Whitening Inverts the Hierarchy: What the Norm of a Whitened Embedding Measures

Anonymous

Abstract

Whitening a foundation-model embedding and reading off its squared norm has become a training-free likelihood surrogate, justified by a single observation: the whitened coordinates are close to standard normal. We show that this observation is guaranteed by the projection central limit theorem, so it says nothing about the joint law, and that the joint law is not Gaussian. Against clones with identical mean and covariance the whitened radius is over-dispersed, for every encoder we tested and across three training objectives, and the published agreement between empirical and theoretical norm statistics is an algebraic identity that in-sample whitening imposes on any distribution.Whitening divides every direction by its eigenvalue and so inverts the encoder's hierarchy, moving the mass of the norm off the semantic subspace and onto the near-degenerate directions the encoder treats as noise, where a single input-dependent scale sets the variance; we identify that scale from two moments and confirm it by predicting, with no free parameter, how strongly disjoint halves of the spectrum move together. The squared whitened norm is therefore a Mahalanobis estimate of semantic atypicality rather than a log-likelihood, and that identification explains both its successes and its failures: it ranks and detects well, agreeing with a nonparametric density estimate and with encoders that share no objective, training data or supervision, while its Gaussian thresholds are wrong by orders of magnitude. We prove that a contrastive objective built on cosine similarity cannot determine the radial law, derive in closed form the group size at which a coordinatewise test can detect a given kurtosis, and give the one-dimensional recalibration that restores nominal tail probabilities.

1 Introduction

Let f be an encoder and $z = f(x) \in \mathbb{R}^d$ a raw, unnormalised embedding. Write μ and Σ for its mean and covariance over a corpus, and

$$w \;=\; \Sigma^{-1/2}(z - \mu) \tag{1}$$

for the whitened embedding. A recent programme (Levi and Gilboa, 2025; Betser et al., 2025, 2026a; Tasker et al., 2026) observes that the coordinates of w, and more generally its low-dimensional projections, are close to standard normal (Diaconis and Freedman, 1984; Sudakov, 1978; Klartag, 2007), and draws a practical conclusion: since $-\log \phi(w) = \frac{1}{2}\|w\|^2 + \frac{d}{2}\log 2\pi$, the squared whitened norm is a training-free likelihood surrogate. It is now used for out-of-distribution scoring, caption quality, and the detection of generated images (Betser et al., 2026b) and video (Ben Hayun et al., 2026). The observation is correct and the score is useful. What has been missing is the joint law, and the joint law turns out to be more informative than the Gaussian would have been: the whitened norm is a density estimate, and knowing what it estimates explains both where it succeeds and where it fails.

Our route is to take the published evidence seriously enough to reproduce it. We embedded $5{,}000$ MS-COCO val2017 image-caption pairs (Lin et al., 2014) with CLIP ViT-L/14 (Radford et al., 2021;

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.

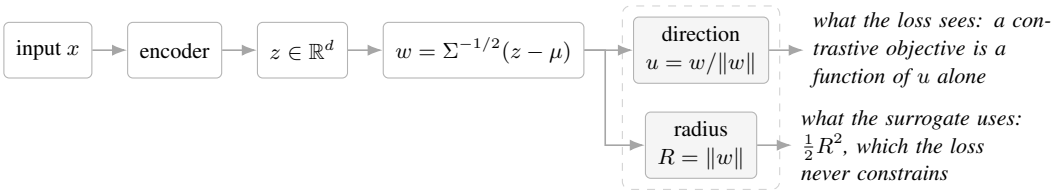


Figure 1: The polar decomposition of a whitened embedding, and the division this paper is about. The directional factor is what a contrastive loss optimises; the radial factor is the whole of the likelihood surrogate, and Theorem 5 shows the objective cannot determine it.

