SAME BEHAVIOR, DIFFERENT COMPRESSION: AUDITING COORDINATE DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a compression audit change its verdict when reference behavior stays fixed? On 2,669 Pets examples previously unevaluated in this project, we test a transformation that doubles the feature components in the reference nullspace while preserving the fixed reference behavior. At the fixed setting of random projection to 64 dimensions, agreement falls from 91.7% to 81.2%. The change from passing to failing occurs in all 20 fresh projection draws. We also compare maps without reducing dimension and maps that retain the same subspace on ViT/CIFAR-10, DINOv2/CIFAR-10, DINOv2/Pets and BERT/TREC. These maps retain the same available linear score functions but can yield different fitted predictions. When the regularization objectives match, separate fits give the same class predictions as the orthonormal-coordinate baseline in all tested pairs. Some logit and probability differences still exceed the declared tolerances. Classical covariance equivalence and whitening principles give explicit invariance conditions for the complete randomized audit. Using verified LayerNorm support greatly improves conditioning. None of the four settings gains a feasible compressed point at the tested $q = 2p/3$ under the ordinary isotropic Gaussian-RP probe. The tests of matching objectives and correcting support use benchmark examples examined earlier. They are not independent confirmation on new data. Compression audits should distinguish reference behavior, retained information and access through a fitted procedure.

1 INTRODUCTION

A compression audit asks whether a low-dimensional representation retains a specified behavior. Its answer can change while that reference behavior stays fixed. On frozen DINOv2 Pets features, an invertible intervention selected on validation data changes both PCA and random projection from passing to failing. This result uses the same behavioral criterion on previously unevaluated project examples. The method and dimension remain fixed. Coordinate choice therefore affects an operational decision, not just a geometric statistic (Figure 1, left).

A full-rank linear change retains every available affine score function, but an isotropically regularized probe can still lose reference agreement. Our controls without dimension reduction show this across four modern encoder/task settings. Independent fits recover the decisions in unit coordinates when we transport the Gaussian penalty to match the objective in those coordinates. This separates information retained by a map from access through a fitted procedure (Figure 1, right).

We test whether the same reference behavior can produce different compression verdicts. We then ask whether the same available linear information can produce different fitted behavior. Finally, we test what a coordinate correction controls and leaves unresolved. Gauge freedom under invertible linear changes, equal-Gram matrix equivalence and whitening are established ideas. We study their consequences for compression decisions on frozen neural representations. We retain cases where numerical correction improves conditioning but does not yield useful compression.

We study vision, language, simple models and a graph diagnostic. Pets provides the main held-out operational test. Earlier CIFAR-10 ViT/DINOv2 controls show ranking reversals, with PCA below the behavioral threshold (Appendix H). TREC provides negative results. CIFAR-100 compares fine/coarse labels on shared frozen embeddings (Appendix L). MNIST/CNN, Breast Cancer/MLP, Cora/GCN and SST-2/BERT provide supporting diagnostics (Appendix Q).

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

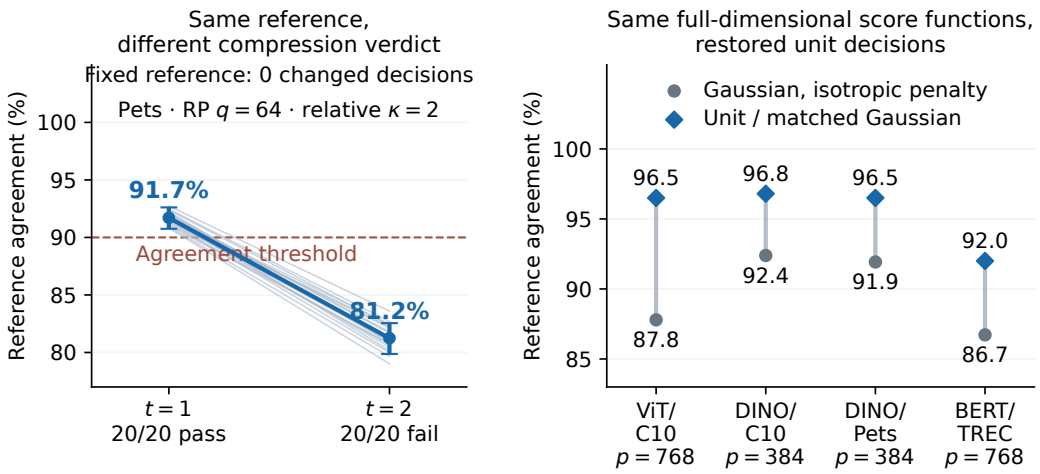


Figure 1: Coordinate choice changes a compression decision; objective matching tests a decoder mechanism. Left: all 20 paired Pets draws pass at $t = 1$ and fail at $t = 2$. No reference decision changes. Faint lines show individual draws. Bars show the approximate simultaneous intervals in Table 1. Passing requires both agreement and relative accuracy to be at least .90. Right: full-dimensional maps retain every available linear score function. Matching the Gaussian penalty recovers the unit-coordinate decisions. Points average 20 test maps. These fits use common-coordinate preconditioning, and continuous numerical exceptions remain. The full-dimensional rows are diagnostics, not compression.

Depth studies use references fitted per layer and fixed final-layer teachers (Appendices R and P). Reference and exposure policies differ across studies.

2 Related Work

Function, coordinates, and representation analysis. Representation similarity studies examine invariance choices (Kornblith et al., 2019; Williams et al., 2021) and how measures relate to model behavior (Ding et al., 2021; Klabunde et al., 2025). Functional and representational similarity can dissociate (Braun et al., 2025). Parameter and stimulus symmetries constrain interpretation (Theiss et al., 2026; Pashakhanloo & Zavatone-Veth, 2026). Cain discusses gauge freedom under invertible linear changes, metric dependence and whitening as a canonical gauge (Cain, 2026). Sevetlidis and Pavlidis use whitening and alignment in a representation holonomy diagnostic (Sevetlidis & Pavlidis, 2026). Remapping in the nullspace of a fixed decoder preserves decoded outputs while changing neural activity (Martín-Sánchez et al., 2025). These studies precede our question about compression.

Whitening, compression, and probing. Whitening's orthogonal ambiguity is classical (Kessy et al., 2018; Chiu, 2019). Whitening and coloring also underpin feature transforms for style transfer (Li et al., 2017). Random projection classifiers and their equivariance conditions also precede this study (Cannings & Samworth, 2017). Thulin's discussion there shows sensitivity to linear feature rescaling. Compressive learning depends on the task and hypothesis class (Kabán & Reeve, 2024). PCA/RP comparisons already evaluate pretrained embeddings (Zhang et al., 2024; Koenig et al., 2023). Control tasks and MDL probing separate predictive quality from memorization and coding cost (Hewitt & Liang, 2019; Voita & Titov, 2020). INLP removes decodable attributes by projection (Ravfogel et al., 2020); our invertible interventions preserve a specified reference. Concept-subspace evaluation also uses separate data for concept estimation, probe training and testing (Naowarat et al., 2026). Related work studies intrinsic dimension and other layer properties (Ansuini et al., 2019; Cheng et al., 2025; Skean et al., 2025; Kanatas et al., 2026), learns nested representations (Kusupati et al., 2022), and translates intermediate states into output predictions (Belrose et al., 2023). We

connect a change in the compression verdict under a fixed reference to objective matching within the same subspace and a separately tested numerical correction. The tools themselves are established.

3 REFERENCE BEHAVIOR AND A DECLARED COMPRESSION PROCEDURE

Let $s(h) = Wh + b$ be a fixed affine reference. Its predictions are those of a declared fitted probe, not necessarily the encoder's original classification head. Under $h' = Th + c$, with T invertible, $W' = WT^{-1}$ and $b' = b - WT^{-1}c$ preserve every logit. For the readout kept unchanged, let D collect pairwise-logit weight differences. The group $\mathcal{G}_D = \{T \in \mathrm{GL}(p) : DT = D\}$ describes linear changes $h' = Th$. An affine change $h' = Th + c$ preserves every score difference for every h exactly when $DT = D$ and $Dc = 0$. Preserving only the winning class is weaker.

In the main controls, students learn task labels using regularized cross entropy. Agreement with the fixed reference is measured after fitting. The fixed-teacher depth study instead uses teacher predictions as targets. Appendix A gives the loss and fitting conventions; earlier diagnostics retain their stated policies.

Operating points measure a procedure. Consider an unsupervised compressor m at dimension q and a newly fitted student. Let $A_{m,q}$ denote agreement with the fixed reference and $a_{m,q}$ denote label accuracy. Relative accuracy is $R_{m,q} = a_{m,q}/a_0$, where $a_0 > 0$ is the reference accuracy on the same examples. A point passes when $A \geq .90$ and $R \geq .90$. Equal accuracy alone does not imply equal decisions. Audit failure can reflect insufficient recoverable target information after compression, the fitting objective, finite samples or optimization; agreement alone does not distinguish these causes. For a K-class affine reference, a basis of $\mathrm{row}(D)$ supplies at most $K-1$ reference-aware coordinates that recover every score difference with the known decoder. An unsupervised compressor plus a fitted student answers a more constrained access question.

For a declared finite grid Q and compressor seeds, select $\widehat{q}_m = \min\{q \in Q : \widehat{\pi}_{m,q} \geq .8\}$ on validation data. Here $\widehat{\pi}_{m,q}$ is the empirical fraction passing the dual criterion. PCA uses one deterministic fit. Missing choices are censored. We check every candidate without assuming monotonicity. We evaluate frozen choices on held-out examples and fresh draws, without reselecting on test results. These quantities depend on the reference, labels, fitting objective, coordinate convention and finite search. Full-dimensional students are reported separately.

4 WHAT COORDINATE CORRECTION CAN GUARANTEE

The complete covariance freedom under a fixed readout. The issue extends beyond independent Gaussian nuisance coordinates. Let the fixed affine logits be $s_k(h) = w_k^\top h + b_k$, and let D have rows $(w_k - w_K)^\top$. Work in an orthonormal basis whose first $r = \mathrm{rank}(D)$ coordinates span $\mathrm{row}(D)$. Write $h = (x, n)$, with $x \in \mathbb{R}^r$ and $n \in \mathbb{R}^{p-r}$. The following characterization applies classical equal-Gram orthogonal equivalence (Arvind & Rattan, 2014, Section 2). Its block construction residualizes the nuisance coordinates, then applies standard whitening and coloring (Kessy et al., 2018; Li et al., 2017).

Proposition 1 (Covariance orbit with an unchanged affine reference)**.** *Assume $0 < r < p$ and h has finite second moments and positive definite centered covariance $\Sigma = \left(\begin{smallmatrix} A & B \\ B^\top & C \end{smallmatrix}\right)$. For a positive definite target Ω, attainability with the readout unchanged is characterized in any coordinates by*

$$\begin{aligned} \exists\, T \in \mathrm{GL}(p): \quad & DT = D, \quad T\Sigma T^\top = \Omega \\ \iff \quad & D\Omega D^\top = D\Sigma D^\top. \end{aligned} \tag{1}$$

Thus, in the declared orthonormal basis, the attainable covariances are exactly

$$\{T\Sigma T^\top : DT = D,\ T \text{ invertible}\} = \{\Omega \succ 0 : \Omega_{xx} = A\}. \tag{2}$$

In particular, for every $G \succ 0$ on the reference-null coordinates, there is such a T with $\mathrm{Cov}(Th) = \mathrm{diag}(A, G)$. Thus the positive eigenvalues and eigenvectors of the null block may be chosen freely, while reference decisions and softmax probabilities remain identical for every h.

No Gaussianity or zero cross-covariance assumption is needed. The reference block A is fixed: the proposition does not permit an arbitrary full covariance spectrum. A scalar covariance-based

recommendation is invariant to all these maps exactly when it depends on Σ only through A, for fixed D and the declared basis. If $p - r \geq r$, the top-r population PCA subspace can contain the entire reference subspace or lie entirely in its nullspace, by placing G below or above the spectrum of A. For Gaussian h, the latter coordinates are independent of the reference decision, so optimal agreement changes from 1 to the majority-reference probability. For non-Gaussian h, uncorrelated null coordinates can still contain predictive information; no analogous agreement conclusion follows from covariance alone. Appendix B gives the construction, proof, and singular covariance extension. These statements concern unrestricted invertible maps and population or exact empirical covariance, not the smaller finite scaling set used by the experiment or the optimum reached by a fitted probe.

A compression consequence. Let $U \in \mathbb{R}$ and $N \in \mathbb{R}^{p-1}$ have independent standard Gaussian components, with $p \geq 2$. Use fixed logits $(U, 0)$ and column representation $h_t = (U, tN)$ with $t > 0$. One-dimensional population PCA has optimal affine-decoder agreement 1 for $t < 1$ and $1/2$ for $t > 1$. A fixed Gaussian projection $(a, b^\top)$, with $a \in \mathbb{R}$ and $b \in \mathbb{R}^{p-1}$, has agreement $1/2 + \pi^{-1} \arctan(|a|/(t\|b\|))$, strictly between them almost surely. Their ordering reverses while the reference outputs stay fixed. Appendix E retains the degenerate boundary, optimal-decoder scope and proof. This example does not predict every fitted empirical compressor.

When the random-projection audit is invariant. Whitening is an established coordinate choice (Kessy et al., 2018; Cain, 2026); orthogonal invariance of random-projection classifiers also predates this work (Cannings & Samworth, 2017). Whitening does not by itself define a new compression algorithm. Its orthogonal ambiguity gives a precise guarantee for a complete randomized compression audit.

Proposition 2 (Invariance in law of the whitened RP audit)**.** *Let the fit features have mean μ and centered covariance $\Sigma \succ 0$. Apply a full whitener estimated only from the fit split $z(h) = F(h - \mu)$, with $F\Sigma F^\top = I$, to every fit and evaluation feature. For an invertible affine change $h' = Th + c$, recompute this whitener on the transformed fit set, obtaining*

$$z'(Th + c) = Qz(h), \qquad Q = F'TF^{-1}, \quad QQ^\top = I,$$

where $F'(T\Sigma T^\top)F'^\top = I$. Suppose the joint RP matrix collection, including all candidate dimensions and seeds, is invariant in law under $R_j \mapsto R_j Q^\top$; independent Gaussian matrices or a shared Gaussian master matrix satisfy this condition. With the same labels, reference predictions, and downstream fitting/selection rule, the whitened RP audit has the same joint distribution of scores and selected dimension in both coordinate systems. Under the explicit coupling $R'_j = R_j Q^\top$, all compressed inputs and audit outputs are identical.

Reference targets remain fixed or are preserved as described in Section 3.

The coupling is not ordinary reuse of an integer random seed. A new independent collection of draws can give a different finite-sample selection, despite the same distribution. Whitening is frozen within the audit; refitting it across phases requires independent phase-specific RP collections or an explicitly compatible coupling. Singular fit covariance additionally requires every evaluated centered feature to lie in the fit span (Appendix C); ridge whitening has no general affine-invariance guarantee. The result supplies a coordinate invariant distribution for this declared RP protocol, not a guarantee of small dimension, accurate compression, or invariance of PCA/UMAP choices.

The empirical assumptions require their own checks. A verified affine LayerNorm interface obeys $\sum_j (h_j - \beta_j)/\gamma_j = 0$ in ideal arithmetic for nonzero gains. Our four inspected interfaces therefore have structural support dimension $k = p - 1$. Their rounded caches can have numerical rank p. The support-aware method constructs the correctly transported affine support and fits covariance only within it. It checks the support of each evaluation split. It uses no ridge or eigenvalue floor. Passing these checks concerns numerical applicability. Projecting a rounded cache can still change reference scores. Neither invariance nor improved conditioning proves useful compression.

5 Experiments and evidence scope

Frozen features and references. The main controls use the actual saved ViT-B/16 (Dosovitskiy et al., 2021), DINOv2-small (Oquab et al., 2024), and BERT-base (Devlin et al., 2019) interfaces on

Table 1: Compression results on the 2,669 newly evaluated Pets examples. Each cell reports a point estimate and its approximate simultaneous interval. The eight scalar intervals reserve a 0.05 family error budget by Bonferroni. RP points average all 20 draws. Their intervals resample paired examples and draw columns. PCA uses paired-example resampling. The feasible and failing endpoints were chosen using validation only.

Method	q	t	Agreement A	Relative accuracy R
PCA	16	1	$0.930\ [0.916, 0.943]$	$0.978\ [0.964, 0.992]$
PCA	16	2	$0.868\ [0.850, 0.885]$	$0.912\ [0.893, 0.931]$
Gaussian RP	64	1	$0.917\ [0.908, 0.926]$	$0.948\ [0.939, 0.957]$
Gaussian RP	64	2	$0.812\ [0.799, 0.825]$	$0.845\ [0.831, 0.858]$

CIFAR-10 (Krizhevsky, 2009), Oxford-IIIT Pets (Parkhi et al., 2012) and TREC (Li & Roth, 2002). Each reference is a logistic probe trained only on the fit split. No encoder is retrained. The appendix documents model settings and the distinct reference policies used by earlier studies.

Fixed-reference intervention. Validation fixed PCA $q = 16$, RP $q = 64$ and reference-null scaling $t = 1 \to 2$. Models and rules were frozen before extracting 2,669 official test examples, disjoint from the earlier 1,000 and the fit/validation pool. Twenty fresh RP draws test the pass-to-fail criterion. The benchmark and earlier results were known; these project examples were previously unextracted and unevaluated. The earlier PCA-versus-RP ranking test remains a separate failed confirmation.

Independent objective matching. For row features Z and a Gaussian map $G = UR$ with orthonormal U, write $X_U = ZU$, $X_G = X_U R$ and $\Theta = W_G R^\top$. Each row of W_G contains one class's coefficients. The ordinary Gaussian penalty is $\|W_G\|_F^2 = \|\Theta R^{-\top}\|_F^2$. The penalty matched to unit coordinates is instead

$$J_{G,\mathrm{match}}(W_G, b) = \overline{\mathrm{CE}}(X_G W_G^\top + b, y) + \frac{\|W_G R^\top\|_F^2}{2Cn} = J_U(\Theta, b). \tag{3}$$

Here C controls inverse regularization strength, n is the fit sample size, and b is an unpenalized intercept. Unit and matched fits use separate zero starts, $C = 1$ and no coordinate scaler. The matched optimizer evaluates the Gaussian design through a triangular solve at every step. We check stationarity and the resulting coefficients, scores and predictions (Appendix A). This tests separately fitted corresponding objectives.

The saved-map comparison uses $q = 2p/3, p$; its thin-support counterpart uses $q = 2p/3, k$. The support control compares full-cache and thin whitening under identity, orthogonal and condition-32 gauges, with direct baselines and fresh Gaussian draws. All schedules were fixed before comparing results. Appendix A gives the objectives, seeds, counts and numerical failures.

Uncertainty and exposure. Intervals resample paired examples and compressor draws, conditional on the frozen fit sample and validation-selected conditions. They exclude encoder-training variation. A pass frequency of .8 does not certify a population pass probability. Numerical follow-ups reuse exposed examples and retain every failed fit, warning and censored choice.

6 Results: behavior, information, and fitted access

6.1 The same reference can receive different compression verdicts

The frozen Pets study confirms the strong criterion for a change from passing to failing for PCA and RP at their fixed dimensions (Table 1). All 20 fresh RP draws jointly pass at $t = 1$ and fail at $t = 2$. The reference-preservation checks pass. This shows that coordinates affect the operational verdict for the tested procedures. It is not an unrestricted claim about representation information.

Numerical resolution of the inherited PCA warning. A separately declared refinement of the same objective checks both $q = 16$ endpoints, using the saved and zero starts with a fixed numerical

Table 2: Full-dimensional diagnostic: mean agreement over the 20 inherited test maps. Gaussian $G = UR$ and unit U span the full space. Matching the Gaussian penalty to the unit objective recovers the unit fitted decisions in every cell. Match counts apply the full predeclared independent-fit gate, including continuous scores and common-coordinate stationarity. The largest tested projection-map condition is $\kappa(G)$. The inherited full-cache covariance remains ill-conditioned. These rows are not dimensional reduction.

Setting	p	Gaussian	Unit U	Matched G	Match	max $\kappa(G)$
ViT / CIFAR-10	768	0.878	0.965	0.965	20/20	1.18×10^4
DINOv2 / CIFAR-10	384	0.924	0.968	0.968	20/20	3.58×10^3
DINOv2 / Pets	384	0.919	0.965	0.965	20/20	3.58×10^3
BERT / TREC	768	0.867	0.920	0.920	20/20	1.18×10^4

budget. Both endpoints reach the common stationarity threshold, and their observed pass/fail verdicts remain unchanged on validation and the exposed complement. The refined $t = 2$ fit changes 4 complement predictions, while agreement remains 0.868. Numerical differences are substantive: maximum probability change is 0.134 and maximum centered-logit change is 7.490. The original frozen confirmation and predictions are preserved. This solver follow-up is not another independent prospective confirmation.

6.2 The same linear information can yield different fitted behavior

The matched fit uses common-coordinate preconditioning. It checks the same regularized objective in corresponding coordinates. It does not test unpreconditioned solver convergence. Table 2 separates retained information from fitted access without reducing dimension. An invertible Gaussian map retains every available linear score function. Its isotropically regularized fitted probe can still lose substantial agreement. Independently matching its penalty to the unit-coordinate objective restores the hard predictions in unit coordinates in all 240 cells using the saved maps, including both tested dimensions and both phases. The complete numerical equivalence gate passes in 216/240 cells. In the remaining 24 cells, BERT / TREC logit or probability differences exceed the tolerances. Both fits satisfy the declared common-coordinate stationarity check, and no hard prediction changes. Those failures preclude a claim of exact equality from the finite solver results.

The same independent comparison in verified thin coordinates has 0 changed prediction instances across 240 cells. The full numerical gate passes in 211/240 cells. These are separately optimized fits of corresponding objectives, not a transported fitted solution. Algebraic coefficient transport is checked separately. This explains the observed contrast in hard decisions under the tested fitting objectives. It is not a new equivalence theorem or a claim that all finite logits coincide. Across both comparisons, maximum centered-logit and probability discrepancies are 4.88×10^{-5} and 4.38×10^{-6}, respectively. All 960 independent unit/matched fits meet the common-coordinate stationarity threshold. At the compressed $q = 2p/3$ point, unit and matched penalties preserve the same task-specific pattern. On Pets, mean test agreement changes from 0.875 to 0.919 in full-cache coordinates and from 0.868 to 0.916 in thin coordinates. Matched fits pass the behavioral criterion in 20/20 and 20/20 draws, respectively. Projecting the Pets reference onto the thin support still fails the probability tolerance. Its behavioral result therefore does not establish utility under all criteria for the corrected audit. On TREC, agreement decreases from 0.756 to 0.749 and from 0.752 to 0.747, respectively. Objective matching is not a universal performance improvement. The earlier energy-matched comparison also improves agreement in three vision settings and worsens it on TREC. Appendix G retains its distinct contrast and adjusted intervals. The result with matching penalties gives a more direct explanation for the tested gap between decisions fitted in Gaussian and unit coordinates. It does not attribute all earlier coordinate effects to a single mechanism.

6.3 Correction improves conditioning without guaranteeing compression

For all four inspected LayerNorm interfaces, the FIT data span the verified $k = p - 1$ support and thin covariance/whitening checks pass. Table 3 shows the large reduction in condition number. No tested setting recovers a qualifying compressed Gaussian operating point at the fixed $q = 2p/3$ can-

Table 3: Support-aware whitening improves conditioning without a passing Gaussian compressed point. We report identity-gauge fit covariance condition, test agreement at $q = 2p/3$, and full/thin passing draws. Passing requires $A, R \geq .90$ before stricter numerical checks. These fresh support-study draws differ from the saved full-cache maps used for objective matching. The comparison also changes the support convention.

Setting	Full condition	Thin condition	Full A	Thin A	Passes full / thin
ViT / CIFAR-10	9.56×10^{14}	3.11×10^{4}	0.782	0.784	0/20; 0/20
DINOv2 / CIFAR-10	6.60×10^{14}	5.59×10^{2}	0.835	0.842	0/20; 0/20
DINOv2 / Pets	1.09×10^{15}	2.53×10^{3}	0.872	0.868	0/20; 0/20
BERT / TREC	3.14×10^{14}	9.63×10^{3}	0.755	0.752	0/20; 0/20

didate. Direct $q = p$ and $q = k$ baselines remain diagnostics: removing one structurally redundant coordinate is not a compression achievement.

Projecting rounded caches onto that support changes 0 observed hard reference decisions. However, 28/36 gauge/split comparisons fail the stricter fixed score/probability tolerances. These failures remain in eligibility calculations. The largest reference probability and pairwise-logit changes are 1.81×10^{-7} and 1.26×10^{-6}, respectively. The explicitly transported audit produces identical hard predictions in 512/512 separately fitted gauge comparisons. Continuous predictions still differ. Thus support-aware whitening improves the numerical prerequisites. It neither guarantees useful compression nor establishes exact finite-precision reference preservation. The original full-whitening grid was censored in all seven tested encoder/task settings. In the CIFAR-100 fine/coarse check using the same features, both tasks are censored through $q = 256$, despite feasible full-dimensional probes (Appendix L). This is not evidence of equal intrinsic task dimension. The original Pets ranking test did not confirm, and TREC selected no qualifying validation pair. Numerical refinement also leaves the Pets interval criterion unmet. The vision reversals persist without postcompression scaling. Both follow-ups use exposed examples (Appendix S). These results remain separate from the later intervention, which changes feasibility while keeping the method fixed.

7 Implications and limitations

State the task, fixed reference, coordinates, probe, fitting rule and budget. Before interpreting probe failure as information loss, evaluate a baseline without dimension reduction under the same fitting convention. Check support, conditioning and convergence, then compare compressed behavior with that baseline. A known coordinate map permits objective transport for a controlled comparison. Fit-only whitening gives the stated conditional RP invariance result, not guaranteed compression. If no candidate qualifies, report no qualifying choice for that protocol and grid, rather than an intrinsic information limit.

The audit does not require the original pretraining coordinates. Calling chosen coordinates canonical does not recover those coordinates or supply an intrinsic interpretation.

