SPECTRAL UNIT TESTS FOR PREDICTIVE REPRESENTATION MODELS: A FAULT ATLAS AND ITS SCOPE CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive representation models are trained and monitored through one scalar loss, which can hide a predictor that loses to the unconditional mean of its targets in some directions. We audit the residual spectrally: whitening its second moment against the target covariance makes any generalized eigenvalue above one a certificate of regret against the mean; a covariance ceiling attributes that regret to the predictor, not to irreducible noise. A LayerNorm-aware rank bound predicts the unit-eigenvalue plateau a normalized affine head forces; five public checkpoints saturate it, and excising it makes the spectrum interpretable. On the official I-JEPA head the scalar and the spectrum disagree: its NMSE is below that of a ridge head refit on the audit data, yet its plateau-excluded regret mass is 2.2 times that of the ridge head, a gap that official-loss refits narrow by only 4–10%. A calibrated gate fires at full power on a five percent rise in scalar error for rank, bias and noise faults, and reaches faults confined to low-variance directions, which no calibrated NMSE gate reaches on this source budget. NMSE-matched injected faults leave distinct spectral signatures consistent with the theory's bounds, and rules fit on one checkpoint transfer blind to two others under conditions we state and measure. The audit also caught the HuggingFace I-JEPA export carrying the online encoder rather than the EMA target.

1 INTRODUCTION

Predicting one representation from another is the shared core of joint-embedding architectures (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025), of self-distillation, and of feature distillation. The predictor is trained, tuned, and monitored through one number: the average prediction error. Recent work likewise questions whether that number tracks representation quality, and redesigns training so that it does (Li et al., 2026). We audit predictors that are already trained.

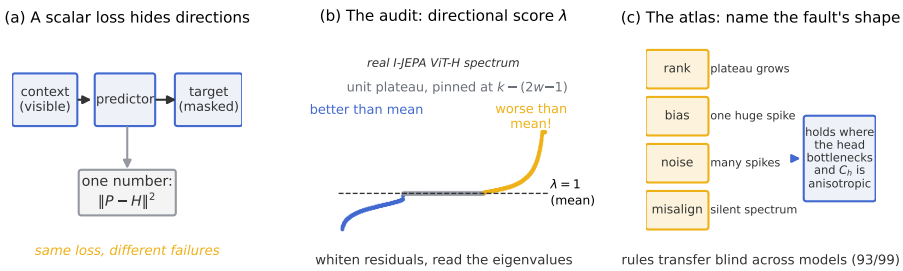

Figure 1: **Overview.** (a) Training compresses all directional structure of the residual into one scalar. (b) The audit whitens the residual against the target covariance and reads generalized eigenvalues: $\lambda > 1$ certifies directions where the predictor loses to the mean, and a unit plateau whose size is determined by the head width through a LayerNorm-aware rank count (Theorem 1, Corollary 1). (c) Faults matched in scalar NMSE leave separable fingerprints, and rules transfer blind where the head has a bottleneck and the target is anisotropic (Appendix L).

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.

Averages conceal direction: a predictor can be excellent where the target carries its variance and semantics and still be *worse than the mean* in a minority of directions (Figure 1).

This paper asks what the residual's second-moment geometry can certify. Writing C_h for the target covariance and $C_e(P)$ for the residual second moment of a predictor P, the generalized spectrum of the pencil (C_e, C_h) scores every direction of representation space against the simplest baseline available, the unconditional mean, one eigenvalue λ per direction: $\lambda < 1$ means the predictor helps, $\lambda = 1$ means it matches the mean, and $\lambda > 1$ means it actively hurts. Two exact facts organize everything that follows. First, a covariance ceiling: for the Bayes predictor, $C_e \preceq C_h$ in the positive-semidefinite order, so $\lambda > 1$ certifies predictor deficiency (capacity, optimization, calibration, or shift), never irreducible noise (Proposition 1). Second, a rank stamp: an affine head of rank r forces at least $k - 2r_a$ of the k readable eigenvalues to equal exactly one, and permits at most r_a on either side, where $r_a \leq r+1$ is the dimension of the span of the head's range and the offset of its bias from the target mean (Theorem 1). The head bottleneck that produces this plateau is a performance-motivated design choice: a width-384 predictor *beats* a full-width one downstream (Assran et al., 2023, Table 14). The theorem makes the bottleneck's algebraic consequence checkable: counting eigenvalues of $C_e - C_h$ against 10^{-6} times the largest, the rank saturates its ceiling $2w-1$ on all five public checkpoints we audit. The bottleneck thus gives the spectrum a known reference structure.

Measuring this spectrum naively is misleading, and much of our contribution is the estimator design that makes it reliable: support projection, plateau excision, invariant masses, source-clustered cross-fitting, and a refit null with the mean's home advantage (§3), which support three families of results.

Held-out regret survives a refit null, and a calibrated gate catches it. The official I-JEPA head exceeds its refit-null by $+7.5$ points of plateau-excluded regret mass on transfer data, while a five-split control finds its NMSE lower. On in-distribution Imagenette images, the regret-mass gap is $+18.1$ points. As a unit test, a two-channel gate calibrated on healthy resplits holds its target false-positive rate on fresh ones and, on resplits disjoint from its calibration, fires on 118 of 120 faulted resplits at severities from $+5$ to $+30\%$ NMSE, every one but two of the faintest misalignment (§4.1).

A LayerNorm-aware bound predicts the unit plateau, and checkpoints attain it. The count of Corollary 1 is verified with exact signature bookkeeping on the I-JEPA cache ($(+384, -383)$, rank $767 = 2w-1$) and agrees exactly with the ambient nullity $d - (2w-1)$ on four further checkpoints spanning ViT-L, ViT-H and ViT-g at the 10^{-6} counting tolerance used in the appendix. It moves as predicted under exact head truncation and frozen capacity sweeps, and it *emerges* during training as the head settles into its low-rank structure (§4.2).

Faults matched in the scalar separate in the spectrum. With severities calibrated so that scalar NMSE is matched across fault types, rank, bias, and noise defects separate in the spectrum, and misalignment-type faults are detected by elimination. A rulebook of five baseline-relative features fit on I-JEPA alone correctly classifies 93 of 99 blind cases on V-JEPA and V-JEPA2, with all errors on two documented edges; the verdicts survive re-matching in the official SmoothL1 and ℓ_1 losses, and the ambient signature recovers the ranks of trained heads where the bound binds (§4.3).

The audit is a unit test, not a benchmark, and §6 states its scope in full.

2 The audit operator and its exact structure

Setup. Let $H \in \mathbb{R}^d$ be the target representation (post target-LayerNorm, the space in which JEPA losses are computed), X the context, $P(X)$ a predictor, and $m = \mathbb{E}[H]$. Define $C_h = \mathrm{Cov}(H)$ and the residual second moment $C_e(P) = \mathbb{E}\big[(P-H)(P-H)^\top\big]$. On the support of C_h (dimension k; LayerNorm makes C_h singular), define the whitened residual operator $D = C_h^{\dagger/2}\, C_e(P)\, C_h^{\dagger/2}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$, and the signature counts $n_{\mathrm{better}}, n_{\mathrm{unit}}, n_{\mathrm{worse}}$ of eigenvalues below, at, and above 1.

Proposition 1 (Covariance ceiling; classical)**.** *For any predictor, with $\mu(X) = \mathbb{E}[H \mid X]$,*

$$C_h - C_e(P) \;=\; \mathrm{Cov}\big(\mu(X)\big) \;-\; \mathbb{E}\big[(P-\mu)(P-\mu)^\top\big].$$

In particular $C_e(\mu) \preceq C_h$: no direction of the Bayes predictor is worse than the mean.

Proof. Expand both sides using $H - m = (H - \mu) + (\mu - m)$ and $\mathbb{E}[H - \mu \mid X] = 0$. This is the law of total covariance combined with the bias–variance decomposition of $P - H$. □

The identity is classical. Its consequence for auditing is not: every $\lambda > 1$ is a certificate of predictor deficiency, never of aleatoric noise. We accordingly call $\lambda - 1 > 0$ *mean-relative spectral regret* (not excess risk, which is measured against Bayes).

Theorem 1 (Rank–signature bound)**.** *Let $P(X) = Wz(X) + b$ with $\mathrm{rank}(W) \le r$, and let $r_a = \dim \mathrm{span}\big(\mathrm{range}(W) \cup \{b - m\}\big) \le r + 1$. Then the quadratic form of $C_e - C_h$ vanishes on an $(d - r_a)$-dimensional subspace, and on a retained support of dimension k,*

$$n_{\mathrm{worse}} \le r_a, \qquad n_{\mathrm{better}} \le r_a, \qquad n_{\mathrm{unit}} \ge \max(k - 2r_a,\, 0).$$

The proof is a short congruence argument, deferred to Appendix A.

Corollary 1 (LayerNorm-aware sharp count)**.** *If the head input is the output of a width-w* `LayerNorm` *(as in every JEPA predictor audited here, where a final normalization feeds the linear projection), the normalized coordinates carry one exact linear constraint, so the* centered *prediction span has dimension $q \le w - 1$. Because the mean-shift term $\delta\delta^\top$, where $\delta = \mathbb{E}[P] - m$, is positive semidefinite, it can raise only the positive inertia: for v in the orthogonal complement of the centered span the form equals $(v^\top \delta)^2 \ge 0$. Hence $n_{\mathrm{worse}} \le q + 1$, $n_{\mathrm{better}} \le q$, and $\mathrm{rank}(C_e - C_h) \le 2q + 1 \le 2w - 1$ as an inequality, not merely a generic prediction. The signature saturates at $(q{+}1,\, q)$, giving $\mathrm{rank}(C_e - C_h) = 2w - 1$ and ambient nullity $d - (2w - 1)$ of the form; the support unit multiplicity $n_{\mathrm{unit}}(D) \ge k - (2w - 1)$ differs from it by the excluded C_h-null directions, where the pencil is undefined. Saturation holds under four genericity conditions: nonzero LayerNorm scales, centered activations of full rank $w{-}1$, a projection injective on the LayerNorm hyperplane, and no accidental cross-covariance cancellation. We verify rather than assume these on the audited caches (e.g., centered activation rank exactly 383 with tail eigenvalue 1.5×10^{-7}). Saturation is an empirical finding under stated conditions, not a theorem. SVD truncation to rank $r < q$ acts after the normalization and restores the unconstrained case, with generic signature $(r{+}1,\, r)$.*

Nominal width w, projection rank $\mathrm{rank}(W)$, centered span q, and the bias-augmented dimension are four different integers that differ by at most one. Every count below states which one it reports.

Definition 1 (Invariant spectral masses)**.** *For threshold τ and plateau band $|\lambda - 1| \le \epsilon_u$ (we use $\epsilon_u = 10^{-3}$ throughout),*

$$M_{\mathrm{sel}}(\tau) = \frac{\sum_{\lambda_i > \tau} \lambda_i}{\sum_i \lambda_i}, \qquad M_{\mathrm{reg}}(\tau) = \frac{\sum_{\lambda_i > \tau} (\lambda_i - 1)}{\sum_i \lambda_i},$$

with plateau-excluded variants restricting the numerator to off-plateau eigenvalues. Both are invariant to orthogonal changes of basis under a fixed support-selection rule. Native-coordinate trace shares are not invariant and are never primary statistics. Only the plateau-excluded variants are interpreted: near $\tau = 1$ the non-excluded M_{sel} is dominated by numerical drift of the rank-forced unit eigenvalues (0.63 against 0.43 on I-JEPA at cutoff 10^{-5}), whereas M_{reg} weights by $\lambda_i - 1$ and is insensitive to it. Only the numerator is plateau-excluded; the denominator keeps a model-dependent share of dimension-forced unit mass, so we compare $M_{\mathrm{reg}}^{\mathrm{ex}}$ levels within a model only.

Finite-sample sign recovery under the off-plateau spectral margin is a deterministic Weyl argument, in Appendix B with its caveats (support alignment between the empirical and population pencils is not addressed).

3 Measuring the spectrum reliably

The naive plug-in estimate of D fails in two ways. LayerNorm makes C_h singular, so fixed-ridge whitening amplifies near-null directions. Theorem 1 forces at least $k - 2r_a$ eigenvalues to *exactly* one, a plateau that numerical noise splits into spurious "$\lambda > 1$" counts. The audit protocol neutralizes both, in six steps (Figure 4, Appendix C):

1. **Cache.** Collect pre-head activations and LayerNormed targets once, bucketed by source image or clip, stored in fp16 from a bf16 forward; every moment and every decomposition downstream is fp64. Every variant and baseline reads the same cache, so differences between them are algebraic, not sampling variation.

Table 1: **Invariant audit across support cutoffs** (plateau-excluded, LayerNorm-aware floors $k-(2w-1)$). Values are cutoff-sensitive and the image/video ordering is not stable. No cell should be quoted without the sweep. Masses here pool the whole cache.

Model	cutoff	k	n_{unit} / floor	$M_{\text{reg}}^{\text{ex}}(1.0)$	$M_{\text{reg}}^{\text{ex}}(1.1)$
I-JEPA	10^{-3}	443	0 / 0	3.35%	2.88%
I-JEPA	10^{-4}	829	65 / 62	11.98%	11.10%
I-JEPA	10^{-5}	1279	512 / 512	13.08%	12.63%
V-JEPA	10^{-3}	840	86 / 73	5.22%	4.18%
V-JEPA	10^{-4}, 10^{-5}	1023	256 / 256	6.83%	5.93%

2. **Moments and support.** Form C_e, C_h; project onto the support of C_h with a *relative* eigenvalue cutoff, and report each mass over the cutoff sweep 10^{-3}–10^{-5} (off-support residual share at most 1.9% at the loosest cutoff, under 0.3% otherwise). No cell is quoted alone.
3. **Whiten, excise, count.** Eigendecompose the whitened pencil; excise the rank-forced plateau *before* interpreting any mass; never use raw direction counts as evidence.
4. **Invariant masses.** Report $M_{\text{sel}}, M_{\text{reg}}$ (Definition 1) over every support cutoff and every τ. Congruence-invariant signature counts of $C_e - C_h$ replace tolerance counts on one cache.
5. **Held-out validation.** Select a subspace V on one source split and read the signed score $\text{tr}(D_V) - \dim V$ on the other, D_V being D compressed to V, with source-cluster bootstrap intervals, against four nulls: the mean, a position-conditioned mean, an intercept-calibrated predictor, and a ridge head refit on the audit domain (the *refit null*).
6. **Calibrated gate.** Detection thresholds are never fixed constants: healthy source-resplit noise exceeds any a-priori gate (its 99.5th percentile relative NMSE gap is 4.1% on our source budget), so the gate is calibrated on healthy resplits at a target false-positive rate and frozen (§4.1). It is two-sided on both channels, and the firing pattern routes a flagged item before any classification (§4.3).

The audit uses second moments throughout, although the audited models are *not* trained with a second-moment objective: the released I-JEPA implementation uses SmoothL1 where the paper writes squared ℓ_2, and V-JEPA and V-JEPA2 train with ℓ_1, whose unconditional optimum is the coordinatewise median, not the mean. We audit the implementations as released. Every $\lambda > 1$ statement certifies mean-relative *quadratic* regret, an auxiliary diagnostic for ℓ_1-trained models rather than regret in their official objective, and every fault arm (one fault type at one severity) is matched in NMSE, not in training loss.

4 Experiments

Setup. Frozen official checkpoints: I-JEPA ViT-H/14 (Assran et al., 2023), V-JEPA ViT-L/16 (Bardes et al., 2024), and V-JEPA2 ViT-L, ViT-H and ViT-g (Assran et al., 2025). The HF V-JEPA2 adapter runs the encoder on the full clip rather than on context tokens alone. That is appropriate for its intended feature-extraction use but not for a masked-prediction audit, so those three checkpoints carry baseline-relative readouts and the stamp count, not the audit itself. V-JEPA2 ViT-L is also re-audited through its *official masked pretraining forward* (official repository and checkpoint), which is the primary V-JEPA2 evidence. Audit data: STL-10 and Imagenette for images; UCF101 and HMDB51 clips for video; official mask distributions throughout. Toy from-scratch models (ViT-S) provide trained-head sweeps and training trajectories. The released I-JEPA sampler never draws the last row or column of the patch grid, so every number from the frozen ViT-H/14 caches rests on 225 of their 256 positions, and the trained fleets lose the same border (Appendix M).

4.1 Detection: held-out regret above a calibrated null

Regret above a refit null. Plateau-excluded regret mass is positive in every setting (Table 1), but positivity alone is trivial: a ridge head *refit on the audit domain* still shows 6.6% on a fixed held-out half of STL-10, the floor set by finite samples and linear capacity. Against that null the official head

gains $+7.5$ points, averaged over random half-splits, and $+18.1$ points on in-distribution Imagenette, which rejects the domain-shift explanation. Refitting the null in the official losses, SmoothL1 or ℓ_1, with its intercept fitted as well, closes only 4–10% of the gap, and on the held-out half the official head's NMSE is *below* the null's on every split (0.564 against 0.571) while its regret mass is 2.2 times larger (Appendix C). The gap is stable, positive in 20 of 20 repeated source half-splits at every support cutoff (10^{-3}, 10^{-4} and 10^{-5}). A two-stage source-level bootstrap (half-split of the 1920 source buckets, then buckets resampled with replacement within each half) includes between-source sampling variance, unlike a resplit band, and is bounded away from zero at all three cutoffs ($[+3.7, +5.2]$ points at 10^{-4}). Resampling leaves about 607 distinct buckets per half and the gap grows with distinct buckets, so the interval is evidence of sign, not of magnitude. The gap needs source diversity: it inverts below half the buckets (Appendix C).

The calibrated gate. The gate of §3 (step 6) was calibrated on 500 healthy resplits and tested, thresholds frozen, on 200 fresh healthy resplits plus 120 faulted resplits, twelve arms at ten each (a resplit is a random halving of the source buckets, baseline on one half and statistic on the other). It realizes 1.5% false positives at the 1% familywise target, each channel taken at its 99.5th healthy percentile, a rate its confidence interval leaves consistent with that target (Appendix C). Power was first measured on ten resplit seeds inside the calibration range, where every one of the 120 faulted cells fires; on ten seeds disjoint from both the calibration and the fresh false-positive test it fires on 118 of them, every arm on every resplit except the $+5\%$ misalignment, which fires on 8 of its 10, near the plug-in estimate from the calibration resplits: that arm leaves the eigenvalue channel alone, so its power is the NMSE channel's, 0.73 at this severity (`p0p_power_fresh.json`). The paired comparisons between gate variants in Appendix C were run on the first set. The looser 5% target exceeds its nominal false-positive rate (Appendix C); we deploy the stricter target.

The scalar channel alone. An NMSE-only gate matched to the union gate's realized false-positive rate misses 1 of the 40 $+5\%$ replicates, a rank truncation, and up to 8 as its threshold is tightened towards the union allocation, 6 of them rank truncations; the eigenvalue channel recovers every one (Appendix C). For *scalar-visible* faults the scalar NMSE gap is therefore the primary channel, and the spectral gain is small and concentrated on rank rather than evidence of blindness in the scalar one. What the spectrum adds there is morphology: the top whitened eigenvalue separates those rank, bias and noise arms at the empirical zero-false-positive point.

What no scalar gate can see. No NMSE gate reaches faults confined to low-variance directions at any calibrated false-positive rate on this source budget. Their scalar displacement ($+0.1$–0.5% NMSE) is an order of magnitude *smaller* than healthy split noise (4.1% at the deployed 99.5th percentile); yet their whitened top eigenvalue jumps by two orders of magnitude, from $10^{0.7}$ to $10^{2.5}$, and regret mass nearly triples, from 14 to 37%. Across ten deployment resplits the scalar gate fires on none of the ten arms and the eigenvalue gate on all of them. For the two injected kinds, a shift and broadband noise inside the pocket, a mean test detects the first and a span test the second; the eigenvalue channel catches both with one statistic (Appendix C). Direction *counts* carry no information (Appendix C); Appendix D brackets the pockets' utility.

