone may miss.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading, improving the clarity and readability of the manuscript, and assisting with the writing, debugging, and refinement of code. We did not use generative AI tools for generating research ideas, designing algorithms, conducting experiments, analyzing results, or producing scientific claims. All AI-assisted text modifications and code were carefully reviewed, verified, and tested by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Samet Akçay, Amir Atapour-Abarghouei, and Toby Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *Proceedings of the Asian Conference on Computer Vision*, 2018.

Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 127–137, 2024. doi: 10.1109/WACV57701.2024.00020.

Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9584–9592, 2019. doi: 10.1109/CVPR.2019.00982.

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008. doi: 10.1214/009053607000000758.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017. doi: 10.1016/j.physrep.2016.10.005.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, pp. 475–489. Springer, 2021.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9727–9736, 2022. doi: 10.1109/CVPR52688.2022.00951.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019. doi: 10.1109/ICCV.2019.00179.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1819–1828, 2022. doi: 10.1109/WACV51458.2022.00188.

Chih-Hui Ho, Kuan-Chuan Peng, and Nuno Vasconcelos. Long-tailed anomaly detection with learnable class names. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12435–12446, 2024.

Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8771–8780, 2021. doi: 10.1109/ICCV48922.2021.00867.

J. Edward Jackson and Govind S. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.

Yoon Gyo Jung, Jaewoo Park, Jaeho Yoon, Kuan-Chuan Peng, Wonchul Kim, Andrew Beng Jin Teoh, and Octavia Camps. TailedCore: Few-shot sampling for unsupervised long-tail noisy anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25539–25548, 2025.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Olivier Ledoit and Michael Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065, 2020. doi: 10.1214/19-AOS1921.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023. doi: 10.1109/CVPR52729.2023.01954.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457–483, 1967.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2805–2813, 2021. doi: 10.1109/CVPR46437.2021.00283.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14298–14308, 2022. doi: 10.1109/CVPR52688.2022.01392.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Margarethe Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Lucas Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4571–4584, 2022.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. FastFlow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. doi: 10.1016/j.patcog.2020.107706.

Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain ViT features for multi-class unsupervised visual anomaly detection. *Computer Vision and Image Understanding*, 253:104308, 2025. doi: 10.1016/j.cviu.2025.104308.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. DeSTSeg: Segmentation guided denoising student–teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3914–3923, 2023a. doi: 10.1109/CVPR52729.2023.00381.

Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 151:103990, 2023b. doi: 10.1016/j.compind.2023.103990.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16489–16498, 2021.

A Protocol and Result Provenance

Paired identity. All paired comparisons share the category, target cluster, seed, coverage split, residual cache, test membership, and patch-to-image aggregation. Mahalanobis and HCM use separate method-specific thresholds selected from the independent full-coverage validation set. Target $\Delta\mathrm{FPR}$ always subtracts the $q = 1$ target FPR for the same method, category, target, and seed.

Spectral identity. The fixed-residual analysis projects the effective metric difference into the saved $q = 1$ eigensystem and retains eigenvector rotation and cross-block terms. The selected leading, middle, and tail bands each contain 15% of ordered directions; the remaining 55% is not part of that band summary. A separate equal-width thirds analysis partitions all ranks, and HCM's additional taper acts on its first 15%. The maximum saved closure error for the metric decomposition is 2.73×10^{-12}.

Inferential summaries. A separate holdout reanalysis produces different point estimates and category-block intervals and is therefore not combined with the selected primary result set. Because the paired category–target–seed rows for the selected primary summary are unavailable, we report descriptive point estimates and omit uncertainty intervals that cannot be reconstructed for the same result set.

Availability. At $q = 0.1$, the three toothbrush target-delete splits are absent because target support is below the preregistered minimum. Such combinations are marked unavailable and excluded from numerical means. Hence $q = 0.1$ has 14 categories, whereas $q \in \{1, 0.5, 0.25\}$ has 15. Other experiment-specific infeasibilities are documented below rather than converted to zeros.