Our positive guarantee concerns a declared randomized audit under explicit covariance, support and coupling assumptions. Equality in joint distribution and equality under the transported coupling are different statements. Ordinary reuse of an integer seed and finite-precision results are also distinct. Regularization, finite optimization, labels, sample size and distribution shift remain important choices. Full whitening changes the probe objective as well as numerical conditioning. Thin whitening also projects rounded caches. The controls test these specified changes and do not explain every earlier whitening deficit or the Pets intervention. Stable coordinates or unchanged hard reference decisions alone do not establish useful support-aware compression.

8 Conclusion

A coordinate change can change a compression verdict while reference behavior stays fixed. Retaining the same available linear information does not guarantee the same fitted decoder when the objective changes. Independent objective matching recovers the tested decisions in unit coordinates.

Some continuous outputs still fail the numerical tolerances. Support-aware whitening improves conditioning but leaves the tested Gaussian compression failure intact. Compression results should state the coordinate convention, decoder, full-dimensional baseline and fitting procedure.

AI USE STATEMENT

AI assisted with code and tests, mathematical analysis, experiment design, result review, reference checks, and language editing. The authors reviewed the AI-assisted work and take responsibility for the final paper.

REPRODUCIBILITY STATEMENT

The anonymous supplement includes experimental configurations, data split information, saved predictions and analysis code. The supplied scripts reproduce key behavioral statistics and the opening figure from saved outputs. Diagnostics of continuous outputs and covariance are included as reported results. The appendix provides model settings, compression methods, evaluation rules and statistical analyses. The workflow does not retrain the encoders.

ETHICS STATEMENT

The study analyzes public benchmark representations and does not provide a clinical decision system. The Breast Cancer diagnostic does not establish clinical utility, fairness or safety. Preserving reference behavior can preserve errors and biases. Dataset sources and licenses are documented. Human authors remain responsible for the final content.

REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html.

V. Arvind and Gaurav Rattan. The complexity of geometric graph isomorphism. Technical Report TR14-070, Electronic Colloquium on Computational Complexity, 2014. URL https://eccc.weizmann.ac.il/report/2014/070/.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112v6. arXiv:2303.08112v6, revised November 2025.

Lukas Braun, Erin Grant, and Andrew M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5355–5382. PMLR, 2025. URL https://proceedings.mlr.press/v267/braun25a.html.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces, 2026. URL https://arxiv.org/abs/2603.06774.

Timothy I. Cannings and Richard J. Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):959–1035, 2017. doi: 10.1111/rssb.12228. URL https://academic.oup.com/jrsssb/article/79/4/959/7040648.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 380–388, 2002. doi: 10.1145/509907.509965. URL https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/CharikarEstim.pdf.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/57568e093cbe0a222de0334b36e83cf5-Abstract-Conference.html.

Tai-Yin Chiu. Understanding generalized whitening and coloring transform for universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4452–4460, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Chiu_Understanding_Generalized_Whitening_and_Coloring_Transform_for_Universal_Style_Transfer_ICCV_2019_paper.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Peter Frankl and Hiroshi Maehara. Some geometric applications of the beta distribution. *Annals of the Institute of Statistical Mathematics*, 42(3):463–474, 1990. doi: 10.1007/BF00049302. URL https://www.ism.ac.jp/editsec/aism/pdf/042_3_0463.pdf.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10929–10974, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Ata Kabán and Henry Reeve. Structure discovery in PAC-learning by random projections. *Machine Learning*, 113:5685–5730, 2024. doi: 10.1007/s10994-024-06531-0. URL https://link.springer.com/article/10.1007/s10994-024-06531-0.

Angelos-Nikolaos Kanatas, Yuexuan Kong, Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. What makes a good layer? assessing the layer-wise intrinsic properties of music foundation models. In *International Society for Music Information Retrieval Conference*, 2026. URL https://arxiv.org/abs/2608.14819v1. Accepted at ISMIR 2026; forthcoming.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. doi: 10.1080/00031305.2016.1277159. URL https://arxiv.org/abs/1512.00809.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.02907.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. ReSi: A comprehensive benchmark for representational similarity measures. In *International Conference on Learning Representations*, pp. 18701–18740, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2eef1f75516b0cfd3e944345e5f88c08-Abstract-Conference.html.

Alexander Koenig, Maximilian Schambach, and Johannes Otterbach. Uncovering the inner workings of STEGO for safe unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3789–3798, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/SAIAD/papers/Koenig_Uncovering_the_Inner_Workings_of_STEGO_for_Safe_Unsupervised_Semantic_CVPRW_2023_paper.pdf.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://leon.bottou.org/publications/pdf/ieee-1998.pdf.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/49182f81e6a13cf5eaa496d51fea6406-Abstract.html.

Guillermo Martín-Sánchez, Christian K. Machens, and William F. Podlaski. Three types of remapping with linear decoders: A population-geometric perspective. *PLOS Computational Biology*, 21(10):e1013545, 2025. doi: 10.1371/journal.pcbi.1013545. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1013545.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://joss.theoj.org/papers/10.21105/joss.00861.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426v3*, 2020. doi: 10.48550/arXiv.1802.03426. URL https://arxiv.org/abs/1802.03426v3.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 574–587. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.conll-main.34. URL https://aclanthology.org/2026.conll-main.34/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/.

Farhad Pashakhanloo and Jacob A. Zavatone-Veth. Stimulus symmetries can confound representational similarity analyses, 2026. URL https://arxiv.org/abs/2605.21324.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: 10.1609/aimag.v29i3.2157. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2157.

Vasileios Sevetlidis and George Pavlidis. Gauge-invariant representation holonomy. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=czJqKToDGq.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 55854–55875, 2025. URL https://proceedings.mlr.press/v267/skean25a.html.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905 of *Proceedings of SPIE*, pp. 861–870. SPIE, 1993. doi: 10.1117/12.148698. URL https://doi.org/10.1117/12.148698.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://pubmed.ncbi.nlm.nih.gov/11125149/.

Marvin Theiss, Lukas Braun, Andrew M. Saxe, and Erin Grant. Parameter symmetries determine representational geometry in overparameterized nonlinear networks. In *ICML 2026 Workshop on Weight-Space Symmetries: From Foundations to Practical Applications*, 2026. URL https://openreview.net/forum?id=CT8TbdqYmn.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355. Association for Computational Linguistics, 2018. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48. PMLR, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6530–6543, 2024. URL https://aclanthology.org/2024.lrec-main.579/.

A DECLARED OBJECTIVE AND SUPPORT FOLLOW-UPS

These studies were motivated by inspection of the earlier results. All validation and benchmark test examples were already exposed. Fresh seeds refresh compressor randomness only. The protocols, dimensions, penalties, numerical thresholds and schedules were locked before the follow-up outcome comparisons. Original favorable and unfavorable outcomes, including failed prospective ranking reversal, remain unchanged. The protocols were fixed before these comparisons, but were not externally preregistered. These comparisons do not provide independent confirmation on new data.

Schedule and numerical support. The support control retains all 744 slots: 720 Gaussian fits at $q = 2p/3$ and 24 direct baselines. Each direct fit is evaluated on both phases. The four settings use identity, orthogonal and condition-32 gauges; validation seeds are 61000–61009 and test seeds 62000–62019. Thin coordinates use verified gains/biases and correctly transported affine supports. Centering and SVD covariance use only the FIT split, without ridge or eigenvalue flooring. Support residuals and fixed-reference score/probability/decision changes are checked independently on all three splits. We report both q/p and q/k.

Targets and loss. The modern encoder controls fit the task labels. Their objective is mean multiclass cross entropy plus $\|W\|_F^2/(2Cn_{\mathrm{fit}})$, with $C = 1$, no class/sample weights and an unpenalized intercept. The Pets intervention uses raw centered compressed inputs and no postcompression scaler. The support and matched-objective controls use their declared whitened coordinates and no such scaler. The fixed-teacher depth study uses teacher predictions as training targets. The earlier core, layer-reference and reversal studies retain their own standardization and train/refit rules, documented in their respective sections.

Coefficient convention. Let $Z \in \mathbb{R}^{n\times p}$, $G = UR \in \mathbb{R}^{p\times q}$, $U^\top U = I_q$ and invertible $R \in \mathbb{R}^{q\times q}$. For K classes use coefficient rows $W_G, \Theta \in \mathbb{R}^{K\times q}$, with $\Theta = W_G R^\top$. Then $X_U = ZU$, $X_G = X_U R$ and $X_G W_G^\top + \mathbf{1}b^\top = X_U \Theta^\top + \mathbf{1}b^\top$. The class count K is distinct from inverse regularization strength C.

Independent matched objectives. With row-model coefficients W_G and $G = UR$, common coefficients are $\Theta = W_G R^\top$. Unit and matched-Gaussian fits minimize mean multiclass cross entropy plus $\|\Theta\|_F^2/(2Cn)$, with $C = 1$, fixed labels, sorted original classes, no weights, and an unpenalized intercept. Each arm starts from zero in a separate L-BFGS-B invocation. The matched arm uses invertible common-Θ preconditioning. It evaluates its Gaussian design on every call through $W_G = \mathrm{solve}(R, \Theta^\top)^\top$. Thus the comparison independently optimizes corresponding objectives and checks the Gaussian computation path. It does not test unpreconditioned native-Gaussian optimization. No matrix inverse is formed. Algebraic transport of fitted unit coefficients is a separate sanity check.

The primary identity control reuses the exact saved H2 maps at $q = 2p/3$ and $q = p$ for all four settings, with seeds 51000–51009 and 52000–52019. This gives 240 paired cells, 240 retained Gaussian fits and 480 new independent unit/matched fits. The thin repetition contains 240 cells at $q = 2p/3$ and $q = k$: 120 compressed Gaussian fits are reused from the support control, 120 full-k Gaussian fits are added, and 480 independent unit/matched fits are added. Its algebraic prerequisites and reference-preservation/utility gates are reported separately.

L-BFGS-B uses maxiter=20000, maxfun=200000, maxls=50, maxcor=10, gtol=10^{-10} and ftol equal to float64 epsilon. The new full-k ordinary Gaussian fits retain the original sklearn convention (tol=10^{-8}, maxiter=20000). Every attempt ends with a saved outcome. Outcomes do not trigger retries or relaxed tolerances.

Predeclared equivalence criteria. Each independent arm must return successfully without warnings, with finite quantities and common coefficient/intercept gradient infinity norms at most 10^{-6}. The paired gate requires scaled objective difference at most 10^{-8}, gauge-fixed relative coefficient difference at most 10^{-5}, maximum centered-logit discrepancy at most 10^{-5}, probability discrepancy at most 10^{-6}, and zero changed hard predictions on FIT and evaluation examples. Class-mean coefficient and intercept shifts fix the softmax comparison gauge. A top-two margin at most 2×10^{-5}

Table 4: All groups with independently fitted matching penalties. Agreement is averaged over the complete declared group. Match counts use every continuous and discrete gate; failed gates remain visible.

Setting	Coordinates	q	Phase	G	U	Match	Gates	max logit
ViT / CIFAR-10	Full	512	Val.	0.784	0.835	0.835	10/10	1.97×10^{-8}
ViT / CIFAR-10	Full	768	Val.	0.888	0.978	0.978	10/10	1.48×10^{-6}
ViT / CIFAR-10	Full	512	Test	0.782	0.833	0.833	20/20	1.37×10^{-6}
ViT / CIFAR-10	Full	768	Test	0.878	0.965	0.965	20/20	7.16×10^{-7}
DINOv2 / CIFAR-10	Full	256	Val.	0.852	0.873	0.873	10/10	9.42×10^{-7}
DINOv2 / CIFAR-10	Full	384	Val.	0.928	0.977	0.977	10/10	4.96×10^{-7}
DINOv2 / CIFAR-10	Full	256	Test	0.839	0.863	0.863	20/20	1.19×10^{-6}
DINOv2 / CIFAR-10	Full	384	Test	0.924	0.968	0.968	20/20	4.54×10^{-7}
DINOv2 / Pets	Full	256	Val.	0.892	0.934	0.934	10/10	3.12×10^{-7}
DINOv2 / Pets	Full	384	Val.	0.938	0.984	0.984	10/10	8.55×10^{-8}
DINOv2 / Pets	Full	256	Test	0.875	0.919	0.919	20/20	4.04×10^{-8}
DINOv2 / Pets	Full	384	Test	0.919	0.965	0.965	20/20	1.51×10^{-8}
BERT / TREC	Full	512	Val.	0.694	0.691	0.691	1/10	4.88×10^{-5}
BERT / TREC	Full	768	Val.	0.814	0.888	0.888	9/10	1.84×10^{-5}
BERT / TREC	Full	512	Test	0.756	0.749	0.749	6/20	2.84×10^{-5}
BERT / TREC	Full	768	Test	0.867	0.920	0.920	20/20	3.06×10^{-6}
ViT / CIFAR-10	Thin	512	Val.	0.779	0.832	0.832	10/10	3.22×10^{-7}
ViT / CIFAR-10	Thin	767	Val.	0.883	0.977	0.977	10/10	9.79×10^{-8}
ViT / CIFAR-10	Thin	512	Test	0.784	0.837	0.837	20/20	5.06×10^{-7}
ViT / CIFAR-10	Thin	767	Test	0.879	0.963	0.963	20/20	2.08×10^{-7}
DINOv2 / CIFAR-10	Thin	256	Val.	0.843	0.869	0.869	10/10	1.76×10^{-6}
DINOv2 / CIFAR-10	Thin	383	Val.	0.925	0.978	0.978	10/10	6.79×10^{-7}
DINOv2 / CIFAR-10	Thin	256	Test	0.842	0.864	0.864	20/20	1.11×10^{-6}
DINOv2 / CIFAR-10	Thin	383	Test	0.922	0.968	0.968	20/20	6.55×10^{-7}
DINOv2 / Pets	Thin	256	Val.	0.897	0.943	0.943	10/10	1.39×10^{-6}
DINOv2 / Pets	Thin	383	Val.	0.940	0.985	0.985	10/10	4.86×10^{-7}
DINOv2 / Pets	Thin	256	Test	0.868	0.916	0.916	20/20	2.09×10^{-6}
DINOv2 / Pets	Thin	383	Test	0.916	0.965	0.965	20/20	4.76×10^{-7}
BERT / TREC	Thin	512	Val.	0.697	0.691	0.691	2/10	3.84×10^{-5}
BERT / TREC	Thin	767	Val.	0.811	0.896	0.896	8/10	2.26×10^{-5}
BERT / TREC	Thin	512	Test	0.752	0.747	0.747	4/20	3.90×10^{-5}
BERT / TREC	Thin	767	Test	0.866	0.920	0.920	17/20	1.36×10^{-5}

defines a near tie; near ties never remove changes from the gate. Common-coordinate stationarity is recomputed independently, since native-coordinate gradients alone can hide poor conditioning.

Validation and exposure. Only the declared $q = 2p/3$ candidate is considered by the support control. Eight of ten validation and sixteen of twenty test seeds must satisfy the dual behavioral predicate. Strict eligibility also enforces numerical and reference gates. Test results never select another dimension. Full-p and full-k rows are diagnostic, not useful compression. Conditional means are withheld whenever a declared cell is missing or failed. Seed summaries condition on the exposed benchmark and fixed training/reference choices. They do not measure variation from encoder training.

The PCA diagnostic uses two bounded starts per endpoint (saved and zero), maxiter=100000, maxfun=200000, maxls=50, maxcor=20, gtol=10^{-10} and float64-epsilon ftol, with a 1800-second limit per attempt. Selection requires solver success, finite values and gradient infinity norm at most 10^{-6}. The smallest objective wins, with the saved start winning an exact tie. All four attempts and both endpoint outcomes are retained.

Table 5: Direct full-coordinate baselines under every fixed gauge. Gauge 1 is the orthogonal control; gauge 32 has condition number 32. All fits and phases are retained. Strict eligibility includes reference-preservation tolerances, and is distinct from the behavioral predicate.

Setting	Coord.	Gauge	q	Phase	A	R	Strict
DINOv2 / CIFAR-10	full	1	384	Val.	0.977	0.984	0
DINOv2 / CIFAR-10	full	32	384	Val.	0.977	0.984	0
DINOv2 / CIFAR-10	full	identity	384	Val.	0.977	0.984	0
DINOv2 / CIFAR-10	thin	1	383	Val.	0.978	0.986	1
DINOv2 / CIFAR-10	thin	32	383	Val.	0.978	0.986	0
DINOv2 / CIFAR-10	thin	identity	383	Val.	0.978	0.986	1
ViT / CIFAR-10	full	1	768	Val.	0.978	0.997	0
ViT / CIFAR-10	full	32	768	Val.	0.978	0.997	0
ViT / CIFAR-10	full	identity	768	Val.	0.978	0.997	0
ViT / CIFAR-10	thin	1	767	Val.	0.977	0.995	0
ViT / CIFAR-10	thin	32	767	Val.	0.977	0.995	0
ViT / CIFAR-10	thin	identity	767	Val.	0.977	0.995	0
DINOv2 / Pets	full	1	384	Val.	0.984	0.997	0
DINOv2 / Pets	full	32	384	Val.	0.984	0.997	0
DINOv2 / Pets	full	identity	384	Val.	0.984	0.997	0
DINOv2 / Pets	thin	1	383	Val.	0.985	0.996	0
DINOv2 / Pets	thin	32	383	Val.	0.985	0.996	0
DINOv2 / Pets	thin	identity	383	Val.	0.985	0.996	0
BERT / TREC	full	1	768	Val.	0.888	0.969	0
BERT / TREC	full	32	768	Val.	0.888	0.969	0
BERT / TREC	full	identity	768	Val.	0.888	0.969	0
BERT / TREC	thin	1	767	Val.	0.896	0.972	0
BERT / TREC	thin	32	767	Val.	0.896	0.972	0
BERT / TREC	thin	identity	767	Val.	0.896	0.972	0

Table 6: Direct full-coordinate baselines under every fixed gauge. Gauge 1 is the orthogonal control; gauge 32 has condition number 32. All fits and phases are retained. Strict eligibility includes reference-preservation tolerances, and is distinct from the behavioral predicate.

Setting	Coord.	Gauge	q	Phase	A	R	Strict
DINOv2 / CIFAR-10	full	1	384	Test	0.968	0.982	0
DINOv2 / CIFAR-10	full	32	384	Test	0.968	0.982	0
DINOv2 / CIFAR-10	full	identity	384	Test	0.968	0.982	0
DINOv2 / CIFAR-10	thin	1	383	Test	0.968	0.982	1
DINOv2 / CIFAR-10	thin	32	383	Test	0.968	0.982	0
DINOv2 / CIFAR-10	thin	identity	383	Test	0.968	0.982	1
ViT / CIFAR-10	full	1	768	Test	0.965	0.986	0
ViT / CIFAR-10	full	32	768	Test	0.965	0.986	0
ViT / CIFAR-10	full	identity	768	Test	0.965	0.986	0
ViT / CIFAR-10	thin	1	767	Test	0.963	0.984	0
ViT / CIFAR-10	thin	32	767	Test	0.963	0.984	0
ViT / CIFAR-10	thin	identity	767	Test	0.963	0.984	0
DINOv2 / Pets	full	1	384	Test	0.965	0.986	0
DINOv2 / Pets	full	32	384	Test	0.965	0.986	0
DINOv2 / Pets	full	identity	384	Test	0.965	0.986	0
DINOv2 / Pets	thin	1	383	Test	0.965	0.986	0
DINOv2 / Pets	thin	32	383	Test	0.965	0.986	0
DINOv2 / Pets	thin	identity	383	Test	0.965	0.986	0
BERT / TREC	full	1	768	Test	0.920	0.987	0
BERT / TREC	full	32	768	Test	0.920	0.987	0
BERT / TREC	full	identity	768	Test	0.920	0.987	0
BERT / TREC	thin	1	767	Test	0.920	0.985	0
BERT / TREC	thin	32	767	Test	0.920	0.985	0
BERT / TREC	thin	identity	767	Test	0.920	0.985	0

Table 7: Fixed-objective PCA solver diagnostic at both endpoints. Each endpoint uses the same saved map, FIT examples, targets, reference and $C = 1$ objective as its original fit. The lowest-objective successful stationary candidate is selected using FIT diagnostics alone.

t	Original J	Refined J	Original $\|g\|_\infty$	Refined $\|g\|_\infty$	Start
1	0.080486844	0.080485369	9.24×10^{-6}	4.40×10^{-9}	zero
2	0.235788296	0.235774062	1.33×10^{-4}	7.23×10^{-8}	zero

Table 8: PCA refined predictions versus the original saved fit. The original confirmation calculations remain unchanged. The complement is now exposed, so this is a numerical follow-up, not another prospective confirmation.

t	Split	A	R	Changes	max prob.	max logit	Verdict changed
1	Val.	0.927	0.958	0	0.010	0.966	False
1	Complement	0.930	0.978	0	0.111	0.995	False
2	Val.	0.868	0.892	1	0.056	7.162	False
2	Complement	0.868	0.912	4	0.134	7.490	False

B COMPLETE COVARIANCE FREEDOM UNDER A FIXED AFFINE REFERENCE

Proposition 1 is an application of classical matrix equivalence, not a new matrix theorem. We give a short equal-Gram proof, followed by a constructive residualization proof that also covers singular covariance. Covariances are centered throughout. The representation mean need not be zero, and the reference intercepts remain unchanged.

Coordinate-free proof from equal-Gram equivalence. Necessity in Equation (1) follows by multiplying $T\Sigma T^\top = \Omega$ by D and $D^\top$ and using $DT = D$. For sufficiency, put $U = D\Sigma^{1/2}$ and $V = D\Omega^{1/2}$. Then $UU^\top = VV^\top$, so classical equal-Gram equivalence gives an orthogonal O with $VO = U$ (Arvind & Rattan, 2014, Section 2). Define $T = \Omega^{1/2} O \Sigma^{-1/2}$. It is invertible and satisfies

$$DT = VO\Sigma^{-1/2} = D, \qquad T\Sigma T^\top = \Omega.$$

This argument does not require full row rank of D. In the chosen basis $D = (D_x, 0)$ with D_x of full column rank, so $D\Omega D^\top = D\Sigma D^\top$ is equivalent to $\Omega_{xx} = A$.

All admissible linear maps. In the orthonormal coordinates used in the main statement, $D = (D_x, 0)$ and D_x has full column rank r. For a block matrix T, the equation $DT = D$ therefore forces its top block row to be $(I_r, 0)$. Hence all and only the admissible invertible maps have

$$T = \begin{pmatrix} I_r & 0 \\ L & M \end{pmatrix}, \qquad L \in \mathbb{R}^{(p-r)\times r}, \quad M \in \mathrm{GL}(p-r). \tag{4}$$

Their inverses have bottom block row $(-M^{-1}L, M^{-1})$. The condition preserves $(w_k - w_K)^\top h + b_k - b_K$ for every input h. It therefore preserves every pairwise logit difference, all softmax probabilities, and decisions under a fixed tie rule. Raw logits can change by an input-dependent scalar common to all classes for each input. Conversely, equality of every softmax probability for every h implies equality of each log probability ratio and hence $DT = D$. Equality of decisions alone is a weaker requirement. In the original coordinates conjugate T by the chosen orthogonal basis.

Necessity and constructive sufficiency for positive definite covariance. Equation (4) leaves x unchanged, so every attainable covariance has top-left block A. Positive definiteness is preserved under invertible congruence. For the converse, let any target covariance with this block be

$$\Omega = \begin{pmatrix} A & K \\ K^\top & C_* \end{pmatrix} \succ 0, \qquad S = C - B^\top A^{-1} B, \qquad G = C_* - K^\top A^{-1} K.$$

The two Schur complements S, G are positive definite. Define

$$M = G^{1/2} S^{-1/2}, \qquad L = K^\top A^{-1} - M B^\top A^{-1}. \tag{5}$$

Symmetric positive definite square roots give $MSM^\top = G$; M and therefore T are invertible. Put $e = n - B^\top A^{-1} x$. Then $\mathrm{Cov}(x, e) = 0$ and $\mathrm{Cov}(e) = S$, without any independence assumption. The transformed null coordinates are $n' = Lx + Mn = K^\top A^{-1} x + Me$. Thus

$$\mathrm{Cov}(x, n') = K, \qquad \mathrm{Cov}(n') = K^\top A^{-1} K + G = C_*.$$

This proves $T\Sigma T^\top = \Omega$ and both inclusions in Equation (2). Taking $K = 0$ proves the block-diagonal claim. For any orthogonal matrix V on the nullspace and any positive numbers $g_1, \ldots, g_{p-r}$, setting $G = V\,\mathrm{diag}(g_j)V^\top$ realizes those null eigenvalues and directions. The eigenvalues of the full block-diagonal covariance are the union of these numbers and those of A. They are not all free, and a scalar scaling T_t generally cannot realize this full orbit when $B \neq 0$.

Exactly which covariance summaries are invariant. Fix D and the basis. If a scalar or discrete recommendation $d(\Sigma)$ is a function only of A, its invariance is immediate. Conversely, any two positive definite covariances with the same A lie in the same orbit by the proposition, so invariance forces $d(\Sigma)$ to be constant on that set. Define $\phi(A) = d(\mathrm{diag}(A, I_{p-r}))$; then $d(\Sigma) = \phi(A)$. This characterizes covariance-only recommendations. A behavioral dimension selected on validation data also depends on labels, higher moments, fitting, and random draws. This corollary alone therefore does not characterize that entire procedure or imply failure of every compressor.