Cross-fitting. We select the subspace on one bucket set, fit the refit null on a second, and score both heads on a third over the same 64 directions. The official head carries 0.544 of regret per direction against the null's $+0.007$ on images and 0.413 against -0.016 on video, with paired intervals bounded away from zero (Appendix C). Source-grouped cross-fitting passes both directions for both checkpoints, and the score survives a position-conditioned mean baseline and intercept calibration. The regret directions have low variance and low measured utility: removing them costs downstream probes less than removing random directions of equal number (Appendix D; Figure 6 shows the content they select). Whether they matter is task-dependent; under input drift the audit channels move before a frozen probe's accuracy does (Appendix F).

4.2 Artifact: the rank stamp, five ways

Exact bookkeeping. On the I-JEPA cache the bookkeeping of Corollary 1 is verified exactly: $d = 1280$, $w = 384$, $\mathrm{rank}(W) = 384$, and the centered pre-projection activations have effective rank $q = 383$, their covariance spectrum dropping six orders of magnitude to a single 1.5×10^{-7} tail eigenvalue. The signature of $C_e - C_h$ is $(+384, -383)$ at every tolerance from 10^{-6} to 10^{-8}: rank

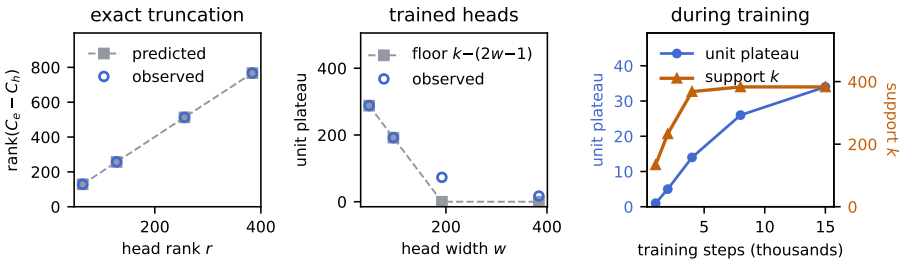

Figure 2: **The stamp moves as predicted.** Three sweeps on separate axes, because they change different things. Here k is the retained support of C_h, the directions the pencil can score at all.

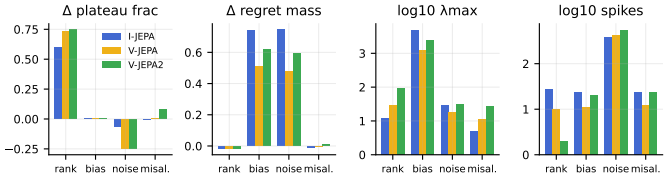

Figure 3: **Fingerprints at matched scalar NMSE** (severity level +30%; three models). Rank grows the plateau; bias plants a single huge spike; noise collapses the plateau, which no affine fault can do (Proposition 2), and floods spike counts (a spike is $\lambda > 2$); misalignment leaves the spectrum statistically unchanged, so it is diagnosed by elimination.

$767 = 2w - 1$, ambient nullity $513 = d - (2w - 1)$ and support plateau $512 = k - (2w - 1)$ at $k = 1279$, all saturating the corollary's generic prediction.

Four further checkpoints. The same count holds for the other checkpoints: V-JEPA ViT-L ($d = 1024$) shows support plateau $256 = 1023 - 767$ exactly. V-JEPA2 ViT-L audited through its official masked pretraining forward (context tokens only) reproduces the full bookkeeping (signature $(+384, -383)$, rank 767, sharp nullity $257 = 1024-767$ exactly at counting tolerance 10^{-6}). The three HF-adapter caches saturate the same ceiling at that tolerance, and their support plateaus attain $k - (2w-1)$ exactly (Appendix G). Three further tests, all in Figure 2, make the stamp causal rather than correlational. Truncation is the cleanest: on the shared cache (fp16 activations, fp64 algebra) a rank-r head has signature exactly $(r+1,\ r)$ with tolerance-free nullity $d - (2r+1)$, exactly as the corollary predicts. The capacity sweep, at $k = 383$, attains the floor exactly where it binds $(288(288), 192(192), 73(0), 17(0)$, floor in parentheses). Training shows the constraint present from initialization and its *visibility* emerging: the plateau grows from 0.7% to 8.9% of the retained support (cutoff 10^{-4}) as activations reach full effective rank.

The two diagnostics disagree, which is why both are kept: widening the head from 48 to 384 drops regret mass from 0.53 to 0.26% while NMSE rises from 0.498 to 0.647, and over training regret falls from 6.10 to 2.06% while NMSE rises from 0.26 to 0.62 (Figure 9), since larger heads converge more slowly and targets grow more informative.

4.3 Diagnosis: A fault atlas that transfers blind

Rig. All arms of a model share one cache of pre-head activations, LayerNormed targets and source identifiers. Faults are exact fp64 operations on the head: rank truncates W to its top r singular directions, bias adds a fixed offset δu along a random unit vector, misalignment re-pairs a fraction f of the predictions with the targets of other samples, and noise adds i.i.d. Gaussian noise of scale σ to every coordinate of every prediction. The first three keep every prediction inside the head's affine span, so Theorem 1 bounds their plateaus from below; the fourth leaves it, and the resulting plateau collapse is diagnostic: a LayerNormed affine head of width w keeps the unit count at $k-(2w-1)$ or above under any fault that stays in its span, and any affine head of width w at $k-2(w+1)$ or above, so a plateau below the second floor certifies a residual component no affine head of that width can

Table 2: **The rulebook.** Rules 2 to 5 were fit on I-JEPA alone, then frozen; the first rule that fires decides. Rule 1 is the calibrated two-channel gate of §3, which also routes a downward NMSE excursion to *easier* and an eigenvalue-only one to *pocket*, neither entering rules 2 to 5. Every quantity is a difference from the same model's healthy baseline: NMSE relative, the plateau as a share of the support, and the plateau-excluded regret mass $M_{\mathrm{reg}}^{\mathrm{ex}}$. Noise is tested before bias because a broadband fault pushes many eigenvalues past threshold while a mean shift pushes one far.

	If	Verdict
1	two-channel gate silent (§3, step 6)	healthy
2	Δplateau share > 0.10	rank
3	spike count more than triples	noise
4	$\Delta \log_{10} \lambda_{\max} > 0.3$ and $\Delta M_{\mathrm{reg}}^{\mathrm{ex}} > 0.2$	bias
5	$\lvert\Delta$plateau share$\rvert < 0.05$ and $\Delta M_{\mathrm{reg}}^{\mathrm{ex}} < 0.1$	alignment
6	otherwise	abstain

produce. On the three primary grids none of the 104 distinct affine-span arms sits below the first floor and all 45 noise arms sit below the second; Proposition 2 (Appendix A) ties their spike count to the target directions whose variance falls below the noise level. Severities are calibrated by binary search to matched NMSE grids (+5% to +50%), so construction makes the fault types indistinguishable by scalar NMSE. Exception: on V-JEPA2 rank truncation saturates below the grid (a rank-one head reaches $+29.8\%$ NMSE), so its higher-severity rank arms are deduplicated, removing one of fifty.

Separability and blind transfer. Rank, bias, noise and misalignment separate in the spectrum of both native models (Figure 3, Table 2), where image-space error maps separate rank and misalignment by eye and leave bias and noise to statistics chosen knowing the fault (Figure 11, Appendix M). Head quantization to bf16 is harmless. The families are separable by construction (truncation lowers the rank of the centered predictions, bias moves the residual mean, noise leaves the affine span, misalignment changes none of these), and two statistics that measure those properties type all 149 distinct arms without whitening or the residual pencil (Appendix I); the pencil reads the four signatures Theorem 1 and Proposition 2 predict from the residual that carries the regret certificate. Fit on I-JEPA only, the rulebook scores $50/50$ on its own model and $93/99$ blind across V-JEPA and V-JEPA2 (149 distinct arms in total), reading V-JEPA2 through the official masked forward the setup calls primary ($89/98$ on the adapter cache, Appendix H). Every blind arm is faulted by construction, so this is morphology accuracy conditional on a fault of a known family, with no specificity term and no claim about cause. Its six errors sit on two documented edges: the $+5\%$ bias arms, where the spike sits below threshold, and the heaviest misalignment arms on 64-frame clips, where aggressive permutation mimics a rank defect. Out of family they transfer unevenly: SimSiam scores $43/50$ and BYOL $32/50$ against a preregistered 35, losing exactly the two rules whose preconditions it fails (Appendix L). A column-mask distribution shift falls in the alignment family: the pencil detects a *task-shift family* by elimination but does not split its members; per-position statistics may.

What "diagnosis" can mean. Throughout, diagnosis is *fault-morphology classification under a specified fault model*, not cause identification: two mechanisms with identical residual first and second moments are indistinguishable to any audit built on them. A second-moment rulebook conflates a constant bias with a sign-symmetric rank-one noise, which the residual first moment separates; adding it as a sixth feature loses three arms on the primary grids (90 against 93 of 99) and the same three on the adapter-cache grid (86 against 89), so the five-feature rulebook is the one scored above.

Controls and scope. Fitted rather than injected (reduced-rank regression, drifted or noisy targets, permuted pairs, few-step SGD), the faults land in the intended class in 40 of 40 arms over 5 seeds and all 10 undertrained heads read as rank-deficient (Appendix I); re-matched in the official losses the rulebook scores $8/8$ and $7/8$, the ambient signature recovers the widths of four trained heads where the bound binds (Appendix G), faults upstream of the head land in the elimination family (Appendix J), and Theorem 1 binds any affine head, SimSiam's and DINOv2's included (Appendix L).

Natural misconfigurations, and one procedure. Injected faults raise the loss by construction. Natural ones need not. Across three fleets of small I-JEPA runs (ViT-S/8 at 3k steps, ViT-S/16 at 40k, ViT-B/16 at 120k; Appendix P), six harmful misconfiguration families, with final-probe drops of 8.9 to 41.0 points, drive NMSE *below* the healthy runs, because the target gets easier to predict while the encoder degrades. The rulebook's original one-sided healthy rule classifies all

six as healthy. The two-sided two-channel gate of §4.1 does not, so we apply the same procedure throughout: the gate decides healthy or not, a deterministic route on its channel pattern sends a downward NMSE excursion to *easier* and an eigenvalue-only excursion to *pocket*, and rules 2 to 5 classify what remains. On the natural fleets, 18 of 18 runs of the six families route to *easier* at the end of training, the gate separates harmful from healthy runs with AUC 1.00 on the 40k fleet, and it misses one family on the 120k fleet: a 6.1-point corruption that is flagged at 10 and 40% of training and is back inside the healthy null by the end (AUC 0.99 then 0.81). Cause and morphology are not one to one; the same learning-rate fault reads as misalignment at 3k steps and *easier* at 40k and 120k.

The held-out test. This composition is post hoc on those fleets, so we pre-registered a held-out test, run once: on 3 families new to this scale and 4 new healthy seeds the gate separates every harmful run from every healthy one (AUC 1.00; permutation $p = 1.4 \times 10^{-3}$ under exchangeable labels of the held-out runs), as its NMSE channel does on its own, and the lower-tail family routes to *easier* in every seed. No run reached stage 2; every harmful family we tried had made the residual easier by then. Detection transfers; morphology of natural faults is not claimed (Appendix P).

What ranks runs. On the same fleets, read two-sided at 40% of training, the rank correlation with the final probe is -0.74, -0.73 and -0.80 for NMSE and -0.48, -0.63 and -0.82 for the raw loss. Selected on the toy fleet and applied unchanged, NMSE is best there and inside the interval of the best spectral channel on both larger fleets (Appendix P). Ranking runs needs a two-sided scalar and no eigendecomposition; normalizing by the target variance helps on two fleets and is neutral on the third, and every channel fails when read one-sided on the two larger fleets because the loss falls as the run degrades. The spectrum adds the morphology of a failing run: on the toy fleet, context leakage barely dents the final probe yet produces the largest regret mass, so predictor-side regret and encoder quality are distinct axes, and the morphology-matched fix at 40% restores a frozen-EMA run to near-healthy probe accuracy where a mismatched control does not. These statistics give negative results for both sweep pruning and repair selection (Appendices Q and N).

4.4 Equivalence of shipped artifacts

A practitioner holding a reference checkpoint and a candidate export asks whether they compute the same function and, if not, what changed. Running both on the same inputs answers the first question; neither baseline alone answers the second. Table 3 applies the gate, calibrated on the cache's own healthy resplits, and the frozen rulebook to eight deployment operations on the I-JEPA cache, severities matched on a disjoint third of the source buckets and every verdict scored on 30 resplits of the rest, next to the two comparisons a practitioner already has: the output difference, and the parameter difference read through an oracle cutoff. The output difference detects every change but types none. The parameter difference types four of the eight and fails on the same four on all 3 caches, and in both directions: a gauge re-parameterization moves every weight ($\|\Delta W\|/\|W\| = 1.000$) and changes the function by 2×10^{-8} in NMSE, since the centered LayerNorm input has rank $w-1$ (Corollary 1); a permuted serving order moves no parameter and changes the function as much as a matched fault. Used together with oracle thresholds, one to decide and the other to type, they name 9 of the 9 rows on every cache. The audit types all four matched faults on every resplit of every cache and needs only a stored healthy fingerprint of the reference, on other inputs, where both baselines need the reference itself on the same ones (Appendix R).

The wild case. In the field, the HuggingFace release of I-JEPA (`facebook/ijepa_vith14_1k`) differs from the original release's *target* encoder by 19.5% NMSE, broadband with a 0.61 mean shift, the fingerprint of EMA drift, while two HF copies agree to machine precision; at the tensor level it matches the *online* encoder bit-for-bit and the EMA target on no tensor (Appendix S), and the model card does not name the branch. A global probe does not register the difference (0.9911 against 0.9906), but the difference concentrates on object boundaries (Figure 13) and persists in dense prediction, where the EMA encoder leads the export in every seed (Table 23); VOC 2012 reproduces both, and the MAE and DINOv2 families yield three distinct verdicts (Appendix S).

Table 3: **Typing a shipped artifact**, I-JEPA head, matched faults at $+10\%$ relative NMSE. Parameter diff: the type assigned by a rank-and-norm reading of $(\Delta W, \Delta b)$ under an oracle cutoff. Output diff: relative disagreement with the reference. Audit: gate route, or rulebook label at stage 2, unanimous over 30 resplits unless counted. Other caches and the synthetic row: Appendix R.

shipped artifact	truth	$\|\Delta W\|$, $\|\Delta b\|$ (rel.)	param. diff	output diff	audit
byte-identical copy	equivalent	0.000 / 0.000	healthy	0.0000	healthy
gauge re-parameterization	equivalent	1.000 / 2.437	rank	0.0002	healthy
bf16 export	benign	0.002 / 0.000	noise	0.0018	healthy
int8 export (per-channel)	benign	0.008 / 0.000	noise	0.0074	healthy
low-rank compression	rank	0.402 / 0.000	rank	0.3572	rank
intercept drift	bias	0.000 / 1.381	bias	0.3411	bias
broadband weight noise	noise	0.355 / 0.000	noise	0.3386	noise
permuted serving order	misalign	0.000 / 0.000	healthy	0.3481	misalign

5 Related work

Joint-embedding prediction. JEPAs (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025) and their variants, including those that condition the predictor on stochastic positions to model location uncertainty (Bar et al., 2024), all treat the residual as a scalar to minimize. We audit its geometry. Littwin et al. (2024) prove an implicit bias toward high-influence features in deep linear self-distillation networks, and our regret pockets are the measured residue of that bias.

What the geometry literature measures. Work on embedding geometry, including isotropic-Gaussian objectives (Balestriero & LeCun, 2025; Kuhn et al., 2026) and the structured alternatives that contest them (Le, 2026; Zhao et al., 2026), shapes the *marginal* of embeddings, where our object is the residual pencil (Appendix L). Isotropization simplifies our whitening. Spectral SSL diagnostics score representations, via effective rank (Garrido et al., 2023) and discriminant rank (Thilak et al., 2024), and are structurally blind to predictor-side faults, as our battery shows. Collapse regularizers (Bardes et al., 2022; Grill et al., 2020) constrain the marginal covariance.

Where the audit comes from. It inherits from multivariate process monitoring (PCA residual charts, Hotelling T^2) and innovation-whiteness testing, differing in that it corrects an architecture-induced eigenstructure first. The linear algebra is standard: for an optimal linear predictor the pencil spectrum is the CCA residual spectrum $1 - \rho_i^2$ (Hotelling, 1936; Izenman, 1975; Kalaitzis & Lawrence, 2012). Per direction $1 - \lambda$ is the classical skill score against a climatological forecast (Nash & Sutcliffe, 1970; Murphy, 1988), and the rank restriction behind Theorem 1 is the reduced-rank regression bound (Anderson, 1951; Reinsel & Velu, 1998). Our additions are the reading as mean-relative regret under a covariance ceiling, the rank-plateau correction that reading needs, and the fault atlas built on both. A per-direction training signal, after signed geometries (Law & Stam, 2020; Xiong et al., 2022) and reversed gradients (Ganin et al., 2016), is future work.

6 Limitations and conclusion

The faults here are injected: 149 arms on three checkpoints whose signatures survive generation by fitting, but not a census of how predictors fail in the wild. The rulebook's original one-sided healthy rule classified 6 harmful natural families as healthy. Its replacement is post hoc on those fleets, and the pre-registered held-out test passed for detection but did not test morphology, which we claim for injected faults only. Two limits are structural: the pencil reaches misalignment and task shift only by elimination, and it reads second moments while the official losses do not, so every certificate is quadratic and auxiliary for an ℓ_1-trained model. Two concern calibration: bias severity moves with the healthy predictor's own spike, and the invariant masses move with the support cutoff, so we report sweeps, never a single cell. Within these limits, predictive models ship with one number and could ship with a unit test; we offer the audit as that, not as a case against any architecture.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: to implement methods, including the audit library, the experiment scripts, the generators of the synthetic data, and the scripts that clean and reformat the public datasets into activation caches; to formulate Proposition 2 and assist in writing its proof, and to assist in repairing the proof of Theorem 1; to propose and refine hypotheses, including the scope conditions of Appendix L; to design and give feedback on experimental protocols, including the pre-registered held-out test and the paired re-measurement of Appendix J; to interpret results; and to translate internal records and code comments into English for the code release. Supporting qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to create figures, to draft and edit the text, to suggest the structure of the appendix, to search for and summarize the literature, to format references, and to review the code and the manuscript against each other. We have reviewed all AI-assisted work: every number quoted through a macro is generated by a script from the released result files and checked by a build gate, and the theoretical claims, the experimental designs and the final text were checked by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every number quoted through a generated macro is derived by a deterministic script from the shipped result files; some literals sit outside that gate, as the release README notes. We release the audit library and experiment scripts, activation-cache builders that import the official repositories at pinned commits, a checkpoint manifest, the JSON result files behind every table and figure, and, for the training fleets, the per-run records, the calibration resplits and the pre-registered held-out protocol with its scoring script. Cache collection is the one unseeded step, since the official mask collators draw from an unseeded generator (Appendix J measures the resulting spread), and the frozen capacity sweep ships as a result file only, its toy heads not having been retained. The tensor-level wild-case comparisons need only the public exports and the official release URLs.

REFERENCES

T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951. doi: 10.1214/aoms/1177729580.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Amir Bar, Florian Bordes, Assaf Shocher, Mahmoud Assran, Pascal Vincent, Nicolas Ballas, Trevor Darrell, Amir Globerson, and Yann LeCun. Stochastic positional embeddings improve masked image modeling. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 2944–2958, 2024. arXiv:2308.00566.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. arXiv:2404.08471.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning (ICML)*, pp. 10929–10974, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Alfredo A. Kalaitzis and Neil D. Lawrence. Residual component analysis: Generalising PCA for more flexible inference in linear-Gaussian models. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012. arXiv:1206.4560.