HCM configuration identity. The fixed probe uses $\gamma = 0.7$, $\beta = 0.75$, $\epsilon = 0.01$, $\alpha = 0$, $h = 0.15$, and $p = 1$. The implementation stabilizes the full spectrum using $\tilde{\lambda}_i = (1 - \alpha)\lambda_i + (\alpha + \epsilon)\bar{\lambda}$. It applies $\gamma_i = \gamma(1 - \beta t_i)$ to the first $H = \lceil hD \rceil$ ranks, clamped to $[1, D]$, and uses $\gamma_i = \gamma$ elsewhere. Non-head coefficients equal the globally stabilized base rather than raw Mahalanobis, and HCM remains active at $q = 1$. Equal-budget tapered-head is a separate control that retains Mahalanobis coefficients outside its intervention band.

Score convention. Patch scores are quadratic energies $S(r) = \sum_i w_i c_i^2$. Image scores aggregate the largest 1% of these patch scores.

Semantic grouping record. The illumination- and texture-grouped tables are retained as exploratory results only. Their grouping and cutoff procedures are not fully documented, including the rules used to set bins and apply them across splits. We therefore do not use these summaries to support semantic-generalization claims or claims of independence from anomaly labels or test information.

B Controls for the Coverage Mechanism

B.1 Targeted deletion versus matched random deletion

Table 12 reports Targeted − Random target $\Delta\mathrm{FPR}$ under the fixed full-coverage threshold. Positive values mean target deletion produces a larger target-$\Delta\mathrm{FPR}$ change than a matched random deletion.

Table 12: Targeted − matched-random target $\Delta\mathrm{FPR}$ under the fixed full-coverage threshold. Positive values indicate a larger effect for targeted deletion; point estimates are descriptive.

Method	q	Mean effect
Mahalanobis	0.50	0.088
Mahalanobis	0.25	0.107
Mahalanobis	0.10	0.134

B.2 FIXED-SUPPORT REWEIGHTING

Fixed-support weighting changes cluster mass without changing sample identity. Table 13 gives the descriptive Targeted − Reweight contrast for Mahalanobis under the fixed threshold.

Table 13: Targeted − fixed-support Reweight target $\Delta\mathrm{FPR}$ for Mahalanobis under the fixed full-coverage threshold. Positive values indicate a larger effect for deletion; point estimates are descriptive.

Method	q	Fixed-threshold effect
Mahalanobis	0.50	0.251
Mahalanobis	0.25	0.423
Mahalanobis	0.10	0.475

B.3 THRESHOLD PROTOCOL

Per-coverage recalibration was evaluated as a diagnostic of threshold transfer by recomputing method-specific thresholds from validation scores at every q. Quantitative recalibration results are omitted because matching paired rows for the selected primary result set are unavailable.

C EQUAL-BUDGET SPECTRAL INTERVENTION

All perturbations in this intervention study have the same quadratic-weight L_1 budget

$$B = \sum_i |w_i - w_i^{\mathrm{Mah}}|. \tag{11}$$

The head intervention uses a tapered reduction; the middle and tail interventions use the same normalized shape at different locations; hard-head uses a constant multiplier. Table 14 shows that only the tapered-head intervention consistently departs from Mahalanobis. It is called *equal-budget tapered-head*: its non-head weights remain Mahalanobis weights, so it is not HCM.

Table 14: Fixed-$q = 1$ target $\Delta\mathrm{FPR}$ for equal-budget interventions. Values are category-block means; "Random" is conditional on budget feasibility.

Method	$q = 0.50$	$q = 0.25$	$q = 0.10$
Mahalanobis	0.2659	0.4790	0.6286
Equal-budget tapered-head	**0.2540**	**0.4591**	**0.5990**
Hard head	0.2675	0.4791	0.6294
Tapered middle	0.2658	0.4789	0.6286
Tapered tail	0.2659	0.4790	0.6286
Feasible random	0.2723	0.4841	0.6286

The grid contains 27,000 requested result combinations: 26,028 are feasible, 450 correspond to protocol-defined unavailable source splits, and 522 random-direction combinations are infeasible under the simultaneous equal-budget and strict-positive-weight constraints. Among random draws with available source splits, 8,589/8,850 (97.05%) are feasible. Random results are therefore a conditional control rather than an unconditional random baseline.