A general Gaussian PCA consequence. Suppose $m = p - r \geq r$ and h is Gaussian with $\Sigma \succ 0$. Choose two target covariances $\mathrm{diag}(A, \epsilon I_m)$ and $\mathrm{diag}(A, \gamma I_m)$, with $0 < \epsilon < \lambda_{\min}(A)$ and $\gamma > \lambda_{\max}(A)$, using the proposition. At $q = r$, the first population PCA projector is exactly the projector onto the reference subspace. A specified affine decoder can therefore recover every original pairwise logit difference and achieve agreement one. Every top-r population PCA subspace of the second covariance lies inside the nullspace, even if its internal eigenspace is non-unique. The transformed signal and null coordinates are jointly Gaussian and uncorrelated, hence independent. The reference decision is a deterministic function of the signal coordinates. No classifier of the retained null coordinates can therefore have agreement greater than $\max_k \Pr\{f(h) = k\}$, attained by a constant affine classifier. The drop is strict when the reference decision is not almost surely constant. This is an existence result with optimal decoding. It does not identify the classifier found by finite-sample regularized logistic regression, or imply a PCA–RP reversal in every Gaussian or real model. For non-Gaussian distributions only the subspace claim is automatic. For example, if $X \sim \mathcal{N}(0,1)$ and $N = \mathbf{1}\{|X| > 1\} - \Pr\{|X| > 1\}$, then $\mathrm{Cov}(X, N) = 0$ but N predicts the three-class affine reference with logits $(-X - 1, 0, X - 1)$. Even zero cross-covariance is not independence.

Singular covariance and rank constraints. If $\Sigma \succeq 0$ is singular, the exact orbit is

$$\{T\Sigma T^\top : DT = D,\ T \text{ invertible}\} = \{\Omega \succeq 0 : \Omega_{xx} = A, \mathrm{rank}(\Omega) = \mathrm{rank}(\Sigma)\}. \tag{6}$$

Here both A and the residual covariance may be singular; do not insert ordinary inverse square roots into Equation (5). To verify the extension, positive semidefiniteness implies $\mathrm{range}(B) \subseteq \mathrm{range}(A)$: if $v \in \ker(A)$, nonnegativity of $(v, tn)^\top \Sigma (v, tn) = 2tv^\top Bn + t^2 n^\top Cn$ for both signs of sufficiently small t forces $v^\top B = 0$. With $A^\dagger$ the Moore–Penrose inverse, residualization yields $S = C - B^\top A^\dagger B \succeq 0$ and a block-diagonal congruence $\mathrm{diag}(A, S)$, so $\mathrm{rank}(\Sigma) = \mathrm{rank}(A) + \mathrm{rank}(S)$. The target likewise has residual covariance $G = C_* - K^\top A^\dagger K \succeq 0$ and $\mathrm{range}(K) \subseteq \mathrm{range}(A)$. The common total rank implies $\mathrm{rank}(G) = \mathrm{rank}(S)$. Two positive semidefinite matrices of equal rank admit an invertible congruence $MSM^\top = G$: align their positive eigenspaces, rescale the positive eigenvalues, and map their nullspaces invertibly. Set $L = (K^\top - MB^\top)A^\dagger$. The residual calculation above now gives the target covariance, using $AA^\dagger B = B$ and $AA^\dagger K = K$. Necessity follows from preservation of x and covariance rank. Thus an invertible map cannot create variance in new independent sample directions. In particular, a sample covariance of rank at most $n - 1 < p$ cannot be made positive definite by these transformations.

Endpoint reference ranks. If $r = p$, then $DT = D$ forces $T = I_p$ and the covariance orbit is a singleton. If $r = 0$, every pairwise weight difference is zero. The fixed reference probabilities are constant in h and every invertible linear map is admissible. Its covariance orbit is all positive definite covariances in the positive definite case, or all positive semidefinite covariances of the same rank otherwise.

Operational scope. The characterization is exact for a given population covariance, or for an empirical covariance used as a deterministic matrix. A transform estimated from a fit split need not decorrelate held-out covariance. Behavioral preservation, in contrast, holds for every input when $DT = D$, independently of sampling. The maps use the fixed reference weights; the complete intervention is therefore reference-informed even when PCA or RP is fit without labels. Their condition numbers are unrestricted in the theorem and may be large. The finite, bounded scaling experiment is a sensitivity audit of a declared subset, not an estimate or certificate of the unrestricted orbit's worst case.

C AN INVARIANT DISTRIBUTION FOR THE WHITENED RP AUDIT

Positive definite fit covariance. The reference premise can be supplied by retaining the original reference targets, compensating the readout, or using an admissible fixed-readout transformation. For the last case an affine change requires both $DT = D$ and $Dc = 0$. Centering alone does not preserve an unchanged reference. Task labels used in the loss remain distinct from the reference targets used for agreement.

Use feature column vectors and the same covariance normalization in both coordinate systems. Any full whitener F satisfying $F\Sigma F^\top = I$ is invertible. Fitting the mean and covariance after $h' = Th$

gives $\mu' = T\mu$ and $\Sigma' = T\Sigma T^\top$. Direct multiplication yields

$$QQ^\top = F'TF^{-1}F^{-\top}T^\top F'^\top = F'T\Sigma T^\top F'^\top = I, \qquad z'(Th) = Qz(h).$$

For symmetric whitening this specializes to $Q = (T\Sigma T^\top)^{-1/2}T\Sigma^{1/2}$. An exact SVD-based PCA whitener also satisfies the premise; no eigenvalue floor, ridge term, or rank truncation is included. The whitener remains fixed, so these equalities hold for arbitrary evaluation inputs, including points outside the fit sample. For affine transformations $h' = Th + c$, the fitted mean also removes c, so the same identity holds. The reference outputs must separately be preserved; for the unchanged reference, $DT = D$ suffices in the linear case; an added translation also requires $Dc = 0$.

In a row-feature implementation write $Z = (H - \mathbf{1}\mu^\top)F^\top$. The corresponding relation is $Z' = ZQ^\top$ and a projection factor $P_j \in \mathbb{R}^{p\times q}$ couples as $P'_j = QP_j$, so $Z'P'_j = ZP_j$. This is the transpose of the column-vector convention, not an additional invariance assumption.

For a compressed feature $R_j z$, set $R'_j = R_j Q^\top$. Then $R'_j z' = R_j Q^\top Qz = R_j z$ for every fit, validation, and test example. Any deterministic downstream processing of these inputs, including a compressed-feature scaler, fitted student, failure rule, and validation selector, therefore gives identical outputs under this coupling. If the student itself uses randomness, include and couple the same independent random variables in both audits. An isotropic Gaussian row has the same distribution after right multiplication by an orthogonal matrix. The complete pipeline requires this invariance for the *joint* indexed matrix collection. Invariance of each matrix's marginal distribution alone is insufficient. It holds for independent Gaussian matrices at every dimension and seed, and for nested rows of an isotropic Gaussian master matrix, with the declared q-dependent normalizations. Thus the coupled matrices have the same joint distribution as the original randomization scheme, proving Proposition 2. A fixed pseudorandom generator with one particular seed is an implementation of a draw, not an exactly rotation-invariant mathematical distribution.

Orthogonally equivariant compressors and isotropic logistic probes. More generally, suppose a declared compressor family satisfies the joint property that replacing every input z by Qz replaces every compressed input u_j by $O_j u_j$, where O_j is orthogonal and fixed across fit and evaluation examples. The property may be exact, or hold under a jointly distribution-preserving coupling of its random variables. Let the student minimize

$$J(B, a) = \sum_i \ell(Bu_i + a, y_i) + \frac{\lambda}{2}\|B\|_F^2, \qquad \lambda > 0,$$

where ℓ is multiclass cross entropy, the intercept is unpenalized, and the data weights and hyperparameters are the same. Under $u'_i = Ou_i$, set $B' = BO^\top$ and $a' = a$. Every logit and the Frobenius penalty are unchanged, and this is a bijection of parameter spaces. Thus it maps global minimizers to global minimizers. If a minimizer exists and prediction uniqueness or equivariant minimizer selection is specified, all evaluation predictions agree. A fixed common-class intercept gauge, such as $\sum_k a_k = 0$, removes the irrelevant common intercept shift. For a fixed finite-sample optimizer, this is an objective-level statement; stopping tolerances and optimizer paths require separate verification.

Consequently an audit assembled from such compressors, probes, fixed reference targets, and deterministic score-based validation selection is invariant, or invariant in law under the joint randomization qualification. This sufficient condition includes the Gaussian RP coupling above with $O_j = I$. It does not license any arbitrary compressor or software implementation. An additional coordinate-wise standardizer can break the generic orthogonal-equivariance premise; the canonical experiment therefore uses the plain isotropic L2 probe.

An ideal distance-based procedure defined only by the fit distance matrix, query-to-fit distances, and fixed random variables sees identical inputs after full whitening of affine-equivalent features. Its outputs are therefore identical under the same random variables, or equivalent under any explicitly specified embedding-coordinate ambiguity. Exact Isomap with distance-based tie handling is an example under a fixed out-of-sample extension. This observation does not prove equivariance of approximate neighbor search, coordinate-dependent initialization, or a particular UMAP/Isomap implementation. For PCA after exact whitening, the entire covariance eigenspace is degenerate; a selected proper leading subspace has no unique variance ranking.

Validation and test phases. The result extends jointly to multiple phases when the whitening fit is frozen, giving one orthogonal Q for all phases, even if a student is refit using more labeled

training examples. If each phase fits a different whitener, it generally has a different rotation Q_a. Independent RP collections across phases can be coupled with $R'_{a,j} = R_{a,j}Q_a^\top$ and preserve their joint distribution. If the same RP matrix is reused across differently whitened phases, these transformed matrices need not satisfy the required equality constraint. Thus phasewise invariance does not prove joint selector-and-test invariance. Fresh test randomization independent of selection is a sufficient remedy. Any validation-selected hyperparameters or stopping criteria must be part of the same input-measurable audit and cannot use held-out test scores.

Thin whitening when the fit covariance is singular. Let the fit covariance have rank $k > 0$, and write its positive eigensystem as $\Sigma = U\Lambda U^\top$, with $U \in \mathbb{R}^{p\times k}$ orthonormal and $\Lambda \succ 0$. Define thin whitening $z(h) = \Lambda^{-1/2}U^\top(h-\mu) \in \mathbb{R}^k$. For the transformed fit covariance use $\Sigma' = U'\Lambda' U'^\top$, which has the same rank. Assume *every* evaluated $h - \mu$ is in $\mathrm{range}(U)$. Then

$$Q = \Lambda'^{-1/2}U'^\top TU\Lambda^{1/2} \quad \text{satisfies} \quad QQ^\top = I_k, \qquad z'(Th+c) = Qz(h).$$

Here the transformed mean is $T\mu+c$. Indeed, $T\,\mathrm{range}(U) = \mathrm{range}(U')$, which gives the covariance identity, and the support assumption permits writing $h - \mu = U\Lambda^{1/2}z(h)$. Gaussian RP in $\mathbb{R}^k$ then has exactly the same guarantee. The condition holds for every fit point by construction; it is an additional assumption for held-out points. If $k = 0$, all fit points coincide and whitening supplies no nontrivial coordinate distribution; the evaluation-support assumption would make all evaluation points coincide with them as well.

The support assumption cannot be dropped. Let the fit covariance be $\mathrm{diag}(1, 0)$ and let $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. The transformed fit covariance is unchanged. Yet for the centered test point $h = (0, 1)^\top$, thin whitening before transformation is zero and thin whitening after transformation is one. No orthogonal coupling can identify them. Neural layer normalization can produce an exact or nearly singular affine support, so a numerical experiment must report its rank threshold, spectrum, condition number, and fit-span residuals on evaluation inputs. An eigenvalue floor or ridge term changes the procedure and invalidates an automatic appeal to this exact theorem.

Features outside the fit support. This elementary identity quantifies the preceding restriction. Use column vectors, $P = UU^\top$, $x = h - \mu$, $e = (I - P)x$, $F = \Lambda^{-1/2}U^\top$, $F' = \Lambda'^{-1/2}U'^\top$ and $Q = F'TU\Lambda^{1/2}$. Since $QF = F'TP$,

$$F'Tx - QFx = F'Te, \qquad \|F'Tx - QFx\|_2 \le \|F'T\|_2\,\|e\|_2. \tag{7}$$

For affine input $Th + c$, the transformed fitted mean $T\mu + c$ removes the translation. The multiplier can be large, so a small raw support residual alone does not guarantee a small whitened error. Separately projecting in both coordinate systems does not generally repair this: for the transformed support projector P', $P'T \neq TP$ outside the fit support in general.

Projecting h to $\mu+Px$ changes fixed logits by $-We$ and score differences by $-De$. Their Euclidean norms are at most $\|W\|_2\|e\|_2$ and $\|D\|_2\|e\|_2$, respectively. A unique winning class is retained if its top-two score margin exceeds $2\|We\|_\infty$; without a margin condition, unchanged hard predictions do not follow. This is a coordinate and reference bound, not a stability theorem for independently refitted finite solvers. The empirical fit span may be smaller than structural support, and rounded caches can have additional directions. Distribution shift may remain within the same structural LayerNorm support; it does not automatically violate the support premise.

What finite randomization can and cannot stabilize. The invariance result concerns a distribution over compressor draws conditional on the fixed data, reference, and fitting protocol. Let p_j be a candidate's true pass probability under that conditional randomization, and let its empirical pass frequency use m_j independent draws. For threshold ρ, suppose every one of the J predeclared candidates satisfies $|p_j - \rho| \ge \gamma_j > 0$. For $p_j > \rho$, an error requires $\widehat{p}_j < \rho$, hence a lower-tail deviation of at least γ_j. For $p_j < \rho$, an error requires $\widehat{p}_j \ge \rho$, hence an upper-tail deviation of at least γ_j. One side suffices for each candidate. Hoeffding's inequality (Hoeffding, 1963) and a union bound give

$$\Pr\{\text{any empirical feasibility decision differs from } \mathbf{1}[p_j \ge \rho]\} \le \min\left(1, \sum_{j=1}^{J} e^{-2m_j\gamma_j^2}\right). \tag{8}$$

The minimum feasible dimension therefore agrees with its ideal probability-threshold counterpart on the complementary event. This includes the case of no feasible candidate. Draws may be dependent across candidates, as with nested Gaussian matrices; independence is required across repeated draws within each candidate. This is a conditional stability statement with unknown margins, not a certificate supplied by ten or twenty observed successes. It neither guarantees stable tie-broken method identities nor controls data or encoder-training uncertainty. Coordinate changes covered by the theorem leave the p_j values unchanged; they do not increase the margins from ρ.

Limits and practical prescription. Exact whitening, frozen fit preprocessing, isotropic joint randomization, and preserved reference targets are substantive assumptions. In floating-point arithmetic, ill-conditioned covariance and transformations can magnify numerical error. A measured coupling residual is therefore needed. Full whitening also removes variance ordering: it is not a justification for treating PCA's arbitrary eigenspace in whitened coordinates as a unique compressor. No UMAP implementation-level equivariance is assumed. We recommend reporting a clearly specified whitened RP audit and testing its coupling, alongside coordinate-sensitive results. This does not replace the restricted finite transformation sensitivity audit or promise improved predictive compression.

D A POST-VALIDATION ILLUSTRATION OF RANDOM COMPRESSION COST

We developed this explanatory calculation after observing the modern whitened-RP validation censoring. It is not a predeclared prediction or a new empirical test. The result combines the classical beta law for a random subspace (Frankl & Maehara, 1990) and the Gaussian angular agreement identity (Charikar, 2002); we claim no novelty for either identity or their consequence here.

Proposition 3 (Population Gaussian binary-teacher illustration)**.** *Let $h \sim \mathcal{N}(0, I_p)$ and fix a unit vector $u \in \mathbb{R}^p$. Let $Y = \operatorname{sign}(u^\top h)$, with zero teacher intercept. Independently draw $R \in \mathbb{R}^{q\times p}$ with independent standard Gaussian entries, where $0 < q < p$. Define*

$$P_R = R^\top (RR^\top)^{-1} R, \qquad S = u^\top P_R u.$$

Almost surely, R has full row rank and $0 < S < 1$. The highest possible population agreement of a measurable decoder of Rh, allowed to know R and u, is

$$A_{\mathrm{Bayes}}(R) := \sup_{g:\mathbb{R}^q\to\{-1,1\}} \mathbb{P}_h\{g(Rh) = Y \mid R\} = \frac{1}{2} + \frac{1}{\pi}\arcsin\sqrt{S}, \qquad S \sim \mathrm{Beta}\left(\frac{q}{2}, \frac{p-q}{2}\right). \tag{9}$$

For any $\tau \in (1/2, 1)$, putting $t_\tau = \sin^2(\pi(\tau - 1/2))$ gives

$$\mathbb{P}_R\{A_{\mathrm{Bayes}}(R) \geq \tau\} = 1 - I_{t_\tau}\left(\frac{q}{2}, \frac{p-q}{2}\right), \tag{10}$$

where I is the regularized incomplete beta function. For $q = 0$ agreement is $1/2$; for a full-rank $q = p$ map it is 1.

Proof. Fix a full-row-rank R. Observing Rh is equivalent to observing $P_R h$, since $P_R h = R^\top(RR^\top)^{-1}Rh$ and $Rh = RP_R h$. Decompose the teacher score into

$$X = u^\top h = M + E, \qquad M = u^\top P_R h, \qquad E = u^\top (I - P_R)h.$$

Conditional on R, joint Gaussianity and $R(I - P_R)u = 0$ imply that E is independent of the entire observation Rh. Its variance is $1 - S$, while $\mathrm{Var}(M \mid R) = S$. Consequently

$$\mathbb{P}(Y = 1 \mid Rh, R) = \Phi\left(\frac{M}{\sqrt{1-S}}\right),$$

so a Bayes decoder is $\operatorname{sign}(M)$, a measurable function of Rh. The standardized pair $(X, M/\sqrt{S})$ has correlation $\sqrt{S}$. For centered jointly Gaussian unit-variance variables with correlation ρ, the signs agree with probability $1 - \arccos(\rho)/\pi$: write them as $a^\top G, b^\top G$ for a standard isotropic planar Gaussian G and unit vectors with angle $\arccos\rho$. The direction of G is uniform, and the two disagreement wedges have total angle $2\arccos\rho$. This proves the agreement formula.

Table 9: Theory-only probability, over an independent Gaussian projection, that the optimal decoder of a zero-intercept isotropic Gaussian binary teacher attains at least 90% population agreement. All values are generated by deterministic incomplete-beta calculations.

p	q	$\mathbb{P}_R(A_{\text{Bayes}} \geq 0.9)$
2	1	0.200000
3	2	0.309017
384	256	9.255×10^{-20}
768	512	1.076×10^{-37}

Rotational invariance makes the row space of an independent Gaussian R uniform among q-dimensional subspaces. Therefore, for independent $G_1, \ldots, G_p \sim \mathcal{N}(0,1)$,

$$S \overset{d}{=} \frac{\sum_{j=1}^{q} G_j^2}{\sum_{j=1}^{p} G_j^2}.$$

The numerator and the remaining sum are independent chi-squares with q and $p-q$ degrees of freedom, respectively, yielding the stated beta law. Monotonicity of the agreement function gives Equation (10). With no observation, the balanced teacher has majority probability $1/2$; an invertible observation recovers the teacher score exactly. These establish the endpoint cases. □

What the fraction q/p does and does not imply. The beta moments give

$$\mathbb{E}_R[S] = \frac{q}{p}, \qquad \text{Var}_R(S) = \frac{2(q/p)(1-q/p)}{p+2}.$$

Thus along $p \to \infty$ with $q/p \to \gamma \in (0,1)$, $S \to \gamma$ in probability and the Bayes agreement converges in probability to $1/2 + \arcsin(\sqrt{\gamma})/\pi$. At $\gamma = 2/3$ this limit is 0.804087. At finite p, applying the agreement formula to $\mathbb{E}[S]$ does not compute $\mathbb{E}[A_{\text{Bayes}}]$, and q/p does not determine the agreement of a particular projection. For $\tau = 0.9$, the exact required signal fraction is $t_{0.9} = (5+\sqrt{5})/8$. Table 9 illustrates the projection probability from Equation (10). These are mathematical probabilities under the stated population model, not empirical success rates or confidence intervals.

Scope and a counterexample to a universal reading. This statement is about an independent, task-agnostic Gaussian projection and an optimal population decoder. A task-aware rank-one map $R = u^\top$ achieves perfect agreement. The proposition therefore supplies no lower bound for all compressors and no claim about the training performance of a finite logistic probe. Nonzero teacher intercepts change the Bayes agreement formula; multiclass teachers are also outside its scope. Whitening an empirical covariance does not make a population Gaussian, and it does not establish the proposition's distributional assumptions. The calculation uses oracle isotropic population coordinates; estimating whitening from a finite fit sample introduces additional covariance and sampling effects absent from this model. In particular, the matched aspect ratios $256/384$ and $512/768$ cannot predict the real DINOv2, ViT, or BERT results. The illustration shows how coordinate invariance can differ from task utility at low dimension. It does not claim that this mechanism explains all observed censoring.

Numerical calculation. The calculation uses no empirical data and fits no model. It compares the conditional Gaussian integral with the angular formula and checks a fixed projection's signal/residual decomposition. It also checks two beta laws with elementary closed forms and compares all reported incomplete-beta tails with separately integrated beta densities.

E BEHAVIOR-PRESERVING CONTROLS

Proposition 4 (Affine reference subspace)**.** *Let a fixed K-class reference probe have logits $s_k(h) = w_k^\top h + b_k$, with a fixed rule for ties. Define $D \in \mathbb{R}^{(K-1)\times p}$ by $D_{k,:} = (w_k - w_K)^\top$, let $r = \text{rank}(D)$, and let $U \in \mathbb{R}^{p\times r}$ have orthonormal columns spanning the row space of D. The linear compressor $z = U^\top h$ and affine decoder*

$$\widetilde{s}_k(z) = (w_k - w_K)^\top U z + b_k - b_K \quad (k < K), \qquad \widetilde{s}_K(z) = 0$$

preserve every reference decision and softmax probability for every h. In particular, $r \le K-1$.

Proof. The row-space identity $DUU^\top = D$ gives $\widetilde{s}_k(U^\top h) = s_k(h) - s_K(h)$. Subtracting the same scalar from all logits preserves their ordering and softmax probabilities, including ties. □

This is a reference-informed diagnostic with a specified decoder. It does not guarantee that a separately retrained probe with different regularization or optimization reproduces that decoder. A binary affine reference admits an exact one-dimensional diagnostic even when the tested unsupervised compressors require many coordinates. The rank bound concerns reference behavior; it does not identify the intrinsic dimension of the representation or exhaust label-relevant information.

Proposition 5 (Invertible reference-null scaling)**.** *With $\Pi = UU^\top$ as above, let $T_t = \Pi + t(I-\Pi)$ for $t > 0$. Then $DT_t = D$, $T_t^{-1} = \Pi + t^{-1}(I-\Pi)$, and*

$$\|T_t(h-h')\|^2 = \|\Pi(h-h')\|^2 + t^2\|(I-\Pi)(h-h')\|^2.$$

The unchanged reference probe therefore has identical decisions, pairwise logit margins, and softmax probabilities at every t, while Euclidean geometry can change.

Proof. Use $\Pi^2 = \Pi$, $\Pi(I-\Pi) = 0$, and $D\Pi = D$. The distance formula follows from orthogonality of the two summands. □

Invertibility preserves all recoverable information and support topology, as well as smooth manifold dimension when defined. It need not preserve covariance, neighbor ranks, or numerical effective-dimension summaries. Raw logits may change by a common scalar; preserving every raw logit instead requires the row space of the full weight matrix. The reference and its preprocessing are frozen across t; the compressors and compressed probes are fit using training data only.

Proposition 6 (A change of compressor advantage with fixed behavior)**.** *Let $U_0 \sim \mathcal{N}(0,1)$ and $N \sim \mathcal{N}(0, I_d)$ be independent, $d \ge 1$, and $h_t = (U_0, tN)$ for $t > 0$. Let the fixed binary reference have logits $(U_0, 0)$ and decision $Y = \mathbf{1}[U_0 \ge 0]$. Both logits, hence softmax probabilities, are unchanged across t. For one-dimensional population PCA followed by an optimal classifier, agreement is 1 if $t < 1$ and $1/2$ if $t > 1$. For a fixed random-projection row $(a, b^\top)$ with $a \ne 0$ and $b \ne 0$, the optimal compressed classifier is affine and has agreement*

$$A_{\mathrm{RP}}(t) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{|a|}{t\|b\|}\right).$$

Thus PCA has greater agreement than RP for $t < 1$ and lower agreement for $t > 1$, although reference behavior, information, and support dimension are unchanged.

Proof. The covariance is $\mathrm{diag}(1, t^2 I_d)$. For $t < 1$ PCA selects the reference coordinate; for $t > 1$ every leading principal direction is independent of Y. Write the RP output as $Z = aU_0 + t\|b\|V$, where V is independent standard normal. The conditional mean of U_0 given Z has the sign of aZ, which is the Bayes decision. Rotational symmetry of (U_0, V) shows that the disagreement probability between the halfplanes $U_0 \ge 0$ and $|a|U_0 + t\|b\|V \ge 0$ equals $\arctan(t\|b\|/|a|)/\pi$, yielding the formula. For finite $t > 0$, it lies strictly between $1/2$ and 1. Finally $h_t = \mathrm{diag}(1, tI_d)h_1$ is an invertible reference-preserving transformation, and every h_t has full support $\mathbb{R}^{d+1}$. □

The population PCA statement excludes $t = 1$, where its leading eigenspace is degenerate. Finite-sample PCA and fitted probes need not achieve the population optima. The construction establishes coordinate dependence of the comparison; it makes no claim that one compressor universally dominates.

Accuracy and agreement. For predictions f and reference f_0 on the same labeled examples, let a, a_0 denote accuracies and A their agreement. Then

$$|a - a_0| \le 1 - A, \qquad \left|\frac{a}{a_0} - 1\right| \le \frac{1-A}{a_0} \quad (a_0 > 0).$$

The correctness indicators can differ only on examples where the predictions differ. Equal accuracy does not imply high agreement.