Lukas Kuhn, Lucas Maes, Giuseppe Serra, Quentin Le Lidec, Yann LeCun, Randall Balestriero, and Florian Buettner. LeVJEPA: Efficient and scalable video pretraining without the heuristics. *arXiv preprint arXiv:2608.27395*, 2026.

Marc T. Law and Jos Stam. Ultrahyperbolic representation learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1668–1678, 2020.

Triet M. Le. UR-JEPA: Uniform rectifiability as a regularizer for joint-embedding predictive architectures. *arXiv preprint arXiv:2606.01443*, 2026.

Xianhang Li, Chen Huang, Chun-Liang Li, Vimal Thilak, Eran Malach, Josh Susskind, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=3cB9243E9i`.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self-distillation networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2407.03475.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, and Nicolas Ballas. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 671–689. Springer, 2026. doi: 10.1007/978-3-032-37369-4_36.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988. doi: 10.1175/1520-0493(1988)116⟨2417:SSBOTM⟩2.0.CO;2.

J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I: A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. doi: 10.1016/0022-1694(70)90255-6.

Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*, volume 136 of *Lecture Notes in Statistics*. Springer, New York, 1998. doi: 10.1007/978-1-4757-2853-8.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2312.04000.

Bo Xiong, Shichao Zhu, Nico Potyka, Shirui Pan, Chuan Zhou, and Steffen Staab. Pseudo-Riemannian graph convolutional networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3488–3501, 2022.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-JEPA: Subspace Gaussian regularization for stable end-to-end world models. *arXiv preprint arXiv:2605.09241*, 2026.

APPENDIX OVERVIEW

The appendices follow the order of the main text.

- **Theory** (§2): the proofs of Theorem 1 and Proposition 2 (Appendix A) and the sign-recovery lemma (B).
- **Measurement and detection** (§3, §4.1): estimator ablations, the refit null and the calibrated gate (C); what the regret directions select and are worth (D); condition sweeps (E); early warning under drift (F).
- **The rank stamp** (§4.2): per-cache bookkeeping, severity inversion and official-loss controls (G).
- **The fault atlas** (§4.3): the two V-JEPA2 grids (H); training-induced faults, the feature ladder and classical baselines (I); encoder-level faults (J); whether the rule constants are tuned (K); out-of-family heads and the scope conditions (L); fault damage in image space and the sampler's reach (M); repair policies (N).
- **Natural misconfigurations** (§4.3): the toy training battery (O); the three fleets, the unified procedure and the pre-registered held-out test (P); pruning a sweep (Q).
- **Equivalence of shipped artifacts** (§4.4): typed equivalence on three caches (R); the exported encoders (S).
- **Cost** of the audit (T).

A PROOF OF THEOREM 1, AND WHAT ADDITIVE NOISE DOES TO IT

Proof. Write $\tilde{P} = P - \mathbb{E}[P]$ and $\tilde{H} = H - m$. Then $C_e - C_h = \mathrm{Cov}(\tilde{P}) - M - M^\top + \delta\delta^\top$ with $M = \mathbb{E}[\tilde{P}\tilde{H}^\top]$ and $\delta = \mathbb{E}[P] - m$. The columns of $\mathrm{Cov}(\tilde{P})$ and of M lie in $\mathrm{range}(W)$, and $\delta \in \mathrm{span}(\mathrm{range}(W), b - m)$. Hence for any $v \perp \mathrm{span}(\mathrm{range}(W), b - m)$: $v^\top \mathrm{Cov}(\tilde{P})v = 0$, $v^\top M v = 0$ (since $Mv \in \mathrm{range}(W)$), and $v^\top \delta = 0$, so $v^\top (C_e - C_h)v = 0$. Writing $U = \mathrm{span}(\mathrm{range}(W), b - m)^\perp$, polarization gives $\Pi_U(C_e - C_h)\Pi_U = 0$, so U is totally isotropic for the form. A maximal totally isotropic subspace of a symmetric form with inertia (n_+, n_-, n_0) has dimension $n_0 + \min(n_+, n_-)$, hence $\max(n_+, n_-) \le \mathrm{codim}\, U = r_a$, i.e. $n_\pm(C_e - C_h) \le r_a$. On the retained support, $D - I = C_h^{\dagger/2}(C_e - C_h)C_h^{\dagger/2}$ is a compression of $C_e - C_h$ by a matrix of rank k, so its inertia can only decrease: $n_\pm(D - I) \le n_\pm(C_e - C_h) \le r_a$. Therefore $n_{\mathrm{worse}} \le r_a$, $n_{\mathrm{better}} \le r_a$, and $n_{\mathrm{unit}} = k - n_+(D - I) - n_-(D - I) \ge \max(k - 2r_a,\, 0)$. □

The proof uses only that $P - b$ lies in $\mathrm{range}(W)$ almost surely; it does not use that P is a function of the paired input. So the bound holds for a faulted head as long as the fault leaves the predictor affine: truncating W, moving b, or re-pairing predictions with other samples' targets all keep $P - b \in \mathrm{range}(W)$, and none can push the plateau below the floor of Corollary 1. The theorem applies to the empirical distribution of a cache as it does to the population, so on a cache its bound is exact up to the plateau tolerance; the LayerNorm floor also relies on the constraint, which the caches hold to fp16 precision (Appendix G). On the three primary dense grids (I-JEPA, V-JEPA, and V-JEPA2 through the official forward), 0 of the 104 distinct rank, bias and misalignment arms fall below the floor, and misalignment at $+50\%$ attains it exactly on I-JEPA, while 45 of the 45 noise arms fall below it, and below Theorem 1's weaker width-w floor $k - 2(w{+}1)$ as well (44 on I-JEPA, 253 on the video models), which needs no LayerNorm constraint.

Additive noise leaves the affine span, so a plateau is no longer forced; the next proposition bounds the remaining unit count and spike counts. A plateau below $k - 2(w{+}1)$ certifies, by Theorem 1 alone and on a cache as in the population, a component that no affine head of width w produces, and a plateau below the LayerNorm floor does the same for LayerNormed heads up to the precision at which the cache stores the constraint; that the noised plateau falls to between 0 and 4 on every noise arm of the three grids, at most 1 on I-JEPA and 4 on the video models against LayerNorm floors of 47 and 256, is observed, not forced. The certificate reads one way: the proposition covers independent isotropic noise, the one fault of this kind we inject, and a contamination that is input-dependent or low-rank also leaves the span but need not empty the band. Write $\lambda_1(C_h) \le \cdots \le \lambda_k(C_h)$ for the retained eigenvalues of C_h, ϵ_u for the plateau tolerance of Definition 1, and n_{unit} here for the count inside that band.

Proposition 2 (Additive noise lifts the forced plateau). *Let $P = A + \varepsilon$, where $A - b \in \mathrm{range}(W)$ almost surely with r_a as in Theorem 1, and ε is independent of (A, H) with $\mathbb{E}\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. Then on the retained support $D(P) = D(A) + \sigma^2 C_h^\dagger$, and for every $\tau > 0$, writing $n_\tau = \#\{i \le k : \lambda_i(C_h) < \sigma^2/\tau\}$,*

$$n_\tau - 2r_a \;\le\; \#\{\lambda(D(P)) > 1 + \tau\} \;\le\; n_\tau + r_a,$$
$$n_{\mathrm{unit}}(D(P)) \;\le\; n_{\mathrm{better}}(D(A)) + m_{\epsilon_u} \;\le\; r_a + m_{\epsilon_u}, \qquad m_{\epsilon_u} = \#\{i \le k : \lambda_i(C_h) \ge \sigma^2/\epsilon_u\}.$$

So at most r_a unit eigenvalues remain once $\sigma^2 > \epsilon_u \lambda_k(C_h)$, and the number of eigenvalues past $1 + \tau$ tracks, to within $2r_a$, the number of target directions whose variance is below σ^2/τ. For an isotropic target, $C_h = \bar{\lambda}\,\Pi$ on its support, with Π the projector onto that support, and $D(P) - I = (D(A) - I) + (\sigma^2/\bar{\lambda})I$: the noise shifts the healthy spectrum rigidly, so no eigenvalue at or below 1 crosses $1 + \tau$ until $\sigma^2 > \tau\bar{\lambda}$, which at $\tau = 1$ is a noise variance exceeding the target variance itself; a nearly isotropic target obeys the same statement up to $\sigma^2 \max_i |\lambda_i(C_h)^{-1} - \bar{\lambda}^{-1}|$ by Weyl's inequality. Conversely, any predictor with $P - b \in \mathrm{range}(W)$ almost surely for some W of rank at most w obeys Theorem 1, so a unit count below $k - 2(w+1)$, or below the floor of Corollary 1 for a LayerNormed head, certifies that the residual carries a component no affine head of that width can produce.

Proof. Independence and $\mathbb{E}\varepsilon = 0$ eliminate the cross terms, so $C_e(P) = C_e(A) + \sigma^2 I$ and, after whitening on the support, $D(P) - I = (D(A) - I) + \sigma^2 C_h^\dagger$. Let N be the null space of $D(A) - I$; by Theorem 1, $\dim N \ge k - 2r_a$. For $v \in N$, $v^\top (D(P) - I)v = \sigma^2 v^\top C_h^\dagger v$. Let E_τ be the span of the eigenvectors of C_h on the support with $\lambda_i(C_h) < \sigma^2/\tau$, so $\dim E_\tau = n_\tau$ and $\sigma^2 v^\top C_h^\dagger v > \tau\|v\|^2$ on E_τ. Then $\dim(N \cap E_\tau) \ge n_\tau - 2r_a$ and the form of $D(P) - I$ exceeds τ on that subspace, so by the Courant–Fischer theorem $D(P) - I$ has at least $n_\tau - 2r_a$ eigenvalues above τ. For the upper bound let V be the span of the eigenvectors of $D(A) - I$ with nonpositive eigenvalue, so $\dim V \ge k - r_a$ by Theorem 1, and let $F_\tau = E_\tau^\perp$, on which $\sigma^2 v^\top C_h^\dagger v \le \tau\|v\|^2$; on $V \cap F_\tau$, of dimension at least $k - r_a - n_\tau$, the form of $D(P) - I$ is at most τ, so at most $n_\tau + r_a$ eigenvalues exceed τ. For the plateau bound write $D(A) - I = X_+ - X_-$ with $X_\pm \succeq 0$ and $\mathrm{rank}\, X_- = n_{\mathrm{better}}(D(A)) \le r_a$. Then $D(P) - I \succeq \sigma^2 C_h^\dagger - X_-$, and subtracting a positive semidefinite matrix of rank ρ lowers each ascending eigenvalue by at most ρ positions, $\lambda_j(\sigma^2 C_h^\dagger - X_-) \ge \lambda_{j-\rho}(\sigma^2 C_h^\dagger)$, so at most $\rho + m_{\epsilon_u}$ eigenvalues of $D(P) - I$ lie at or below ϵ_u; equality needs the affine part to cancel the noise lift exactly on a better-than-mean direction. The isotropic identity is the whitening computed with $C_h^\dagger = \bar{\lambda}^{-1}\Pi$, and the nearly isotropic case is Weyl's inequality applied to $\sigma^2(C_h^\dagger - \bar{\lambda}^{-1}\Pi)$. The converse follows by taking the contrapositive of Theorem 1. □

The noise bounds of Proposition 2 are population statements. Its converse holds on a cache in the form of Theorem 1, whose bound the empirical distribution satisfies exactly; the sharper LayerNorm floor also needs the cached activations to satisfy the LayerNorm constraint, which they do to fp16 precision (Appendix G). On a cache of n tokens the sample second moment of ε differs from $\sigma^2 I$ and its sample cross-covariance with $A - H$ is not zero. For fixed dimension both deviations are $O_p(n^{-1/2})$ before whitening, but whitening multiplies them by up to $\lambda_1(C_h)^{-1}$, the same amplification that makes low-variance pockets visible, so they are not small against the band: for one noise draw at the $+5\%$ scale on each primary cache their whitened spectral norm is 0.10–0.37, against $\epsilon_u = 10^{-3}$. Restricted to the healthy plateau it is 0.04–0.10, and on each cache below that cache's smallest population lift there (0.07–1.02 across caches), so the sampling error does not return the lifted plateau to the band for these draws (`review_0923_measure.json`); other draws are not bounded by the measurement. The bounds hold on a cache only up to this perturbation, and the spike and plateau counts measured on the noise arms (§4.3) are observations that the proposition explains, not consequences it certifies.

B Deterministic sign-recovery lemma

Let $\gamma = \min\{|\lambda_i - 1| : \lambda_i \ne 1\}$ be the off-plateau spectral margin. If $\|\hat{D} - D\|_2 < \gamma$ on a common support, Weyl's inequality preserves the classification of every off-plateau eigenvalue. The n_{unit} exactly-unit eigenvalues of Theorem 1 have zero margin and may fall on either side of 1 (on

I-JEPA at cutoff 10^{-5}, 644 raw against 384 plateau-excluded $\lambda > 1$ counts, so 260 of 512 unit eigenvalues drift above one), which is why the plateau excision of step 3 in §3 is mandatory rather than optional. This is a fixed-support perturbation statement, not a finite-sample guarantee: the empirical support-selection step is not covered, and token correlation makes i.i.d. effective-sample arguments optimistic.

C ESTIMATOR ABLATIONS

The following cross-fitting controls are weaker than the third-bucket refit null of §4.1. Signed subspace scores are $+72.1/+74.7$ on images and $+48.5/+48.4$ on video with cluster-bootstrap 95% intervals bounded away from zero; they remain positive against a position-conditioned mean ($+81.7/+84.6$, $+53.3/+53.0$) and under intercept calibration ($+69/+71$, $+45/+45$). The position-conditioned totals are larger only because that rung selects a wider subspace, so it is a control the statistic passes rather than a strengthening.

Is the gap a loss mismatch? The refit null is a ridge head, a quadratic fit like the criterion the audit reads, with its intercept fixed at the fit half's target mean, while the official I-JEPA head was trained with SmoothL1. Refitting a linear head on the same features and half in SmoothL1 or in ℓ_1, the V-JEPA loss, and reading every head on the held-out half with the same estimator (`loss_mismatch_null.py`, five splits) moves the null's plateau-excluded regret mass from 6.5 to 6.8 and 7.3%, against 14.1% for the official head: the change of objective, with the intercept fitted as well, closes 4–10% of the gap, and the comparison does not separate the two changes. On the same halves the official head's NMSE is below the ridge null's on every split (0.564 against 0.571), so the scalar prefers the head whose regret mass is 2.2 times larger.

The pocket arms against elementary tests. The low-variance arms of §4.1 are of two kinds: constant shifts along low-variance target directions, and independent noise in the 64 lowest-variance ones. Under the same deployment protocol, a residual-mean test calibrated like the gate at the 99.5th percentile of healthy resplits fires on every shift arm and on no noise arm, and a span test, centered prediction rank above w, fires on every noise arm (rank $447 = 383 + 64$) and on no shift arm (`pocket_elementary.py`). Each kind can be detected without the spectrum by its corresponding test; the eigenvalue channel reaches both with one statistic.

One statistic instead of two. Collapsing the two channels into a single scale-standardized statistic instead of unioning them gives $0/200$ false positives (exact binomial 95% CI $[0, 1.8]\%$) at the same full power on all 120 faulted resplits of the calibration-range seeds, which on these paired resplits is not distinguishable from the union gate (McNemar exact $p = 0.25$). It is a simplification, not a repair.

The source-diversity floor. The detection gap of §4.1 needs source diversity to exist at all. With only 25% of the source buckets it inverts (-3.3 points) and turns positive only from 50% upward, reaching $+3.7$, $+6.3$ and $+7.5$ points at a half, three quarters and all of them.

The gate realizes 8.5% false positives at the looser 5% target ($17/200$, exact binomial 95% CI $[5.0, 13.3]\%$), and exceeds that target for a reason documented in the artifact: all of the excess sits in the eigenvalue channel, whose tail five hundred healthy resplits still under-estimate, and the two channels never co-fire, so the Bonferroni union is not the source. The single statistic is $T = \max_j |x_j|/s_j$, where x_1 is the relative NMSE gap of a resplit and x_2 its top-whitened-eigenvalue gap, thresholded at a calibration percentile of T. Its false-positive count depends on the normalizer, so the artifact reports every candidate scale on the same arrays: unstandardized gives $5/200$, the calibration standard deviation $2/200$, the per-channel 1% thresholds $3/200$, and only the calibration maximum reaches $0/200$ at both the 99th and 99.5th percentile.

Levels behind the gaps. Direction counts do not separate the heads: 0.467 against 0.457 of the support carries $\lambda > 1$ for the official and refit heads. The paired cross-fitting differences of §4.1 are bracketed at $[0.511, 0.578]$ on images and $[0.395, 0.485]$ on video. The in-distribution comparison reads 22.96% for the official head against 4.83% for the refit null. Its two-stage bootstrap leaves 63% of the source buckets distinct in each half. The gate's realized false-positive rate carries an exact binomial 95% CI of $[0.3, 4.3]\%$ and is consistent with its 1% target at one-sided binomial $p = 0.32$.

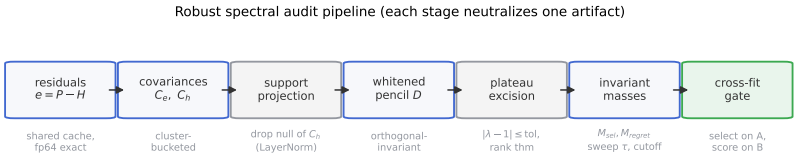

Figure 4: **The audit pipeline.** Each stage neutralizes one artifact; the six steps of §3 give the constants.

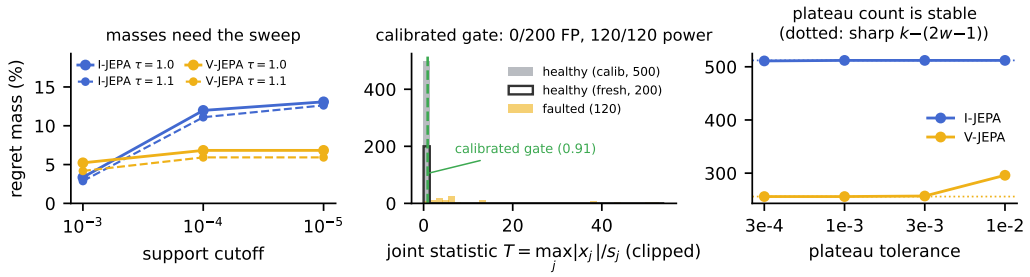

Figure 5: **Estimator ablations and the calibrated gate.** Left: regret masses vary with the support cutoff, so the sweep is mandatory. Middle: healthy resplits (gray/black) and faulted arms (amber) on the single statistic T of this appendix, the alternative to the deployed union gate, thresholded at its 99th calibration percentile ($0/200$ fresh false positives; power $120/120$ on the calibration-range resplit seeds plotted here and $118/120$ on seeds disjoint from calibration; the nearest faulted mass is the $+5\%$ severity tier). The deployed union gate of §4.1 realizes $3/200$ and fires on 118 of 120 on the disjoint seeds. Right: plateau counts sit on the sharp $k-(2w-1)$ line over the tighter decade of the plateau band ϵ_u and leave it at the loosest setting, where V-JEPA reads 296 against 256.

What the scalar channel alone would catch. An NMSE-only gate calibrated on the same healthy resplits misses part of the $+5\%$ grid, and how much depends on where its threshold sits. Matched to the union gate's realized false-positive rate, at the 98th percentile of the healthy NMSE gap in absolute value and so at 3 of 200 fresh false positives, it misses 1 of the 40 replicates. At the 99th percentile, a 1% nominal target giving one false positive in 200, it misses 3. At the 99.5th percentile the union gate allocates each channel, it misses 8, of which 6 are rank truncations and one each is a bias and a noise replicate. The eigenvalue channel recovers every miss at every threshold, so the growth is a property of the threshold rather than of the scalar channel's reach. These counts are derived from the per-arm channel values the release stores (`p0p_fpr500.json`, `power_arms`).