Ilharco et al., 2021) and reproduced ten of the twelve published summary statistics to within five percent, which fixes the protocol precisely, down to the basis in which the coordinates were tested. From there the paper establishes four things. The published agreement between empirical and theoretical norm statistics is an identity that in-sample whitening imposes on any distribution; evaluated against the genuine χ_d prediction the same table is off by $5.57\times$ and $8.09\times$, and the published grand averages sit $+62$ and $+126$ standard deviations from the null of the protocol itself (Section 2). Against clones with identical mean and covariance the joint law is neither Gaussian nor spherical, at 8 to 64 standard deviations, for all six encoders we tested (Section 3). The reason is that whitening inverts the encoder's hierarchy, moving the share of the norm carried by the k leading directions from 66-69 percent down to k/d, or 9-27 percent: the norm measures the near-degenerate tail, where one per-input scale, identified from two moments, sets the variance (Section 4). That scale is semantic atypicality, $\rho = 0.80$ against a nonparametric density estimate and 0.62 shared across encoders, and the Mahalanobis out-of-distribution score is its tail term to three decimals (Section 5). Section 6 proves that a contrastive objective built on cosine similarity cannot determine the radial law and gives the recalibration.

That contrastive training shapes the directional distribution is well understood (Wang and Isola, 2020; HaoChen et al., 2021; Balestriero and LeCun, 2022), as is the modality gap (Liang et al., 2022), and whitening as a fix for anisotropy predates this literature in NLP (Li et al., 2020; Su et al., 2021; Ethayarajh, 2019). Levi and Gilboa (2025) introduce *conformity*, the mean cosine similarity of an instance to the corpus, as a per-instance typicality measure on the unwhitened embedding. Conformity is directional and our per-input scale is radial, so we measure the relation rather than assume it: conformity tracks the *raw* norm closely ($\rho = +0.86$ on CLIP ViT-L/14) and runs *against* the whitened radius (-0.66), the same inversion Section 5 reports. On the detection side, Mahalanobis distance on deep features (Lee et al., 2018) and k-nearest-neighbour distance (Sun et al., 2022) are both established out-of-distribution scores. What is new here is an account of why they rank alike and calibrate differently, and a demonstration that the Mahalanobis score is carried entirely by the spectral tail.

2 Whitening and the published evidence

Given N embeddings, $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance and $w_i = \hat{\Sigma}^{-1/2}(z_i - \hat{\mu})$. When $\hat{\Sigma}$ is estimated in sample, three identities hold exactly,

$$\frac{1}{N}\sum_i w_i = 0, \qquad \frac{1}{N}\sum_i w_i w_i^\top = \mathrm{I}_d, \qquad \frac{1}{N}\sum_i \|w_i\|^2 = d, \tag{2}$$

and in-sample whitening also compresses the radius: a standard Gaussian sample of $N = 5{,}000$ points in $d = 768$ dimensions has radial standard deviation 0.649 under this protocol rather than the population

value 0.707 of χ_d, because d/N of the variance directions are fitted away. Every comparison below is against a covariance-matched clone through the identical pipeline, so this and every other pipeline effect cancels. In polar coordinates $w = R\,u$ with $R = \|w\|$ and $u \in \mathbb{S}^{d-1}$, a spherically symmetric law with radial density p_R has

$$f(w) = \frac{p_R(r)}{S_{d-1}\, r^{\,d-1}}, \qquad S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \tag{3}$$
$$-\log f(w) = -\log p_R(r) + (d-1)\log r + \log S_{d-1},$$

and the Gaussian is the case $p_R = \chi_d$, for which this collapses to $\frac{1}{2}r^2 + \frac{d}{2}\log 2\pi$. Equivalently, $\mathcal{N}(0, \mathrm{I}_d)$ places almost all its mass near the sphere of radius $\sqrt{d}$: $\mathbb{E}\|W\| \sim \sqrt{d}$ with fluctuations of constant order, so the relative radial width is $O(d^{-1/2})$ (Vershynin, 2018, Thm. 3.1.1 and Rem. 3.1.2). That is the thin spherical shell of a high-dimensional Gaussian; whether whitened embeddings concentrate the same way is measured below. Two consequences of eq. (2) govern what the published protocol could and could not see.