F FOLLOW-UP OPERATIONAL CONFIRMATION ON PREVIOUSLY UNEVALUATED PETS EXAMPLES

Hypothesis, selection, and exposure. The H1 declaration preceded the new validation scan and the complement extraction. It screened raw-coordinate PCA and RP within the saved fixed-reference null-scaling panel, requiring a feasible endpoint and a failure by at least 0.02 at the same method and dimension, with relative condition number at most four. RP required at least eight of ten validation draws at each endpoint. The rule prioritized smaller relative condition number, larger minimum endpoint margin, then smaller dimension and scales. Each method retains its primary candidate; stronger-margin diagnostic alternatives do not replace it. Both Pets candidates use $t = 1$ and $t = 2$, so relative condition number is 2. The transform is defined from the known affine reference and preserves its score differences in exact arithmetic. The evaluation uses exactly 2,669 official Pets test IDs, disjoint from 1,000 previously evaluated test IDs and all fit/validation IDs. The official test pool contains 3,669 examples. Full annotation parsing and earlier project results were already known. These are newly extracted and evaluated project examples, not an unseen dataset or external preregistration. No earlier test examples enter H1 selection or confirmation. This does not erase historical knowledge.

Frozen numerical procedure. The original fit sample, fit-only teacher, raw-coordinate transforms and centering remain fixed. The saved PCA map and fitted student are reused without changes. Gaussian RP uses the 20 declared seeds 14000–14019, with the same projection paired across the two transformations. All new RP maps and students are prepared and fixed before complement extraction. Students use isotropically L2-regularized logistic regression with $C = 1$, tolerance 10^{-5}, at most 5,000 iterations, and no postcompression coordinate standardizer. No encoder or reference is retrained. The reused PCA student at $t = 2$ retains a source convergence warning. 0/40 new RP fits have recorded convergence warnings.

Confirmation and uncertainty. At the feasible endpoint both A and R must be at least 0.90. At the failing endpoint at least one must be below 0.90. RP additionally requires at least 16/20 paired draws to exhibit both conditions. Strong confirmation also requires both simultaneous lower bounds at the feasible endpoint to be at least 0.90 and at least one simultaneous upper bound at the failing endpoint to be below 0.90. The stricter failure-margin check replaces that upper bound by 0.88 and separately counts jointly passing/robustly failing draws. Twenty thousand bootstrap replicates use PCG64 seed 20270921. One paired example resample is shared across all endpoints. RP also resamples its 20 paired seed columns. Each relative-accuracy replicate divides by the reference accuracy on that same resampled example set. Eight scalar interval slots remain reserved: two methods, two endpoints and two quantities. Per-interval two-sided alpha is $0.05/8$, using percentile quantiles with linear interpolation. A missing fit/prediction slot prevents strong confirmation for that method. No seed is dropped. Any zero reference-accuracy denominator in the observed data or bootstrap is an explicit analysis failure. The intervals are approximate and conditional on the frozen encoder, fit data, reference and validation-selected conditions. Observed 16/20 recurrence does not certify an underlying pass probability of 0.8.

Complete outcomes. PCA strong confirmation: confirmed; Gaussian RP strong confirmation: confirmed. 20/20 fresh RP draws jointly pass at $t = 1$ and fail at $t = 2$. The separate 0.02-margin confirmation is not met for PCA and met for RP. The PCA agreement estimate is 0.868, but its simultaneous upper endpoint is 0.885. It meets the point-estimate margin of 0.88, not the stricter interval rule. There are 0/42 failed prediction slots and 0/20,000 zero-denominator bootstrap replicates. Table 1 retains all four endpoint estimates and all eight interval slots.

Table 10: Pets point-estimate draw counts and the ordinary interval confirmation rule. The stricter 0.02-margin interval rule fails for PCA.

Method	Paired draws	Point-estimate pass/fail	Point-estimate pass/$\leq$ .88	Interval confirmation
PCA	1	1	1	yes
Gaussian RP	20	20	20	yes

Table 11: H2 primary outcomes. ΔA is energy-matched orthogonal minus Gaussian agreement at $q = 2p/3$, in percentage points (approximate adjusted 98.75% crossed-bootstrap interval). Mechanism requires $\Delta A \geq 2$, a positive lower bound, and stationary fits. Unit-orthogonal utility freezes q using validation. Confirmation requires $\geq 16/20$ joint test passes plus all-fit numerical/stationarity gates. No-choice and unresolved outcomes remain visible.

Setting	ΔA [98.75% CI]	Mechanism	Utility q/p	Joint /20	Utility
CIFAR-10 / ViT	$5.05\ [3.99, 6.14]$	met	–	–	no choice
CIFAR-10 / DINOv2	$2.29\ [1.50, 3.08]$	met	–	–	no choice
Pets / DINOv2	$4.28\ [3.43, 5.17]$	met	256/384	20	confirmed
TREC / BERT	$-1.03\ [-1.84, -0.21]$	not met	–	–	no choice

Reference-preservation checks. At $t = 1$, 0 reference decisions change, the maximum probability discrepancy is 0, and the maximum pairwise-logit discrepancy is 0. The declared numerical check passes. At $t = 2$, 0 reference decisions change, the maximum probability discrepancy is 9.33×10^{-15}, and the maximum pairwise-logit discrepancy is 6.04×10^{-14}. The declared numerical check passes. The decision-change tolerance is zero, with probability tolerance 10^{-8} and pairwise-logit tolerance 10^{-7}. A violation prevents strong confirmation. Numerical preservation concerns this fixed fitted affine reference; it does not assert preservation of the encoder's original classification head.

Interpretation and reproduction. H1 asks whether a fixed compressed dimension crosses a practical operating criterion under a reference-preserving coordinate intervention. It does not test a PCA-versus-RP family winner flip, isolate geometry from regularization and finite optimization, or establish a universal necessary dimension. The original ranking-reversal failures remain separate outcomes. The saved experimental outputs include the declared protocol, complete validation scan and candidate choice, complement IDs, fit warnings, maps, students, per-example predictions and bootstrap samples. Features and fitted models allow prediction recomputation. Saved predictions and bootstrap samples support recomputation of metrics and interval quantiles without fitting or extracting images.

G Follow-up same-subspace controls and complete outcomes

Exposure and fixed scope. All fit/validation/test examples were previously exposed. This follow-up mechanism study uses fresh compressor draws only. The study reuses four previously evaluated feature settings and their fixed fit-only references. CIFAR-100 is outside H2; its earlier completed paired result remains part of the original evidence. It adds no encoder training, new benchmark examples, alternative solver, outcome-driven seed replacement, or relaxed tolerance. Validation uses seeds 51000–51009 and test uses 52000–52019. Dimension fractions are $1/2, 2/3, 7/8, 1$; the three-gauge panel (identity and the saved $\kappa = 1, 32$ generic transforms) applies only at $1/2, 2/3$. The larger fractions are identity diagnostics. Full dimension is never a compressed recommendation.

Table 12: Sample sizes and dimensions for the same-subspace controls.

Setting	p	Fit	Validation	Test
CIFAR-10 / ViT	768	2,400	600	1,000
CIFAR-10 / DINOv2	384	2,400	600	1,000
Pets / DINOv2	384	2,944	736	1,000
TREC / BERT	768	4,361	1,091	500

Maps and known penalty identities. For one Gaussian master draw, positive-diagonal QR gives $G_q = U_q R_q \in \mathbb{R}^{p \times q}$, $U_q^\top U_q = I_q$, with the $1/\sqrt{q}$ scaling included in R_q. The three maps are G_q, $\alpha_q U_q$, and U_q, where $\alpha_q = \|G_q\|_F / \sqrt{q}$. Their column spans agree, and the first two have equal Frobenius norm. For row features $Z \in \mathbb{R}^{n \times p}$, use coefficient rows $W_G, \Theta \in \mathbb{R}^{K \times q}$ and $\Theta = W_G R_q^\top$, with K classes. Scores are $Z G_q W_G^\top + \mathbf{1} b^\top = Z U_q \Theta^\top + \mathbf{1} b^\top$. The ordinary Gaus-

sian penalty is $\|\Theta R_q^{-\top}\|_F^2$, the energy-matched penalty is $\|\Theta\|_F^2/\alpha_q^2$, and the unit penalty is $\|\Theta\|_F^2$. Matching the Gaussian objective to unit coordinates instead uses $\|W_G R_q^\top\|_F^2 = \|\Theta\|_F^2$. The Hessian of the Gaussian quadratic penalty with respect to the common-coordinate coefficients has condition number $\kappa(R_q)^2$. This does not determine the condition number of the complete loss Hessian or force every native-coordinate solver to stall. These elementary change-of-coordinate identities are not a new random-projection construction. The common span fixes the available linear score functions. The penalty and finite optimization can still change the fitted function. The multinomial objective is mean cross entropy plus $\|W\|_F^2/(2Cn_{\text{fit}})$, with an unpenalized intercept. Here $C = 1$, tolerance 10^{-8}, maximum iterations 20,000, and the independently replayed full-gradient infinity norm must be at most 10^{-6} with no convergence warning. Whitening and the transported gauge maps operate on the saved rounded caches. The ideal final LayerNorm representation has an affine support constraint; a full numerical cache rank does not establish ideal positive-definite population covariance.

Primary rules and uncertainty. The identity-gauge mechanism endpoint at $q = 2p/3$ averages energy-matched minus Gaussian agreement across all 20 paired test draws. Its 20,000 crossed bootstrap resamples examples and paired draw columns using seed 53000. The four settings reserve four intervals at two-sided alpha $0.05/4$. The intervals are approximate and conditional on the frozen fit data, representation, reference, and design. A meaningful mechanism gain requires a mean of at least 0.02, a positive interval lower endpoint, and stationary fits in both arms. A failed slot prevents a complete contrast. Nonstationarity leaves optimization unresolved even when point estimates are favorable. For each arm, validation selects the smallest of $p/2, 2p/3$ with at least 8/10 joint draws across all three gauges. Each draw must reach both agreement and relative accuracy of 0.90 in every gauge and preserve coupled hard predictions. Every fit in the profile must also satisfy stationarity and numerical criteria. The primary utility arm is unit orthogonal. The frozen choice requires 16/20 test joint passes and all-fit numerical/stationarity gates. Test is not used to reselect dimension. The pointwise binomial interval describes draw recurrence and does not certify an underlying pass probability of 0.8.

Primary verdicts and complete failure accounting. The mechanism criterion is met in 3/4 settings and strict utility is confirmed in 1/4. There are 960 validation and 1,920 test logistic slots, 480 teacher-score LS diagnostics, and four full-dimensional logistic baseline fits. Of 2880 terminal logistic slots, 0 fail, 0 return without stationarity, 0 return with convergence warnings, and 0 return without numerical eligibility. These overlapping warning/failure categories are not added together. No failed slot is discarded from a primary denominator.

All dimensions and all three arms. The following identity-gauge tables retain every scheduled dimension and arm. A is reference agreement and R is label accuracy divided by fixed-reference accuracy. Means describe the returned fits only and are labeled by available/declared counts; they are not successful-only primary contrasts. E counts eligible fits among all declared slots. $F/S/N$ counts failed fits, returned nonstationary fits, and returned numerical failures. High-dimension rows are diagnostics, including $q = p$; they cannot establish compressed utility.

Table 13: Complete CIFAR-10/ViT same-subspace results.

CIFAR-10 / ViT

q	Arm	Val. A/R	Test A/R	Returned V/T	E V/T	$F/S/N$ V	$F/S/N$ T
384	Gaussian	0.656/0.676	0.664/0.684	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
384	Energy-matched	0.672/0.692	0.679/0.699	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
384	Unit orthogonal	0.678/0.698	0.684/0.704	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Gaussian	0.784/0.805	0.782/0.804	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Energy-matched	0.833/0.857	0.832/0.855	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Unit orthogonal	0.835/0.858	0.833/0.856	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
672	Gaussian	0.858/0.881	0.853/0.878	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
672	Energy-matched	0.942/0.967	0.938/0.962	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
672	Unit orthogonal	0.943/0.967	0.938/0.962	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
768	Gaussian	0.888/0.911	0.878/0.903	10/10; 20/20	1/10; 0/20	0/0/0	0/0/0
768	Energy-matched	0.978/0.997	0.965/0.986	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
768	Unit orthogonal	0.978/0.997	0.965/0.986	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0

Table 14: Complete CIFAR-10/DINOv2 same-subspace results.

CIFAR-10 / DINOv2

q	Arm	Val. A/R	Test A/R	Returned V/T	E V/T	$F/S/N$ V	$F/S/N$ T
192	Gaussian	0.725/0.740	0.717/0.733	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
192	Energy-matched	0.716/0.732	0.714/0.729	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
192	Unit orthogonal	0.730/0.746	0.725/0.740	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
256	Gaussian	0.852/0.868	0.839/0.855	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
256	Energy-matched	0.871/0.888	0.862/0.878	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
256	Unit orthogonal	0.873/0.890	0.863/0.880	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
336	Gaussian	0.917/0.938	0.907/0.924	10/10; 20/20	10/10; 17/20	0/0/0	0/0/0
336	Energy-matched	0.956/0.974	0.949/0.966	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
336	Unit orthogonal	0.956/0.974	0.949/0.966	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Gaussian	0.928/0.949	0.924/0.940	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Energy-matched	0.977/0.984	0.968/0.982	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Unit orthogonal	0.977/0.984	0.968/0.982	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0

Table 15: Complete Pets/DINOv2 same-subspace results.

Pets / DINOv2

q	Arm	Val. A/R	Test A/R	Returned V/T	E V/T	$F/S/N$ V	$F/S/N$ T
192	Gaussian	0.836/0.854	0.808/0.834	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
192	Energy-matched	0.868/0.888	0.838/0.865	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
192	Unit orthogonal	0.869/0.889	0.841/0.869	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
256	Gaussian	0.892/0.913	0.875/0.902	10/10; 20/20	4/10; 0/20	0/0/0	0/0/0
256	Energy-matched	0.933/0.952	0.918/0.945	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
256	Unit orthogonal	0.934/0.953	0.919/0.945	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
336	Gaussian	0.929/0.947	0.907/0.936	10/10; 20/20	10/10; 14/20	0/0/0	0/0/0
336	Energy-matched	0.972/0.989	0.954/0.981	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
336	Unit orthogonal	0.972/0.989	0.954/0.981	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Gaussian	0.938/0.956	0.919/0.947	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Energy-matched	0.984/0.997	0.965/0.986	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0
384	Unit orthogonal	0.984/0.997	0.965/0.986	10/10; 20/20	10/10; 20/20	0/0/0	0/0/0

Table 16: Complete TREC/BERT same-subspace results.

TREC / BERT

q	Arm	Val. A/R	Test A/R	Returned V/T	E V/T	$F/S/N$ V	$F/S/N$ T
384	Gaussian	0.627/0.721	0.694/0.768	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
384	Energy-matched	0.629/0.724	0.693/0.768	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
384	Unit orthogonal	0.629/0.724	0.694/0.769	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Gaussian	0.694/0.792	0.756/0.830	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Energy-matched	0.685/0.779	0.746/0.817	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
512	Unit orthogonal	0.691/0.787	0.749/0.822	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
672	Gaussian	0.777/0.874	0.836/0.907	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
672	Energy-matched	0.793/0.885	0.861/0.927	10/10; 20/20	0/10; 1/20	0/0/0	0/0/0
672	Unit orthogonal	0.794/0.885	0.862/0.928	10/10; 20/20	0/10; 1/20	0/0/0	0/0/0
768	Gaussian	0.814/0.906	0.867/0.935	10/10; 20/20	0/10; 0/20	0/0/0	0/0/0
768	Energy-matched	0.888/0.969	0.920/0.987	10/10; 20/20	0/10; 20/20	0/0/0	0/0/0
768	Unit orthogonal	0.888/0.969	0.920/0.987	10/10; 20/20	0/10; 20/20	0/0/0	0/0/0

Recorded map scales and conditioning. For each fixed dimension the ranges below use all 30 declared master draws. Gaussian and energy-matched maps have equal Frobenius norm by construction, while only the Gaussian map retains the anisotropic singular values. The spectral condition number is a map diagnostic, not a fitted-student success criterion.

Table 17: Map scales and condition numbers across the declared draws.

Setting	q	α_q range	Gaussian condition range
CIFAR-10 / ViT	384	[1.410, 1.417]	[5.52, 6.26]
CIFAR-10 / ViT	512	[1.222, 1.227]	[9.32, 9.97]
CIFAR-10 / ViT	672	[1.067, 1.070]	[25.94, 31.61]
CIFAR-10 / ViT	768	[0.998, 1.001]	[786.47, 24890.27]
CIFAR-10 / DINOv2	192	[1.407, 1.424]	[5.36, 5.99]
CIFAR-10 / DINOv2	256	[1.220, 1.231]	[8.80, 10.39]
CIFAR-10 / DINOv2	336	[1.065, 1.073]	[24.42, 32.62]
CIFAR-10 / DINOv2	384	[0.996, 1.005]	[507.25, 35896.61]
Pets / DINOv2	192	[1.407, 1.424]	[5.36, 5.99]
Pets / DINOv2	256	[1.220, 1.231]	[8.80, 10.39]
Pets / DINOv2	336	[1.065, 1.073]	[24.42, 32.62]
Pets / DINOv2	384	[0.996, 1.005]	[507.25, 35896.61]
TREC / BERT	384	[1.410, 1.417]	[5.52, 6.26]
TREC / BERT	512	[1.222, 1.227]	[9.32, 9.97]
TREC / BERT	672	[1.067, 1.070]	[25.94, 31.61]
TREC / BERT	768	[0.998, 1.001]	[786.47, 24890.27]

Complete three-gauge validation profiles. The numerical and stationarity columns are strict all-fit gates. Joint successes alone do not qualify a profile. The selected dimension for each arm remains frozen even if later diagnostics favor another dimension.

Table 18: Three-gauge validation pass counts and all-fit criteria.

Setting	Arm	q	Joint /10	Numerical	Stationary	Qualifies	Selected q
CIFAR-10 / ViT	Gaussian	384	0	yes	yes	no	–
CIFAR-10 / ViT	Gaussian	512	0	yes	yes	no	–
CIFAR-10 / ViT	Energy-matched	384	0	yes	yes	no	–
CIFAR-10 / ViT	Energy-matched	512	0	yes	yes	no	–
CIFAR-10 / ViT	Unit orthogonal	384	0	yes	yes	no	–
CIFAR-10 / ViT	Unit orthogonal	512	0	yes	yes	no	–
CIFAR-10 / DINOv2	Gaussian	192	0	yes	yes	no	–
CIFAR-10 / DINOv2	Gaussian	256	0	yes	yes	no	–
CIFAR-10 / DINOv2	Energy-matched	192	0	yes	yes	no	–
CIFAR-10 / DINOv2	Energy-matched	256	0	yes	yes	no	–
CIFAR-10 / DINOv2	Unit orthogonal	192	0	yes	yes	no	–
CIFAR-10 / DINOv2	Unit orthogonal	256	0	yes	yes	no	–
Pets / DINOv2	Gaussian	192	0	yes	yes	no	–
Pets / DINOv2	Gaussian	256	4	yes	yes	no	–
Pets / DINOv2	Energy-matched	192	0	yes	yes	no	256
Pets / DINOv2	Energy-matched	256	10	yes	yes	yes	256
Pets / DINOv2	Unit orthogonal	192	0	yes	yes	no	256
Pets / DINOv2	Unit orthogonal	256	10	yes	yes	yes	256
TREC / BERT	Gaussian	384	0	yes	yes	no	–
TREC / BERT	Gaussian	512	0	yes	yes	no	–
TREC / BERT	Energy-matched	384	0	yes	yes	no	–
TREC / BERT	Energy-matched	512	0	yes	yes	no	–
TREC / BERT	Unit orthogonal	384	0	yes	yes	no	–
TREC / BERT	Unit orthogonal	512	0	yes	yes	no	–

Full-dimensional logistic baselines. These fit-only unit-coordinate baselines have the same $C = 1$ objective in the full whitened cache dimension. The same model is evaluated on validation and test; the $q = p$ rotated arms are separate fixed diagnostics. Failure or nonstationarity is retained.

Table 19: Direct full-dimensional logistic baselines.

Setting	Status	Val. acc./A/R	Test acc./A/R	$\|\nabla L\|_\infty$	Stationary	Warnings
CIFAR-10 / ViT	ok	0.960/0.978/0.997	0.956/0.965/0.986	6.62×10^{-9}	yes	0
CIFAR-10 / DINOv2	ok	0.948/0.977/0.984	0.958/0.968/0.982	7.52×10^{-9}	yes	0
Pets / DINOv2	ok	0.963/0.984/0.997	0.936/0.965/0.986	6.70×10^{-9}	yes	0
TREC / BERT	ok	0.820/0.888/0.969	0.892/0.920/0.987	8.05×10^{-9}	yes	0

Fit-only teacher-score recovery diagnostic. Least squares targets the fixed affine teacher's class-score differences on the fit sample using the same random span. It is reference-aware, trained without evaluation labels, and tests recovery of one fitted score function. It is neither a Bayes bound nor a classification upper bound. Full-dimension score recovery follows from linear representability and is not novel. The fit problem is independently replayed from the working arrays; score transport

into Gaussian and energy-matched coordinates is checked without a new classifier fit. Evaluation outcomes do not choose the target, dimension, or solver. Each row combines the fixed validation/test seed families. F is failed LS diagnostics out of 30; means over available diagnostics retain these failure denominators. Fit RMSE is shown as its range over returned diagnostics, and the normal-equation column gives the maximum recorded relative residual.

Table 20: Fixed-teacher score recovery by fit-only least squares.

Setting	q	Val. A/R	Test A/R	Fit RMSE range	Normal residual	F/30
CIFAR-10 / ViT	384	0.726/0.749 (10/10)	0.735/0.756 (20/20)	$[4.23 \times 10^{0}, 4.42 \times 10^{0}]$	7.37×10^{-16}	0
CIFAR-10 / ViT	512	0.855/0.877 (10/10)	0.854/0.877 (20/20)	$[3.36 \times 10^{0}, 3.63 \times 10^{0}]$	6.11×10^{-16}	0
CIFAR-10 / ViT	672	0.947/0.966 (10/10)	0.950/0.971 (20/20)	$[1.91 \times 10^{0}, 2.40 \times 10^{0}]$	6.86×10^{-16}	0
CIFAR-10 / ViT	768	1.000/1.000 (10/10)	1.000/1.000 (20/20)	$[1.01 \times 10^{-14}, 1.06 \times 10^{-14}]$	7.36×10^{-16}	0
CIFAR-10 / DINOv2	192	0.762/0.780 (10/10)	0.754/0.770 (20/20)	$[4.13 \times 10^{0}, 4.48 \times 10^{0}]$	5.63×10^{-16}	0
CIFAR-10 / DINOv2	256	0.878/0.899 (10/10)	0.873/0.890 (20/20)	$[3.32 \times 10^{0}, 3.72 \times 10^{0}]$	6.33×10^{-16}	0
CIFAR-10 / DINOv2	336	0.965/0.980 (10/10)	0.959/0.974 (20/20)	$[1.97 \times 10^{0}, 2.41 \times 10^{0}]$	6.82×10^{-16}	0
CIFAR-10 / DINOv2	384	1.000/1.000 (10/10)	1.000/1.000 (20/20)	$[1.03 \times 10^{-14}, 1.10 \times 10^{-14}]$	7.26×10^{-16}	0
Pets / DINOv2	192	0.773/0.794 (10/10)	0.747/0.773 (20/20)	$[3.41 \times 10^{0}, 3.65 \times 10^{0}]$	5.60×10^{-16}	0
Pets / DINOv2	256	0.889/0.912 (10/10)	0.879/0.902 (20/20)	$[2.73 \times 10^{0}, 2.99 \times 10^{0}]$	6.13×10^{-16}	0
Pets / DINOv2	336	0.968/0.984 (10/10)	0.965/0.982 (20/20)	$[1.64 \times 10^{0}, 1.93 \times 10^{0}]$	6.80×10^{-16}	0
Pets / DINOv2	384	1.000/1.000 (10/10)	1.000/1.000 (20/20)	$[8.88 \times 10^{-15}, 9.16 \times 10^{-15}]$	7.26×10^{-16}	0
TREC / BERT	384	0.603/0.676 (10/10)	0.634/0.688 (20/20)	$[8.37 \times 10^{0}, 9.11 \times 10^{0}]$	7.61×10^{-16}	0
TREC / BERT	512	0.703/0.785 (10/10)	0.752/0.812 (20/20)	$[6.71 \times 10^{0}, 7.44 \times 10^{0}]$	6.37×10^{-16}	0
TREC / BERT	672	0.850/0.921 (10/10)	0.886/0.939 (20/20)	$[3.99 \times 10^{0}, 4.62 \times 10^{0}]$	7.12×10^{-16}	0
TREC / BERT	768	1.000/1.000 (10/10)	1.000/1.000 (20/20)	$[2.19 \times 10^{-14}, 2.26 \times 10^{-14}]$	7.32×10^{-16}	0

All coupled numerical outcomes. The table combines validation and test within each setting and arm at $q = p/2, 2p/3$. Changed instances are compared against the identity prediction, so an example may contribute more than once across dimensions, draws, and gauges. Discrepant vectors and failed comparisons retain their own denominators. The minimum margin is the identity top-two score difference among changed examples; undefined means no changed example was recorded.

Table 21: Coupled numerical differences and retained failure counts.

Setting	Arm	Changed / instances	Discrepant / vectors	Failed	Max. score difference	Min. margin
CIFAR-10 / ViT	Gaussian	0/104000	0/120	0	1.19×10^{-4}	undefined
CIFAR-10 / ViT	Energy-matched	0/104000	0/120	0	1.06×10^{-4}	undefined
CIFAR-10 / ViT	Unit orthogonal	0/104000	0/120	0	6.72×10^{-7}	undefined
CIFAR-10 / DINOv2	Gaussian	0/104000	0/120	0	1.03×10^{-4}	undefined
CIFAR-10 / DINOv2	Energy-matched	0/104000	0/120	0	1.48×10^{-4}	undefined
CIFAR-10 / DINOv2	Unit orthogonal	0/104000	0/120	0	4.51×10^{-5}	undefined
Pets / DINOv2	Gaussian	0/109440	0/120	0	1.68×10^{-4}	undefined
Pets / DINOv2	Energy-matched	0/109440	0/120	0	8.66×10^{-5}	undefined
Pets / DINOv2	Unit orthogonal	0/109440	0/120	0	2.95×10^{-7}	undefined
TREC / BERT	Gaussian	0/83640	0/120	0	1.78×10^{-3}	undefined
TREC / BERT	Energy-matched	0/83640	0/120	0	1.33×10^{-3}	undefined
TREC / BERT	Unit orthogonal	0/83640	0/120	0	9.61×10^{-4}	undefined

Saved-output analyses and limitations. The saved outputs cover 2,880 logistic fits, 480 least-squares comparisons and the frozen choices. Working arrays and coefficients allow recomputation of predictions, scores, gradients, map/QR identities, fixed-reference outputs and least-squares transport. These checks do not retrain encoders, repeat whitening SVDs or optimize classifiers. Saved predictions also support recomputation of hard-prediction metrics, choices and the primary crossed bootstrap. Prediction-only analyses do not repeat the score, gradient or map checks. Numerical agreement alone does not establish scientific utility. Negative and unresolved results remain part of H2. Shared-span differences support a metric/penalty explanation only to the extent optimization and numerical checks succeed. They do not isolate a universal cause of earlier failure or erase the earlier prospective negative results.