Power by scale. At the 99th percentile, power is $120/120$ for every scale except the unstandardized one (119). At the 99.5th the per-channel scale also drops a cell (119) and the unstandardized one falls to 117.

Free constants. Figure 5 sweeps the estimator's two free constants: regret masses vary with the support cutoff, while plateau counts are stable across an order of magnitude of the plateau band ϵ_u at the 10^{-5} cutoff.

D Downstream utility of regret directions

Regret is concentrated where the representation is least used downstream. Two readouts on I-JEPA/STL-10 show this, in two different bases. How the same quantities move along horizon, stride and context–target distance is a separate sweep, reported in Appendix E.

Native coordinates. Score each of the 1280 embedding axes two ways: by its diagonal error ratio $\mathrm{diag}(C_e)_i/\mathrm{diag}(C_h)_i$, and by its influence on a linear probe, taken as the ℓ_1 row norm of a standardized ridge probe's weights times that axis's own standard deviation. The two orderings run

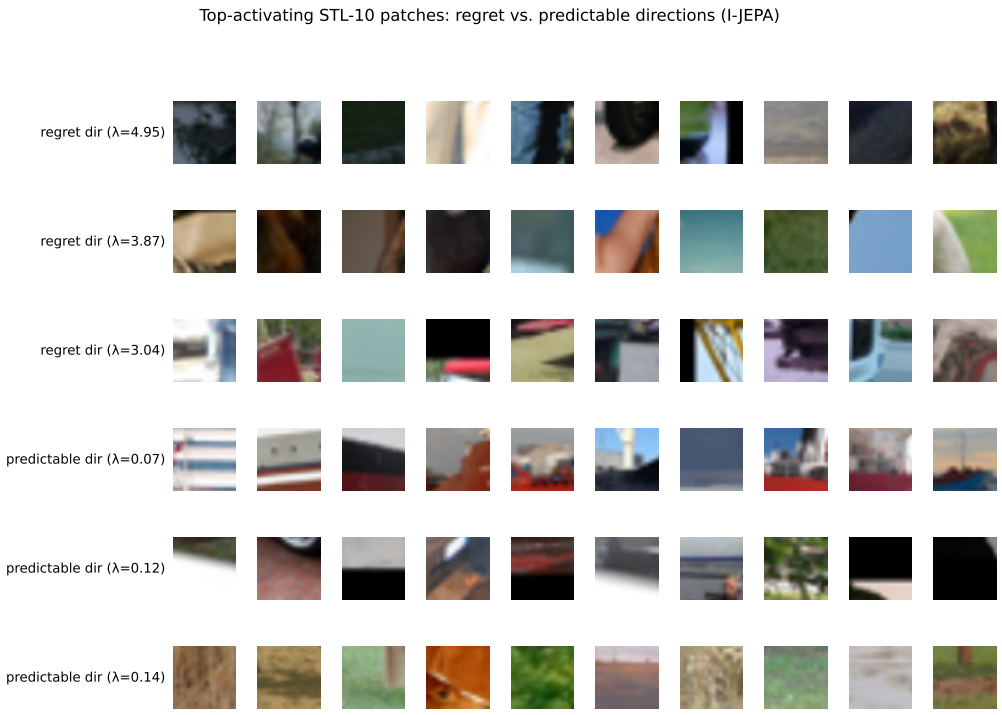

Figure 6: **What the regret directions select.** Top-activating STL-10 patches for the three most regretful vs. three most predictable audit directions of I-JEPA. Regret directions select heterogeneous boundary/occlusion fragments, content the context cannot locate exactly, while predictable directions select extended surfaces and object-color schemes. This grounds the finding: mean-relative regret concentrates in locally ambiguous, utility-light content, which is why strong downstream performance and directional regret coexist.

opposite (Spearman -0.770). This is a per-axis diagnostic and deliberately not one of the invariant masses of Definition 1: it is basis-dependent, and we report it only because the axes are the basis a probe actually reads. It also shows why the diagonal is not enough on its own. Not one axis has a diagonal ratio above one, while 32% of whitened directions carry $\lambda > 1$ on the same cache. The per-coordinate view finds no defect where the pencil flags a third of the directions.

Pencil directions. Here the evidence is causal rather than correlational. Projecting out the 409 flagged directions costs low-shot probes less than projecting out the same number of random directions (0.959 against 0.949 at 10-shot, from a 0.961 full-feature baseline).

How much the pockets are worth. Re-weighting each direction's excess by its raw target variance shrinks regret mass from 12.0 to 2.5% on images and from 6.8 to 3.1% on video, which is what the implicit bias toward high-influence features (Littwin et al., 2024) predicts of its residue. They are not automatically negligible either: weighting by linear-probe salience yields 19.8%, because ridge probes place non-trivial weight on low-variance directions. The two weightings bracket the engineering significance of the pockets rather than settling it.

What the regret directions select. Figure 6 shows the patches that most activate the most regretful and the most predictable audit directions of I-JEPA (§4.1).

E Condition sweeps and synthetic mechanism

Scalar normalized MSE rises monotonically with prediction horizon on V-JEPA ($0.656 \rightarrow 0.798$ over four horizons; causal masks are out-of-distribution for the pretraining mask family, so only within-family gradients are claimed) and is flat across temporal stride (0.712–0.715). Along

context–target distance on I-JEPA it is not monotone: it rises over the first four gaps ($0.518 \to 0.706$) and falls back at the widest (0.661). Regret mass moves differently on each axis. It falls strictly with horizon ($16.7 \to 15.0\%$ of residual energy in the flagged directions), is non-monotone with distance, and is flat with stride.

Under the support-projected, plateau-excluded estimator the off-plateau $\lambda > 1$ share of the support is near-flat over from-scratch toy training (46.3% at 1k steps, 41.5% at 15k) while the regret mass those directions carry decays (§4.2).

In a synthetic regime with common aleatoric contamination and a capacity-constrained encoder, a transiently signed per-direction objective beats magnitude-matched positive weighting under ten paired seeds ($\Delta R^2 = +0.213$, $p = 0.0008$). The effect vanishes exactly under equal-variance construction, where the two arms agree on every seed, and attenuates rather than disappears under random rotation of the target coordinates ($\Delta R^2 = +0.039$, still positive under the same paired test). For diagonal weightings, the mechanism is therefore strongest in axis-aligned high-variance regimes, but it also occurs outside them. Legacy targeted-intervention experiments on natural data, run before the corrected estimator under low-powered and partially failed protocols, showed no detectable downstream gain; a corrected-protocol rerun found the unstable toy video regime collapses in 3–4 of 5 seeds in *every* arm, leaving intervention efficacy unresolved rather than answered.

F Early warning under drift

The regret directions carry little downstream utility, and they are still worth monitoring because the audit moves *first*. Under a Gaussian input-corruption ramp ($\sigma = 0 \to 0.8$ in normalized units), a linear probe trained on clean features and deployed frozen, which is the deployment scenario, has covered only 1% of its eventual accuracy collapse at $\sigma = 0.1$ and 3% at $\sigma = 0.2$, while the audit channels (residual NMSE, top whitened eigenvalue) have already moved 13–17% and 30–37% of their full range: the deployed probe's accuracy remains nearly flat before declining sharply, the audit's rises smoothly from the first severity step. (A probe *retrained* at each severity degrades more gracefully (0.906 vs. 0.760 final accuracy), which gives the probe an advantage; we report the frozen deployment curve as primary and disclose both.) Regret mass itself is a *late* channel here; the early channels are the scalar NMSE and the top of the whitened spectrum, which localizes the failing directions.

G The rank stamp, one row per audited cache

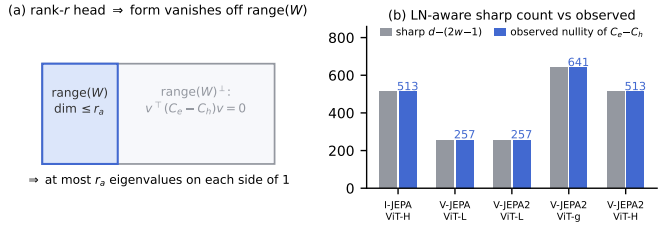

Figure 7: **The rank–signature bound.** (a) Geometry of Theorem 1. (b) Nullity of $C_e - C_h$ on five public checkpoints against the LayerNorm-aware prediction of Corollary 1, at counting tolerance 10^{-6}; Appendix G tabulates every row and the support plateaus alongside.

Table 4 is the per-checkpoint bookkeeping behind Figure 7(b) and behind the counts quoted in §4.2. Every cache returns the same signature $(+384, -383)$, so $\mathrm{rank}(C_e - C_h) = 767 = 2w - 1$ throughout and the nullity of the form is $d - (2w - 1)$ on each one. The centered pre-projection activations have rank $383 = w - 1$ on all 6 caches, across three ambient dimensions, so the LayerNorm hyperplane that the corollary assumes is present in every audited predictor rather than in the one where it was first checked.

Counting tolerance. Eigenvalues of $C_e - C_h$ are counted against a relative threshold, τ times the largest absolute eigenvalue, taken over the whole cache. The count is 767 on all six caches

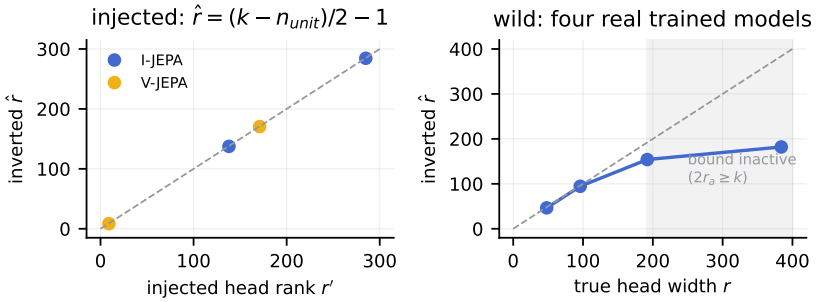

Figure 8: **Severity inversion.** Injected ranks are recovered to within a few percent; the widths of independently trained heads are recovered exactly in the regime where the bound constrains the plateau.

at $\tau = 10^{-6}$. Loosening it adds two modes: one cache reaches 769 at 10^{-7} and five do at 10^{-8} (Table 4), past the ceiling Corollary 1 proves for an exact LayerNorm constraint. Wherever they appear the signature becomes $(+(w{+}1), -w)$, which is $(q{+}1, q)$ at $q = w$, the saturated signature of the same head without that constraint. That is what the caches hold. The LayerNorm feeding the head runs in float32 under bf16 autocast, so the constraint holds to float32 round-off until the cache stores the activations in fp16; along the LayerNorm null direction the stored head input varies by 1.1×10^{-4} to 3.2×10^{-4} of its mean across the six caches, the order of fp16's unit roundoff $2^{-11} \approx 4.9 \times 10^{-4}$ rather than bf16's 2^{-8}.

Reprojection. Projecting every cached activation onto the constraint hyperplane, taken as the near-null direction of the centered activations, restores rank 767 at every tolerance on each of the 5 caches where the extra modes appear (`review_0923_measure.json`). The two modes therefore come from the variance the fp16 cache leaves in the one direction LayerNorm forbids, which we interpret as a violation of the constraint at the cache's precision rather than as model structure; the storage and decomposition contributions are not separated further. 10^{-6} is the column we read because it counts the constraint the architecture imposes. The I-JEPA cache holds 767 at every tolerance. Counts taken on an evaluation half rather than the whole cache move one boundary eigenvalue across the 10^{-7} threshold on the official-forward V-JEPA2 cache, which is the difference between this column and the half-split figures reported alongside that model's own audit.

Support plateau. The plateau column counts $|\lambda{-}1| \leq \epsilon_u$ in the whitened spectrum at support cutoff 10^{-4}, and attains its floor $k - (2w - 1)$ on the five video-family caches. It exceeds the floor on the I-JEPA cache, where that cutoff retains 825 of 1280 directions: equality needs the compression onto the retained support to preserve the rank of the form, and a cutoff that discards directions the form lives on loses inertia instead. At the tighter 10^{-5} cutoff of Table 1 the same checkpoint attains it exactly ($512 = 1279 - 767$). The two passes bucket sources differently, so k here and k in Table 1 differ slightly at the same nominal cutoff; the ambient counts do not depend on the cutoff at all.

Table 4: **Rank stamp per audited cache.** Signature $(+384, -383)$ and rank 767 on every row at counting tolerance 10^{-6}. Nullity is of the form $C_e - C_h$ in the ambient space; plateau is the unit count of the whitened pencil on the retained support. The last column is the rank at the two looser tolerances. V-JEPA2 is read through the official masked forward and through the HuggingFace adapter (HF).

Cache	d	k	nullity / $d{-}(2w{-}1)$	plateau / $k{-}(2w{-}1)$	rank at 10^{-7}, 10^{-8}
I-JEPA ViT-H/14	1280	825	513 / 513	61 / 58	767, 767
V-JEPA ViT-L/16	1024	1023	257 / 257	256 / 256	767, 769
V-JEPA2 ViT-L (official)	1024	1023	257 / 257	256 / 256	767, 769
V-JEPA2 ViT-L (HF)	1024	1023	257 / 257	256 / 256	767, 769
V-JEPA2 ViT-g (HF)	1408	1405	641 / 641	638 / 638	767, 769
V-JEPA2 ViT-H (HF)	1280	1279	513 / 513	512 / 512	769, 769

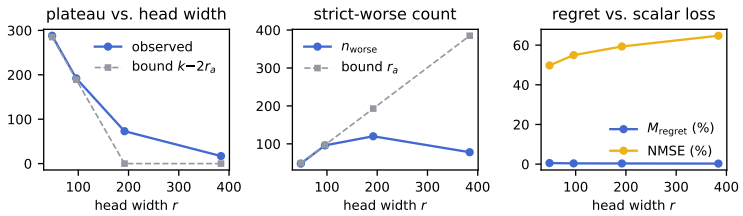

Figure 9: **Frozen capacity sweep** (fixed final heads, fresh collection). Plateau tracks $\max(k - 2r_a, 0)$; strict-worse counts respect r_a; regret mass and scalar NMSE move in opposite directions.

SEVERITY INVERSION AND OFFICIAL-LOSS CONTROLS

Severity inversion. The ambient signature counts of $C_e - C_h$ are basis-independent and need no whitening or tolerance. Because $C_h^{\dagger/2}$ is singular, the whitened form $D - I$ is a compression rather than a congruence of $C_e - C_h$, so these counts upper-bound n_{worse} and n_{better}; they match on the shared cache at the 10^{-5} cutoff. There $\hat{q} = \lceil \text{rank}(C_e - C_h)/2 \rceil - 1$ recovers the effective centered rank with zero error, and on four independently *trained* heads the rules flag the rank-limited widths and invert them where the bound binds, degrading gracefully where $2r_a \geq k$ leaves the plateau unconstrained (Figure 8).

Official-loss controls. Re-calibrating severities so the *official* scalar is matched (SmoothL1 for I-JEPA, ℓ_1 for V-JEPA) leaves the frozen rulebook intact ($8/8$ and $7/8$, the one miss being the documented heavy-misalignment edge; two video arms undershoot their +30% loss target, at 1.18 and 1.26). Because official-loss matching does not equalize NMSE across arms, separability is a property of the faults rather than of the matching protocol. The regret is real under an ℓ_1 reading too: on the top 20 regret directions, selected on one half of the evaluation sources and scored on the other, the predictor's ℓ_1 error exceeds that of the ℓ_1-optimal constant, the directional median, on 20 of 20 for I-JEPA and 20 of 20 for V-JEPA.

H THE TWO V-JEPA2 GRIDS

V-JEPA2 is audited through two caches, and the fault grid exists on both. The official masked pretraining forward runs the encoder on context tokens as the model was trained; §4 names it the primary V-JEPA2 evidence and the blind transfer number of §4.3 is computed on it. The HuggingFace adapter runs the encoder on the full clip, which suits feature extraction and not a masked-prediction audit, and this paper uses it for baseline-relative readouts and the structural stamp count. An earlier version of this work scored the fault grid on the adapter cache, which was an inconsistency with its own setup rather than a choice. Rebuilding the grid on the official forward, with the same severities, the same deduplication and the same frozen rules, moves the V-JEPA2 row from $42/48$ to $46/49$ and the blind total from $89/98$ to $93/99$. Both grids are in the release.

The differing denominators arise from cache-dependent deduplication. Rank truncation saturates on this head, so arms that request a severity the head cannot reach collapse onto an already present predictor and are deduplicated. The ceiling is $+19.2\%$ NMSE on the adapter cache and $+29.8\%$ on the official forward, so the adapter grid loses two arms to deduplication and the official grid one. The extra arm is not an extra test case; it is one fewer duplicate under a cache-dependent saturation geometry.

The official-forward cache reduces the errors while preserving their pattern. On the adapter cache the frozen rules miss six arms: three abstentions at $\times 1.3$ misalignment and three $\times 1.5$ misalignment arms read as rank. On the official forward the three abstentions become correct and the three misclassifications remain, on the same arms and in the same direction. Nothing in the grid was selected or reweighted; every V-JEPA2 arm was regenerated and rescored.

I TRAINING-INDUCED FAULTS, THE FEATURE LADDER, AND CLASSICAL BASELINES

This appendix reports three controls on the fault atlas of §4.3: the same faults generated by fitting rather than injection, a ladder of cheaper feature sets for morphology, and classical process-monitoring baselines on the loss-matched arms, with a counterexample for representation-quality scores.

Training-induced faults, in full. To test that fingerprints reflect mechanisms rather than injection formulas, we generate the same faults by *fitting*: reduced-rank regression, heads fit on mean-drifted or noisy targets, on partially permuted pairs, and heads undertrained by few-step SGD, with directions drawn independently of everything used while developing the rules. Over 5 seeds at two severities the frozen six-feature rulebook places 40 of 40 trained arms in the intended class, and all 10 of 10 undertrained heads read as rank-deficient, which is correct: gradient descent acquires dominant singular directions first. The trained noise arms are a different mechanism from the injected ones. A head fitted to noisy targets is still affine, so its plateau sits on the LayerNorm floor rather than below it (47 of 814 on I-JEPA) and a span test does not flag it; the rulebook names it noise through its spike count (306 and 351 against a healthy 0), the variance inflation of an overfitted affine head, which Proposition 2 does not describe.

Do cheap statistics suffice for morphology? We train one nearest-centroid classifier on I-JEPA arms and evaluate it on the held-out models, giving every rung of a feature-set ladder the same healthy-baseline access. Percent correct on the two held-out models climbs from scalar NMSE alone (19–22) through residual-mean, per-coordinate and unwhitened spectral summaries (31–78) to the whitened-pencil rulebook features (88 and 87.5; this ladder reads V-JEPA2 through the adapter cache).

Rank and mean alone. The ladder omits the statistics used to construct the faults, which suffice on their own. A rule fixed in advance on the numerical rank of the centered predictions (above w: noise; below the healthy rank: rank) and the residual-mean shift against its sampling error ($z > 5$: bias; otherwise misalignment) types every arm of the three primary grids, 149 distinct arms, with margins that leave nothing to tune: rank arms reach at most rank 332 against a healthy 383, noise arms at least 1024, and bias arms at least $z = 91.7$ against at most 0.74 for noise and exactly zero for misalignment, whose permutation preserves the mean (`elementary_typing.py`). Used this way the first moment types every bias arm, which the six-feature rulebook of §4.3 does not show. On injected faults the whitening is therefore not what makes the families separable; it is what lets one residual pencil carry all four signatures together with the regret certificate.

Representation-quality scores. RankMe (Garrido et al., 2023) is unchanged under every head fault, by construction, since it reads only encoder embeddings; that is why predictor-side unit tests are needed alongside representation-quality scores (Thilak et al., 2024). The gap is not only about head faults. Permuting one label-carrying coordinate of a representation across samples keeps the multiset of its values, so its variance is preserved exactly, and holding the predictor fixed leaves the same map on both sides; $\mathrm{tr}(C_h)$ and NMSE then agree to the last bit of double precision while the linear probe falls from 0.99 to 0.50, and RankMe returns the same value to within 2×10^{-6} (`rankme_counterexample.py` in the release). No statistic that reads only these second-moment summaries can order the two, which is why the audit reports the residual's directions rather than a single quality score.