Lemma 1 (in-sample whitening)**.** *Let $w_i = R_i u_i$ be in-sample whitened, with $\hat{s}^2 = \frac{1}{N}\sum_i (R_i - \hat{m})^2$. Then (a) for any distribution $\hat{s}^2 = d - \hat{m}^2$; and (b) $\frac{1}{N}\sum_i R_i^2 \langle u_i, v\rangle^2 = 1$ for every unit vector v. In particular the radius-weighted directional energy in (b) is constant in v by construction, so it carries no information about a radius-direction coupling at any sample size.*

Proof. Claim (a) is the third identity of eq. (2), $\frac{1}{N}\sum_i R_i^2 = d$, combined with $\hat{s}^2 = \frac{1}{N}\sum_i R_i^2 - \hat{m}^2$; claim (b) is the second identity, $v^\top \big(\frac{1}{N}\sum_i w_i w_i^\top\big) v = 1$, written in polar form. □

Betser et al. (2025) report empirical and theoretical norm statistics side by side, from their Eq. (8),

$$\mathbb{E}[S] = \sqrt{2}\,\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}, \qquad \mathrm{Std}(S) = \sqrt{d - \mu_S^2}\,, \tag{4}$$

which are the correct χ_d formulas and give $\mathbb{E}[S] = 27.7038$ and $\mathrm{Std}(S) = 0.7070$ at $d = 768$. The standard deviation depends on d alone, so any two rows sharing d must share it, and the published rows do not: the theoretical standard deviation is 3.96 for images and 6.60 for captions, while the theoretical mean is 27.7 in both rows, correctly depending on d alone. One formula of eq. (4) was evaluated theoretically and the other was not. Substituting the empirical mean $\hat{m}$ for μ_S in the second reproduces both entries and both reported deviations:

	$\hat{m}$	$\sqrt{\lvert d - \hat{m}^2\rvert}$	published "theory"	published deviation		recomputed
				mean	sd	sd
images	27.43	$\sqrt{+15.60} = 3.949$	3.96	0.98%	0.55%	0.23%
captions	28.49	$\sqrt{-43.68} = 6.609$	6.60	2.85%	13.24%	13.45%

By Lemma 1(a), $\sqrt{d - \hat{m}^2}$ is the sample standard deviation recomputed from the sample mean, so its agreement with the sample standard deviation is arithmetic and would hold for uniform vectors, for a scale mixture, for anything. The caption row shows the substitution a second way: there $d - \hat{m}^2 = -43.68$ is negative, so the published entry can only be $\sqrt{\lvert d - \hat{m}^2\rvert}$, and eq. (4) read literally returns an imaginary standard deviation. Against the genuine prediction the same table reads very differently. The χ_{768} standard deviation is 0.7070; the reported values 3.94 and 5.72 exceed it by 5.57 and 8.09 times. The formula is

Table 1: Published summary statistics against the same statistics on our own embeddings, in-sample whitening at the published group size 250. "Gaussian" is an exact Gaussian of the same shape through the same pipeline, its radial sd the mean of 30 draws, spread ± 0.006.

	images			captions		
	published	measured	Gaussian	published	measured	Gaussian
radial mean	27.43	27.43	27.7038	28.49	27.07	27.7038
radial sd	3.94	3.92	0.649	5.72	5.18	0.649
AD pass	98.3%	98.7%	100.0%	90.1%	92.3%	100.0%
mean A^2	0.4890	0.4951	0.3885	0.5926	0.5802	0.3885
DP pass	99.3%	99.3%	100.0%	99.2%	99.2%	100.0%
mean p	0.3624	0.3630	0.5066	0.2568	0.2741	0.5066

right and the evaluation is not, and the lesson is the protocol adopted throughout this paper: a distributional claim needs a matched null.

The protocol splits the $d = 768$ whitened coordinates of $5{,}000$ embeddings into 20 groups of 250 and submits each coordinate of each group to an Anderson-Darling (Anderson and Darling, 1954) and a D'Agostino-Pearson (D'Agostino and Pearson, 1973) test at the estimated-parameter critical value $A^2_{\text{crit}} = 0.752$ (Stephens, 1986, Table 4.7). Run on our embeddings (Table 1), every image statistic reproduces to within 1.2% and all three caption test statistics to within 2.4%, which fixes the protocol down to the basis: principal axes give a mean Anderson-Darling statistic of 0.4951 against the published 0.4890, the symmetric square root $\hat{\Sigma}^{-1/2}$ gives 0.4366. The caption radial law is the one row that does not reproduce, low by 5.0% on the mean and 9.4% on the standard deviation, and that row fails the identity that fixes the rest. In-sample whitening forces $\hat{m}^2 + \hat{s}^2 = d$; ours gives 759.8 against 768 (-1.07%, the compression noted above), the published row 844.4 ($+9.95\%$), which no in-sample whitening of a single sample can produce. It is the same row that needs a negative number under the square root.