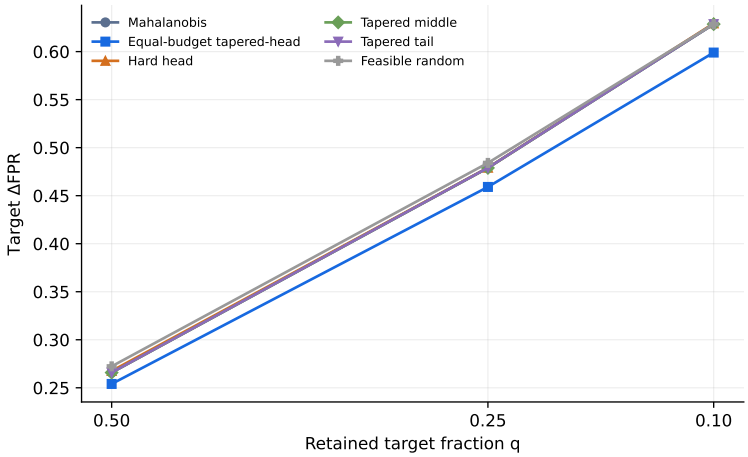


Figure 4: Target ΔFPR for equal-budget spectral interventions under the fixed $q = 1$ threshold; values are category-block means.

D Alternative Metrics and Covariance Models

This comparison uses per-coverage recalibrated thresholds and reports the FPR and detection metrics for each covariance baseline.

Table 15: Alternative covariance models at $q = 0.5$ with recalibrated thresholds. Values are category-block means.

Method	Target FPR	FPR gap	TPR	Image AUROC	AUPRO
Mahalanobis	.193	.048	.893	.951	.325
HCM	.194	.065	.801	.925	.295
Cluster-shared	.128	.010	.567	.806	.212
Gaussian-mixture OAS	.075	−.049	.560	.804	.212
Diagonal covariance	.147	.063	.532	.794	.213
Ledoit–Wolf	.148	.060	.504	.767	.208
Ridge ($\alpha = .01$)	.143	.056	.499	.768	.209

E Cluster-Count Sensitivity

The minimal profile uses seed 0, $q \in \{1, 0.1\}$, a fixed-$q = 1$ threshold, and all targets for $K \in \{2, 3, 4, 5\}$. Assignments, calibrators, scores, and thresholds are computed separately for each K. Table 16 shows how the intervention partition changes conditional results. Here K defines the coverage intervention and group-wise evaluation; HCM does not use a cluster count at scoring time.

Table 16: Cluster-count sensitivity at $q = 0.1$ under the fixed $q = 1$ threshold. Values are category–target macro means; unavailable target/test combinations remain missing.

K	Mah. ΔFPR	HCM ΔFPR	Gain	HCM TPR	HCM AUROC
2	0.933	0.928	0.005	1.000	0.311
3	0.640	0.366	**0.274**	0.922	**0.913**
4	0.944	0.869	0.075	0.919	0.317
5	0.878	0.784	0.094	0.829	0.336

Six toothbrush/$K = 3$/$q = 0.1$ rows are unavailable by the source-split protocol. Twenty additional target metrics are unavailable because the relevant target cluster has no normal test member; these values remain missing rather than being set to zero.

F HCM Parameter Sensitivity

The HCM configuration is $(\gamma, \beta, \epsilon, \alpha, h, p) = (0.7, 0.75, 0.01, 0, 0.15, 1)$. For the main representation $D = 768$ and $H = \lceil 0.15D \rceil = 116$. These values are held fixed across categories, targets, coverage levels, and seeds. The summarized sensitivity grid contains 27,000 rows: 26,550 complete and 450 unavailable toothbrush/$q = 0.1$ rows. It describes parameter sensitivity under the listed protocol.

Table 17: Parameter sensitivity at $q = 0.1$ with recalibrated thresholds. Entries are variant minus HCM category-block means; a positive target-ΔFPR difference is worse, whereas positive detection differences are better.

Parameter	Value	Target ΔFPR	TPR	Image AUROC	AUPRO
β	0	−0.0212	0.0686	0.0204	0.0125
	0.25	−0.0227	0.0677	0.0203	0.0124
	0.50	−0.0173	0.0607	0.0186	0.0106
	0.75 (default)	0	0	0	0
	1.0	−0.0013	−0.1626	−0.0894	−0.0381
h	0.05	0.0093	0.0038	0.0007	0.0024
	0.10	−0.0018	0.0014	0.0000	0.0001
	0.15 (default)	0	0	0	0
	0.20	0.0066	−0.0005	0.0001	0.0007
	0.30	0.0052	−0.0013	−0.0001	0.0004
p	0.5	0.0018	−0.0009	−0.0003	−0.0003
	1.0 (default)	0	0	0	0
	2.0	0.0029	0.0021	0.0004	0.0011
ϵ	0.001	−0.0159	0.0776	0.0214	0.0215
	0.01 (default)	0	0	0	0
	0.10	0.0496	−0.1380	−0.0695	−0.0461