The blindness reaches further than second moments. The coordinates of this construction are independent, so permuting one of them across samples leaves the joint law untouched, and with it every statistic of the embedding *distribution*. That family includes the collapse-preventing objectives that score how close an embedding batch is to an isotropic Gaussian (Balestriero & LeCun, 2025): a sketched normality test of that kind separates our pair by 0.7%, against 1.9% of its own seed-to-seed spread, so it does not separate them at all. Such an objective guarantees a property of the embedding it shapes; the audit measures the predictor.

Classical baselines. On the SmoothL1-matched image arms of the official-loss control (eight arms: rank, bias, noise, misalignment at two severities; same cache, $169{,}600$ held-out tokens), we scored five classical detectors: the PCA-residual SPE/Q statistic (302 components for 90% of healthy residual variance), Hotelling T^2 on the residual means, the top eigenvalue of the unwhitened C_e, canon-

Table 5: **Classical baselines on the loss-matched image arms** (correct fault labels out of 8; interval rules fit in-sample per channel).

Channel	1 cut	2 cuts	3 cuts (ref.)
PCA residual SPE/Q	$4/8$	$5/8$	$6/8$
Hotelling T^2 (residual means)	$4/8$	$6/8$	$8/8$
top eigenvalue of C_e	$4/8$	$6/8$	$8/8$
canonical variate analysis	$4/8$	$6/8$	$8/8$
$\|C_e - C_h\|_F$	$4/8$	$6/8$	$7/8$
$\|C_e - C_e^{\text{healthy}}\|_F$	$4/8$	$6/8$	$7/8$
spectral rulebook (frozen)		$8/8$	

ical variate analysis (CVA), and $\|C_e - C_h\|_F$, plus $\|C_e - C_e^{\text{healthy}}\|_F$ as a supplement. Each scalar received the *best* interval rule fit in-sample on the same eight values, which is deliberately generous (Table 5). With one cut every channel names $4/8$ arms; with two cuts, 5–$6/8$; even the ideal one-dimensional rule (three in-sample cuts, one interval per fault class) reaches $8/8$ only for T^2, the top eigenvalue and CVA. CVA is the closest classical relative of the audited pencil and the strongest of these baselines, but it is structurally blind to one of the faults: it centers both blocks, so the two bias severities return channel values identical to twelve significant digits, and its rank on that family comes from the in-sample cuts rather than from the statistic. The frozen rulebook, whose thresholds predate these arms, scores $8/8$. Classical deviation scores can order these arms, but classifying the fault requires either oracle in-sample thresholds or the whitened morphology.

J ENCODER-LEVEL FAULTS

Faults *upstream* of the head, zeroing a full transformer block of the context encoder (depths $8/16/24$), zeroing one of the target encoder (depth 16), or 50% magnitude pruning of every linear layer, never receive a rank, bias or noise verdict. A single run of this battery cannot serve as a measurement: the mask collator samples block locations from the global RNG and hands out per-batch seeds through a counter that the loader workers race for, so 5 healthy draws of the same model span 1.42% in NMSE, the size of the gaps being tested. Seeding the global RNG with a single-process loader makes a collection reproducible to the last digit, so we pair the comparison: within a draw the baseline and every arm see identical masks, and we report the mean over 3 draws. Pairing cuts the spread of each gap by a factor of between two and three hundred, depending on the arm.

Two arms clear the gate's NMSE channel, the block skip at depth 8 and the pruning ($+7.1$ to $+11.1\%$), and both read as the elimination family: scalar up, plateau and regret mass unchanged. The block skip at depth 24 sits just below it ($+3.65 \pm 0.33\%$ against a 4.1% threshold, below it in every paired draw); unpaired, its spread is ± 1.15 points and it routes both ways across draws, as expected for a borderline arm without a paired comparison. Block skip at depth 16 ($+1.5\%$), the target-encoder skip (-0.73%) and the perturbations the 630M-parameter encoder simply absorbs (qkv weight noise at 0.05–0.1 standard deviations, whole-encoder int8 weight quantization, $+0.02$ to $+0.22\%$) sit below that channel.

On the eigenvalue channel exactly one of them fires: the target-encoder skip, at $\Delta \log_{10} \lambda_{\max} = 0.0765 \pm 0.0004$ against a threshold of 0.0495, routing to `pocket` in every draw, while the largest among the others reaches only 0.0232. Shifting the targets moves the residual into directions the targets barely span without making the residual larger, which is what that route is for. This is the intended division of labor: positive fingerprints certify *head* mechanisms; upstream damage that the audit sees lands in the fourth family, so the audit localizes a fault to "not the head" rather than misattributing it.

K ARE THE RULEBOOK'S CONSTANTS TUNED?

The rules transfer across checkpoints, which the blind score measures, but every fault family was visible when the rule constants were chosen (the four swept here from rules 2 to 5, and the 2% healthy constant of the original first rule, since replaced by the calibrated gate; rule 5's plateau

window is swept in Appendix L), so a reader may reasonably ask whether each of the five constants is fitted to a single fault family. A leave-one-family-out fit cannot answer it, because removing a family removes the positive class its own threshold is fitted against. We instead use two measurements that do not require this holdout.

Margin. Sweeping each threshold and holding the others fixed, we record the widest interval over which every verdict on both held-out checkpoints is unchanged; the sweep reads the two held-out 50-arm dense grids the blind-transfer number is scored on (V-JEPA, and V-JEPA2 through the official masked forward), and the severity holdout below re-derives its constants from the I-JEPA grid. Three of the five constants sit inside wide windows: the bias spike threshold may be anything in $[0.000, 0.975]$ against a shipped 0.3, the spike-count constant behaves the same way, and the original healthy constant's window, $[0, 4.95\%]$ of relative NMSE, contains the calibrated gate's 4.1%, consistent with the gate changing no grid verdict when it replaced that constant (checked arm by arm in the release). The spike-count window carries a condition the others do not: the rule is a ratio against the healthy baseline, and the guard that protects it from dividing by zero turns it into an absolute count whenever that baseline has no super-unit eigenvalues. It was measured on batteries whose healthy runs carry 23 and 11 of them; on the natural battery, whose healthy runs carry none, the same constant misfires on two families that the other four rules place correctly.

The other two are tighter. The regret-mass constant holds only over $[0.152, 0.212]$ against a shipped 0.2, and the plateau constant sits near the upper edge of its window ($[0.006, 0.111]$ against a shipped 0.10) rather than in the middle. We read this as three constants that reflect the mechanism and two whose placement is sensitive on this battery.

Severity holdout. Re-deriving the plateau, spike-count and eigenvalue-spike thresholds mechanically from I-JEPA's $+20/+30/+50\%$ arms alone, where a fault is obvious (the regret-mass constant does not separate the families there and the healthy constant is kept), and scoring the $+5/+10\%$ arms of the held-out checkpoints, the re-derived rules recover 22 of 40 against the shipped rules' 37. The shipped constants were therefore chosen with the low-severity regime in view, which is consistent with the detection floor near $+10\%$ NMSE that §4.3 already reports, and is a limit on the rules rather than a property of the spectrum.

L BEYOND JEPA: SIMSIAM AND DISTILLATION AUDITS

Nothing in the audit is specific to masked prediction since Theorem 1 binds *any* predictor whose final layer is affine. Two out-of-family tests confirm this. *(i) SimSiam* (Chen & He, 2021), augmentation-based SSL with an MLP predictor (bottleneck 512): the signature of $C_e - C_h$ saturates at $(+512, -511)$, recovering the augmented rank $r_a = 512$ exactly, and, instructively, in the *other* bias-alignment case of the theorem than every JEPA head we audited ($(r_a, r_a{-}1)$ vs. $(q{+}1, q)$ saturation). That count is read at $\tau = 10^{-7}$ rather than the 10^{-6} of Appendix G, and the direction of the correction is the reason to trust it: on the JEPA caches loosening τ pushes the count *past* the ceiling, into modes the exact LayerNorm constraint does not allow (Appendix G), whereas here the tighter threshold leaves the count *below* it (330 at 10^{-6}) and loosening lands on the bound exactly. In both cases the proven ceiling, not the tolerance, decides which count to read.

The unchanged fault battery transfers: rank, bias, and mild misalignment classify correctly; the two noise arms initially read as bias because the frozen rule keyed on an *absolute* spike magnitude and SimSiam's healthy predictor already carries a $\lambda \sim 10^3$ spike (itself an audit finding); making that one feature baseline-relative, which is the rulebook's own stated design principle, yields $7/8$ with identical decisions on all three JEPA grids, and we ship the corrected rule. Scored on the full dense grid rather than that eight-arm battery, the frozen rules reach $43/50$ on SimSiam, with 6 abstentions and one misclassification. The abstentions are the heaviest misalignment arms: SimSiam's healthy predictor already carries the plateau and the $\lambda \sim 10^3$ spike that the alignment rule expects to find unchanged, so at high severity the rulebook declines rather than guessing, which is the behavior the abstain rule exists for. The thresholds here are still the ones fit on I-JEPA, so they are blind; the noise-before-bias ordering was set after looking at the eight-arm battery above, so on this grid the ordering is frozen but not blind, and we count SimSiam outside the headline transfer number for that reason. The reordering changes no verdict on any JEPA grid: the earlier rule order returns the same label on all 200 cells of the four dense grids, checked cell by cell in the release (`e12_rulebook_v1_identity.json`), so the blind counts above do not depend on it.

(ii) Distillation quality. For each officially distilled DINOv2 student, we audit the best affine map to its ViT-g teacher: rank-truncated heads saturate the theorem's bound exactly in this new domain ($n_{\text{worse}} = 65/257$ at $r = 64/256$, bounds $65/257$); the ViT-S student's plateau sits at 0.500, exactly the $1-(2\cdot384-1)/1535$ floor forced by its dimension, so the spectrum reads off the structural deficit of a too-small student, while capacity buys *reachability* (NMSE $0.746 \to 0.610 \to 0.485$ for S/B/L) while the plateau-excluded regret mass stays between 34 and 39%: what remains unreachable carries persistent per-direction regret, a distinction invisible to any scalar distillation loss.

(iii) BYOL, and where the rulebook stops. BYOL ResNet-50 (mmselfsup, 200 epochs on ImageNet-1k) is the only audited head whose thresholds and whose rule order were both fixed before its cache existed. The rulebook fails on it: 32 of 50 distinct arms against a preregistered bar of 35. The failure is not spread over the atlas. Alignment is read $15/15$ and bias $11/15$ with four abstentions, while rank is $0/5$ and noise $6/15$. The BYOL audit requires a scale convention. BYOL trains a cosine loss, which fixes no scale between the predictor output and the target projection, so the audit runs on the raw affine output against the raw target projection with a single global scalar fixing that freedom ($s = 0.00818$), chosen once on the healthy pair and held across every arm, which puts the healthy NMSE at 0.427; severities are then matched in NMSE as everywhere else. Read after that scaling, the theorem's own bookkeeping confirms by measurement what the architecture implies: the signature of $C_e - C_h$ is $(+2, -254)$ at every counting tolerance, so the head's rank reaches $256 = d$ and the forced plateau is empty rather than merely small. Both losses are structural, and both follow from quantities measurable on the healthy cache alone, before a fault is injected and without labels.

The rank rule reads a change in the unit plateau, and this head offers none. Corollary 1 forces a plateau of $\max(0, d - (2r_a - 1))$, and BYOL's final affine maps into $d = 256$ with no bottleneck below that width, so nothing is forced at baseline. Matching severities to the same NMSE grid selected truncation ranks from 245 down to 146, every one above $d/2$, so no arm forces one either, and the measured change in plateau share is exactly zero on all five. The four heads where the rule works sit on the other side of that line by a wide margin: their smallest truncation rank is at most 0.06 of $d/2$, against 1.14 here. The regret in those arms instead appears as super-unit eigenvalues, at spike ratios of 10 to 29, and the noise rule fires first.

The noise rule fails in the mirror image. The ℓ_2-normalized BYOL target is close to isotropic: its largest eigenvalue is 39 times the median against 293 to 5275 on the four heads where the rules work, and 0.21 of directions lie below one percent of the largest against 0.71 to 0.90. The rule presupposes an anisotropic target for a reason that Proposition 2 makes precise: isotropic noise of variance σ^2 is guaranteed to lift eigenvalues past $1 + \tau$ only in proportion to the number of target directions whose variance is below σ^2/τ, and on an isotropic target it shifts the healthy spectrum rigidly by $\sigma^2/\bar{\lambda}$, so nothing at or below the plateau crosses $1+\tau$ until $\sigma^2 > \tau\bar{\lambda}$. Here injected broadband noise raises no eigenvalue past 2 until $+30\%$ NMSE, and below that the nine mild noise arms present an unchanged plateau with small regret mass, which is the alignment signature, so rule 5 claims them.

Table 9 and Figure 10 read both quantities off every audited head, and they separate BYOL from the other four on both axes. The alignment errors on the other heads have a separate explanation that does not add a third condition. Rule 5 fires only inside $|\Delta\text{plateau}| < 0.05$, and across all five grids an alignment arm is lost exactly when it leaves that window: the three heads that lose none never move the plateau by more than 0.005, while V-JEPA2 reaches 0.11 and SimSiam 0.08 at their strongest severities. The sign decides which way it fails. V-JEPA2's excursion is positive and large enough to trip the rank rule first, so those arms are misclassified; SimSiam's is negative, so nothing fires and the rulebook abstains. These results do not provide a quantity measurable in advance: V-JEPA and V-JEPA2 have the same healthy plateau share to four decimals, and misalignment moves it by nothing on one and by 0.11 on the other, so how far an arm travels is a property of the head and the fault together rather than of the healthy cache. It is the mechanism behind the two documented edges, not a precondition alongside the other two.

The two out-of-family losses are therefore of different kinds. BYOL's are structural: two rules have nothing to read on that head, and no choice of threshold would give them something. SimSiam's are a calibration edge: every rule can read what it needs, and the arms are lost because misalignment there walks out of one rule's fixed window. The rulebook then abstains rather than guessing, and we leave the window where it was fit. One is a limit on where the method applies, the other a limit on how finely a frozen constant can be set, and only the first belongs in the scope conditions.

Table 6: **The alignment rule's plateau window, swept.** Distinct arms named correctly as the window of rule 5 moves, with every other constant held. Four grids are constant across the whole range, so the value is not determined by the data the rules were fit on; the six arms it decides are all on SimSiam, and all of them are abstentions rather than misclassifications. The shipped value is the row in bold and was not chosen with this table in view.

Window	I-JEPA	V-JEPA	V-JEPA2	SimSiam	BYOL
0.02	50/50	47/50	46/49	40/50	32/50
0.04	50/50	47/50	46/49	40/50	32/50
0.05	**50/50**	**47/50**	**46/49**	**43/50**	**32/50**
0.08	50/50	47/50	46/49	47/50	32/50
0.09	50/50	47/50	46/49	49/50	32/50
0.15	50/50	47/50	46/49	49/50	32/50

We quantify sensitivity to that second limit because the governing constant was not among the five swept in the margin appendix. Sweeping it from 0.02 to 0.15 and holding the rest, 4 of the five grids do not move at all: the constant is unconstrained by the model the rules were fit on and by two of the three it was tested on. Only SimSiam moves, and it reaches $49/50$ at a window of 0.09 (Table 6). We did not take those six arms. The thresholds were fit once on I-JEPA and frozen. Widening the window after seeing a held-out model would invalidate that claim, even if it recovered six arms. The number generator refuses to build if the window ever leaves the value it was frozen at.

None of this makes the audited architecture a historical one, and it is worth being precise about what an isotropic target does and does not cost us. The certificate remains valid. Whitening against an isotropic target is a scaling, so $\lambda > 1$ still says that the predictor loses to the unconditional mean along that direction, and the covariance ceiling that makes it a statement about the predictor is a fact about the Bayes predictor rather than about the shape of C_h. What it costs is the separation between morphologies: with no anisotropy to disturb, broadband noise and alignment damage leave the same trace, which is the second scope condition read from the other side.

The line we audit is also still being extended (Mur-Labadia et al., 2026), and it is the line where both conditions hold. More to the point, the two programmes treat the same complaint differently rather than one replacing the other. This paper documents a scalar that cannot be trusted on its own; a collapse-preventing objective answers it by changing training until the scalar can be, and we answer it by diagnosing, without retraining, checkpoints and predictive stacks that already exist. Where retraining is unavailable, which includes verifying a published export against its own protocol, the second answer is the only one available. Where conditional prediction is the product, the predictor remains: world models built on the same isotropy regularizer keep one (Zhao et al., 2026), and the audit surface persists with it.

Table 7 tests the claim as directly as possible without a new head. Dropping the rank rule from the chain costs four or five arms on each of the heads whose bottleneck binds and costs *nothing at all* on BYOL, because there it never fires; dropping the noise rule costs fifteen on each of those four and six on BYOL. The two rules whose preconditions BYOL fails are the two that are not doing its work, and the two that survive are worth as much there as anywhere. Dropping the healthy rule changes nothing on any grid, since an injected grid contains no healthy arms for it to catch, and we report this to avoid implying that the healthy rule contributes on these grids. Table 8 says the same thing from the other side, one rule at a time: rank is at or near perfect on the four heads whose bottleneck binds and is $0/5$ on the one whose does not, noise is perfect on all four and $6/15$ there, and the two rules that survive on BYOL are the two that give ground elsewhere, on the video and SimSiam alignment arms this appendix and §4.3 already document.

Together they suggest a scope condition: the rank rule needs a head whose bottleneck binds, and the noise rule needs an anisotropic target. The second half is the isotropic case of Proposition 2, read with the Weyl slack a target this far from isotropic carries; the first, and the pairing of the two as the boundary of the rulebook, we report as exploratory rather than settled. The quantities were computed after BYOL had been scored, and four successes and one failure do not establish a general rule. What the episode does settle is that this was not a tuning failure. No threshold and no rule order was changed in response to it, and the two rules that survive are the two the argument

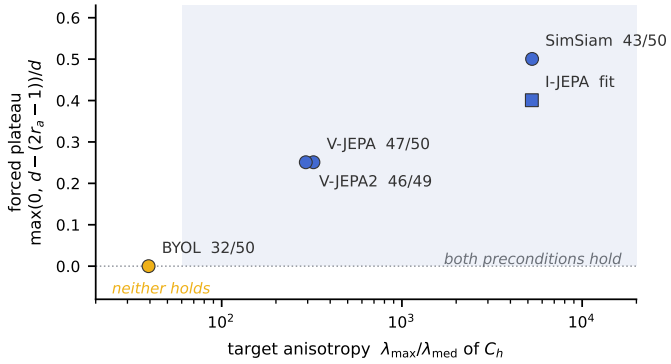

Figure 10: **Where the rulebook holds.** Both axes are measured on the healthy predictor and its targets, with no fault injected and no labels: the anisotropy of C_h and the plateau Corollary 1 forces. Marker labels give the blind score on that head's dense grid; the square is the head the rules were fit on. The shaded region is where both preconditions hold, drawn as a quadrant rather than a fitted boundary because nothing here is fitted. BYOL is outside on both axes, and is the only head where the rulebook misses its bar.

above leaves untouched. Whether the five constants are five memorizations is a separate question, and Appendix K answers it separately.