H Complete fixed-reference reversal audit

ViT and DINOv2 use CIFAR-10; BERT uses SST-2. The endpoint table reports $q = 4$ for every row.

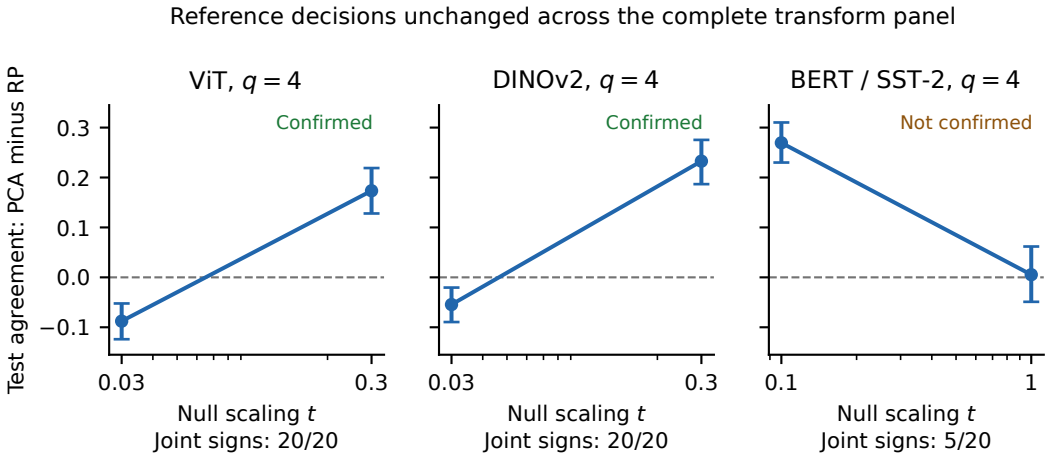


Figure 2: The same fixed reference supports opposite fixed-dimension agreement rankings in two vision encoders on CIFAR-10. At each model's frozen $q = 4$ and validation-selected scales, points show held-out PCA-minus-mean-RP agreement; bars are crossed example/draw bootstrap intervals adjusted over 12 primary/secondary endpoints. All reference decisions remain unchanged. ViT and DINOv2 confirm both opposite signs on all 20 fresh draws; BERT / SST-2 does not confirm. The vision pair has relative condition number 10 and maximum endpoint condition number 33.3. Both winning PCA endpoints remain below the operating threshold.

The declared scale grid is $.03, .1, .3, 1, 3, 10, 30$, with doubling dimensions below the ambient dimension. It extends the earlier $.1, 1, 10$ control approximately logarithmically and was fixed before this study fitted candidates. The first sweep operates in the fixed fit-reference scaler coordinates. It differs from the later raw/axis-standardized/whitened comparison in both input convention and probe policy. Its students retain the original postcompression scaler; the canonicalized experiment uses plain isotropic logistic probes.

One fit-only teacher, PCA fit, and training set remain fixed across validation and test. RP matrices are paired across scales, with ten validation seeds 200–209 and twenty new test seeds 2000–2019. A validation candidate must have opposite mean agreement differences of magnitude at least .02 and at least 8/10 correctly signed draws at each endpoint. The primary rule maximizes the smaller endpoint margin, then prefers smaller q, narrower scale ratio, and lexicographic scales. One pair per model, all dimension recommendations and the complete test union freeze globally before test access.

The minimum-distortion extension was specified after the primary validation run had begun and before any new test evaluation. It preserves the original primary choices. It selects the least relative-condition-number qualifying pair, followed by endpoint condition number, margin, q, and scales. Combined claims use two-sided crossed bootstrap intervals adjusted over twelve endpoint slots, including duplicate endpoints, and a six-comparison adjustment for the conditional seed-sign test. Repeating DINOv2's primary pair in this extension is not another replication. The bootstrap draws paired example indices and paired seed indices independently, using the same resamples at both endpoints. It is conditional on the fixed model, fit sample, references, selected conditions and historically exposed benchmark; it is not a finite-sample confidence certificate.

Table 22: Fixed endpoint agreement at $q = 4$. Primary and secondary pairs retain their original scales, references and test draws. BERT uses SST-2.

Model/task	Panel	t	PCA A	RP A	PCA pass	RP passes
ViT	Primary	0.03	0.815	0.903	no	11/20
ViT	Primary	0.3	0.815	0.642	no	0/20
DINOv2	Primary	0.03	0.820	0.874	no	3/20
DINOv2	Primary	0.3	0.828	0.595	no	0/20
BERT / SST-2	Primary	0.1	0.966	0.696	yes	0/20
BERT / SST-2	Primary	1	0.588	0.583	no	0/20
ViT	Secondary	0.1	0.820	0.852	no	1/20
ViT	Secondary	0.3	0.815	0.642	no	0/20
DINOv2	Secondary	0.03	0.820	0.874	no	3/20
DINOv2	Secondary	0.3	0.828	0.595	no	0/20
BERT / SST-2	Secondary	0.1	0.966	0.696	yes	0/20
BERT / SST-2	Secondary	0.3	0.641	0.607	no	0/20

What the reversal establishes. Across the vision endpoints, RP agreement falls substantially and PCA agreement changes little. These values establish a ranking reversal for the tested method. They do not prove that PCA eigenspaces remain unchanged or that null-variance dilution is the sole cause. The Gaussian illustration has different distributional assumptions and does not identify the mechanism in these representations. Differences are computed before rounding: DINOv2's first PCA-minus-RP difference is -0.05450.

I Earlier full-coordinate and ranking summaries

Figure 2 shows a held-out ranking reversal in ViT and DINOv2 under the frozen primary rule. At $q = 4$, changing t from 0.03 to 0.3 changes PCA-minus-RP agreement from -8.8 to $+17.3$ points for ViT and from -5.5 to $+23.3$ points for DINOv2. All 20 fresh projection draws show both opposite signs for each encoder. The multiplicity-adjusted crossed intervals exclude zero at both endpoints. The BERT validation candidate does not confirm: its second test difference is slightly positive, with an interval spanning zero.

Every evaluated reference decision is unchanged, with maximum probability discrepancy 2.03×10^{-13}. The confirmed pair has relative condition number $\kappa(T_{0.3}T_{0.03}^{-1}) = 10$; its endpoint condition numbers relative to the original coordinates are 33.3 and 3.33. These quantities must not be conflated. A separately frozen ViT pair with relative condition number 3 has the expected mean signs and 16/20 joint successes, but its first adjusted interval includes zero. It therefore fails the confirmation criterion. We do not claim a confirmed lower-distortion reversal or a global minimum over untested transformations.

The comparison concerns the ordering of agreement at a fixed q, not a reversal of the selected dimension or of dual-threshold feasibility. At these vision endpoints PCA agreement ranges from 0.815 to 0.828, below the .90 criterion. Thus a method can win a compression comparison while both procedures fail the stated operating requirement. Complete validation curves, failed candidates, task accuracy, and finite-grid dimension envelopes appear in Appendix H. Across 3 historical model/task pairs and 19 declared gauges each, the explicitly coupled whitened coordinate audit passes its full numerical/prediction/selection endpoint. There are 0 changed predictions in 6,300 gauge/map/phase comparisons, with maximum relative compressed-coordinate error 1.10×10^{-8}. This tests the implemented coupling and complete selector on the declared panel. The theorem supplies the separate idealized distributional statement.

Table 23 and Figure 3 show the accompanying compression cost. 3 of 3 historical pairs have no qualifying whitened compressed point at any declared gauge. Universal censoring is absence of an established compressed operating point, not useful dimension stability. The uncompressed whitened student is reported separately, so a failed compressed grid is not repaired by silently inserting the ambient dimension. The finite range depends on the declared grid, reference and decoder.

At the fixed $q = 8$ control, reusing the same nominal Gaussian seed produces changed predictions in 1,620/1,620 comparisons. These draws are not the transported coupling. Pathwise variation here

Table 23: Coordinate sensitivity and compression cost. Raw/standardized cells give the range of validation-selected PCA/RP dimensions over the entire declared gauge panel. Whitened cells give the coupled-RP range. ∅: no qualifying grid point; †: at least one gauge censored. Test draws use the identity-gauge frozen whitened choice; n/a means no such choice. Full A gives validation/test agreement of an uncompressed whitened student with the same reference, not reference accuracy.

Model	Raw PCA/RP	Std. PCA/RP	White RP	Test draws	Full A (val/test)
ViT/CIFAR-10	8–16/8–256†	8–32/8–512	∅	n/a	0.975/0.965
DINOv2/CIFAR-10	8–64/8–256†	8–64/8–256†	∅	n/a	0.977/0.968
BERT/SST-2	1–512†/8–256†	1–512†/16–256†	∅	n/a	0.917/0.914
DINOv2/Pets	8–256†/16–256†	8–256/16–256†	∅	n/a	0.984/0.963
BERT/TREC	8–512†/32–512†	8–512†/64–512†	∅	n/a	0.889/0.924

is compatible with invariance in law. Prewhitening fit covariances reach condition numbers up to 6.18×10^{17} before normalization, so the small floating-point singular directions and their amplification remain part of the numerical audit. DINOv2/Pets does not confirm its frozen $q = 64$ pair: agreement differences are +13.1 and +0.5 points, with 8/20 draws jointly supporting the two prescribed signs. BERT/TREC has no qualifying validation pair and contributes no confirmation. The prospective coupled canonical endpoint has 0 failing gauge/phase numerical checks, 0 failing coordinate comparisons and 253 changed prediction instances on validation. DINOv2/Pets has 3 changed instances across 2,200 gauge/map/phase comparison vectors. BERT/TREC has 250 changed instances across 2,400 gauge/map/phase comparison vectors. The held-out whitened RP check covers only the frozen $q = 8$ control, not the complete validation dimension grid. Every declared whitened recommendation is grid-censored. Unchanged censoring does not establish an operating point or useful compression stability. BERT/TREC's full-dimensional whitened probe already fails the validation dual predicate ($A = 0.889$; raw full-probe $A = 0.937$). Its compressed-grid censoring cannot be attributed solely to dimension reduction. The declared coordinate convention and isotropic L2 probe already change the fitted full-dimensional behavior. DINOv2/Pets has convergence warnings in 2/2 PCA source fits, 19/20 validation RP fits and 36/40 fresh test RP fits; PCA models are reused at test. The result concerns the frozen finite-fit procedure. The cause of non-confirmation is not uniquely identified. The reversal analysis reserves all four endpoint slots for 98.75% crossed intervals and both model slots for the joint-seed test, even when a validation candidate is absent. No alternate test pair is substituted. Table 23 reports whether any compression survives the correction; full prospective intervals, negative outcomes and coordinate controls appear in Appendix K.

Paired CIFAR-100 labels after full whitening. Fine and coarse labels share the exact DINOv2 features, fit split, full whitening, and Gaussian maps. Neither task has a passing compressed dimension on the declared grid through $q = 256$: zero of 10 validation draws pass the dual criterion at every dimension for both tasks. This is joint censoring, not evidence of equal intrinsic task dimension. Full-dimensional whitened probes remain feasible at $p = 384$, with held-out reference agreement 0.913 (fine) and 0.934 (coarse). The result separates full-dimensional accessibility from successful random compression. The earlier fine/coarse dimension contrast used a different compressor family, probe convention, and selection rule. Its absence here cannot be attributed to whitening alone. Coordinate correction does not remove compressor randomness. At the identity gauge, per-draw whitened dimension recommendations are ViT: ∅ (10/10 censored); DINOv2: ∅ (10/10 censored); BERT: ∅ (10/10 censored); Pets: 256† (5/10 censored); TREC: ∅ (10/10 censored). These minima use the same fixed teacher, training sample and dimension grid. They measure variation over compressor draws, not encoder runs. Fresh-draw pass counts and pointwise intervals in the appendix evaluate frozen recommendations without test reselection; their empirical frequency threshold is not a certified lower probability bound.

J Canonicalization and finite-panel dimension audit

Let $F_{g,q,s}$ be the dual-threshold pass indicator for gauge g, dimension q and compressor draw s; failed fits have indicator zero. Per-gauge recommendations minimize q over the declared grid at empirical pass frequency at least ρ. The maximum of these minima, $q_{\max}$, need not be feasible

at one common dimension. We separately compute $q_{\text{uniform}} = \min\{q : \forall g,\ \widehat{\pi}_g(q) \geq \rho\}$ and $q_{\text{joint}} = \min\{q : S^{-1} \sum_s \prod_g F_{g,q,s} \geq \rho\}$. No monotonicity is assumed. For deterministic PCA the marginal and joint rules coincide. A missing value denotes no qualifying point on the tested grid; if any per-gauge minimum is missing, the full-panel maximum is reported as censored rather than the largest observed finite minimum.

The historical null scales are 0.03, 0.1, 0.3, 1, 3, 10, 30; the general-linear panel has condition numbers 1, 2, 4, 8, 16, 32, each with 2 independently declared orientations. The validation and fresh-test compressor counts are 10 and 20. Only the fit sample determines all centering, scales, full-SVD whitening, maps and teacher parameters. No postcompression scaler is used. Plain L2 students and the same fit covariance persist through test; fresh RP draws change the compressor, not the training sample. Full canonical uncompressed students are accessibility/cost baselines, not candidate compressed grid points.

The following tables retain every declared gauge, all method-specific minima, marginal/joint envelopes, censoring and the unchanged identity-gauge recommendations transferred to fresh test draws. The accompanying machine-readable frequency tables contain every evaluated q, gauge and method, including failed fits. Binomial intervals are pointwise Clopper–Pearson intervals over fresh compressor draws conditional on the test sample and frozen choice. They do not account for unknown training-sample or encoder variation or establish a population pass-probability certificate. Numerical agreement alone does not establish the stricter scientific canonicalization endpoint. Same nominal seeds and explicitly rotated maps are reported separately.

J.1 ViT/CIFAR-10

The fit split has 2400 examples, validation has 600, ambient dimension is 768, and the complete panel contains 19 gauges.

Table 24: ViT/CIFAR-10: dimension ranges and censored gauge choices.

Convention/method	$q_{\min}$	$q_{\max}$	q_{uniform}	q_{joint}	Censored
Raw/PCA	8	16	16	16	0/19
Raw/RP	8	∅	∅	∅	2/19
Standardized/PCA	8	32	32	32	0/19
Standardized/RP	8	512	512	512	0/19
Whitened/RP	∅	∅	∅	∅	19/19

Table 25: ViT/CIFAR-10: minimum qualifying dimensions for each gauge.

Gauge	Raw PCA	Raw RP	Std. PCA	Std. RP	White RP
Null $t = 0.03$	8	8	8	8	∅
Null $t = 0.1$	8	8	8	8	∅
Null $t = 0.3$	8	16	8	32	∅
Null $t = 1$	8	128	8	128	∅
Null $t = 3$	16	256	16	256	∅
Null $t = 10$	16	∅	32	512	∅
Null $t = 30$	16	∅	32	512	∅
GL $\kappa = 1$, seed 4101	8	128	8	128	∅
GL $\kappa = 2$, seed 4101	8	128	8	128	∅
GL $\kappa = 4$, seed 4101	8	128	8	128	∅
GL $\kappa = 8$, seed 4101	8	128	8	128	∅
GL $\kappa = 16$, seed 4101	8	128	8	128	∅
GL $\kappa = 32$, seed 4101	8	128	8	128	∅
GL $\kappa = 1$, seed 4102	8	128	8	128	∅
GL $\kappa = 2$, seed 4102	8	128	8	128	∅
GL $\kappa = 4$, seed 4102	8	128	8	128	∅
GL $\kappa = 8$, seed 4102	8	128	8	128	∅
GL $\kappa = 16$, seed 4102	8	128	8	128	∅
GL $\kappa = 32$, seed 4102	8	128	8	128	∅

Table 26: ViT/CIFAR-10: frozen identity choices evaluated with fresh test draws.

Convention/method	Frozen identity q	Test passes [95% CI]	Test A	Test accuracy	Full A
Raw/PCA	8	1/1	0.941	0.943	0.993
Raw/RP	128	20/20 [0.83,1.00]	0.950	0.943	0.993
Standardized/PCA	8	1/1	0.961	0.955	1.000
Standardized/RP	128	20/20 [0.83,1.00]	0.943	0.936	1.000
Whitened/RP	∅	n/a	n/a	n/a	0.965

Table 27: ViT/CIFAR-10: direct full-dimensional baselines.

Full-dimensional probe	Validation A	Validation dual	Test A	Test dual
Raw	0.993	pass	0.993	pass
Standardized	1.000	pass	1.000	pass
Whitened	0.975	pass	0.965	pass

These deterministic full-dimensional students retain the same fitted reference, fit sample and fixed isotropic L2 setting. Both feasibility thresholds are applied. Passing or failing at the full dimension does not imply monotonicity along the compressed grid.

Across both evaluation phases, maximum reference probability error is 1.07×10^{-13}, maximum reference score-difference error is 7.57×10^{-13}, and the reference decision-change total is 0. Maximum relative compressed-coordinate discrepancy is 5.25×10^{-9}. Raw orthogonal controls have 14 changed predictions across 388 method/map/phase comparisons; this total includes separately fitted PCA and uncompressed students.

J.2 DINOv2/CIFAR-10

The fit split has 2400 examples, validation has 600, ambient dimension is 384, and the complete panel contains 19 gauges.

Table 28: DINOv2/CIFAR-10: dimension ranges and censored gauge choices.

Convention/method	$q_{\min}$	$q_{\max}$	q_{uniform}	q_{joint}	Censored
Raw/PCA	8	64	64	64	0/19
Raw/RP	8	∅	∅	∅	2/19
Standardized/PCA	8	64	64	64	0/19
Standardized/RP	8	∅	∅	∅	2/19
Whitened/RP	∅	∅	∅	∅	19/19

Table 29: DINOv2/CIFAR-10: minimum qualifying dimensions for each gauge.

Gauge	Raw PCA	Raw RP	Std. PCA	Std. RP	White RP
Null $t = 0.03$	8	8	8	8	∅
Null $t = 0.1$	8	8	8	8	∅
Null $t = 0.3$	8	16	8	16	∅
Null $t = 1$	8	128	8	128	∅
Null $t = 3$	32	256	32	256	∅
Null $t = 10$	64	∅	64	∅	∅
Null $t = 30$	64	∅	64	∅	∅
GL $\kappa = 1$, seed 4101	8	128	8	128	∅
GL $\kappa = 2$, seed 4101	8	128	8	128	∅
GL $\kappa = 4$, seed 4101	8	128	8	128	∅
GL $\kappa = 8$, seed 4101	8	128	16	128	∅
GL $\kappa = 16$, seed 4101	16	128	16	128	∅
GL $\kappa = 32$, seed 4101	16	128	16	128	∅
GL $\kappa = 1$, seed 4102	8	128	8	128	∅
GL $\kappa = 2$, seed 4102	8	128	8	128	∅
GL $\kappa = 4$, seed 4102	8	128	8	128	∅
GL $\kappa = 8$, seed 4102	8	128	16	128	∅
GL $\kappa = 16$, seed 4102	16	128	16	128	∅
GL $\kappa = 32$, seed 4102	16	128	32	128	∅

Table 30: DINOv2/CIFAR-10: frozen identity choices evaluated with fresh test draws.

Convention/method	Frozen identity q	Test passes [95% CI]	Test A	Test accuracy	Full A
Raw/PCA	8	1/1	0.938	0.930	0.996
Raw/RP	128	20/20 [0.83,1.00]	0.941	0.935	0.996
Standardized/PCA	8	1/1	0.941	0.928	1.000
Standardized/RP	128	20/20 [0.83,1.00]	0.943	0.939	1.000
Whitened/RP	∅	n/a	n/a	n/a	0.968

Table 31: DINOv2/CIFAR-10: direct full-dimensional baselines.

Full-dimensional probe	Validation A	Validation dual	Test A	Test dual
Raw	0.992	pass	0.996	pass
Standardized	1.000	pass	1.000	pass
Whitened	0.977	pass	0.968	pass

These deterministic full-dimensional students retain the same fitted reference, fit sample and fixed isotropic L2 setting. Both feasibility thresholds are applied. Passing or failing at the full dimension does not imply monotonicity along the compressed grid.

Across both evaluation phases, maximum reference probability error is 9.68×10^{-14}, maximum reference score-difference error is 7.71×10^{-13}, and the reference decision-change total is 0. Maximum relative compressed-coordinate discrepancy is 1.10×10^{-8}. Raw orthogonal controls have 31 changed predictions across 370 method/map/phase comparisons; this total includes separately fitted PCA and uncompressed students.

J.3 BERT/SST-2

The fit split has 2400 examples, validation has 600, ambient dimension is 768, and the complete panel contains 19 gauges.

Table 32: BERT/SST-2: dimension ranges and censored gauge choices.

Convention/method	$q_{\min}$	$q_{\max}$	q_{uniform}	q_{joint}	Censored
Raw/PCA	1	∅	∅	∅	14/19
Raw/RP	8	∅	∅	∅	16/19
Standardized/PCA	1	∅	∅	∅	11/19
Standardized/RP	16	∅	∅	∅	16/19
Whitened/RP	∅	∅	∅	∅	19/19

Table 33: BERT/SST-2: minimum qualifying dimensions for each gauge.

Gauge	Raw PCA	Raw RP	Std. PCA	Std. RP	White RP
Null $t = 0.03$	1	8	1	16	∅
Null $t = 0.1$	4	64	4	64	∅
Null $t = 0.3$	64	256	64	256	∅
Null $t = 1$	∅	∅	512	∅	∅
Null $t = 3$	∅	∅	∅	∅	∅
Null $t = 10$	∅	∅	∅	∅	∅
Null $t = 30$	∅	∅	∅	∅	∅
GL $\kappa = 1$, seed 4101	∅	∅	∅	∅	∅
GL $\kappa = 2$, seed 4101	∅	∅	512	∅	∅
GL $\kappa = 4$, seed 4101	∅	∅	512	∅	∅
GL $\kappa = 8$, seed 4101	512	∅	∅	∅	∅
GL $\kappa = 16$, seed 4101	512	∅	512	∅	∅
GL $\kappa = 32$, seed 4101	∅	∅	∅	∅	∅
GL $\kappa = 1$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 2$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 4$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 8$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 16$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 32$, seed 4102	∅	∅	512	∅	∅

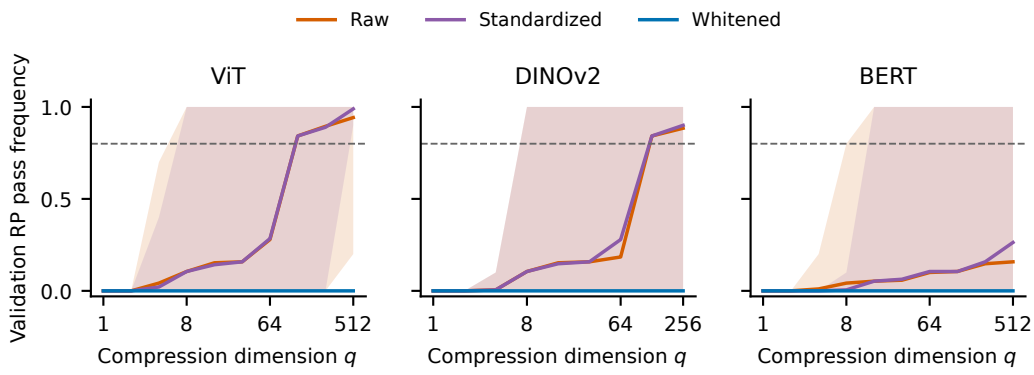


Figure 3: Complete validation RP profiles under three coordinate conventions. Lines average the empirical pass frequency over all declared gauges; shaded envelopes give its minimum and maximum across gauges, not confidence intervals. The dashed line is the frozen seed-frequency selection threshold. Whitened maps use explicit transport; all student probes use plain isotropic L2 regularization.

Table 34: BERT/SST-2: frozen identity choices evaluated with fresh test draws.

Convention/method	Frozen identity q	Test passes [95% CI]	Test A	Test accuracy	Full A
Raw/PCA	∅	n/a	n/a	n/a	0.921
Raw/RP	∅	n/a	n/a	n/a	0.921
Standardized/PCA	512	1/1	0.912	0.800	1.000
Standardized/RP	∅	n/a	n/a	n/a	1.000
Whitened/RP	∅	n/a	n/a	n/a	0.914

Table 35: BERT/SST-2: direct full-dimensional baselines.

Full-dimensional probe	Validation A	Validation dual	Test A	Test dual
Raw	0.922	pass	0.921	pass
Standardized	1.000	pass	1.000	pass
Whitened	0.917	pass	0.914	pass

These deterministic full-dimensional students retain the same fitted reference, fit sample and fixed isotropic L2 setting. Both feasibility thresholds are applied. Passing or failing at the full dimension does not imply monotonicity along the compressed grid.

Across both evaluation phases, maximum reference probability error is 9.39×10^{-13}, maximum reference score-difference error is 4.71×10^{-12}, and the reference decision-change total is 0. Maximum relative compressed-coordinate discrepancy is 2.61×10^{-9}. Raw orthogonal controls have 0 changed predictions across 354 method/map/phase comparisons; this total includes separately fitted PCA and uncompressed students.