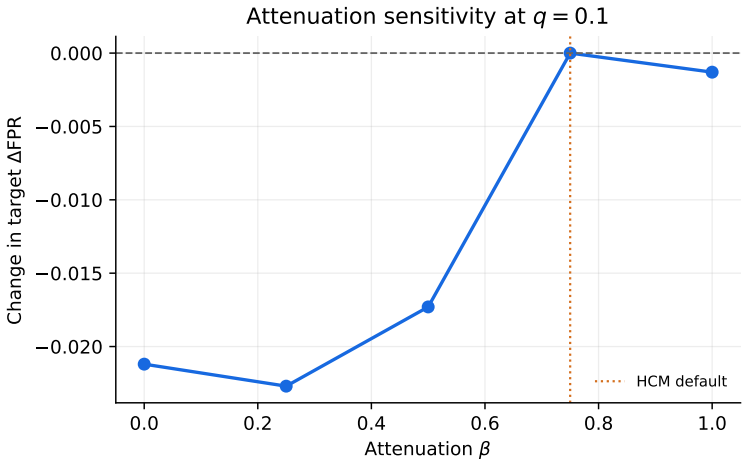


Figure 5: Attenuation sensitivity at $q = 0.1$ with recalibrated thresholds. Entries are variant-minus-default point estimates for target ΔFPR and detection metrics.

G Component Ablation

Table 18 lists the target ΔFPR and detection metrics for the tested HCM component variants at $q = 0.5$. These are descriptive component results.

Table 18: Component comparison at $q = 0.5$ under the fixed full-coverage threshold. Values are category-block means. "Spectral only" and "Regularization only" share the stabilized-base path; "HCM without regularization" and "Taper only" share the taper-only path; "HCM without spectral reweighting" and HCM use the same implemented head modulation. Identical rows follow from these component definitions.

Method	Target $\Delta\mathrm{FPR}$	TPR	Image AUROC	Pixel AUROC	AUPRO
Mahalanobis	0.2670	0.9697	0.9514	0.8030	0.3251
HCM	0.1313	0.8045	0.9061	0.7758	0.2761
Spectral only	0.1651	0.9484	0.9484	0.7999	0.3130
Taper only	0.1481	0.8735	0.9251	0.7875	0.2949
Regularization only	0.1651	0.9484	0.9484	0.7999	0.3130
HCM without taper	0.1527	0.9429	0.9474	0.7963	0.3035
HCM without regularization	0.1481	0.8735	0.9251	0.7875	0.2949
HCM without spectral reweighting	0.1313	0.8045	0.9061	0.7758	0.2761
Centered ℓ_2	0.0438	0.4954	0.7653	0.7114	0.2097

H Controlled Cross-Dataset Results

The main text reports controlled target-cluster score inflation at each tested coverage. This appendix provides the available detection summaries for the same four datasets.

Table 19: Cross-dataset top-1% full-coverage detection and $q = 0.1$ target-FPR inflation where available. Values are descriptive means; a dash indicates that image AUROC or target FPR was not defined for that dataset.

Dataset	Method	AUROC ($q = 1$)	TPR ($q = 1$)	Target $\Delta\mathrm{FPR}$ ($q = 0.1$)
Camelyon16	Mahalanobis	0.671	0.053	0.155
	HCM	0.542	0.017	0.094
MURA	Mahalanobis	0.554	0.110	0.180
	HCM	0.476	0.053	0.069
PlantVillage	Mahalanobis	–	0.888	–
	HCM	–	0.875	–
SDNET2018	Mahalanobis	–	0.139	–
	HCM	–	0.062	–

Across the four datasets, controlled coverage reduction is accompanied by target-cluster score inflation. The accompanying detection summaries show that outcomes vary across residual domains and are reported alongside the score changes.

I Cross-Backbone Results

Coverage sensitivity was additionally evaluated across DINOv2 ViT-B/14 and DINOv3 ViT-S/16. Table 20 summarizes conditional-FPR changes and full-coverage AUROC. The exact leading-direction concentration result reported in the main analysis remains specific to DINOv3 ViT-B/16; no cross-backbone spectral-location claim is made.