The prediction the condition makes, that a head trained to an isotropic target loses the rank and noise rules while keeping bias and alignment, is stated here so that it can be checked against a head we have not audited. Kuhn et al. (2026) is the natural test, and it fails both conditions for reasons that are architectural rather than incidental. Its objective is already a squared error between two vectors in $\mathbb{R}^{256}$, a local view's projector embedding against the global view's, so the audit runs on it unchanged. That projector ends in an affine map $2048 \to 256$, which is BYOL's geometry and forces no plateau, and SIGReg drives the embedding distribution towards a standard Gaussian, so the target is isotropic by objective rather than by accident. The conditions therefore predict that the rank and noise rules fail there while bias and alignment transfer. We registered that prediction, with its refutation conditions, before any cache was built, and it cannot be run from the public release: the checkpoint carries the encoder alone, its model code stating that the training loss and the projection head were removed, so the two vectors the prediction is about are not constructible from the public weights. Retraining a projector ourselves would audit our head rather than theirs. The registration is therefore a test that anyone holding the trained head can run against fixed bars, and we report it here in that form.

The released encoder still allows a separately registered measurement, in a different space, and it shows where the isotropy is. Reading a local view of a clip against the global one on that encoder's own output, 0.96 of the directions of C_h sit below one percent of the largest and the ratio to the median is 1.0×10^5, which makes it the most anisotropic target in this paper, against 0.71 on I-JEPA and 0.21 on BYOL. The guarantee is a property of the space the release discards, not of the space downstream uses. On the same pass, 55 of the 239 retained directions carry $\lambda > 1$, so the certificate remains informative where the registered test cannot run. This is not that test and settles none of its bars. An independent measurement already addresses one half of the prediction: Le (2026) report the projected spectrum of a LeJEPA model as near-flat, with a top-to-bottom ratio of at most 3.6, against a four to five order-of-magnitude drop for the rectifiable target they propose instead. That is an independent measurement of the condition our noise rule needs, taken on the family before we audit it, and it is why the isotropy half of the prediction is stated with more confidence than the rule-level halves.

Table 7: **Leave one rule out.** Distinct arms named correctly when one rule is skipped in the ordered chain, so its arms fall through to whatever fires next. Nothing is retuned and no threshold moves. The rank row summarizes the scope claim: removing that rule is free on the head with no bottleneck and costs four or five arms everywhere else.

Rulebook	I-JEPA	V-JEPA	V-JEPA2	SimSiam	BYOL
as shipped	50	47	46	43	32
without 1 healthy	50	47	46	43	32
without 2 rank	45	42	42	39	32
without 3 noise	35	32	31	28	26
without 4 bias	35	35	31	28	21
without 5 alignment	35	32	34	34	17

Table 8: **Which rule fails where.** Distinct arms named correctly by the frozen rulebook, one fault family at a time, on every dense grid. The pattern is what the scope conditions predict: rank collapses only on the head with no bottleneck, noise only on the head with a near-isotropic target, and the alignment losses elsewhere are the documented edges rather than a scope failure.

Head	rank	bias	noise	alignment
I-JEPA (fit)	$5/5$	$15/15$	$15/15$	$15/15$
V-JEPA	$5/5$	$12/15$	$15/15$	$15/15$
V-JEPA2	$4/4$	$15/15$	$15/15$	$12/15$
SimSiam	$4/5$	$15/15$	$15/15$	$9/15$
BYOL	$0/5$	$11/15$	$6/15$	$15/15$

Table 9: **Two healthy-cache quantities, and where the rulebook holds.** Both are read from the healthy predictor and its targets, with no fault injected and no labels. The plateau column is what Corollary 1 forces, $\max(0, d - (2r_a - 1))$, as a share of d: the rank rule has nothing to read when it is zero. The anisotropy columns are the largest eigenvalue of C_h over its median, and the share of directions below one percent of the largest. Blind scores are on distinct arms; I-JEPA is the model the rules were fit on.

Head	forced plateau	$\lambda_{\max}/\lambda_{\text{med}}$	share below 1%	blind
I-JEPA	0.401	5259	0.71	(fit)
V-JEPA	0.251	323	0.79	$47/50$
V-JEPA2	0.251	293	0.73	$46/49$
SimSiam	0.500	5275	0.90	$43/50$
BYOL	0.000	39	0.21	$32/50$

M FAULT DAMAGE IN IMAGE SPACE, AND WHICH PATCHES ARE READ

The rulebook does not use these results. They show the effects of the injected faults on images and identify the image positions used for every measurement.

Damage maps. Figure 11 shows the per-patch prediction error each fault *adds* to the healthy head on two STL-10 images, each patch averaged over the mask draws (of 16) that predicted it. The four arms are the $+30\%$ arms of the audit-set grid, matched there to within 0.001 of the $1.30\times$ healthy target in absolute NMSE (the rank arm is integer-valued and lands at $+30.17\%$). On these two images the four arms do not reach identical local NMSE, because bias and noise add a near-constant per-token error and rank and misalignment do not, so the maps share one linear scale and each row states its coefficient of variation across patches: 0.63 for rank, 1.17 for misalignment, 0.11 for bias and 0.04 for noise.

Spatial statistics. Spatial structure is read with Moran's I over the four-neighbor graph of observed patches and a one-sided permutation test for positive autocorrelation ($2{,}000$ permutations). Rank truncation leaves clustered damage (I 0.38 and 0.32 on the two images, $p = 0.0005$) whose pattern differs between the images; misalignment destroys a random subset of patches (I -0.02, 0.04). Bias and noise cannot be distinguished visually on the photo, and their maps correlate at -0.02. The maps alone differ on these two images: bias shows positive spatial autocorrelation (I 0.20, 0.16, $p = 0.0005$), no significant positive autocorrelation was detected for noise (I -0.05, -0.09), and the coefficients of variation are 0.11 against 0.04. Two images cannot show that noise damage has no spatial structure, and whether such image-space statistics classify faults on new images or checkpoints is not evaluated here. The spectral rulebook's separation of these two, one dominant spike against hundreds, does not depend on a photo.

Which patches are ever targets. The released I-JEPA sampler draws the top-left corner of every block, context and target alike, as `torch.randint(0, H - h)` and `torch.randint(0, W - w)` (`src/masks/multiblock.py`, lines 89 and 90 at the commit the release pins). The upper bound is exclusive, so no block ever covers the last row or column; an upstream pull request that would add one to the bound is open and unmerged.[1] On the 16×16 grid of the ViT-H/14 caches that leaves 225 of the 256 positions, and each map in Figure 11 has 38 gray patches: 31 in the last row and column and 7 that none of the 16 masks happened to draw. The trained fleets inherit the exclusion on their own grids: the ViT-S/8 runs at 96 pixels import the same sampler and can reach at most 121 of their 144 positions, and the 40k ViT-S/16 fleet at 160 pixels, whose trainer subclasses it, 81 of 100, and the 120k ViT-B/16 fleet at 176 pixels, whose trainer subclasses it the same way, 100 of 121. The video samplers of V-JEPA and V-JEPA2 draw their corner from $[0, H-h]$ and exclude nothing. Theorem 1 does not depend on which tokens are audited; every covariance does, and the excluded positions are not evaluated here.

[1] `https://github.com/facebookresearch/ijepa/pull/40`

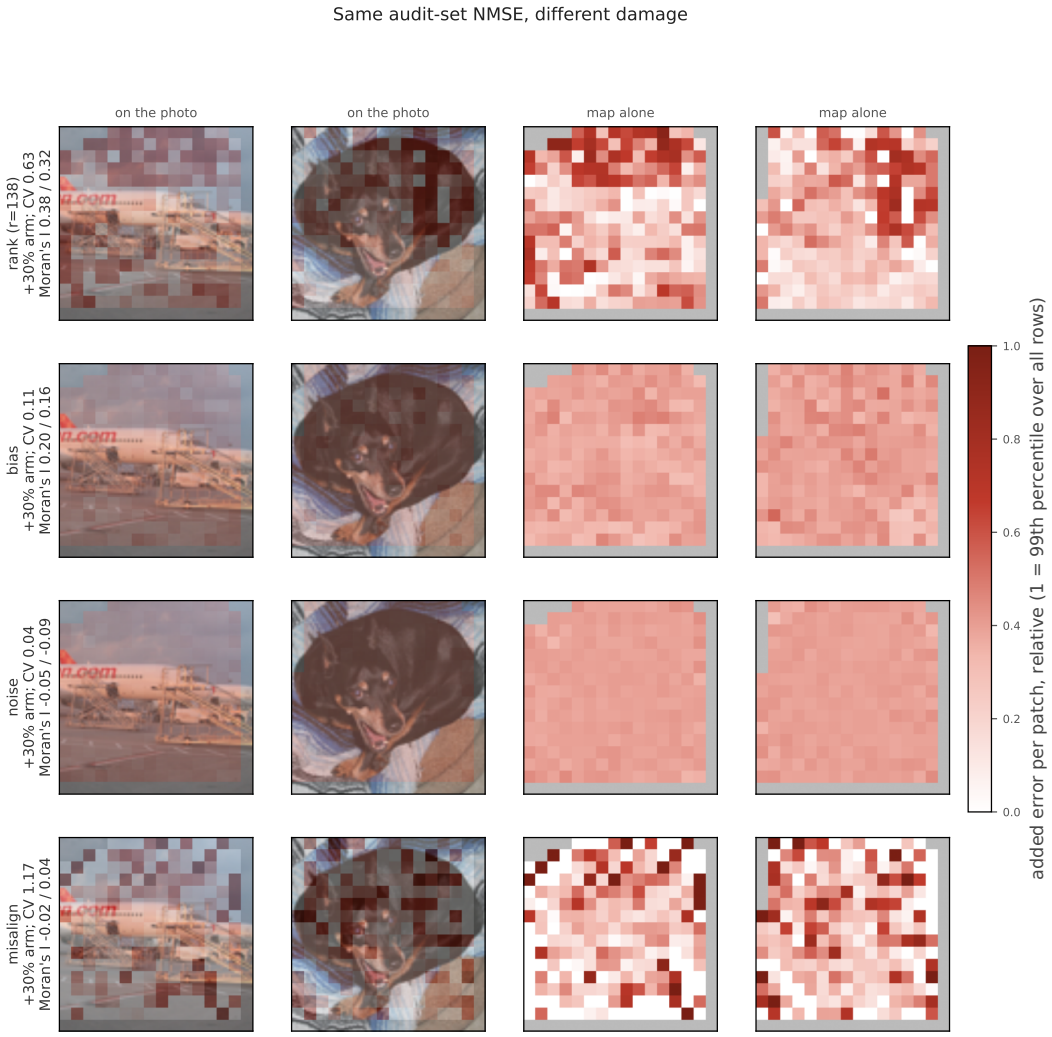

Figure 11: **Same scalar, different damage, on real images.** Per-patch prediction error added by each fault relative to the healthy head at the $+30\%$ arms, on two STL-10 images, composited on the photo (left pair) and alone (right pair) on one linear scale. Row labels give each map's coefficient of variation across patches and its Moran's I on the two images. Gray patches were never a prediction target.

N REPAIR-POLICY COMPARISON

We compare five repair policies on the injected batteries. Each repair is fit on one half of the sources and scored on the other: always refit the head; a prespecified map from the rulebook's verdict to one repair (rank and alignment to a refit, bias to an intercept recalibration, noise to shrinkage, an abstention to a refit, and no repair when the gate is silent); a scalar-only nested selection that fits every repair and keeps the cheapest within 1% of the best validation NMSE; the spectrum-gated shortlist, which lets the verdict restrict the candidates (bias to recalibration or refit, noise to shrinkage or refit, the rest to a refit) before the same nested selection; and the per-case oracle. Table 11 scores them on both models.

What this benchmark can and cannot show. Two of its properties bound the claim. First, a full refit equals the per-case oracle on all 108 rows of the multi-draw version below, so no selector can beat it. The shortlist's agreement with the oracle therefore reflects a choice that was never open. Second, the shortlist does not avoid the refit it is meant to save. It fits the refit in 72 of 72 cases and a second, cheaper candidate in 36 of them, so it pays the full refit cost plus the cost of some cheaper candidates. An earlier version of this work read the comparison as evidence that the spectrum's value is candidate pruning. This benchmark does not support that reading, which we withdraw. Figure 12 shows the morphological finding that the cheap repair is the right one exactly where the diagnosis says so; the budget-dependent characterization follows below.

Multiple fault draws. Resplits resample the split and not the fault, so the comparison was repeated over three independent fault directions on three checkpoints, with nested fit and selection splits and thresholds never touching test. On the image and video checkpoints the scalar-only selector falls $+3.7$ and $+5.4$ points behind at twice the fits, entirely on the alignment family, which never appears on its own selection split, while the shortlist and always-refit both match the oracle. That symmetry does not hold on the third checkpoint once V-JEPA2 is read through the official masked forward: there the shortlist trails the oracle by $+3.5$ points against the scalar selector's $+6.4$, because three of the nine alignment rows read as abstentions on that cache and open the shortlist to the full candidate set. The gap is a property of the diagnosis, not of the repair set.

Two conditions the comparison depends on. The scalar selector's deficit depends on where the fault is injected. In the control that injects misalignment into the fit and selection splits as well as the evaluation stream, its excess is *exactly* zero on all three checkpoints: with the fault visible on its own selection split a held-out scalar loss is sufficient, and the advantage the section reports exists only when the fault is confined to the deployment stream. The second condition is the candidate set. It does not contain "do not repair", which is available in any deployment and costs nothing to evaluate. Allowing it, the shortlist and the scalar selector return *exactly* the same excess on the image and video checkpoints, differing only in fits per arm. The shortlist matched the oracle because it forced the refit into the candidate set; selection quality did not account for the match.

Label budget. The comparison above gives every policy the whole fit split. Table 10 sweeps the share of it a repair may use, against the best single fixed policy at that budget, which is the appropriate comparator for a user without a diagnosis. The shortlist wins on 2 of three checkpoints and only in a window between a tenth and a half of the split, by 0.5 to 1.8 points, and the gain is almost entirely the noise family taking a cheap shrink where a refit would be fit on too little data. Below the window the shortlist is the worst policy in the comparison, losing as much as $+324$ points because it prescribes a refit of $491,904$ parameters on as few as $1,224$ tokens. The regime is identifiable without the spectrum, by comparing a refit and a shrink once on the selection split, and any deployment of the shortlist should do that first.

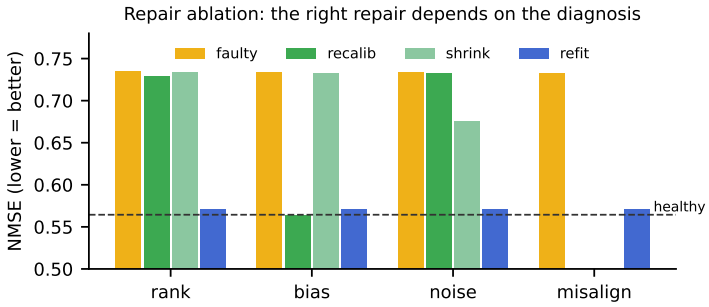

Figure 12: **Repair ablation** (image battery, $+30\%$ severity). Only the diagnosis-matched repair recovers cheaply.

Table 10: **Repair policy under a label budget.** Excess NMSE over the per-case oracle, in points, pooled over three fault draws, four families and three severities, against the best of always-refit, always-recalibrate and always-shrink at the same budget. Budget is the share of the fit split the repair may use; the refit has $491,904$ parameters. Negative is the shortlist winning.

Budget (share of fit split)	I-JEPA	V-JEPA	V-JEPA2
0.01	$+324.3$	$+57.7$	$+5.0$
0.02	$+151.2$	$+35.5$	$+0.4$
0.05	$+43.7$	$+13.4$	$+2.6$
0.10	$+11.1$	$\mathbf{-1.80}$	$+2.9$
0.25	$\mathbf{-0.50}$	$\mathbf{-0.09}$	$+3.3$
0.50	$\mathbf{-0.02}$	0.00	$+3.4$
1.00	0.0	0.0	$+3.5$

Table 11: **Repair-policy comparison** (72 cases; excess NMSE over the per-case oracle, in points, and repair candidates fit per arm). The pruned shortlist matches the oracle on I-JEPA and is $+0.13$ on V-JEPA, where always-refit is already optimal; on the official V-JEPA2 forward it trails always-refit by 3.5 points.

Policy	I-JEPA excess	V-JEPA excess	fits/arm
oracle (upper bound)	0.00	0.00	n/a
spectrum-gated shortlist	$\mathbf{+0.00}$	$\mathbf{+0.13}$	1.5
always refit	$+0.21$	$+0.00$	1.0
prespecified fault→repair map	$+2.45$	$+2.89$	1.0
scalar-only nested selection	$+3.80$	$+5.26$	3.0

O THE TOY TRAINING BATTERY AND PRESCRIPTIONS

This appendix gives the protocol of the from-scratch training battery behind the run-ranking and prescription results of §4.3, and its per-condition outcomes (Table 12). How the unified procedure treats each natural misconfiguration, on this battery and on two larger fleets, is Appendix P.

Training protocol. Every run pretrains a ViT-S/8 I-JEPA from scratch on the unlabeled split of STL-10 at 96 px: batch size 128, 3,000 steps, AdamW with learning rate 2×10^{-4} for the encoder and twice that for the predictor, weight decay 0.05, linear warmup over the first 500 steps, gradient norms clipped to 1.0 separately for encoder and predictor, bf16 autocast with the loss accumulated in fp32. The target encoder is an EMA copy with momentum 0.996; targets are LayerNormed token embeddings; the loss is the mean squared error on masked tokens. Masks follow the official multi-block sampler: one context block at scale $(0.85, 1.0)$ and four prediction blocks at scale $(0.15, 0.2)$, aspect ratios $(0.75, 1.5)$, no overlap. The predictor has depth 6 and width 192.

Audit and final metric. The audit set is fixed once for all runs and all conditions: $1{,}024$ held-out STL-10 test images with masks drawn once at seed 999 from the *healthy* mask distribution, so the audit task never varies with the misconfiguration. Each run is audited at 20% and 40% of training (steps 600 and 1,200): residual and target moments are accumulated in fp64 on the audit set, and the four channels are the target-normalized residual $\operatorname{tr} C_e / \operatorname{tr} C_h$ (NMSE), the top whitened eigenvalue $\log_{10} \lambda_{\max}$, the plateau fraction ($|\lambda - 1| \le 10^{-3}$), and the plateau-excluded regret mass $M_{\mathrm{reg}}^{\mathrm{ex}}$, all through the support-projected estimator of §3. The final metric is a ridge probe on the *deployed* artifact: frozen target-encoder features, mean-pooled after the target LayerNorm, standardized, one-hot ridge regression (penalty $10^{-2}n$) on the $5{,}000$ labeled STL-10 training images, scored on the $8{,}000$-image test split.

Conditions. Six healthy seeds (0–5) and nine misconfiguration families at three seeds each (0–2), 33 runs in total (prescription and width-sweep runs are excluded from this pool). Exact parameters: `lr_hi` and `lr_lo` set the learning rate to 2×10^{-3} and 2×10^{-5}; `ema_frozen` and `ema_fast` set the target momentum to 1.0 and 0.8; `width16` shrinks the predictor width to 48 (divisible by its 6 attention heads; the name is historical); `mask_tiny` shrinks the prediction-block scale to $(0.05, 0.08)$, i.e. 7–11 patches per block against the healthy 21–28; `corrupt30` replaces 30% of every training batch with unit Gaussian noise images; `no_target_ln` disables the target LayerNorm; `leak` sets the context scale to $(0.99, 1.0)$ with overlap allowed, so the context contains the prediction targets.

Prescription protocol. For the three conditions with a morphology-matched fix we intervene once, at step $1{,}200$ (the 40% audit): `ema_frozen` restores the momentum to 0.996, `width16` swaps in a freshly initialized width-192 predictor (depth unchanged, added to the optimizer), and `lr_lo` multiplies every group's learning rate by 10. The control is a *mis-prescription*: the same three actions rotated one position (`lr_lo` gets the EMA fix, `ema_frozen` the wider predictor, `width16` the learning-rate raise), so every control receives a real intervention, just the wrong one. Three seeds per cell; 17 of 18 runs completed (the seed-0 frozen-EMA control is missing, so that cell averages two seeds).