K Prospective confirmation

Oxford-IIIT Pets and TREC were chosen from the available models and datasets before pilot compression results. The official training pools are split stratifiably using the locked seed; all transformation and dimension decisions across both tasks freeze before test annotation parsing or feature extraction. This is not external preregistration or a claim of absence from encoder pretraining. The covariance normalization was reconciled with the full-SVD denominator before any compression fit. This did not change extraction, dataset selection, split, randomization or inference rules.

The raw-null ranking analysis uses exactly the validation-selected common dimension and pair. Confirmation requires the two preassigned signs, both 98.75% crossed bootstrap intervals excluding zero, and at least sixteen of twenty fresh RP draws jointly exhibiting those signs. Twenty thousand replicates resample paired example rows and paired RP-seed columns, with common weights at

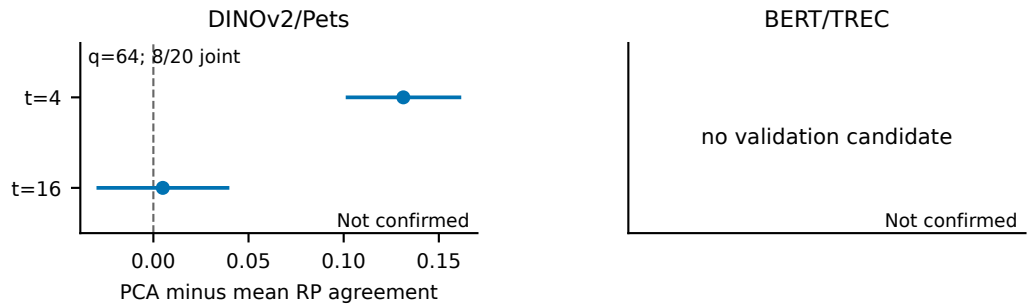


Figure 4: The two prospective primary slots. Points and horizontal bars show frozen-pair agreement differences and 98.75% crossed intervals. Absent candidates and failed test fits remain explicit non-confirmations.

both endpoints. Four endpoint slots and two model-level conditional binomial tests remain reserved irrespective of missing candidates. The binomial null concerns joint-sign probability at most one half conditional on the observed test set. The crossed intervals are approximate; neither statistic by itself proves a population reversal.

Table 36: The original prospective ranking test, including the failed Pets confirmation.

Task	q	First difference [CI]	Second difference [CI]	Joint draws	Adjusted p
DINOv2/Pets	64	+13.1 [+10.2, +16.1]	+0.5 [-2.9, +3.9]	8/20	1
BERT/TREC	∅	n/a	n/a	n/a	n/a

DINOv2/Pets has convergence warnings in 2/2 PCA source fits, 19/20 validation RP fits and 36/40 fresh test RP fits; PCA models are reused at test. The result concerns the frozen finite-fit procedure. The cause of non-confirmation is not uniquely identified.

Differences and intervals above are percentage points. Full canonical per-gauge recommendations and identity-gauge fresh-draw transfer follow; their pass-frequency intervals are pointwise and distinct from the family-adjusted ranking intervals.

K.1 DINOv2/Pets

The fit split has 2944 examples, validation has 736, ambient dimension is 384, and the complete panel contains 21 gauges.

Table 37: DINOv2/Pets: dimension ranges and censored gauge choices.

Convention/method	$q_{\min}$	$q_{\max}$	q_{uniform}	q_{joint}	Censored
Raw/PCA	8	∅	∅	∅	2/21
Raw/RP	16	∅	∅	∅	2/21
Standardized/PCA	8	256	256	256	0/21
Standardized/RP	16	∅	∅	∅	1/21
Whitened/RP	∅	∅	∅	∅	21/21

Table 38: DINOv2/Pets: minimum qualifying dimensions for each gauge.

Gauge	Raw PCA	Raw RP	Std. PCA	Std. RP	White RP
Null $t = 0.0625$	8	16	8	16	∅
Null $t = 0.125$	8	16	8	16	∅
Null $t = 0.25$	8	16	8	16	∅
Null $t = 0.5$	8	32	8	32	∅
Null $t = 1$	16	64	16	64	∅
Null $t = 2$	64	128	32	128	∅
Null $t = 4$	256	256	128	256	∅
Null $t = 8$	∅	∅	256	256	∅
Null $t = 16$	∅	∅	256	∅	∅
GL $\kappa = 1$, seed 4101	16	64	16	64	∅
GL $\kappa = 2$, seed 4101	16	64	16	64	∅
GL $\kappa = 4$, seed 4101	16	64	16	64	∅
GL $\kappa = 8$, seed 4101	16	64	16	64	∅
GL $\kappa = 16$, seed 4101	32	64	32	64	∅
GL $\kappa = 32$, seed 4101	32	64	32	64	∅
GL $\kappa = 1$, seed 4102	16	64	16	64	∅
GL $\kappa = 2$, seed 4102	16	64	16	64	∅
GL $\kappa = 4$, seed 4102	16	64	16	64	∅
GL $\kappa = 8$, seed 4102	16	64	16	64	∅
GL $\kappa = 16$, seed 4102	16	64	16	64	∅
GL $\kappa = 32$, seed 4102	32	64	16	64	∅

Table 39: DINOv2/Pets: frozen identity choices evaluated with fresh test draws.

Convention/method	Frozen identity q	Test passes [95% CI]	Test A	Test accuracy	Full A
Raw/PCA	16	1/1	0.937	0.919	0.996
Raw/RP	64	20/20 [0.83, 1.00]	0.924	0.900	0.996
Standardized/PCA	16	1/1	0.947	0.929	1.000
Standardized/RP	64	20/20 [0.83, 1.00]	0.928	0.903	1.000
Whitened/RP	∅	n/a	n/a	n/a	0.963

Table 40: DINOv2/Pets: direct full-dimensional baselines.

Full-dimensional probe	Validation A	Validation dual	Test A	Test dual
Raw	0.997	pass	0.996	pass
Standardized	1.000	pass	1.000	pass
Whitened	0.984	pass	0.963	pass

These deterministic full-dimensional students retain the same fitted reference, fit sample and fixed isotropic L2 setting. Both feasibility thresholds are applied; passing or failing at the full dimension does not imply monotonicity along the compressed grid.

Across both evaluation phases, maximum reference probability error is 7.70×10^{-14}, maximum reference score-difference error is 5.84×10^{-13}, and the reference decision-change total is 0. Maximum relative compressed-coordinate discrepancy is 9.47×10^{-9}. Raw orthogonal controls have 285 changed predictions across 452 method/map/phase comparisons; this total includes separately fitted PCA and uncompressed students.

K.2 BERT/TREC

The fit split has 4361 examples, validation has 1091, ambient dimension is 768, and the complete panel contains 21 gauges.

Table 41: BERT/TREC: dimension ranges and censored gauge choices.

Convention/method	$q_{\min}$	$q_{\max}$	q_{uniform}	q_{joint}	Censored
Raw/PCA	8	∅	∅	∅	7/21
Raw/RP	32	∅	∅	∅	17/21
Standardized/PCA	8	∅	∅	∅	9/21
Standardized/RP	64	∅	∅	∅	17/21
Whitened/RP	∅	∅	∅	∅	21/21

Table 42: BERT/TREC: minimum qualifying dimensions for each gauge.

Gauge	Raw PCA	Raw RP	Std. PCA	Std. RP	White RP
Null $t = 0.0625$	8	32	8	64	∅
Null $t = 0.125$	8	128	8	128	∅
Null $t = 0.25$	32	256	32	256	∅
Null $t = 0.5$	128	512	128	512	∅
Null $t = 1$	512	∅	512	∅	∅
Null $t = 2$	∅	∅	∅	∅	∅
Null $t = 4$	∅	∅	∅	∅	∅
Null $t = 8$	∅	∅	∅	∅	∅
Null $t = 16$	∅	∅	∅	∅	∅
GL $\kappa = 1$, seed 4101	512	∅	512	∅	∅
GL $\kappa = 2$, seed 4101	512	∅	512	∅	∅
GL $\kappa = 4$, seed 4101	512	∅	512	∅	∅
GL $\kappa = 8$, seed 4101	512	∅	512	∅	∅
GL $\kappa = 16$, seed 4101	512	∅	∅	∅	∅
GL $\kappa = 32$, seed 4101	512	∅	∅	∅	∅
GL $\kappa = 1$, seed 4102	512	∅	512	∅	∅
GL $\kappa = 2$, seed 4102	512	∅	512	∅	∅
GL $\kappa = 4$, seed 4102	512	∅	512	∅	∅
GL $\kappa = 8$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 16$, seed 4102	∅	∅	∅	∅	∅
GL $\kappa = 32$, seed 4102	∅	∅	∅	∅	∅

Table 43: BERT/TREC: frozen identity choices evaluated with fresh test draws.

Convention/method	Frozen identity q	Test passes [95% CI]	Test A	Test accuracy	Full A
Raw/PCA	512	1/1	0.942	0.916	0.950
Raw/RP	∅	n/a	n/a	n/a	0.950
Standardized/PCA	512	1/1	0.952	0.882	1.000
Standardized/RP	∅	n/a	n/a	n/a	1.000
Whitened/RP	∅	n/a	n/a	n/a	0.924

Table 44: BERT/TREC: direct full-dimensional baselines.

Full-dimensional probe	Validation A	Validation dual	Test A	Test dual
Raw	0.937	pass	0.950	pass
Standardized	1.000	pass	1.000	pass
Whitened	0.889	fail	0.924	pass

These deterministic full-dimensional students retain the same fitted reference, fit sample and fixed isotropic L2 setting. Both feasibility thresholds are applied; passing or failing at the full dimension does not imply monotonicity along the compressed grid. BERT/TREC's full-dimensional whitened probe already fails the validation dual predicate ($A = 0.889$; raw full-probe $A = 0.937$). Its compressed-grid censoring cannot be attributed solely to dimension reduction: the declared coordinate convention and isotropic L2 probe already alter the fitted full-dimensional behavior.

Across both evaluation phases, maximum reference probability error is 5.98×10^{-13}, maximum reference score-difference error is 2.71×10^{-12}, and the reference decision-change total is 0. Maximum relative compressed-coordinate discrepancy is 2.51×10^{-9}. Raw orthogonal controls have 0 changed predictions across 432 method/map/phase comparisons; this total includes separately fitted PCA and uncompressed students.

L CANONICALIZED PAIRED-TASK AUDIT

The two tasks use identical images and splits, identity-gauge full SVD whitening, and the same Gaussian matrices for every dimension and draw. Task-specific labels, fixed references, and students differ. This identity-only comparison is a task control, not a test of nontrivial affine invariance. Both validation choices freeze before the held-out phase. Because both choices are censored, the frozen test union contains the declared $q = 8$ diagnostic and full-dimensional comparator; no test-grid search is performed.

On held-out examples the full-dimensional fine/coarse label accuracies are 0.811/0.864. All 284 prediction records were replayed numerically; 0 fits failed and 0 convergence warnings were recorded.

Table 45: Complete validation grid for the paired-task full-whitening/RP procedure. Entries give mean agreement and dual-criterion passing draws. The final row uses a separately fitted full-dimensional logistic student.

	Fine labels		Coarse labels	
Dimension	agreement	passes	agreement	passes
1	0.018	0/10	0.065	0/10
2	0.022	0/10	0.075	0/10
4	0.041	0/10	0.089	0/10
8	0.081	0/10	0.117	0/10
16	0.162	0/10	0.175	0/10
32	0.304	0/10	0.253	0/10
64	0.444	0/10	0.401	0/10
128	0.650	0/10	0.612	0/10
256	0.804	0/10	0.799	0/10
384 (full)	0.929	1/1	0.944	1/1

This task comparison supplies no positive selected-dimension gap because both grid choices are censored.

M Structural support of the normalized encoder outputs

The exact-rank condition is substantive for these encoder interfaces. This is a known consequence of affine LayerNorm (Ba et al., 2016), not a new representation theorem. For pre-normalization coordinates $x \in \mathbb{R}^p$, fixed gains $\gamma_j \neq 0$ and biases β_j, write

$$h_j = \gamma_j \frac{x_j - \bar{x}}{\sqrt{p^{-1}\sum_k (x_k - \bar{x})^2 + \epsilon}} + \beta_j.$$

Then $\sum_j (h_j - \beta_j)/\gamma_j = 0$ in exact arithmetic, including when $\epsilon > 0$. All outputs therefore lie in a fixed affine hyperplane; after centering, covariance rank is at most $p-1$. This statement concerns the ideal arithmetic operation. Finite-precision normalization and stored rounding can break the exact relation.

The actual ViT and DINOv2 models apply a final affine LayerNorm before the recorded CLS extraction. The recorded BERT CLS is taken after its native final LayerNorm. Each inspected gain is finite and nonzero. The model documentation records normalization parameters, pooling policies and the implemented forward computations. This inspection makes no claim that the rounded feature caches have exact rank $p - 1$.

Passing a float64 full-column-rank test on those caches therefore does not establish positive-definite covariance for the ideal encoder representation. Our full-p whitening results describe the declared procedure on numerical caches, including directions supported by roundoff. The all-$p - 1$ float32 rank rejections in the stress study are consistent with this structural limitation. We retain the failed numerical cases and the full-whitening results. Neither precision promotion nor a small coordinate residual proves useful or stable compression. The separately declared support-aware thin-whitening control checks these structural and held-out support conditions (Appendix A). It improves conditioning while retaining the failed compressed Gaussian operating points and the strict reference-projection tolerance failures. The original full-cache experiments remain unchanged.

N Numerical and parameter-level controls

Validation-only numerical stress. The locked panel crosses fit sizes 1200, 2400, float64/float32 arithmetic, two fixed orientations, and condition numbers $\kappa \in \{1, 4, 32\}$, plus identity. Fit subsets are deterministic prefixes of the existing fit-index order; the same 600 validation examples are used throughout. No test array, dimension selection, transformation selection, or learner selection is involved. Each successful case fits plain L2 logistic regression at $C \in \{0.1, 1.0, 10.0\}$ and an L2 squared-hinge linear SVM at $C = 1.0$. There is no post-whitening coordinate scaler. The float64

Table 46: Complete validation-only stress summary. Maxima span identity and all six GL transformations. Accuracy ranges span the four declared learners and concern the fixed $q = 8$ draw. Each float64 group preserves all predictions; every float32 group is rejected.

Encoder	n_{fit}	float32 rank	covariance error	projected error	accuracy
ViT	1200	767/768	6.88×10^{-9}	1.75×10^{-9}	0.150–0.163
ViT	2400	767/768	1.07×10^{-8}	1.43×10^{-9}	0.147–0.148
DINOv2	1200	383/384	1.50×10^{-8}	2.71×10^{-9}	0.142–0.148
DINOv2	2400	383/384	1.33×10^{-8}	2.53×10^{-9}	0.192–0.195
BERT	1200	767/768	1.13×10^{-8}	1.59×10^{-9}	0.582–0.582
BERT	2400	767/768	8.46×10^{-9}	1.37×10^{-9}	0.552–0.555

implementation uses full SVD with covariance denominator $n - 1$ and no eigenvalue floor, ridge, or truncation. Both fit sizes exceed every feature dimension, and every float64 fit matrix passes the declared numerical full-column-rank test. The exact theorem remains conditional on positive definiteness of the mathematical fit covariance. This floating-point rank check applies to the rounded cache. It does not certify positive definiteness of the ideal normalized encoder representation (Appendix M). No thin-whitening extension is invoked. Float32 explicitly changes centering, SVD, matrix multiplication, and RP coupling arithmetic, then tests rank against $\epsilon_{32} \max(n, p) s_{\max}$. All such cells have numerical rank $p - 1$, so no float32 student is fitted. A rejection under this precision-dependent criterion is not a claim that the float64 stored matrix has exact rank $p - 1$. Learner solvers may internally promote arithmetic. The precision label refers to canonicalization and projected features. Across float64 cases, the maximum covariance-identity error is 1.50×10^{-8} and maximum normalized projected validation error is 2.71×10^{-9}. Every independently fitted student's validation prediction agrees with its identity-gauge counterpart, with 0 recorded warnings. These comparisons are within each fit size and learner. Accuracy can change with fit size or learner, and the low fixed-q accuracies show that invariance is separate from compression utility. The single RP draw supplies no seed-reliability inference. The numerical checks recompute all rejected ranks from saved singular values and check compressed-array residuals, student coefficients and predictions. They do not rerun every full SVD. The full-rank float32 rejection is consistent with the LayerNorm support constraint. It does not show that float32 destroys task information or causes downstream solver failure; those students were not fitted.

Actual copied-parameter experiment. For a positive diagonal S, we change the saved hidden-layer weights and bias to SW_2, Sb_2, and the classifier to $W_3 S^{-1}, b_3$. Evaluation-mode ReLU homogeneity gives $h' = Sh$ and identical logits. Input BatchNorm precedes the transformed layers; dropout is disabled. The reference is the original neural head, separately from the modern-encoder logistic references. The original saved seed-7 model configuration includes input noise with standard deviation 0.05; we reproduce its loader and noise order. All 12 transformations and $q = 4$ were fixed before the new validation panel. We fit on 291 examples, validate on 73, and use the historical test set of 114; this is historical test reuse. The complete validation panel was fixed before test access. Test draws use fresh RP seeds, with the same matrix paired across transformations. The seeds 611 and 977 permute the assignment of geometrically spaced positive scales to the 32 hidden coordinates. The resulting maps are diagonal rescalings, not pure permutations. Two identity labels represent the same control, not independent replicates. Numerical comparisons cover 384 compressed students and 9,228 copied-parameter forward examples, with zero reference-decision changes. Repeated examples in numerical checks are not independent observations. Saved outputs include parameters, projections, coefficients, predictions and example IDs.

O Reference-preserving empirical controls

Figure 5 evaluates the reference-preserving construction. Across 10 independent Gaussian train/test draws, the mean PCA-minus-RP agreement changes from $+0.341$ at $t = 0.3$ to -0.019 at $t = 3$.

For the real controls, the reference's training-fitted scaler defines the coordinates before intervention; thus $t = 1$ is a standardized-coordinate control, distinct from the primary raw-coordinate compressors. The reference is fixed across t within each phase, with a separate full-training refit for test.

Table 47: Actual MLP parameter symmetry, complete historical test panel. Agreement uses the original model head. The three small negative mean differences were absent on validation and do not establish a robust reversal.

Scale-assignment seed	κ	PCA	mean RP	PCA−RP	RP passes
611	1	1.0000	0.9842	+0.0158	20/20
611	2	1.0000	0.9842	+0.0158	20/20
611	4	0.9912	0.9908	+0.0004	20/20
611	8	0.9912	0.9921	−0.0009	20/20
611	16	0.9912	0.9925	−0.0013	20/20
611	32	0.9912	0.9921	−0.0009	20/20
977	1	1.0000	0.9842	+0.0158	20/20
977	2	1.0000	0.9842	+0.0158	20/20
977	4	0.9912	0.9825	+0.0088	20/20
977	8	0.9912	0.9807	+0.0105	20/20
977	16	0.9825	0.9820	+0.0004	20/20
977	32	0.9825	0.9816	+0.0009	20/20

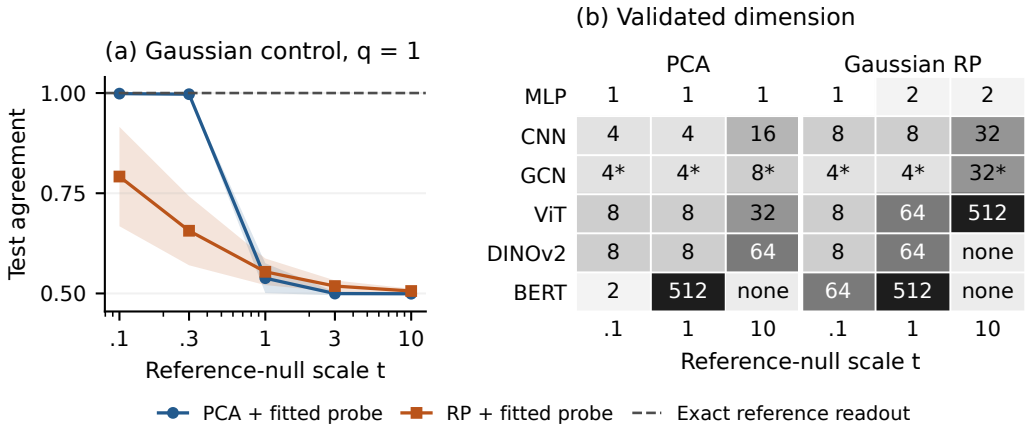


Figure 5: Reference-preserving controls. Left: mean agreement and one standard deviation across 10 independent train/test and RP draws, with 4,000 training and 10,000 test examples per draw. Right: method-specific validation-selected dimensions on six fixed representations (canonical RP seed 0). An asterisk marks test failure after refitting; `none` means no validation-feasible grid point. The decoder control is specified exactly, whereas PCA/RP use separately fitted probes. These secondary evaluations reuse historical benchmark test sets.

PCA and RP each select q on validation with seed 0; additional RP draws use only frozen dimensions. All reference decisions remain unchanged, and the specified subspace decoder has agreement 1 throughout. For BERT, PCA selects $q = 2 \to 512 \to$ none at $t = 0.1, 1, 10$; the first two choices achieve test agreement 0.920 and 0.907. All 27 selected configurations outside the graph diagnostic pass the test criteria; all 6 Cora choices fail. These changes concern access through the fitted compression/probe procedure, despite constant reference information; they do not identify intrinsic dimension.

P Fixed-teacher and seed-stability details

Table 48 holds the target decision rule fixed across depth. At every sampled early block (3, 6, and 9), all three models' uncompressed students fall below .90 teacher agreement on validation, and neither compression family has a feasible choice. This makes a compression-only explanation insufficient: the declared uncompressed decoder also fails to reproduce the target behavior at the threshold. Only the sampled final block admits selected compressed students. This is a diagnostic

Table 48: One frozen block-12 teacher per model. A_{full} is agreement of a separately fitted uncompressed student with that teacher; validation and test are shown separately. Dimensions are validation-selected; remaining agreements are original test results. P: PCA; U: UMAP. * marks failure of either test criterion. `none` is a grid outcome, not an impossibility claim.

Model	Block	$A_{\text{full}}^{\text{val}}$	$A_{\text{full}}^{\text{test}}$	q_{RP}^{*}	q_{adapt}^{*}	A_{RP}	A_{adapt}
ViT	3	0.597	0.639	none	none	–	–
ViT	6	0.745	0.749	none	none	–	–
ViT	9	0.883	0.897	none	none	–	–
ViT	12	1.000	0.988	64	2 (U)	0.935	0.921
BERT	3	0.733	0.711	none	none	–	–
BERT	6	0.765	0.758	none	none	–	–
BERT	9	0.820	0.796	none	none	–	–
BERT	12	0.920	0.924	512*	none	0.893	–
DINOv2	3	0.543	0.593	none	none	–	–
DINOv2	6	0.658	0.678	none	none	–	–
DINOv2	9	0.823	0.821	none	none	–	–
DINOv2	12	1.000	0.994	64	2 (U)	0.923	0.906

of fitted-probe accessibility, not a claim that earlier representations lack the information or that the transition occurs exactly at block 12.

The distinction changes the reading of the earlier layer-specific-reference study. ViT block 9 previously admitted RP-512 and PCA-256 for its own reference; it admits neither for the fixed final-layer teacher here. The earlier comparison held labels fixed while changing the reference rule. This control fixes the reference and trains students on teacher targets. These protocols answer different questions. Appendix P reports teacher accuracies and all fresh selected-map draws; Appendix R retains the original layer-specific trajectories.

At block 12, ViT and DINOv2 each select RP-64 versus UMAP-2 and both original choices pass on test. BERT selects RP-512, but its test agreement is .893 and it fails the criterion despite the uncompressed student reaching .924. Fresh fixed-teacher test passes are 10/10 RP and 6/10 UMAP for ViT, 7/10 RP and 9/10 UMAP for DINOv2, and 1/10 RP for BERT. These draws belong to the fixed-teacher protocol, separately from the refitted-reference stability study below.

Fixed-teacher accuracy and fresh compressor draws. The teacher is fitted once on 2,400 original fit examples and is never refitted. Its test accuracy is constant across layers of a model. Students and selected maps refit on all 3,000 training examples with hard targets from that same teacher. The table separates original seed zero from ten disjoint fresh stochastic seeds (10000–10009); deterministic maps have one outcome.

Table 49: Fixed-teacher accuracy and fresh compressor-draw pass counts.