Table 20: Cross-backbone top-1% image scoring. Values are target $\Delta\mathrm{FPR}$ at the indicated coverages; AUROC is shown at full coverage.

Backbone	Method	AUROC ($q = 1$)	$q = 0.5$	$q = 0.25$	$q = 0.1$
DINOv2 ViT-B/14	Mahalanobis	0.971	0.022	0.047	0.088
	HCM	0.542	0.005	0.007	0.012
DINOv3 ViT-S/16	Mahalanobis	0.950	0.060	0.111	0.167
	HCM	0.937	0.047	0.116	0.187

The cross-backbone table places full-coverage detection metrics beside target $\Delta\mathrm{FPR}$ for each representation. On DINOv3 ViT-S/16, HCM target $\Delta\mathrm{FPR}$ is lower at $q = .5$ and higher than Mahalanobis at $q = .25$ and .1. These results are used as representation-level boundary tests, not as evidence that one spectral location transfers across backbones.

J Primary Fixed-Threshold Results

The primary paired robustness gains are .136, .223, and .311 at $q = .50, .25, .10$, averaged over available category–target–seed pairs before rounding. Detection metrics for the same primary protocol appear alongside conditional-FPR results in Table 9. Category-level uncertainty intervals are omitted because the matching paired rows are unavailable for this selected result set.

K Exploratory Semantic-Group Records

These tables report the available illumination- and texture-grouped results. Because the grouping procedure is not fully documented (Appendix A), they are treated as exploratory and are not used as evidence for a core semantic-generalization claim. No category-level significance claim is made.

Table 21: Global spectral changes under semantic target deletion. Values are paired distances relative to full coverage; illumination covers 15 categories and texture covers five categories. Per-q effective condition counts are unavailable. Descriptive summaries only.

Mode	q	D_{spec}	D_{cov}	target support	retained energy
Illumination	.50	.0788	.1029	.9796	90.67%
Illumination	.25	.1268	.1449	.9603	90.67%
Illumination	.10	.1485	.1328	.9920	90.61%
Texture	.50	.0963	.0768	1.0003	90.73%
Texture	.25	.1572	.1066	.9992	90.73%
Texture	.10	.2004	.1314	1.0076	90.73%

The available outputs include spectral distances, eigenvalue shifts, and projected variances, but not fixed-residual direction attribution or semantic-group target $\Delta\mathrm{FPR}$. Uniform deletion produces larger global spectral distances than target deletion (for example, illumination $D_{\mathrm{spec}} = 1.2791$ versus .1485 at $q = .1$), while fixed-support weighting produces a smaller change (.0481). This illustrates why a global covariance distance alone does not measure the target-specific score effect.

L Fixed-Residual Direction Attribution

For each category, target, seed, and coverage, the full-coverage center and test residuals are held fixed. Let U_1 be the common $q = 1$ basis, $c = U_1^\top r$, and $\Delta M = U_1^\top (M_q - M_1) U_1$. The symmetric rank attribution

$$A_i = c_i^2 \Delta M_{ii} + \sum_{j \neq i} c_i c_j \Delta M_{ij} \tag{12}$$

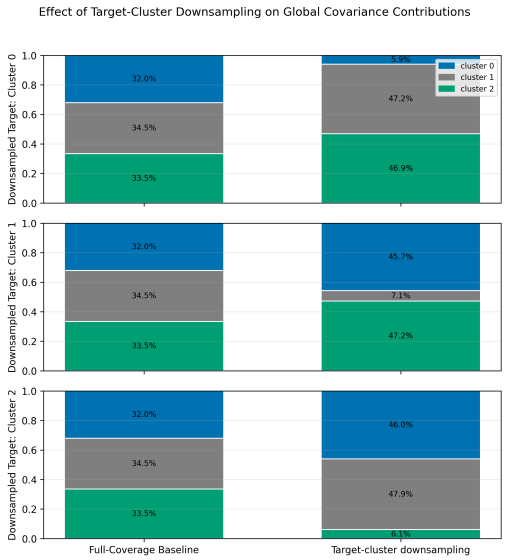
includes the cross terms and satisfies $\sum_i A_i = c^\top \Delta M c$. The DINOv3 ViT-B/16 run completed 2,700 jobs; 60 toothbrush split rows are unavailable by the source-split protocol. The maximum absolute and relative closure errors are 1.46×10^{-10} and 1.55×10^{-14}, respectively.