Prescription results. The frozen-EMA repair lifts the final probe from 0.419 to 0.539 against the healthy 0.541, a $+12.0$-point gain, while its mismatched control stays at 0.420: with momentum 1.0 the deployed target encoder never updates, so a wider predictor cannot reach it, and the control probes match the unrepaired seeds exactly. The under-width repair gains $+2.0$ points (0.469 to 0.489) against $+1.0$ for its control (0.479). The learning-rate arm needed a harness correction before it tested anything: the warmup scheduler recomputed each group's rate from its stored base rate, so the $\times 10$ raise survived exactly one optimizer step. With the scheduler fixed and the cells re-run, the matched prescription lifts the final probe from 0.510 to 0.519 while the rotated control stays at 0.509, and every prescribed seed ends above every control seed. The gain is small compared with the frozen-EMA repair's gain, given the late intervention and short schedule. The `width16` control received the uncorrected learning-rate action, so it remains a no-intervention control.

What changes at inference. The probe reads the deployed target encoder, the artifact that ships. With momentum 1.0 that artifact is the frozen initialization: its final target-variance trace on the audit set is 126.5, against 309.6 for healthy runs. After the step-$1{,}200$ repair the deployed encoder recovers a trace of 317.3 and a final audit indistinguishable from healthy (NMSE 0.345 vs. 0.354,

Table 12: **Training-run battery** (33 runs). Per-condition mean final probe accuracy and mean audit channels, the latter at the 40% checkpoint (step 1,200). The two benign families (`mask_tiny`, `no_target_ln`) match healthy probes, which are true negatives; `leak` pairs the largest regret mass with mild probe damage, and `lr_lo` spikes the top eigenvalue over a nearly healthy encoder.

Condition	change	n	probe	NMSE	$\log_{10} \lambda_{\max}$	$M_{\text{reg}}^{\text{ex}}$
`healthy`	(base config)	6	0.541	0.252	0.18	0.042
`mask_tiny`	pred. scale $(0.05, 0.08)$	3	0.549	0.240	0.75	0.115
`no_target_ln`	target LN off	3	0.549	0.250	0.18	0.041
`lr_lo`	lr 2×10^{-5}	3	0.510	0.401	1.26	0.339
`leak`	ctx. scale $(0.99, 1.0)$, overlap	3	0.510	0.628	1.74	0.595
`corrupt30`	30% noise images	3	0.488	0.297	0.43	0.079
`width16`	pred. width 48	3	0.469	0.348	0.65	0.131
`ema_fast`	momentum 0.8	3	0.429	0.614	0.42	0.125
`ema_frozen`	momentum 1.0	3	0.419	0.425	0.41	0.045
`lr_hi`	lr 2×10^{-3}	3	0.394	0.572	0.13	0.027

plateau fraction 0.021 vs. 0.018): the repaired run's encoder has healthy-level feature diversity. The $+12.0$-point probe gap is therefore an inference-time property of the deployed artifact, not a training-curve cosmetic. The training-time scalar points the wrong way, since the frozen-EMA runs' raw audit error at 40% (52.8–55.3) is *lower* than healthy (76.0–82.5) because static targets are easier to hit.

P Natural misconfigurations at three scales, and the unified procedure

This appendix collects the three natural-misconfiguration fleets, the two-sided repair of the rulebook's original first rule, the unified procedure with its pre-registered held-out test, and the metric ladder for ranking runs. The toy fleet is the training-run battery of Appendix O; the two larger fleets are described below.

Natural morphologies on the toy fleet. We ran the unified procedure of §4.3 over the 40% audits of all 33 battery runs, with the null taken from the six healthy runs' audits at the same step (leave-one-out, as on every fleet; a healthy seed sits inside its own null by construction, so the healthy runs define the table and are not rows in it) and with the spike-count rule unable to fire because these audits store no spike counts. Table 13 gives the verdicts. The gate separates every harmful family from the healthy seeds (AUC 1.00). Of the two benign families, `no_target_ln` stays inside the null in one seed, fires the eigenvalue channel alone in a second and both channels in the third, and `mask_tiny` routes to `easier` in all three: its residual is 4.9% easier than healthy, its top eigenvalue moves as well, and its probe is unchanged (+0.8 points), which is why `easier` is a route and not a harm verdict. `leak` and `lr_lo` match the bias rule's dominant-spike signature; most dynamics-level misconfigurations land in the alignment family or abstain, and the genuinely rank-starved `width16` reads as rank in only one of three seeds.

A re-collection clarifies one potentially misleading row. `corrupt30` corrupts the input rather than the head, and it reads as alignment; recording the missing spike counts and re-scoring with all five rules leaves that verdict unchanged, because neither the arm nor the healthy baseline carries enough super-unit eigenvalues for the ratio rule to fire. The injected noise arm, which adds broadband energy to the head's *output*, reaches a ratio of 16.3 on the same rule. The two share a cause but differ in shape. This arm has the alignment shape, so calling it noise would misread its spectrum. What this battery supports is that the atlas names the *shape* of a natural run's residual and that distinct root causes can share a shape; the held-out test below could not determine whether those shapes correctly describe natural faults, so morphology is claimed for injected faults only.

Two larger fleets, and a blind spot in the original first rule. That battery is one scale. Two larger fleets of natural misconfigurations let the same question be asked where the runs are longer and the data harder: 24 runs of a ViT-S/16 on ImageNet-100 for 40k steps, and 23 runs of a ViT-B/16 on the same data for 120k steps, each with healthy seeds and a linear probe (Table 17). Two things follow, one about the fleets and one about the rulebook.

Table 13: **Verdicts of the unified procedure on natural misconfigurations** (40% audits, leave-one-out healthy null, spike-count rule unavailable). An independent re-collection of this battery, scored with the rulebook as first frozen, reproduced four of that table's ten family cells and every qualitative statement in the text, including `width16` reading as rank in one seed of three. These runs are trained rather than replayed, and a bf16 forward with non-deterministic kernels moves a seed-matched rerun by a fraction of the between-seed spread, so single cells here are less stable than the reading they support.

Condition	verdicts
`mask_tiny`	easier 3/3
`no_target_ln`	easier 1, healthy 1, pocket 1
`leak`	bias (spike) 3/3
`lr_lo`	bias (spike) 3/3
`lr_hi`	alignment 2, rank 1
`corrupt30`	alignment 3/3
`ema_fast`	alignment 2, abstain 1
`ema_frozen`	alignment 2, abstain 1
`width16`	abstain 1, alignment 1, rank 1

The first is that where the spike-count feature exists at these scales, it fires as designed: on the 40k fleet the two leakage-type families carry 15 to 22 super-unit eigenvalues against a healthy zero and read as noise, where the toy battery could not test that rule at all.

The second is a defect. The rulebook's original first rule asked whether relative NMSE had *risen* by more than 2%, so a failure that leaves the fixed audit residual easier than healthy passed it: 6 families whose downstream probe fell by 8.9 to 41.0 points were all called healthy. The first repair needed no new constant. Make that rule two-sided, with the threshold τ_f taken the way the gate of §3 takes its own, as the leave-one-out maximum of $|\Delta\mathrm{NMSE}|$ over the fleet's healthy seeds (0.048, 0.014 and 0.045 on the three fleets), and add the mirror of the alignment rule: a residual that got easier with the spectrum's shape unchanged is a fifth morphology, which we call `easier` after the shape and not the cause, since the families it collects here have three different causes.

The amendment's history limits what it can count as evidence for. After the two fleets motivated the two-sided rule, our preregistered 2% version produced healthy false alarms on the toy fleet (1/6) and on the 120k fleet (1/5); after observing those outcomes we replaced that threshold with the leave-one-out healthy-seed maximum already used by our detection gate. All real-fleet results for this amendment are therefore retrospective rather than held-out validation, although every injected-fault-grid verdict remains unchanged.

Read that way, the amended rule names 5 of the 6 families `easier` and abstains on the sixth, whose plateau does move, so the mirror rule correctly declines it. It raises no healthy false alarm on any of the three fleets, including the two where the original rule raised one. It costs one cell: `corrupt30` on the 120k fleet, a 6-point probe drop with a 2% residual rise, falls inside the widened threshold and is called healthy where the original rule called it alignment in one seed of three. Jackknifing τ_f over the healthy seeds moves no harmful family's verdict and raises no healthy false alarm. On every injected grid the original, the amended and the unified rule below return the same verdict on every arm, because the smallest NMSE rise on any arm of the six grids is 4.8% and all three rules send such an arm to rules 2 to 5; the release checks this arm by arm (`atlas_rulebook_history.py`), and no number in the body depends on the choice.

The unified procedure, and what is post hoc about it. The body now runs one procedure (§4.3): the two-sided two-channel gate of §3 decides, a deterministic route on the channel pattern sends a downward NMSE excursion to `easier` and an eigenvalue-only excursion to `pocket`, and rules 2 to 5 classify what reaches them. One rulebook (`atlas_diagnose.py` in the release) scores the injected grids, the natural fleets and the equivalence bench; the held-out scorer (`heldout_score.py`) was frozen at pre-registration with the same rules, and the release checks the two against each other on every run of the three natural fleets and of the held-out fleet at every audit step. They agree on every route. They differ in one guard: when the healthy null carries no super-unit eigenvalue, the scorer disables the spike-count rule, where the rulebook reads the count

as absolute (Appendix K), which changes the stage-2 label of the two leakage-type families on the 40k fleet (noise under the rulebook, bias under the scorer), no route, and no held-out run.

What is pre-registered. Its parts have different standing. The gate thresholds were frozen on 500 healthy I-JEPA resplits on 2026-08-13, before any natural fleet existed, and rules 2 to 5 were fit on I-JEPA alone; those are pre-registered. The composition, the per-fleet leave-one-out null, and the two routes were fixed after the fleets above had been scored, so every natural-fleet number in Table 15 is post hoc. The leave-one-out maximum over n healthy seeds is exceeded by a fresh healthy seed with probability $1/(n+1)$ per channel by exchangeability, $1/6$ at $n = 5$, which is why the pre-registered bars are rank-based on the continuous gate score rather than verdict counts. The held-out protocol, its family predictions and its bars are in `notes/HELDOUT_PREREGISTRATION.md`, committed before any run, and the fleet was scored once with the committed script (Table 14).

The held-out result. B1 passed: over 9 harmful runs and 9 healthy seeds the gate score has no overlap, the largest healthy score being 1.94 and the smallest harmful one 17.7 (AUC 1.00, and the same at 10 and 40% of training).

The permutation test. The registered scorer reports a rank-sum p of 2×10^{-5} on those scores, but each score is computed against a null built from the runs labeled healthy, so the scores depend on the labels and counting arrangements of fixed scores is not a valid permutation test. Relabeling the runs, rebuilding the null from those labeled healthy and rescoring every run (`heldout_perm_check.py`), under the null that the run labels are exchangeable, gives $p = 8.2 \times 10^{-5}$ over all 18 runs of the registered composition, and $p = 1.4 \times 10^{-3}$, the smallest attainable and the same at every step, over the 13 held-out runs with the 5 40k-fleet seeds kept in the null. The main text quotes the second: as registered, those 5 seeds come from the fleet on which the composition was fixed. The nine harmful runs are three families of three seeds, and with the family as the exchangeable unit the same test gives $p = 0.03$, the smallest a complete separation of three families from four seeds can reach. The null the pre-registration specified contains the seed being scored, which pulls healthy scores toward zero; scoring each healthy seed leave-one-out against the other eight raises the largest healthy score to 2.98 against the same smallest harmful score, and the AUC is 1.00 either way at every step (`heldout_loo.json`).

What separates the runs. This separation comes from the scalar channel: the NMSE channel alone separates the held-out runs at every step (AUC 1.00), the eigenvalue channel alone reaches 0.93 at the end, and at the end of the 120k fleet the two-channel gate (0.81) falls below the NMSE channel alone (0.92), because the maximum admits eigenvalue-channel noise from healthy runs (`critical_review_checks.py`). Nor is the test hard: the harmful families' probe drops (4.6 to 26.9 points) lie well outside the 1.6-point spread of the nine healthy seeds, and no family near the two-point harm bar was run.

The other bars. B3 passed: `ema_frozen` routes to `easier` in 3 of 3 seeds at the end of training, having routed to stage 2 in 3 of 3 at step 4000 with NMSE above healthy, which is the sign reversal the prediction file anticipated for it. B4 was vacuous: the two families predicted on the upper tail from their 3k-step behavior (`ema_fast`, `width16`) are on the lower tail at 40k, so no held-out run reached the classifier and the pre-registered decision applies, restricting morphology classification to injected faults. The three families are harmful by the registered criterion (probe drops of 14.2, 26.9 and 4.6 points). Two choices made before the first run and recorded in `notes/HELDOUT_RUN_DECISIONS.md`: the lower-tail family is `ema_frozen`, absent from this fleet but present in two others, because no never-run configuration was both lower-tail and harmful in the toy records; and `width16` is a 48-wide predictor, as its toy namesake was, since 16 does not divide across the predictor's heads.

Table 15 reports the route of every natural family at each audit step. Two things are visible that a single-step table hides. The six families the original first rule passed route to `easier` at every step from 40% of training on, and the gate separates harmful from healthy runs with AUC 1.000 at the reported step on the 40k fleet. On the 120k fleet the separation is 0.99 at 40% and 0.81 at the end, because `corrupt30` (6.1 points) is flagged at 10 and 40% and has returned inside the healthy null by the end of training; and `lr_lo` changes route from `easier` to `pocket` to stage 2 across the three steps. Morphology is a property of the pair (fault, step), which is the reason the reported step is fixed in advance. The low-variance pocket arms of §4.1 route to `pocket` on all 30 of 30 deployment resplits under the frozen I-JEPA thresholds.

Table 14: **The pre-registered held-out fleet**, ViT-S/16 at 40k steps, scored once. Routes under the unified procedure at the three audit steps (E easier, S stage 2, H healthy), family-mean probe change against the pooled nine healthy seeds, and the pre-registered bars.

Family	registered as	Δprobe	10%	40%	end
ema_fast	upper tail, misalign	-14.2	E3	E3	E3
width16	upper tail, rank	-4.6	E3	E3	E3
ema_frozen	lower tail, easier	-26.9	S3	E3	E3
new healthy seeds (4)	healthy		H4	H4	H4
gate AUC, harmful vs healthy	B1 ≥ 0.90		1.000	1.000	1.000

Table 15: **The unified procedure on the two larger natural fleets, by audit step.** Route counts over three seeds (one for leak on the 40k fleet): E easier, S stage 2, P pocket, H healthy. Family-mean probe change in points, and the gate's AUC for harmful families (probe drop above 2 points) against the fleet's healthy seeds. Post hoc on these fleets; the held-out test is pre-registered.

Fleet	Family	Δprobe	10%	40%	end
ViT-S/16, 40k	leak	-22.7	S1	S1	S1
ViT-S/16, 40k	mask_overlap_mild	-22.1	S3	S3	S3
ViT-S/16, 40k	wd_hi	-11.6	P1 H1 E1	E3	E3
ViT-S/16, 40k	lr_hi	-11.1	E2 H1	E3	E3
ViT-S/16, 40k	warmup_none	-10.2	E2 S1	E3	E3
ViT-S/16, 40k	ema_slow	-8.9	E3	E3	E3
ViT-S/16, 40k	pred_wide	-5.0	S3	S3	S3
ViT-S/16, 40k	gate AUC, harmful vs healthy		0.958	1.000	1.000
ViT-B/16, 120k	ema_frozen	-41.0	E3	E3	E3
ViT-B/16, 120k	leak	-37.4	S3	S3	S3
ViT-B/16, 120k	lr_hi	-16.4	E3	E3	E3
ViT-B/16, 120k	lr_lo	-8.6	E3	P2 E1	S3
ViT-B/16, 120k	corrupt30	-6.1	E3	E2 H1	H3
ViT-B/16, 120k	width96	-0.4	E3	E3	P2 E1
ViT-B/16, 120k	gate AUC, harmful vs healthy		1.000	0.987	0.813

What ranks runs: the ladder. Table 16 is the comparison the main text summarizes: every channel on every fleet, read two-sided against the fleet's healthy mean at 40% of training, as a Spearman rank correlation with the final probe, with Fisher-z 95% intervals. The channel is selected on the toy fleet and applied unchanged to the other two; NMSE is best there and inside the interval of the best spectral channel on both larger fleets. The last column is the same NMSE read one-sided, which fails on both larger fleets because the loss falls as the run degrades. The per-coordinate diagnostics that reach 67% on the morphology ladder of Appendix I are $\mathrm{diag}(C_e)/\mathrm{diag}(C_h)$ summaries; the gap from there to the whitened rulebook features reflects the eigendecomposition's contribution to morphology, not to ranking.

Why rank correlations and not AUROCs. A detection AUROC needs a sign per score, and taking that sign from the very runs being scored flatters whichever statistic is closest to null. Scored that way on the toy fleet the raw validation loss reaches 0.77 and 0.61 at the two audit points, past the audit's 0.73 at 20% though not its 0.88 at 40%. The two-sided rank correlations of Table 16 need no sign, which is why the section reports them; by their own Fisher-z intervals the raw loss and NMSE are tied on every fleet.

Q PRUNING A SWEEP: A NEGATIVE RESULT FOR THE SPECTRUM

Successive halving is a standard method for stopping poorly performing runs early, and its premise is that a run doing well at a rung does well at the end. On these fleets the premise fails in a way an inversion over time would not describe: the quantity the schedule ranks by, the training loss, is anticorrelated with the probe at every horizon (§4.3), so pruning on it discards good runs at every rung by design. The schedule is not what needs fixing. We keep the rungs at 10, 20 and 40% of training and the reduction factor at two, and change only the score.

Table 16: **Metric ladder for ranking runs**, Spearman with the final probe at 40% of training, two-sided deviation from the healthy mean, with 95% Fisher-z intervals. The toy fleet stores no covariance traces. Last row: NMSE read one-sided.

Channel, two-sided	ViT-S/8, 3k (n=33)	ViT-S/16, 40k (n=24)	ViT-B/16, 120k (n=23)
raw loss	-0.48 [-0.71, -0.16]	-0.63 [-0.82, -0.30]	-0.82 [-0.92, -0.62]
NMSE	-0.74 [-0.86, -0.53]	-0.73 [-0.87, -0.46]	-0.80 [-0.91, -0.58]
$\log_{10} \lambda_{\max}$	-0.21 [-0.52, +0.14]	-0.70 [-0.86, -0.42]	-0.91 [-0.96, -0.79]
$M_{\mathrm{reg}}^{\mathrm{ex}}$	-0.19 [-0.50, +0.17]	-0.58 [-0.80, -0.23]	-0.83 [-0.93, -0.64]
plateau share	-0.54 [-0.74, -0.24]	-0.45 [-0.72, -0.06]	-0.41 [-0.70, +0.01]
NMSE, one-sided	-0.78	-0.09	+0.30

Table 17: **Natural misconfigurations at two larger scales** (family-mean probe change against the fleet's healthy mean, in accuracy points; three seeds per family unless noted). Original is the rulebook as first frozen, whose first rule was a one-sided 2% NMSE rise, with the spike-count rule where the feature exists, which is the 40k fleet only. Amended makes that rule two-sided and adds `easier`, as described above. The rows in bold are the ones the original rule reads as healthy. The unified procedure's routes for the same families are in Table 15.