Model	Block	Teacher accuracy	RP fresh passes	Adaptive fresh passes
ViT	3	0.970	–	–
ViT	6	0.970	–	–
ViT	9	0.970	–	–
ViT	12	0.970	10/10	6/10
BERT	3	0.783	–	–
BERT	6	0.783	–	–
BERT	9	0.783	–	–
BERT	12	0.783	1/10	–
DINOv2	3	0.976	–	–
DINOv2	6	0.976	–	–
DINOv2	9	0.976	–	–
DINOv2	12	0.976	7/10	9/10

Complete original and empirical-stability outcomes. Counts below use the new draws for both policies. The original choice was selected using the earlier seed-zero validation result, not these

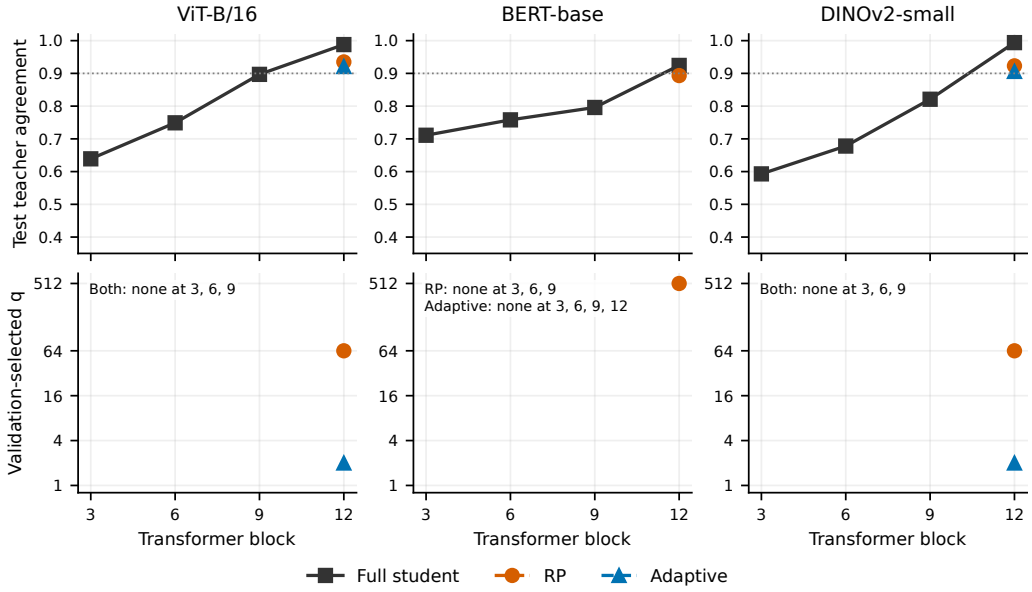


Figure 6: Fixed-teacher depth control. Top: original test agreement of uncompressed and selected compressed students with one unchanged block-12 teacher. Bottom: validation-selected dimensions; text annotations record absent choices without placing them on the numerical dimension axis. Lines join measured layers only.

new validation counts. Intervals describe pointwise stochastic test feasibility conditional on the fixed full-training and test data. They do not provide simultaneous post-selection coverage or a population guarantee for the empirical selector. Deterministic fits have no compressor-frequency interval. Validation draws use seeds 100–119 (RP) and 100–109 (UMAP); test seeds are 1000–1019 and 1000–1009.

Table 50: Original and fresh-draw results for the frozen validation choices.

Model	Policy	Family	q	Val.	Test	Mean test A	Test freq. 95% CI
MNIST	Original	RP	8	20/20	20/20	0.946	[0.832,1.000]
MNIST	Original	Adapt.	2 (U)	10/10	10/10	0.986	[0.692,1.000]
MNIST	$\rho = .8$	RP	8	20/20	20/20	0.946	[0.832,1.000]
MNIST	$\rho = .8$	Adapt.	2 (U)	10/10	10/10	0.986	[0.692,1.000]
ViT	Original	RP	64	18/20	19/20	0.918	[0.751,0.999]
ViT	Original	Adapt.	2 (U)	7/10	7/10	0.909	[0.348,0.933]
ViT	$\rho = .8$	RP	64	18/20	19/20	0.918	[0.751,0.999]
ViT	$\rho = .8$	Adapt.	4 (U)	10/10	10/10	0.957	[0.692,1.000]
DINOv2	Original	RP	64	12/20	18/20	0.912	[0.683,0.988]
DINOv2	Original	Adapt.	2 (U)	9/10	5/10	0.897	[0.187,0.813]
DINOv2	$\rho = .8$	RP	128	20/20	20/20	0.950	[0.832,1.000]
DINOv2	$\rho = .8$	Adapt.	2 (U)	9/10	5/10	0.897	[0.187,0.813]
BERT	Original	RP	512	3/20	8/20	0.896	[0.191,0.639]
BERT	Original	Adapt.	none	–	–	–	–
BERT	$\rho = .8$	RP	none	–	–	–	–
BERT	$\rho = .8$	Adapt.	none	–	–	–	–

Even ten successes in ten independent compressor draws give a pointwise exact 95% binomial interval of approximately $[.692, 1]$. Requiring an empirical fraction of .80 is therefore weaker than certifying a population success probability of .80. These intervals condition on fixed examples and do not represent encoder-training, split, or distribution-shift uncertainty.

Table 51: Separate validation-selected family dimensions and original test agreement. U: UMAP; P: PCA; I: Isomap. †: graph diagnostic. An asterisk marks a test failure of either behavioral criterion. `none` means no feasible adaptive grid point; the BERT gap is undefined, not zero. Grid halvings are $\log_2(q^*_{\mathrm{RP}}/q^*_{\mathrm{adapt}})$.

Representation	q^*_{RP}	q^*_{adapt}	Grid halvings	A_{RP}	A_{adapt}
BC / MLP	2	1 (I)	1	0.956	1.000
C10 / DINOv2	64	2 (U)	5	0.920	0.901
C10 / ViT	64	2 (U)	5	0.936	0.920
Cora / GCN†	4*	1 (U)*	2	0.643	0.495
MNIST / CNN	8	2 (U)	2	0.965	0.988
SST-2 / BERT	512	none	–	0.904	–

Q Protocol and original-core diagnostics

Q.1 How many dimensions does adaptation save?

Table 51 separates the two family choices. In the original validation split, Breast Cancer uses RP at two dimensions versus Isomap at one; MNIST uses RP at eight versus UMAP at two. ViT and DINOv2 use RP at 64 versus UMAP at two: five grid halvings. Both frozen family choices pass the test criteria in these four cases. These are conditional operating points, not universal sufficiency statements. The graph diagnostic has an apparent validation saving but both family choices fail on test. BERT has no validated adaptive candidate through 512 coordinates; its original RP point at 512 narrowly passes. In the original test-seed panel (seeds 0–9), only 2/10 RP draws pass there, and two additional core validation partitions find no feasible configuration. In that same original panel, ViT UMAP at two dimensions passes for 7/10 draws, whereas MNIST and DINOv2 pass for 10/10. These original draws are distinct from the fresh test seeds 1000–1019 (RP) and 1000–1009 (UMAP) reported in the separate seed-stability study. Thus a dimension saving and robust transfer across fitting randomness are separate findings.

Datasets and implementations. The image studies use CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). CIFAR-100 supplies fine and coarse labels for each image. The supporting language and depth studies use binary SST-2 (Socher et al., 2013), distributed through GLUE (Wang et al., 2018); its public validation split supplies the reported evaluation examples. The digit study uses MNIST (LeCun et al., 1998). The tabular study uses scikit-learn's 569-example, 30-feature Wisconsin Diagnostic Breast Cancer dataset (Street et al., 1993). The graph diagnostic uses Cora (Sen et al., 2008) through the Planetoid loader and public split (Yang et al., 2016), with normalized node features. Its encoder is a graph convolutional network (Kipf & Welling, 2017). This is a transductive diagnostic.

Fixed representation metadata. The following dimensions and sample counts are read directly from the saved representation metadata. Original training and held-out identifiers are included. Saved supervised model seed 7 is a model-training seed; the ten projection seeds are compressor draws. Fixed pretrained models do not supply independent training replicates.

Table 52: Representation dimensions and sample counts.

Representation	p	Training pool	Test examples
BC / MLP	32	364	114
MNIST / CNN	128	3000	1000
C10 / ViT	768	3000	1000
C10 / DINOv2	384	3000	1000
SST-2 / BERT	768	3000	872
Cora / GCN†	64	140	1000

Reference versus original head. Where a compatible task head was stored, its accuracy and agreement with the full-dimensional reference are reported below. The pretrained modern backbones do not supply a compatible original task head in this evaluation. Similar accuracy does not imply matching per-example predictions.

Table 53: The fitted reference compared with the original neural head.

Representation	Reference accuracy	Head accuracy	Reference–head A
BC / MLP	0.965	0.965	1.000
Cora / GCN†	0.550	0.623	0.745
MNIST / CNN	0.988	0.989	0.991

Scope of the original selector. The diagnostics below retain the original global selector over random coordinates, sparse RP, Gaussian RP, PCA, UMAP (McInnes et al., 2020; 2018), and Isomap (Tenenbaum et al., 2000). This differs from the separately selected RP/adaptive families in Table 51 and the main paper. The retrospective grid index is descriptive and is never a validation-selected operating point.

Metric sensitivity. The primary criterion combines relative accuracy and agreement, each at least 0.90. The absolute alternative allows at most 0.10 accuracy degradation and also requires agreement at least 0.90. The absolute condition is redundant when agreement is at least 0.90, by $|\Delta a| \leq 1 - A$. Its equality with the agreement-only policy is an exact sanity check. Relative-accuracy-only selection can omit behavioral fidelity. All alternative choices below use the original validation rows and seed, with the original dimension and tie rules. They are secondary analyses, not substitutions for the original held-out operating point.

Table 54: Choices under the stated validation metrics.

Representation	Criterion	Selected method	q
BC / MLP	Relative + agreement	Isomap	1
BC / MLP	Absolute + agreement	Isomap	1
BC / MLP	Agreement only	Isomap	1
BC / MLP	Relative accuracy only	Isomap	1
C10 / DINOv2	Relative + agreement	UMAP	2
C10 / DINOv2	Absolute + agreement	UMAP	2
C10 / DINOv2	Agreement only	UMAP	2
C10 / DINOv2	Relative accuracy only	UMAP	2
C10 / ViT	Relative + agreement	UMAP	2
C10 / ViT	Absolute + agreement	UMAP	2
C10 / ViT	Agreement only	UMAP	2
C10 / ViT	Relative accuracy only	UMAP	2
Cora / GCN†	Relative + agreement	UMAP	1
Cora / GCN†	Absolute + agreement	UMAP	1
Cora / GCN†	Agreement only	UMAP	1
Cora / GCN†	Relative accuracy only	UMAP	1
MNIST / CNN	Relative + agreement	UMAP	2
MNIST / CNN	Absolute + agreement	UMAP	2
MNIST / CNN	Agreement only	UMAP	2
MNIST / CNN	Relative accuracy only	UMAP	2
SST-2 / BERT	Relative + agreement	Gaussian RP	512
SST-2 / BERT	Absolute + agreement	Gaussian RP	512
SST-2 / BERT	Agreement only	Gaussian RP	512
SST-2 / BERT	Relative accuracy only	PCA	16

Retrospective grid index. For transparency, the table below scans the test grid at original compressor seed zero. It uses the same feasibility and tie rules but is a descriptive test-selected index,

never an input to selection. Missing entries mean no qualifying tested compression, not impossibility outside the grid.

Table 55: Descriptive test-grid choices, distinct from validation choices.

Representation	q^*_{sel}	q^*_{grid}	Test-grid method
BC / MLP	1	1	Isomap
C10 / DINOv2	2	2	UMAP
C10 / ViT	2	2	UMAP
Cora / GCN†	1	32	Coordinates
MNIST / CNN	2	2	UMAP
SST-2 / BERT	512	512	PCA

Selection stability. These three partitions use the same full candidate set, original compressor seed zero, and original policy. All decisions use validation data only; the test set is not used to choose among partitions. The partitions overlap, so their outcomes are descriptive stability checks, not independent model trials.

Table 56: Validation choices across the three overlapping split partitions.

Representation	Split seed	Method	q
BC / MLP	2027	Isomap	1
BC / MLP	2028	PCA	1
BC / MLP	2029	Sparse RP	1
MNIST / CNN	2027	UMAP	2
MNIST / CNN	2028	UMAP	2
MNIST / CNN	2029	UMAP	2
C10 / ViT	2027	UMAP	2
C10 / ViT	2028	UMAP	2
C10 / ViT	2029	UMAP	4
C10 / DINOv2	2027	UMAP	2
C10 / DINOv2	2028	UMAP	2
C10 / DINOv2	2029	UMAP	2
SST-2 / BERT	2027	Gaussian RP	512
SST-2 / BERT	2028	–	–
SST-2 / BERT	2029	–	–
Cora / GCN†	2027	UMAP	1
Cora / GCN†	2028	UMAP	2
Cora / GCN†	2029	UMAP	2

Uncertainty and fitting variation. Paired intervals use $2{,}000$ bootstrap repetitions with seed 202709. Stochastic seed zero participated in selection and is excluded from matched-effect uncertainty. The other nine draws are used for evaluative-core stochastic compressors. All draws remain in descriptive curves and pass counts. For each repetition, compressor draws are resampled within each method independently, and the same resampled example indices are applied to both methods. A shared numeric seed does not itself create a scientific pairing between different random algorithms. Deterministic compressors contribute one fitted map; their intervals do not quantify refitting on new training samples. No Cora interval is estimated: only two non-selection UMAP draws are available and test nodes share a graph. Selection is held fixed, so these are neither post-selection coverage guarantees nor intervals over model training.

Numerical reference-subspace implementation. The SVD retains singular values greater than $\max(10^{-12}, 10^{-12}\sigma_1)$, where σ_1 is the largest singular value. The row-space factorization residual is checked against 10^{-8}. Mathematical exactness refers to the stated full row space; the numerical implementation additionally checks saved examples. Across the real controls there are 0 reference-decision changes; maximum logit-difference and probability errors are 4.81×10^{-13} and 9.37×10^{-14}, respectively. These are roundoff checks, not generalization bounds.

Table 57: Fixed-$q = 8$ MNIST depth diagnostic on identical examples. Gaussian RP is averaged over ten draws; PCA uses one prespecified randomized-SVD fit. The full-dimensional reference is fitted separately at each layer. No layer or dimension is selected from these test outcomes.

Layer	p	a_{full}	A_{RP}	A_{PCA}
features 1	6272	0.950	0.468	0.821
features 2	3136	0.979	0.614	0.919
embedding	128	0.988	0.944	0.988
logits	10	0.990	0.989	0.993

Relation to other probe controls. INLP repeatedly fits classifiers and projects onto their nullspaces, generally reducing rank. It aims to remove decodable attributes. Our invertible intervention preserves a specified reference. MDL probing measures coding cost under a declared model and learning procedure. It is not automatically invariant to coordinate changes. These are complementary questions, not substitute baselines for the intervention.

Depth at a fixed compression budget. A secondary MNIST experiment holds $q = 8$ and the training/test examples fixed across four saved layers. Agreement increases from early features to the final embedding and logits (Table 57). Each layer has its own full-dimensional reference, so the comparison does not hold the decision rule fixed as the nullspace intervention does. It describes one trained model and is not a claim about all learning trajectories.

Table 58: Matched core effects: here Δa and ΔA mean the selected method minus Gaussian RP at the same q. Original selected configurations are BC: Isomap-1; MNIST, ViT and DINOv2: UMAP-2; Cora: UMAP-1; BERT: Gaussian RP-512. BERT therefore compares the same procedure and has identically zero effects. The selected method and RP use the same displayed q. This is not a comparison with RP at its separately selected feasible dimension. The Draws column counts selected-method fits; RP uses nine non-selection draws. Stochastic seed zero is excluded. Remaining draws and paired examples are resampled; Cora has no interval because only two non-selection UMAP draws and dependent nodes are available. Intervals exclude encoder-training and selection randomness.

Representation	Method	q	Contrast	Mean effect	95% interval	Draws
BC / MLP	Isomap	1	Δa	$+0.045$	[+0.014, +0.086]	1
BC / MLP	Isomap	1	ΔA	$+0.057$	[+0.023, +0.096]	1
C10 / DINOv2	UMAP	2	Δa	$+0.700$	[+0.674, +0.726]	9
C10 / DINOv2	UMAP	2	ΔA	$+0.701$	[+0.676, +0.725]	9
C10 / ViT	UMAP	2	Δa	$+0.660$	[+0.611, +0.704]	9
C10 / ViT	UMAP	2	ΔA	$+0.665$	[+0.615, +0.709]	9
Cora / GCN†	UMAP	1	Δa	$+0.176$	–	2
Cora / GCN†	UMAP	1	ΔA	$+0.201$	–	2
MNIST / CNN	UMAP	2	Δa	$+0.422$	[+0.384, +0.458]	9
MNIST / CNN	UMAP	2	ΔA	$+0.425$	[+0.387, +0.461]	9
SST-2 / BERT	RP	512	Δa	$+0.000$	[+0.000, +0.000]	9
SST-2 / BERT	RP	512	ΔA	$+0.000$	[+0.000, +0.000]	9

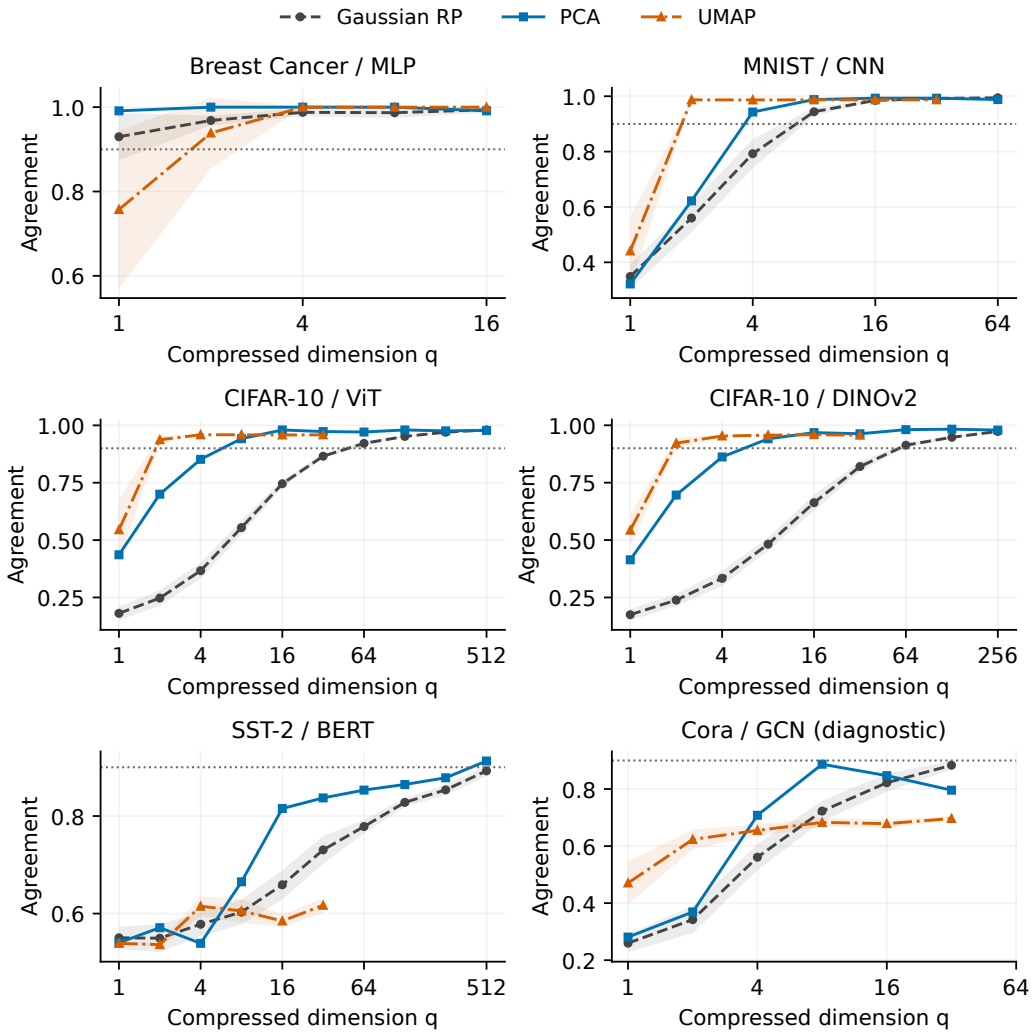


Figure 7: Descriptive core agreement curves for Gaussian RP, PCA, and UMAP. Means and bands of one compressor standard deviation use ten RP draws, three UMAP grid draws, and one PCA fit. The dotted line is the agreement criterion; it does not choose the operating point.

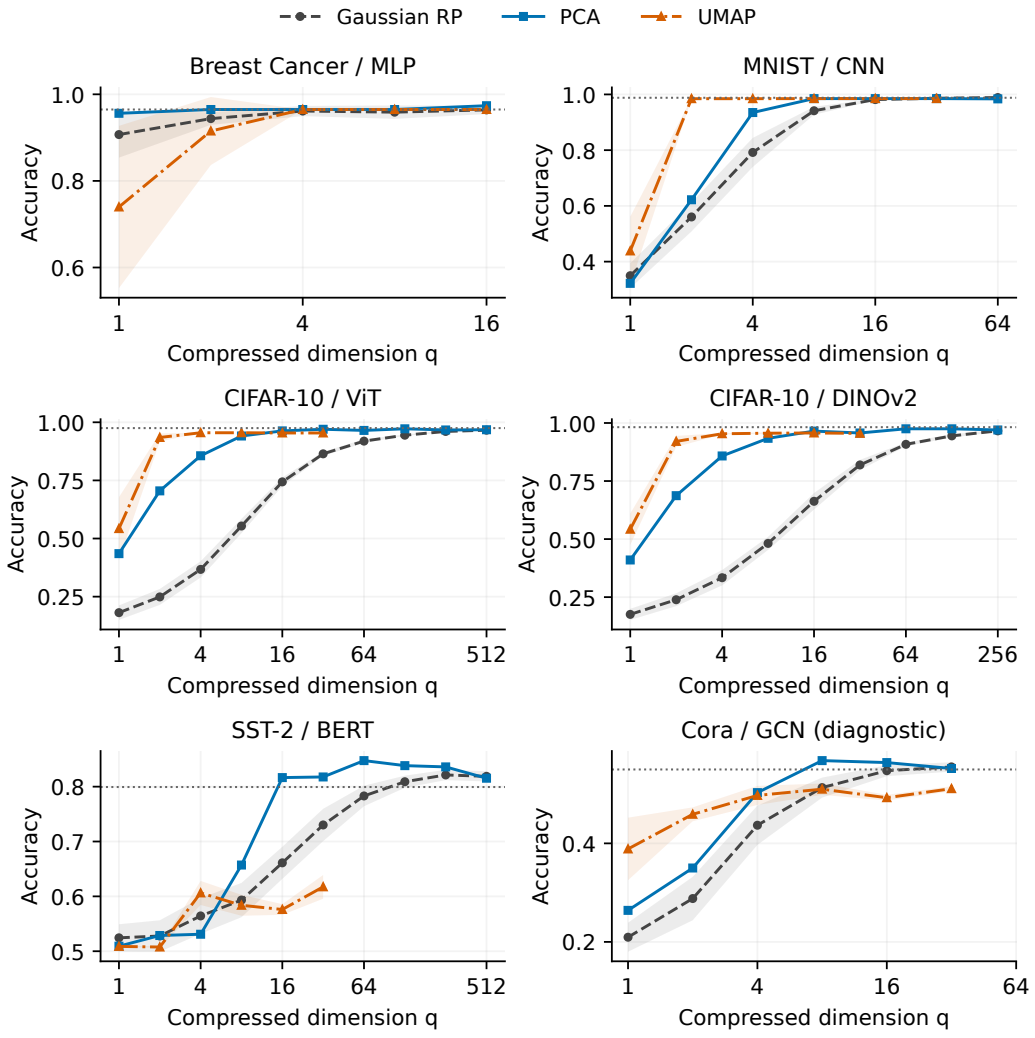


Figure 8: Descriptive core label accuracy for the same three methods; dotted lines give full-reference accuracy.

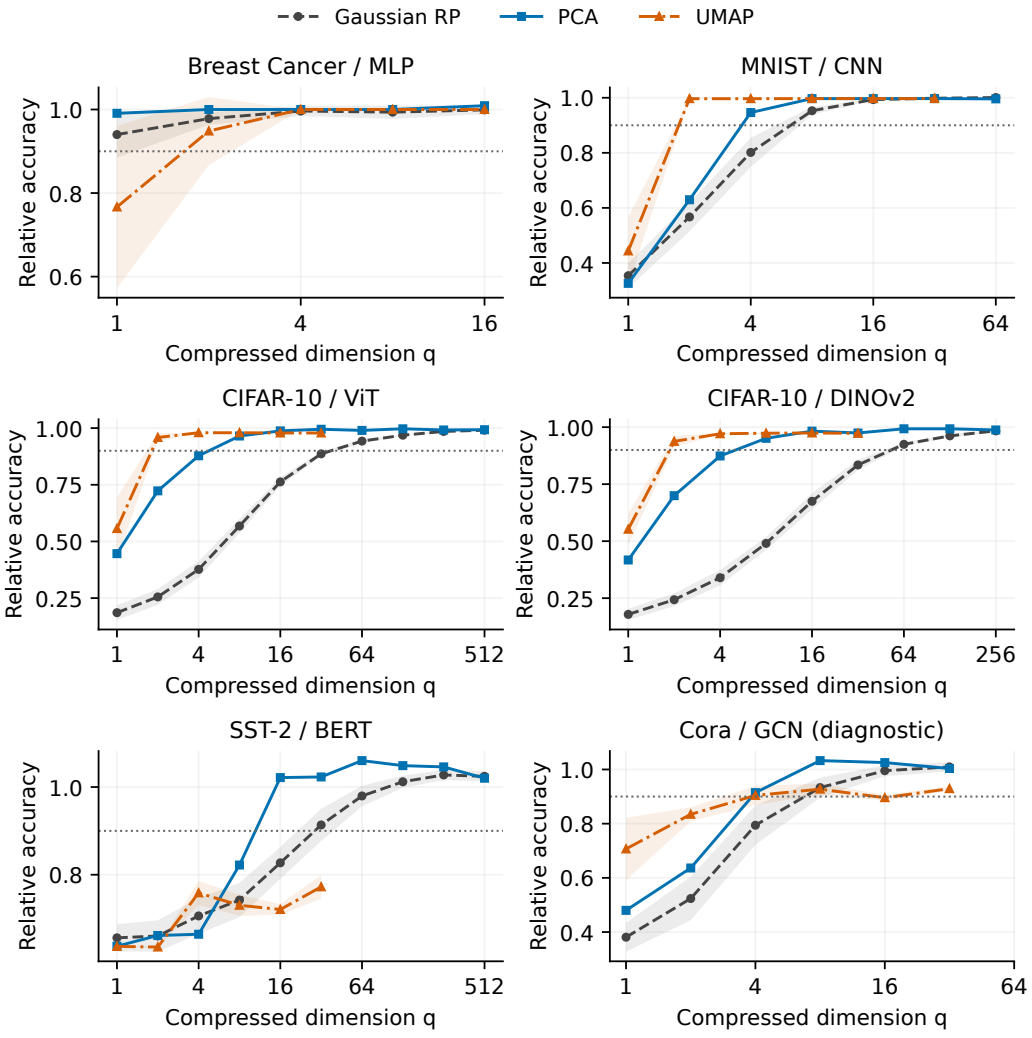


Figure 9: Descriptive core relative accuracy, which may exceed one. Agreement is measured separately.