Table 22: Attribution totals for the fixed-residual analysis. Values are category–target–seed means; equal-width diagnostic bands include diagonal and cross-direction terms, while the HCM head uses the first 15% of ranks.

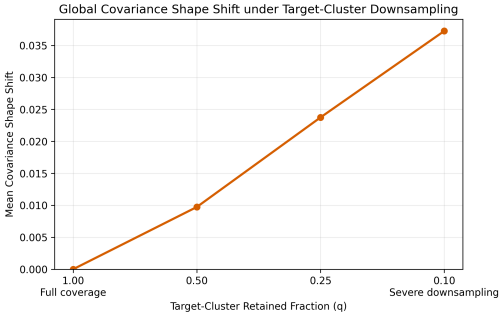
Band	energy	diagonal	cross	total	share
Head (equal-width third)	9240.80	6883.72	813.26	7696.98	86%
Middle (equal-width third)	12.41	502.95	210.62	713.57	8%
Tail (equal-width third)	7.78	474.32	70.12	544.44	6%
HCM head (first 15%)	–	–	–	5550.17	–
Non-head	–	–	–	3404.82	–

The equal-width thirds partition all directions for diagnosis. The selected 15% bands in the metric-only summary leave 55% of directions unreported, and the HCM taper is a separate first-15% intervention. In this DINOv3 ViT-B/16 attribution, net contribution is concentrated in the leading equal-width third; this finding does not establish the same spectral location in other representations.

M ADDITIONAL MECHANISM VISUALIZATIONS



(a) Cluster covariance contributions.



(b) Global covariance-shape shift.

Figure 6: Reducing one normal cluster changes mixture contributions and the global covariance shape.