Table 2: Published grand averages against the null law of the protocol itself, obtained by simulating the entire protocol on exact Gaussians 24 times, with our measured values alongside.

statistic	modality	published	exact Gaussian null	z	measured
mean Anderson–Darling A^2	images	0.4890	0.3891 ± 0.0016	$+62$	0.4951
	captions	0.5926	0.3891 ± 0.0016	$+126$	0.5802
mean D'Agostino–Pearson p	images	0.3624	0.5052 ± 0.0021	-67	0.3630
	captions	0.2568	0.5052 ± 0.0021	-117	0.2741

Averaging over 768 coordinates and 20 groups makes the sampling error of the protocol's grand averages very small, and Table 2 shows what the published averages say once that null is in hand: the published statistics reject normality at $+62$ and $+126$ standard deviations. The pass rates are outside the null too, degenerately, since an exact Gaussian passes 100.0% of coordinates on both tests in all 24 replications. How a 98% pass rate coexists with a 62σ rejection is a matter of group size, and the answer is exact.

Proposition 2 (power of a coordinatewise test)**.** *Let $X = \sqrt{S}\,G$ with $G \sim \mathcal{N}(0,1)$ independent of $S > 0$, $\mathbb{E}S = 1$, $\mathrm{Var}S = V$, so that X has unit variance and excess kurtosis $\gamma_2 = 3V$. As $V \to 0$ with $\mathbb{E}|S-1|^3 = o(V)$, the Anderson-Darling discrepancy of X from the normal, $\Delta^2 = \int (F-\Phi)^2/(\Phi(1-\Phi))\,\mathrm{d}\Phi$, satisfies*

$\Delta^2 = \kappa V^2 + o(V^2)$ *with*

$$\kappa \;=\; \frac{1}{64}\int_{\mathbb{R}} \frac{x^2(3-x^2)^2\,\phi(x)^3}{\Phi(x)\,(1-\Phi(x))}\,\mathrm{d}x \;=\; 0.01508. \tag{5}$$

The proof is in Appendix A.

To turn κ into a group size we use the growth of the mean statistic under a fixed alternative. Since $A_n^2/n = \int (F_n - F_0)^2/[F_0(1-F_0)]\,\mathrm{d}F_0$ and $F_n \to F$ uniformly almost surely, Glivenko-Cantelli gives $A_n^2/n \to \Delta^2$ almost surely whenever Δ^2 is finite, which in the perturbative regime is part of Proposition 2; so $\mathbb{E}[A_n^2] \;\simeq\; \mu_0 + n\Delta^2$, with μ_0 the null mean and the asymptotic null theory of the statistic that of Anderson and Darling (1952). The critical value $A^2_{\mathrm{crit}} = 0.752$ is the estimated-parameter one (Stephens, 1986, Table 4.7). We take $\mu_0 = 0.3891$, simulated for this protocol rather than read from a table, because the protocol averages over groups. Solving for n at $V = \gamma_2/3$,

$$n^\star(\gamma_2) \;\approx\; \frac{9\,(A^2_{\mathrm{crit}} - \mu_0)}{\kappa\,\gamma_2^2} \;=\; \frac{216}{\gamma_2^2}. \tag{6}$$

Equation (6) composes two expansions taken in different limits ($n \to \infty$ at fixed Δ^2, and $V \to 0$ at fixed n), so it is calibrated rather than derived, and we measure its accuracy. A coordinate's mean statistic stays below the critical value at group size n when its kurtosis is below $3\sqrt{(A^2_{\mathrm{crit}} - \mu_0)/(\kappa n)}$, which is 0.93 at $n = 250$ and 0.21 at $n = 5{,}000$. The whitened image coordinates have per-coordinate kurtosis of mean $+0.441$ and root mean square 0.576 in principal axes, the captions $+0.658$ and 0.749, so the group size at which the average coordinate fails is $n^\star = 1{,}113$ and 499, 4.5 and 2.0 times the 250 used. Applying the threshold to the measured kurtosis of each coordinate predicts the pass rate at every group size:

group size	threshold $\lvert\gamma_2\rvert$	images		captions	
		predicted	measured	predicted	measured
125	1.32	97.3%	99.2%	96.6%	99.2%
250	0.93	89.1%	98.7%	81.0%	92.3%
500	0.66	79.0%	90.8%	49.5%	57.3%
1000	0.47	64.6%	71.1%	32.7%	31.0%
2500	0.29	40.0%	40.2%	13.9%	10.8%
5000	0.21	27.0%	26.4%	6.0%	2.6%

The leading-order formula under-predicts the pass rate below $n = 1000$, where the truncation of the Anderson-Darling weight at the resolution of the extreme order statistics lowers the effective discrepancy, and from $n = 2500$ it is accurate to within three and a half points. At $n = 5{,}000$ the published evidence collapses: tested in one group rather than twenty, the same $5{,}000$ embeddings pass at 26.4% and 2.6% while the Gaussian control passes at 96.4%. The published protocol was therefore run at a group size where eq. (6) predicts a pass, and its grand averages, which carry the information the pass rates discard, reject.

3 The joint law: neither Gaussian nor spherical

Lemma 3 (polar characterisation; classical)**.** *For $W \in \mathbb{R}^d$ with $W \neq 0$ a.s., $R = \|W\|$, $U = W/R$: (i) W is spherically symmetric iff $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $U \perp R$; (ii) $W \sim \mathcal{N}(0, \mathrm{I}_d)$ iff additionally $R \sim \chi_d$. (see Fang et al., 1990, §2.1)*

The lemma has two hypotheses, and each can fail on its own. A *scale mixture* $W = S\,G$, with $S > 0$ independent of $G \sim \mathcal{N}(0, \mathrm{I}_d)$, keeps U uniform but breaks $U \perp R$'s consequence for the radial law: it is spherical and not Gaussian, so $\|w\|^2$ still ranks by likelihood and only the calibration is wrong. A *sign-type law* $W_i = \varepsilon_i |V_i|$, with ε Rademacher independent of an equicorrelated V, has exactly $\mathcal{N}(0,1)$ coordinates, correlated squares, an over-dispersed radius, and is not spherical, and there $\|w\|^2$ is not a likelihood ranking at all. By Lemma 3 these are the two ways out, one per hypothesis, and there is no third: a rotation-invariant law on $\mathbb{R}^2$ with independent coordinates is Gaussian by Maxwell's theorem (Bryc, 1995, Thm. 0.0.1), applied here to each coordinate pair.

Neither is visible to a coordinatewise test, but for different reasons, and the difference matters. The sign-type law is invisible *exactly*: its marginals are standard normal, so no test of any power rejects them. The scale mixture is invisible only *to the protocol as run*: its coordinates are not normal, and by Proposition 2 their excess kurtosis $3V$ becomes detectable at group size $n^\star(3V)$, which at the measured spreads exceeds the 250 used. Coordinatewise normality is thus evidence against neither.

Deciding between them needs a control matching the data's covariance, since a spiked spectrum estimated from a finite reference half is noisier than an isotropic one. We build two clones of each dataset: a Gaussian clone with the same mean and covariance, and a spherical clone with the same mean and covariance carrying the data's own whitened radial law on uniform directions. Data and clones go through an identical split-half pipeline with principal axes from the reference half, so nothing is fitted on the data being scored, which Lemma 1(b) makes necessary: in-sample whitening puts the coupling at 0.02 to 0.06 where split-half whitening of the same embeddings gives -0.4533 ± 0.0229 and -0.2698 ± 0.0122.