R LAYER-SPECIFIC REFERENCES AND PAIRED-TASK DIAGNOSTICS

R.1 WHERE ALONG NETWORK DEPTH DOES A SAVING APPEAR?

ViT on CIFAR-10 has RP choices $(none, none, 512, 64)$ and adaptive choices $(none, none, 256, 2)$ at blocks $(3, 6, 9, 12)$, respectively. BERT on SST-2 has RP choices $(none, none, none, 512)$ and adaptive choices $(512, 512, none, none)$ at blocks $(3, 6, 9, 12)$, respectively. DINOv2 on CIFAR-10 has RP choices $(none, none, none, 64)$ and adaptive choices $(none, none, none, 2)$ at blocks $(3, 6, 9, 12)$, respectively. The missing choices are reported explicitly; they are not interpolated into a monotone trajectory. Each reference changes with the readout layer, so its test accuracy is shown alongside agreement. The visual models develop a large measured adaptive advantage at their final block. BERT changes which family qualifies. This is evidence about these fixed models and protocols, not a law of network depth. 8 of 9 frozen family choices pass both test criteria. Failed transfers: BERT / SST-2 block 3 adaptive. Fit-only spectral summaries and their relation to these choices are tabulated in Appendix R. Censoring and only four depths per model preclude a strong predictive conclusion. No intrinsic-dimension estimator, correlation significance, or universal failure of geometric prediction is inferred.

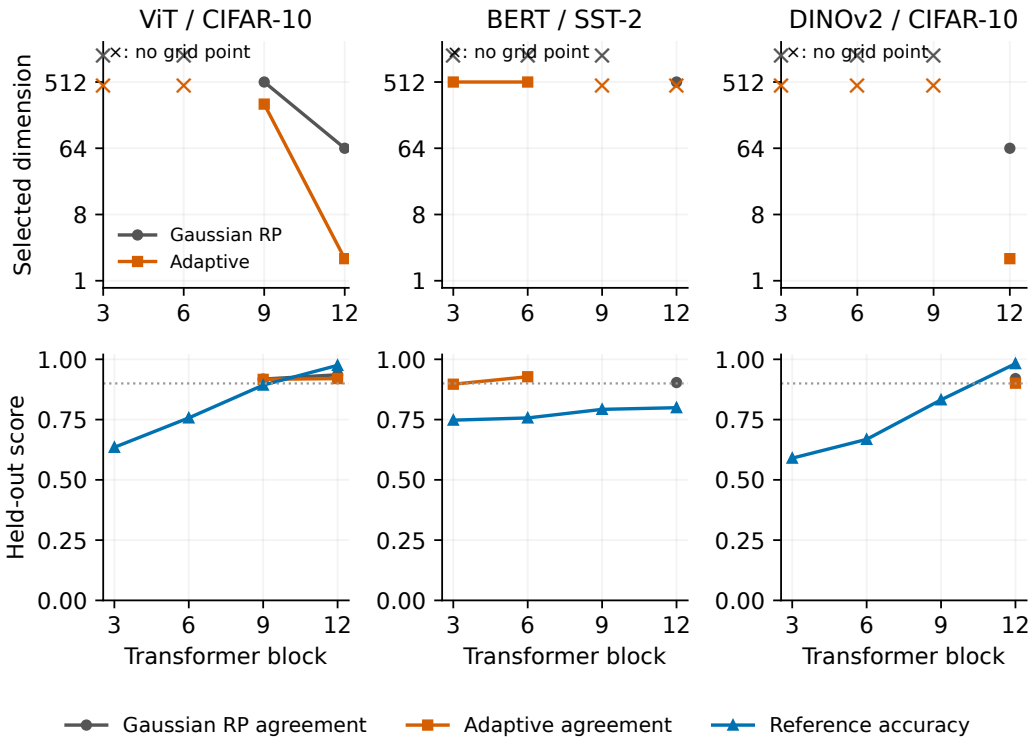


Figure 10: Behavioral compression across depth. Top: separate validation-selected dimensions; colored crosses in the annotation area mean no feasible grid point, not an imputed numerical dimension. Bottom: original selected-probe agreement and full-reference label accuracy on the same held-out examples. The reference is fitted independently at each layer. Lines join measured depths only.

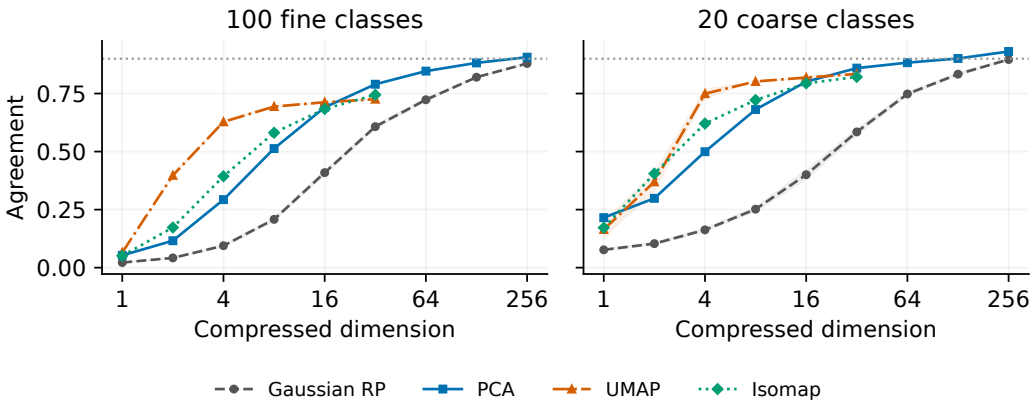


Figure 11: Official CIFAR-100 fine/coarse labels on the same feature matrix and maps, under the original axis-standardized protocol. Lines are descriptive test-grid agreement means; bands are one compressor standard deviation (ten RP draws, three UMAP grid draws, one PCA/Isomap fit). Task-family choices were frozen before test evaluation.

Complete depth operating points. All entries below come from the frozen original family choices. `none` records no feasible tested dimension. Agreement and reference accuracy condition on the indicated layer-specific probe.

Table 59: Layer-specific reference accuracy and frozen family operating points.

Model	Block	a_{full}	q^*_{RP}	q^*_{adapt}	A_{RP}	A_{adapt}
ViT	3	0.635	none	none	–	–
ViT	6	0.757	none	none	–	–
ViT	9	0.893	512	256	0.918	0.917
ViT	12	0.975	64	2	0.936	0.920
BERT	3	0.748	none	512	–	0.897
BERT	6	0.757	none	512	–	0.928
BERT	9	0.792	none	none	–	–
BERT	12	0.799	512	none	0.904	–
DINOv2	3	0.590	none	none	–	–
DINOv2	6	0.668	none	none	–	–
DINOv2	9	0.832	none	none	–	–
DINOv2	12	0.982	64	2	0.920	0.901

Spectral definitions and scope. For the retained numerical singular values s_j of the centered fit representations, let $r_j = s_j / \sum_k s_k$ and $v_j = s_j^2 / \sum_k s_k^2$. We report a centered adaptation of RankMe, $\exp(-\sum_j r_j \log r_j)$ (Garrido et al., 2023), covariance entropy effective rank $\exp(-\sum_j v_j \log v_j)$, and participation ratio $1/\sum_j v_j^2$. The adaptation discards singular values at or below the numerical rank tolerance and omits RankMe's additive epsilon. The tolerance is machine precision times the largest singular value times the larger matrix dimension. These are distinct spectral summaries. The slope regresses log covariance eigenvalues on log rank for the first 128 values. Saved spectra reproduce every reported scalar. The original correlation rule requires at least three uncensored depths for each model and target. None of the 36 model/target/metric combinations meets that rule, so no correlation coefficient is reported. Usable/censored depth counts for $(q^*_{\text{RP}}, q^*_{\text{adapt}}, G)$ are $(2/2, 2/2, 2/2)$ for ViT, $(1/3, 2/2, 0/4)$ for BERT/SST-2 and $(1/3, 1/3, 1/3)$ for DINOv2. Here G is grid halvings as defined in Table 51. No correlation significance or predictive accuracy is claimed.

Table 60: Centered fit-spectrum summaries at the four fixed depths.

Model	Block	Centered RankMe	Cov. entropy	Participation	Slope
ViT	3	150.3	13.8	7.4	−1.957
ViT	6	275.5	31.1	11.1	−1.552
ViT	9	480.4	86.3	21.9	−1.107
ViT	12	543.7	177.7	60.8	−0.871
BERT / SST-2	3	339.2	12.2	3.3	−1.245
BERT / SST-2	6	424.7	42.9	8.2	−1.122
BERT / SST-2	9	479.2	85.6	23.1	−1.106
BERT / SST-2	12	510.1	103.4	23.7	−1.025
DINOv2	3	69.5	9.6	4.9	−2.394
DINOv2	6	128.5	21.2	10.1	−1.790
DINOv2	9	211.2	44.5	17.3	−1.370
DINOv2	12	316.7	170.2	84.0	−0.746

Depth uncertainty differs from the core diagnostic. Depth tables include 2,000 paired-example bootstrap intervals for a single original frozen/refit map. They exclude compressor variation and provide no coverage for selected dimension, encoder training, or split randomness. Other compressor draws are separate descriptive checks. The original-core matched-effect intervals instead exclude stochastic seed zero and resample remaining compressors plus paired examples. These distinct scopes are not pooled.

Table 61: Conditional paired-example intervals for frozen depth choices.

Model	Block	Family	Metric	Estimate	95% interval
ViT	9	rp	A	0.918	[0.900, 0.935]
ViT	9	rp	R	0.970	[0.952, 0.989]
ViT	9	adaptive	A	0.917	[0.900, 0.933]
ViT	9	adaptive	R	0.971	[0.953, 0.988]
ViT	12	rp	A	0.936	[0.920, 0.951]
ViT	12	rp	R	0.950	[0.935, 0.964]
ViT	12	adaptive	A	0.920	[0.903, 0.936]
ViT	12	adaptive	R	0.941	[0.923, 0.957]
BERT	3	adaptive	A	0.897	[0.877, 0.916]
BERT	3	adaptive	R	0.988	[0.958, 1.016]
BERT	6	adaptive	A	0.928	[0.909, 0.944]
BERT	6	adaptive	R	1.008	[0.984, 1.033]
BERT	12	rp	A	0.904	[0.884, 0.922]
BERT	12	rp	R	1.006	[0.981, 1.032]
DINOv2	12	rp	A	0.920	[0.903, 0.937]
DINOv2	12	rp	R	0.931	[0.914, 0.947]
DINOv2	12	adaptive	A	0.901	[0.882, 0.919]
DINOv2	12	adaptive	R	0.918	[0.899, 0.936]

Paired task contrast. The secondary coarse-minus-fine agreement comparison uses the coarse task validation-selected adaptive map at original seed zero, with the same map evaluated on fine labels. We resample 20 paired images within each of the 100 fine classes for 2,000 repetitions, preserving the evaluation sampling design. The pointwise percentile interval conditions on both task probes, the map, representation and selected dimension. It covers none of their fitting or selection variation. It is a secondary benchmark analysis, not a preregistered confirmation. The selected map is PCA at $q = 128$. Fine/coarse reference agreement is 0.882/0.901. Coarse-minus-fine agreement is $+0.0190$, with a conditional 95% interval $[+0.0030, +0.0355]$. This earlier adaptive-map comparison differs from the full-whitening/RP control in Appendix L.

CIFAR-100 sampling and control. The two labelings use the official CIFAR-100 hierarchy (Krizhevsky, 2009). Fifty training and twenty held-out images per fine class give 5,000 training-pool and 2,000 evaluation examples. The common fine-stratified split gives 4,000 fit and 1,000 validation examples. The native DINOv2 preprocessing uses 518-pixel input and post-normalization CLS features. Both task probes receive exactly the same map outputs for every phase, method, dimension, and compressor seed; the official label hierarchy is independently checked. The linear grid ends at 256 because the representation has 384 coordinates; nonlinear maps stop at 32. Fitting failures, if any, remain explicit.

Secondary probe sensitivity on fixed CIFAR-100 maps. We retain the primary PCA dimensions (fine: 256; coarse: 128) and fit standardized logistic probes with $C_{\mathrm{LR}} \in \{0.1, 1, 10\}$ and a one-versus-rest linear SVM with squared-hinge loss and $C = 1$. Each has its own full-dimensional same-class reference. All use the same frozen map outputs, examples and training-only scalers. This secondary analysis was specified after the primary results; it does not reselect dimensions or claim replication across trained encoders. Complete predictions also include PCA at both dimensions and RP at 256. The $C_{\mathrm{LR}} = 1$ rerun exactly reproduces the original predictions. Test pass/fail is descriptive at the original thresholds.

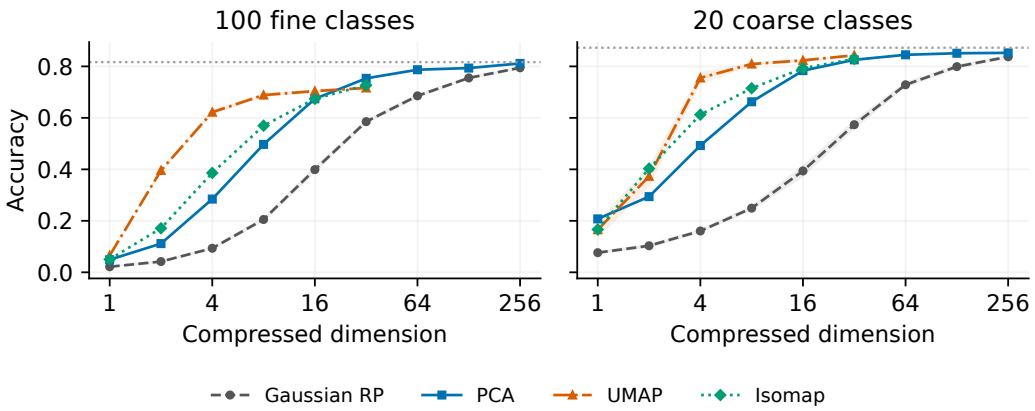


Figure 12: Label accuracy for the same fixed CIFAR-100 maps and tasks as Figure 11. Dotted lines give task-specific full-reference accuracy.

Table 62: Secondary probe sensitivity on the same CIFAR-100 maps.

Task	Probe	a_{full}	a	A	R
Fine	LR_0.1	0.828	0.819	0.914	0.989
Fine	LR_1	0.817	0.811	0.906	0.994
Fine	LR_10	0.817	0.801	0.905	0.982
Fine	LinearSVC_1	0.757	0.747	0.811	0.987
Coarse	LR_0.1	0.876	0.865	0.920	0.987
Coarse	LR_1	0.873	0.851	0.901	0.975
Coarse	LR_10	0.865	0.832	0.870	0.962
Coarse	LinearSVC_1	0.842	0.860	0.851	1.021

S FURTHER NUMERICAL AND FITTING-POLICY CHECKS

S.1 NUMERICAL FOLLOW-UP ON THE EARLIER PETS RANKING COMPARISON

This follow-up retains the earlier fixed $q = 64$ pair, $t = 4$ and $t = 16$. It uses the same maps, labels, reference and fit examples. The examples had already been evaluated. There are 62 distinct fits: two PCA fits and 60 RP fits. The PCA fits are reused at test; RP uses validation seeds 300–309 and test seeds 3000–3019. All 62 fits are refined, including the five without an original convergence warning.

We use fit-only QR preconditioning and retain the original objective: mean multiclass cross entropy plus $\|W\|_F^2/(2n)$, with an unpenalized intercept. The penalty is transported by triangular solves. Each fit starts from its saved coefficients and has one attempt, at most 15,000 iterations and 120 seconds. Evaluation scores do not control stopping.

All 62 fits finish without a warning and satisfy the declared 10^{-6} gradient threshold in both native and orthonormal map coordinates. None of the original fits meets this stricter threshold. The maximum native gradient falls from 3.88×10^{-3} to 6.26×10^{-7}. Every fit has a lower objective after refinement. Table 63 reports the complete endpoint means.

Split	t	Fit	PCA A	RP A	ΔA	PCA accuracy	RP accuracy
Validation	4	Original	0.80571	0.70380	$+0.10190$	0.79076	0.69973
Validation	4	Refined	0.80707	0.70204	$+0.10503$	0.79212	0.69796
Validation	16	Original	0.69158	0.71630	-0.02473	0.68071	0.71318
Validation	16	Refined	0.68478	0.71495	-0.03016	0.67527	0.71196
Test	4	Original	0.81100	0.67965	$+0.13135$	0.79200	0.66960
Test	4	Refined	0.81000	0.68025	$+0.12975$	0.79400	0.67030
Test	16	Original	0.71000	0.70505	$+0.00495$	0.69700	0.69490
Test	16	Refined	0.67300	0.70325	-0.03025	0.66900	0.69350

Table 63: All earlier Pets ranking endpoints and their numerical refinements at $q = 64$. A is agreement with the fixed reference and $\Delta A = A_{\mathrm{PCA}} - A_{\mathrm{RP}}$. RP means include all 10 validation or 20 test draws. Every endpoint remains below the behavioral qualification threshold. The validation and test sets contain 736 and 1,000 examples.

The test mean differences become $+0.12975$ and -0.03025, with the required opposite signs in 19 of 20 paired draws. The original crossed bootstrap rule still fails: the $t = 16$ interval includes zero (Table 64). Thus numerical stationarity does not make this comparison satisfy the full reversal criterion. This follow-up does not change the original prospective non-confirmation.

Fit	$t = 4$: ΔA [98.75% interval]	$t = 16$: ΔA [98.75% interval]	Joint signs
Original	$+0.13135\,[+0.10205, +0.16080]$	$+0.00495\,[-0.02890, +0.03895]$	8/20
Refined	$+0.12975\,[+0.09955, +0.15985]$	$-0.03025\,[-0.06495, +0.00545]$	19/20

Table 64: The original uncertainty rule applied to both sets of saved Pets predictions. The 20,000 crossed resamples use seed 20270913, shared example and draw resamples across endpoints, and four reserved endpoint slots. The two-model adjusted conditional binomial value is 1.0 for the original fits and 4.01×10^{-5} after refinement. Both sets fail the full criterion because the $t = 16$ interval includes zero.

S.2 SENSITIVITY TO POSTCOMPRESSION SCALING

Fixed comparisons. This follow-up uses the already evaluated $q = 4$ comparisons on CIFAR-10 with ViT-B/16 and DINOv2-small, and SST-2 with BERT-base. It keeps both the primary and secondary endpoint pairs, all maps, task labels, fit examples and fixed references. Shared endpoints are evaluated once. Each dataset uses 2,400 fit examples and 600 validation examples. The test sets contain 1,000 examples for each vision encoder and 872 for BERT. RP uses the original validation seeds 200–209 and test seeds 2000–2019. Each PCA fit is reused for validation and test.

Three fitting rules. Arm A replays the original finite fit with postcompression standardization. Arm B removes that standardizer and retains ordinary isotropic L2 with $C = 1$. Arm C removes the

standardizer and transports its objective. If $u = S^{-1}(x - \mu)$, then $W_{\text{raw}} = W_{\text{scaled}}S^{-1}$ and $b_{\text{raw}} = b_{\text{scaled}} - W_{\text{raw}}\mu$. The raw matched penalty is $\|W_{\text{raw}}S\|_F^2/(2Cn)$. Intercepts are unpenalized. The loss is mean multiclass cross entropy for vision and mean binomial cross entropy for SST-2, using its original single coefficient row. The saved fit-only scaler provides S and μ. No selected endpoint has a zero-variance fit column; a constant-coordinate synthetic check verifies the scale-one convention.

Each new fit starts from zero. Arm B retains the original L-BFGS settings. Arm C evaluates raw scores and optimizes in the corresponding scaled coefficient and intercept parameters. This controls the objective and uses common-coordinate preconditioning; it does not test unpreconditioned optimization in raw coordinates. Both use a gradient tolerance of 10^{-5}, at most 5,000 iterations and a 240-second fit budget, with no retries. The panel contains 248 original fits and 496 new fits, with 256 validation/test evaluations per arm.

Table 65: Fixed $q = 4$ test endpoints for the scaling follow-up. A uses the original postcompression standardizer; B uses no standardizer and ordinary isotropic L2. Each agreement value for RP averages 20 original paired maps. P/R gives the PCA pass indicator and number of RP draws passing both behavioral criteria. The independently fitted matched arm C has the same class predictions and behavioral values as A in every validation and test evaluation. All vision PCA endpoints remain below the agreement threshold.

Model / task	t	PCA A	RP A	PCA B	RP B	P/R A	P/R B
ViT / CIFAR-10	0.03	0.8150	0.9027	0.8130	0.9039	no, 11/20	no, 13/20
ViT / CIFAR-10	0.1	0.8200	0.8519	0.8130	0.8508	no, 1/20	no, 1/20
ViT / CIFAR-10	0.3	0.8150	0.6416	0.8210	0.6413	no, 0/20	no, 0/20
DINOv2 / CIFAR-10	0.03	0.8200	0.8745	0.8250	0.8772	no, 3/20	no, 6/20
DINOv2 / CIFAR-10	0.3	0.8280	0.5952	0.8350	0.5958	no, 0/20	no, 0/20
BERT / SST-2	0.1	0.9656	0.6961	0.9656	0.6960	yes, 0/20	yes, 0/20
BERT / SST-2	0.3	0.6411	0.6069	0.6411	0.6069	no, 0/20	no, 0/20
BERT / SST-2	1	0.5883	0.5830	0.5894	0.5831	no, 0/20	no, 0/20

Paired outcomes. Without postcompression scaling, the primary PCA-minus-RP agreement differences are -9.095 and $+17.970$ percentage points for ViT, and -5.225 and $+23.920$ points for DINOv2. Both retain the original opposite signs in all 20 paired test draws. BERT retains its negative result: the two differences are $+26.955$ and $+0.636$ points, with only 4/20 original joint signs. Its original arm has 5/20. In the secondary ViT pair, the differences are -3.775 and $+17.970$ points, with 17/20 joint signs rather than the original 16/20. The secondary BERT pair has $+26.955$ and $+3.412$ points and 3/20 joint signs. The secondary DINOv2 pair duplicates its primary pair. These descriptive follow-ups do not replace the original confirmation decisions or intervals.

Table 66: Paired $q = 4$ test comparisons for every original primary and secondary pair. Entries are PCA-minus-mean-RP agreement in percentage points. A replays the original scaled fit; B removes scaling with ordinary isotropic L2; C removes scaling with the transported penalty and independently fits in common coordinates. Joint counts retain the original two sign directions. No new intervals or confirmation decisions are assigned to these exposed-example follow-ups. P and S denote the primary and secondary pairs. The secondary DINOv2 pair duplicates the primary pair.

Model / task	Pair	Arm	First endpoint	Second endpoint	Joint signs
ViT / CIFAR-10	0.03, 0.3 (P)	A	-8.775	$+17.345$	20/20
ViT / CIFAR-10	0.03, 0.3 (P)	B	-9.095	$+17.970$	20/20
ViT / CIFAR-10	0.03, 0.3 (P)	C	-8.775	$+17.345$	20/20
ViT / CIFAR-10	0.1, 0.3 (S)	A	-3.185	$+17.345$	16/20
ViT / CIFAR-10	0.1, 0.3 (S)	B	-3.775	$+17.970$	17/20
ViT / CIFAR-10	0.1, 0.3 (S)	C	-3.185	$+17.345$	16/20
DINOv2 / CIFAR-10	0.03, 0.3 (P)	A	-5.450	$+23.275$	20/20
DINOv2 / CIFAR-10	0.03, 0.3 (P)	B	-5.225	$+23.920$	20/20
DINOv2 / CIFAR-10	0.03, 0.3 (P)	C	-5.450	$+23.275$	20/20
DINOv2 / CIFAR-10	0.03, 0.3 (S)	A	-5.450	$+23.275$	20/20
DINOv2 / CIFAR-10	0.03, 0.3 (S)	B	-5.225	$+23.920$	20/20
DINOv2 / CIFAR-10	0.03, 0.3 (S)	C	-5.450	$+23.275$	20/20
BERT / SST-2	0.1, 1 (P)	A	$+26.950$	$+0.533$	5/20
BERT / SST-2	0.1, 1 (P)	B	$+26.955$	$+0.636$	4/20
BERT / SST-2	0.1, 1 (P)	C	$+26.950$	$+0.533$	5/20
BERT / SST-2	0.1, 0.3 (S)	A	$+26.950$	$+3.412$	3/20
BERT / SST-2	0.1, 0.3 (S)	B	$+26.955$	$+3.412$	3/20
BERT / SST-2	0.1, 0.3 (S)	C	$+26.950$	$+3.412$	3/20

Numerical checks. All 248 fits in each arm have common-coordinate gradient infinity norm below 10^{-5}. The maximum norms are 9.98×10^{-6} for A and C and 9.99×10^{-6} for B; B also meets the tolerance in its raw solver coordinates. There are no fit failures or convergence warnings. Direct transport of the original parameters gives maximum score error 1.07×10^{-14}, objective error 2.22×10^{-16} and gradient error 6.05×10^{-17}. The independently fitted C arm changes no class predictions across all 256 evaluations. Its maximum logit discrepancy from A is 8.63×10^{-8}, its maximum probability discrepancy is 8.54×10^{-9} and its maximum objective difference is 1.16×10^{-13}. Finite numerical agreement is not a theorem about arbitrary solver runs.

Interpretation. The observed vision rankings persist without the standardizer for these frozen endpoints and draws. Removing it changes the effective regularization, as shown by the changed behavioral values. Matching the penalty and intercept recovers the original outcomes to the reported numerical precision. These results isolate a fitting-policy sensitivity while retaining the BERT negative case and subthreshold vision PCA endpoints. They do not establish unchanged PCA eigenspaces, identify a unique mechanism for the RP agreement drop, or provide new independent confirmation. The supplement includes every validation/test endpoint and paired draw result.