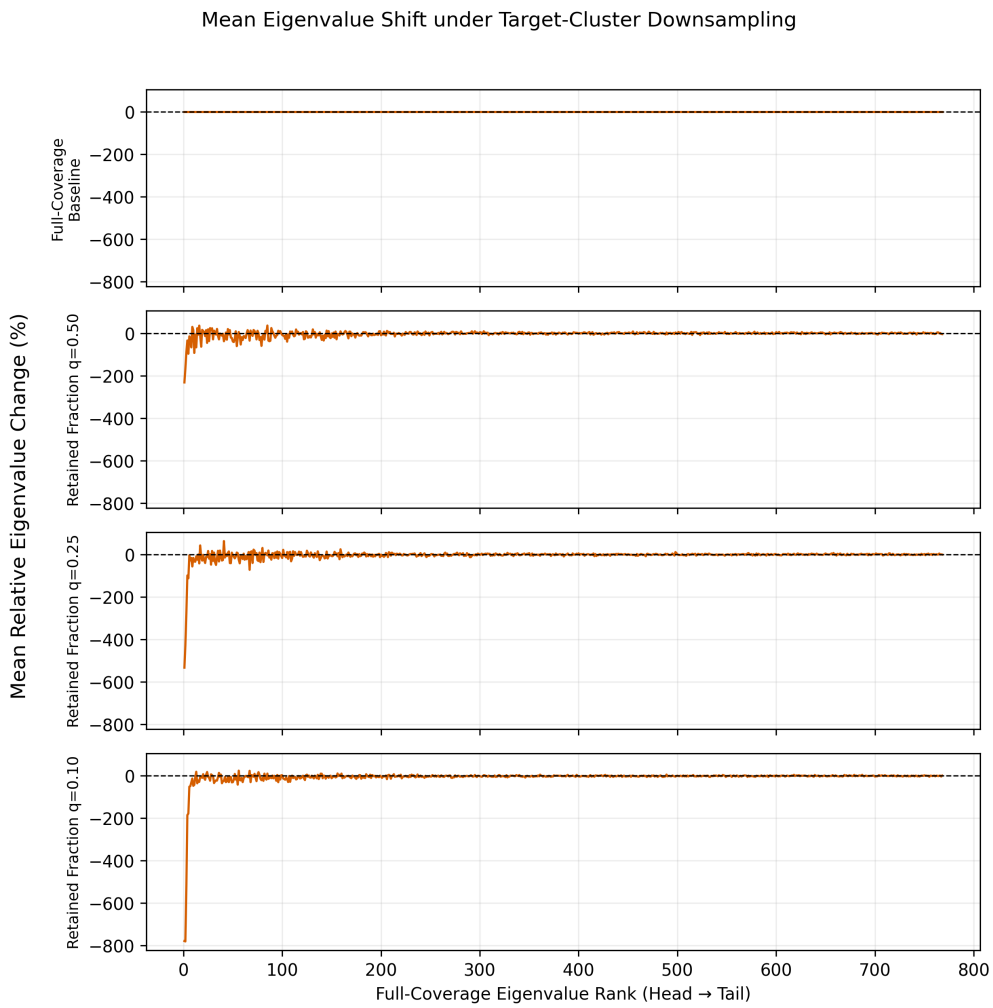


Figure 7: Ordered eigenvalue changes under target-cluster downsampling. This visualization is descriptive; the formal mechanism evidence comes from the fixed-residual metric decomposition.

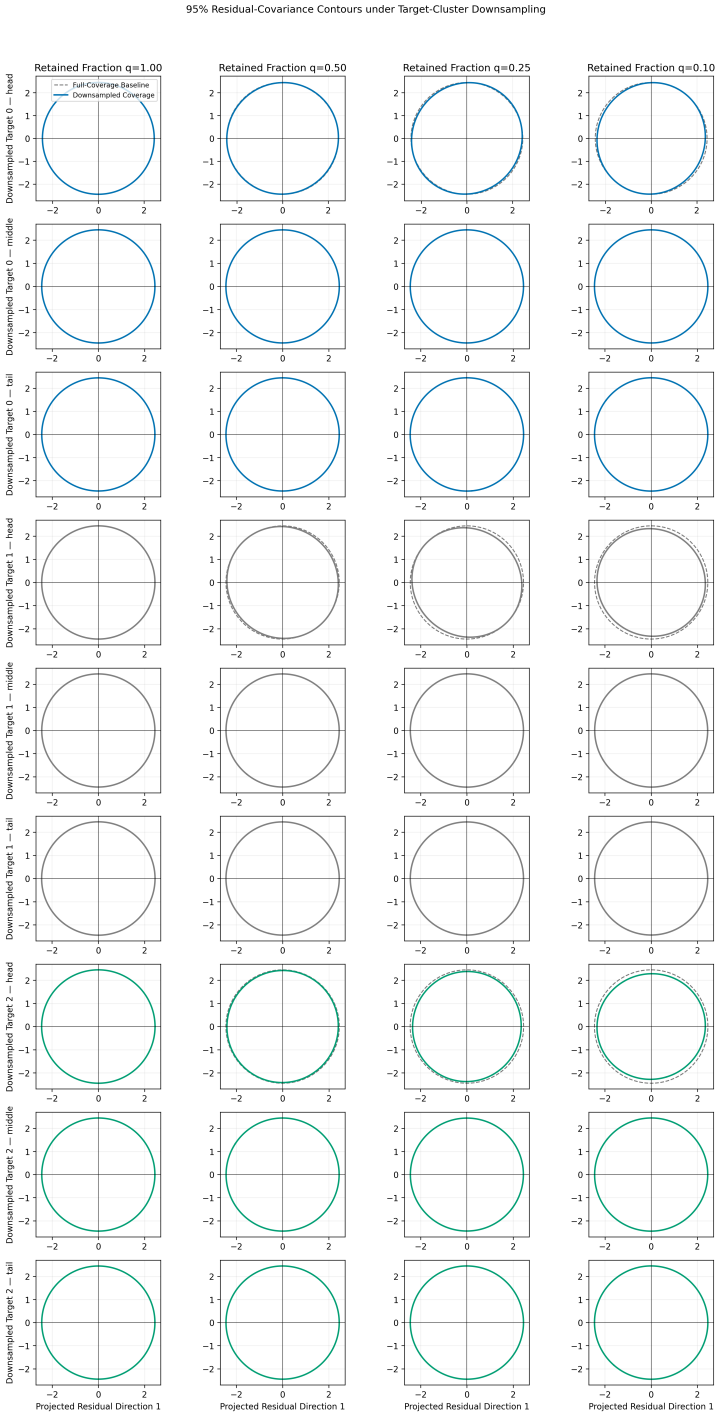


Figure 8: Two-dimensional covariance projections under decreasing target coverage. The panels illustrate changes in covariance geometry and are not used as formal mechanism evidence.