Fleet	Condition	Δprobe	Original	Amended
ViT-S/16, 40k	`leak` (one seed)	-22.7	noise	noise
	`mask_overlap_mild`	-22.1	noise $3/3$	noise $3/3$
	`wd_hi`	-11.6	healthy $3/3$	easier $3/3$
	`lr_hi`	-11.1	healthy $3/3$	abstain $3/3$
	`warmup_none`	-10.2	healthy $3/3$	easier $3/3$
	`ema_slow`	-8.9	healthy $3/3$	easier $3/3$
	`pred_wide`	-5.0	abstain $3/3$	abstain $3/3$
	`healthy`	0.0	healthy $5/5$	healthy $5/5$
ViT-B/16, 120k	**`ema_frozen`**	-41.0	healthy $3/3$	easier $3/3$
	`leak`	-37.4	bias $3/3$	bias $3/3$
	`lr_hi`	-16.4	healthy $3/3$	easier $3/3$
	`lr_lo`	-8.6	alignment $3/3$	alignment $3/3$
	`corrupt30`	-6.1	alignment 1, healthy 2	healthy $3/3$
	`width96`	-0.4	healthy $3/3$	healthy 2, rank 1
	`healthy`	0.0	healthy 4, alignment 1	healthy $5/5$

Table 18 reports the regret of the surviving best against the sweep's best, averaged over $4{,}000$ sweeps of eight runs with three misconfigured, on two fleets: the 40k ViT-S/16 fleet restricted to its four misconfiguration families, and the 120k ViT-B/16 fleet restricted to the four that are realistic misconfigurations rather than constructions (`leak` and `ema_frozen` are excluded there, being close to an identity map and to a frozen random teacher). Every row prunes to the same 0.27 of a full sweep, so the comparison is at matched compute.

Under the standard schedule with its usual signal, pruning on the loss is 3.5 and 1.4 times worse than pruning at random. Reading that same loss two-sidedly, as a deviation from the median of the runs still alive, recovers most of the damage at no cost. The channel this paper selects on the toy fleet (§4.3), NMSE, then improves on the two-sided loss on one fleet and is worse on the other, and the strongest single signal on both fleets is the target variance $\mathrm{tr}(C_h)$, which does not involve the predictor at all. An earlier draft reported the best audit channel per fleet; that is a selection on the outcome, and the per-channel rows replace it. The results show that pruning needs a two-sided normalized scalar and gains nothing from the whitened spectrum here.

The caveat is the fleet size. Each pool holds five healthy runs and 12 misconfigured ones, so the sweeps are resampled from a small population and the regret differences carry that dependence. The simulation reads only stored per-run curves and audits, so it reruns from the release in seconds, and the audits it reads cost what Appendix T measures.

Where the same statistic does not help. Ranking runs differs from predicting what a model is for, and we ran a preregistered go/no-go on the harder question. In a latent world model used for planning, over 8 arms in each of two environments, the audit's target-normalized residual ranks the

Table 18: **Successive halving with the pruning signal swapped.** Regret of the best surviving run against the sweep's best, in probe accuracy, averaged over $4{,}000$ sweeps of eight runs with three misconfigured; lower is better and every row spends the same compute. The deviation rows score $|x - \mathrm{median}(x_{\text{others still alive}})|$, which needs no labels and no held-out set. NMSE is the channel selected on the toy fleet; the other two audit channels are shown so the per-fleet winner is visible rather than chosen.

Pruning signal	ViT-S/16, 40k	ViT-B/16, 120k
loss, one-sided (standard)	0.1744	0.0522
random	0.0504	0.0370
loss deviation (two-sided)	0.0250	0.0103
NMSE deviation (selected on the toy fleet)	0.0075	0.0194
$\mathrm{tr}(C_h)$ deviation	0.0022	0.0048
top-eigenvalue deviation	0.0272	0.0042

arms against planning success exactly as the validation loss does: Spearman 0.74 and 0.89, equal to the loss's own to every digit we report. The best of the six pencil channels clears the loss by less than the 0.15 the preregistration required in one environment (0.84 against 0.74) and not at all in the other (0.89). Both gates fail, and we report it as registered. The reading we take is the one the scope conditions already imply: second moments of the residual say how well the prediction is made, not what the prediction is for.

R TYPED EQUIVALENCE TESTING ON THREE CACHES

Table 3 (§4.4) is the I-JEPA cache. The same script, with no constant changed, runs on the V-JEPA and V-JEPA2 official-forward caches; the tables below carry every cache with resplit means and standard deviations, the combined parameter distance, linear CKA, and the rule condition that produced each stage-2 label together with its margin over the constant. The protocol: the source buckets are split into a matching third and a scoring two-thirds; the three faults and the permutation are bisected on the matching third to $+10\%$ relative NMSE and the construction directions are drawn there; each of 30 resplits of the scoring pool reads the reference on one half and the candidate on the other; the gate thresholds are the 99.5th percentile of 500 healthy resplits of the same pool, which on I-JEPA gives 0.0407 for the NMSE channel against the frozen deployment value 4.1%.

The parameter-diff baseline reads $(\Delta W, \Delta b)$: nothing changed is *healthy*, only the intercept is *bias*, a low-rank ΔW is *rank*, a full-rank one is *noise*, with the rank cutoff chosen per cache to maximize its own score. Even so, it types 5 of the 9 rows on every cache (counting the synthetic pocket row, which it labels *bias*); it misses the same four on each cache: the gauge (rank one, so it reads as compression), the two rounding exports (full rank, so they read as corruption), and the permuted serving order (no parameter moved). Used together they close that gap. The output difference decides, at a threshold chosen between the labeled benign and faulted rows, and the parameter difference types, with a moved output and unmoved parameters read as a serving change; combined, they name 9 of the 9 rows on every cache, since the benign rows' output differences (at most 0.010) sit far below the faults' (at least 0.059; `critical_review_checks.py`). What the audit offers instead is its requirement: the reference enters only through the healthy fingerprint of its own half, so a candidate can be typed against a stored fingerprint, on other inputs and without the reference's weights. This bench is also the only setting in which a parameter diff is defined at all; the held-out fleets of Appendix P are independently trained runs.

On I-JEPA, 8 of 9 rows are unanimous over the 30 resplits; the output difference is 0.0002 for the gauge, whose intercept moves by $\|\Delta b\|/\|b\| = 2.437$, 0.34 for each matched fault, and 0.0587 for the synthetic pocket row; the resplit noise floor of the NMSE channel is 0.0159 (its standard deviation on the identical row), and the recalibrated eigenvalue threshold is 0.0597 against the frozen 0.0495.

†Synthetic: this row is the scalar-invisible construction of Section 4.1, not an operation anyone ships. It is kept here because it is the one arm whose disagreement in outputs is small and whose type is still recovered. Linear CKA centers both inputs, so the intercept row's 1.00000 is translation invariance by definition rather than a measurement; the column is reported for completeness and

Table 19: **I-JEPA, all rows.** Resplit means with standard deviations in parentheses; the binding column is the rule condition that produced the label and its relative margin over the constant.

operation	severity	$\Delta\text{NMSE}_{\text{rel}}$	$\Delta \log_{10} \lambda_{\max}$	output diff	CKA	label	binding condition
byte-identical copy	–	$+0.0017 \pm 0.0159$	-0.0056 ± 0.0227	0.0000 ± 0.0000	1.00000	healthy (30/30)	–
gauge re-parameterization	–	$+0.0017 \pm 0.0159$	-0.0056 ± 0.0227	0.0002 ± 0.0000	1.00000	healthy (30/30)	–
bf16 export	–	$+0.0017 \pm 0.0159$	-0.0055 ± 0.0226	0.0018 ± 0.0000	1.00000	healthy (30/30)	–
int8 export (per-channel)	–	$+0.0018 \pm 0.0159$	-0.0050 ± 0.0227	0.0074 ± 0.0000	0.99999	healthy (30/30)	–
low-rank compression	288	$+0.0985 \pm 0.0156$	$+0.2875 \pm 0.0201$	0.3572 ± 0.0005	0.99319	rank (30/30)	$\Delta p = 0.233$ vs 0.1 (133%)
intercept drift	8.27299	$+0.1011 \pm 0.0167$	$+2.5421 \pm 0.0255$	0.3411 ± 0.0014	1.00000	bias (30/30)	$\Delta M = 0.591$ vs 0.2 (196%)
broadband weight noise	6.05e-03	$+0.1000 \pm 0.0160$	$+1.2531 \pm 0.0215$	0.3386 ± 0.0002	0.99425	noise (30/30)	$s = 7.925$ vs 3 (164%)
permuted serving order	0.0606564	$+0.1001 \pm 0.0158$	-0.0059 ± 0.0226	0.3481 ± 0.0007	0.88256	misalign (30/30)	$\Delta p = 0.002$ vs 0.05 (97%)
low-variance drift†	1.42493	$+0.0047 \pm 0.0159$	$+1.5004 \pm 0.0277$	0.0587 ± 0.0002	1.00000	pocket (29/30)	–

Table 20: **V-JEPA, all rows.**

operation	severity	$\Delta\text{NMSE}_{\text{rel}}$	$\Delta \log_{10} \lambda_{\max}$	output diff	CKA	label	binding condition
byte-identical copy	–	-0.0023 ± 0.0100	$+0.0027 \pm 0.0134$	0.0000 ± 0.0000	1.00000	healthy (30/30)	–
gauge re-parameterization	–	-0.0023 ± 0.0100	$+0.0027 \pm 0.0134$	0.0002 ± 0.0000	1.00000	healthy (30/30)	–
bf16 export	–	-0.0023 ± 0.0100	$+0.0025 \pm 0.0134$	0.0015 ± 0.0000	1.00000	healthy (30/30)	–
int8 export (per-channel)	–	-0.0022 ± 0.0100	$+0.0038 \pm 0.0133$	0.0077 ± 0.0000	1.00000	healthy (30/30)	–
low-rank compression	170	$+0.0995 \pm 0.0094$	$+0.3145 \pm 0.0157$	0.4726 ± 0.0027	0.99617	rank (30/30)	$\Delta p = 0.416$ vs 0.1 (316%)
intercept drift	7.91842	$+0.0971 \pm 0.0104$	$+1.5485 \pm 0.0103$	0.4604 ± 0.0021	1.00000	bias (30/30)	$\Delta M = 0.252$ vs 0.2 (26%)
broadband weight noise	0.0100288	$+0.0979 \pm 0.0100$	$+1.1542 \pm 0.0132$	0.4612 ± 0.0015	0.99657	noise (30/30)	$s = 5.471$ vs 3 (82%)
permuted serving order	0.114741	$+0.0977 \pm 0.0095$	$+0.0045 \pm 0.0133$	0.4790 ± 0.0006	0.78367	misalign (30/30)	$\Delta p = 0.000$ vs 0.05 (100%)
low-variance drift†	1.37221	$+0.0008 \pm 0.0100$	$+0.6903 \pm 0.0126$	0.0798 ± 0.0004	1.00000	pocket (30/30)	–

carries no weight in the argument. The gauge row's $\|\Delta W\|/\|W\| = 1.000$ is by construction: u spans the null direction of the centered head input, so $W + vu^{\top}$ with the compensating intercept computes the same function for any magnitude of v, and we chose $\|v\|$ to make the weight change as large as the weights themselves.

The severity sweep states the scope. From $+10\%$ upward every matched fault is typed on every resplit of every cache. At $+5\%$ one cell abstains, the V-JEPA intercept row, because rule 4's second condition sits closest to its constant there ($\Delta M_{\text{reg}}^{\text{ex}}$ at $+26\%$ over 0.20 in the margin column), and at $+2\%$ the gate mostly routes to *pocket*: something changed, and the type is not resolved. The rulebook abstains rather than guesses where its conditions do not clear, and the margin column says in advance where that will happen. Out of the 18 stage-2 verdicts on the two caches the rules were not fit on, 16/18 are the ground-truth type on every resplit; the remaining two are $29/30$.

S THE EXPORTED ENCODERS: THE I-JEPA CASE, PORT FAMILIES, AND A SECOND DENSE TASK

The HuggingFace I-JEPA export. Figure 13 and Table 23 give the image-space and dense-task evidence for the export comparison of §4.4.

Port families. The I-JEPA case is not isolated: across further port pairs, the same equivalence test found that the HuggingFace MAE ports ship non-canonical weights *family-wide*: `vit-mae-base` and `vit-mae-large` are bit-exact copies of MAE's *visualization demo* checkpoints ($150/150$ and $294/294$ encoder tensors at $\max|\Delta| = 0$; separate runs trained with `norm_pix_loss=False`) rather than the canonical pretraining checkpoints behind the paper's results (1 tensor matching in each case, the deterministic sin-cos position table, and 250% residual NMSE between the two base encoders). The `vit-mae-huge` port matches *no* canonical release we could locate (1 of 390 tensors vs. the pretraining checkpoint; no visualization-demo huge was ever published). None of the model cards states which training run the weights come from.

The audit's applicability depends on the architecture rather than the release cycle: the V-JEPA2.1 checkpoint published in the official repository (dense predictive loss, deep supervision, multi-modal tokenizer) still routes its predictor through a width-384 LayerNorm into an affine projection (now $384 \to 1664$, plus a second context head), so Corollary 1 stamps it with the same $2w-1$ prediction; we verified the head structure directly in the released weights and leave its full masked-forward audit to the artifact release. The DINOv2 ports differ from the originals *only* through their position-embedding interpolation path. They are bitwise identical at native resolution across ViT-S/B/L, 0.1–0.5% NMSE at 224 px, traceable to a single `scale_factor`-vs.-`size` line. Across these port families, the audit yields three distinct verdicts: a different-run export, a run we could not identify from public artifacts, and a benign one-line divergence.

Table 21: **V-JEPA2 (official masked forward), all rows.**

operation	severity	$\Delta\text{NMSE}_{\text{rel}}$	$\Delta \log_{10} \lambda_{\max}$	output diff	CKA	label	binding condition
byte-identical copy	–	$+0.0012 \pm 0.0094$	$+0.0036 \pm 0.0256$	0.0000 ± 0.0000	1.00000	healthy (30/30)	–
gauge re-parameterization	–	$+0.0012 \pm 0.0094$	$+0.0036 \pm 0.0256$	0.0005 ± 0.0000	1.00000	healthy (30/30)	–
bf16 export	–	$+0.0011 \pm 0.0094$	$+0.0036 \pm 0.0256$	0.0018 ± 0.0000	1.00000	healthy (30/30)	–
int8 export (per-channel)	–	$+0.0014 \pm 0.0094$	$+0.0049 \pm 0.0255$	0.0097 ± 0.0000	1.00000	healthy (29/30)	–
low-rank compression	132	$+0.1080 \pm 0.0085$	$+0.2599 \pm 0.0227$	0.6519 ± 0.0020	0.96948	rank (30/30)	$\Delta p = 0.491$ vs 0.1 (391%)
intercept drift	8.32661	$+0.1025 \pm 0.0100$	$+1.2434 \pm 0.0163$	0.5791 ± 0.0039	1.00000	bias (30/30)	$\Delta M = 0.350$ vs 0.2 (75%)
broadband weight noise	0.0332915	$+0.1053 \pm 0.0097$	$+0.8392 \pm 0.0171$	0.5878 ± 0.0022	0.99420	noise (30/30)	$s = 4.648$ vs 3 (55%)
permuted serving order	0.192451	$+0.1053 \pm 0.0086$	$+0.0036 \pm 0.0256$	0.6201 ± 0.0007	0.65261	misalign (30/30)	$\Delta p = 0.000$ vs 0.05 (100%)
low-variance drift†	1.36916	$+0.0042 \pm 0.0095$	$+0.4804 \pm 0.0204$	0.0952 ± 0.0006	1.00000	pocket (29/30)	–

Table 22: **Severity sweep**, label agreement over 30 resplits at four matched severities, one block per cache.

operation	0.02	0.05	0.1	0.2
low-rank compression	pocket (0/5)	rank (4/5)	rank (5/5)	rank (5/5)
intercept drift	pocket (0/5)	bias (4/5)	bias (5/5)	bias (5/5)
broadband weight noise	pocket (0/5)	noise (4/5)	noise (5/5)	noise (5/5)
permuted serving order	healthy (0/5)	misalign (4/5)	misalign (5/5)	misalign (5/5)

operation	0.02	0.05	0.1	0.2
low-rank compression	pocket (0/5)	rank (5/5)	rank (5/5)	rank (5/5)
intercept drift	pocket (0/5)	abstain (0/5)	bias (5/5)	bias (5/5)
broadband weight noise	pocket (0/5)	noise (5/5)	noise (5/5)	noise (5/5)
permuted serving order	healthy (1/5)	misalign (5/5)	misalign (5/5)	misalign (5/5)

operation	0.02	0.05	0.1	0.2
low-rank compression	pocket (0/5)	rank (5/5)	rank (5/5)	rank (5/5)
intercept drift	pocket (0/5)	bias (5/5)	bias (5/5)	bias (5/5)
broadband weight noise	abstain (0/5)	noise (4/5)	noise (5/5)	noise (5/5)
permuted serving order	misalign (3/5)	misalign (5/5)	misalign (5/5)	misalign (5/5)

A second dense task. The provenance finding's downstream consequence is reported on Oxford-IIIT Pet trimap segmentation, one task with three classes, so we repeated the protocol unchanged on Pascal VOC 2012 segmentation: twenty-one classes, a different image distribution, $1{,}449$ validation images, the same two encoders, identical linear probes and the same 3 head seeds.

The direction of the effect is consistent across the two tasks. The EMA target encoder leads the shipped online export by $+0.09$ to $+0.18$ accuracy points, with all 3 paired bootstrap intervals excluding zero. The stratification by audited feature gap also reproduces: the two probes disagree on 13.6% of the patches in the top decile of audited feature gap against 3.0% on the low-gap half. The audit's gap map predicts where dense predictions differ on a task it was not derived from.

On both tasks the effect is small in absolute terms: the claim is that two encoders shipped under one name differ measurably and in a spatially structured way, not that the difference is large.

Table 23: **The two exported encoders differ measurably on a dense task.** Figure 13 shows where on the image the two encoders disagree. Per-patch trimap segmentation with identical linear probes; paired per-image accuracy difference (EMA $-$ online) with bootstrap 95% confidence intervals (CIs); three head seeds. The comparison is paired by head seed because two seeds on the *same* encoder differ by as much as 0.67 points, which is larger than the encoder effect being measured. The EMA encoder leads in all three seeds by $+0.24$ to $+0.31$ points, and the heads' predictions disagree at 15.7% on top-decile-gap patches against 1.5% on the low-gap half, so the gap map locates where dense predictions differ.

Seed	online (HF) acc	EMA (official) acc	Δ (pts)	95% CI
0	0.8122	0.8147	$+0.25$	$[+0.20, +0.31]$
1	0.8057	0.8082	$+0.24$	$[+0.19, +0.30]$
2	0.8118	0.8149	$+0.31$	$[+0.25, +0.36]$



Figure 13: **Where the two shipped I-JEPA encoders disagree.** Per-patch feature gap between the HF export (online encoder) and the official EMA target on real images: structured and boundary-concentrated, not numerical noise.

T WHAT THE AUDIT COSTS

We test the claims that the audit is cheap enough to run as a unit test and that its linear algebra does not need the training hardware. The per-arm cost is one float64 eigendecomposition of a $k \times k$ matrix plus the support projection that produces it. We timed it on a single core of a laptop CPU with synthetic moments at the audited dimensions, so the measurement needs no checkpoint, cache or GPU:

Table 24: **Cost of one audit, on one CPU core in float64.** The calibration column is the deployed gate's budget of 500 healthy resplits at the same per-arm cost.

Model	d	k	per arm (ms)	500 resplits (s)
I-JEPA ViT-H/14	1280	1263	271	135
V-JEPA ViT-L/16	1024	1011	153	76
toy ViT-S head	384	380	14	7

Table 24 reports both. A single arm is a quarter of a second at $d = 1280$, and the gate's whole calibration budget is about two minutes of one core. `audit_cost.py` reproduces the table.

We use a GPU here to build the activation caches, one forward pass over the audit set per checkpoint, paid once rather than per arm. Those passes fit in 24 GB for every model audited except DINOv2 ViT-g and V-JEPA2, which want 40 GB. On disk the frozen checkpoints come to about 20 GB, the datasets to about 7 GB, and the caches to a few more.

The compute-intensive part is the from-scratch battery: 33 ViT-S/8 runs of 3,000 steps at batch 128 and 96 px, plus the prescription and width-sweep runs. We did not instrument their wall clock, so we report the configuration rather than a GPU-hour figure we did not measure.