Table 3: Data against covariance-matched clones; split-half whitening, axes from the reference half, five splits. A spherical law has the same coordinate kurtosis in every basis and a radius independent of direction. Both clones do; the data does neither.

	statistic	data	Gaussian clone	spherical clone
images	radial sd	5.99	1.02	4.78
	kurtosis, principal axes	$+0.5423 \pm 0.0516$	-0.0020 ± 0.0028	$+0.2514 \pm 0.0149$
	kurtosis, encoder's own axes	$+0.0216 \pm 0.0031$	$+0.0025 \pm 0.0045$	$+0.2476 \pm 0.0132$
	$\mathrm{corr}(R, \lvert\langle u, v_1\rangle\rvert)$	-0.4533 ± 0.0229	$+0.0036 \pm 0.0110$	-0.0005 ± 0.0179
captions	radial sd	8.09	1.01	6.39
	kurtosis, principal axes	$+0.9532 \pm 0.0609$	-0.0024 ± 0.0026	$+0.4354 \pm 0.0136$
	kurtosis, encoder's own axes	-0.0116 ± 0.0046	-0.0026 ± 0.0053	$+0.4274 \pm 0.0186$
	$\mathrm{corr}(R, \lvert\langle u, v_1\rangle\rvert)$	-0.2698 ± 0.0122	$+0.0050 \pm 0.0123$	-0.0183 ± 0.0180

Two properties in Table 3 are impossible for a spherical law. The kurtosis is basis-dependent: both clones show the same coordinate kurtosis in the encoder's own axes as in the principal axes, because a spherical excess is isotropic, whereas the data has coordinates indistinguishable from normal in the encoder's basis and strongly leptokurtic in principal axes, $-16.6\,\sigma$ and $-22.9\,\sigma$ from the spherical clone. And the radius depends on the direction: under sphericity $R \perp U$, yet alignment with the reference half's leading principal axis predicts the radius at -0.4533 ± 0.0229 for images against $+0.0036 \pm 0.0110$ for the Gaussian clone, -18.0 and -15.6 standard deviations away.

Table 4 repeats the battery on six encoders. Every row gives the same verdict, the weakest departure across the six standing at $8\,\sigma$ on kurtosis anisotropy, $16\,\sigma$ on the coupling and $18\,\sigma$ on over-dispersion. This is not a property of CLIP, of InfoNCE, or of image-text training: it holds for a sigmoid pairwise loss and for a self-distillation objective with no text at all. The over-dispersion spans only a factor of

Table 4: Six encoders, the same $5{,}000$ images, the same battery. "raw cv" is the coefficient of variation of $\|z\|$ before whitening; "over" is the whitened radial sd divided by a Gaussian clone's under the split-half protocol of Table 3, which is why CLIP-L/14 reads $5.90\times$ here and $6.04\times$ in Table 1.

encoder	d	ends in	raw cv	over	γ_2 raw	γ_2 PC	$\mathrm{corr}(R, \lvert u_1 \rvert)$
CLIP ViT-L/14, openai	768	linear projection	0.0444	5.90×	+0.022	+0.542	-0.453
CLIP ViT-B/32, openai	512	linear projection	0.0528	6.30×	+0.009	+1.523	-0.435
CLIP ViT-B/32, LAION-2B	512	linear projection	0.0505	6.64×	-0.008	+1.438	-0.421
CLIP ViT-B/32, DataComp	512	linear projection	0.0442	6.62×	-0.012	+1.494	-0.413
SigLIP ViT-B/16, WebLI	768	attention pool	0.0710	7.55×	+0.039	+0.922	-0.512
DINOv2 ViT-B/14	768	LayerNorm	0.0199	6.42×	+0.007	+0.687	-0.505

1.28 across the six ($5.90\times$ to $7.55\times$) while the pre-whitening norm dispersion spans 3.56, and the two are uncorrelated ($\rho = +0.37$, $p = 0.47$), a dissociation the next two sections explain.

4 The mechanism: whitening inverts the hierarchy

Write the embedding as a low-rank semantic head plus a residual whose scale depends on the input,

$$z \;=\; A\,c(x) \;+\; \sigma(x)\,\varepsilon, \tag{7}$$

with $A \in \mathbb{R}^{d\times k}$ carrying the leading eigenvalues, k the effective rank, and ε isotropic. Measured effective ranks are 95 of 768 for CLIP ViT-L/14 and 205 for DINOv2, so $k \ll d$ throughout. The encoder's leading k directions carry 66 to 69% of the raw norm; whitening divides every direction by its eigenvalue, and what that does to the norm is exact.

Proposition 4 (whitening inverts the hierarchy)**.** *Let $\Sigma = \sum_j \lambda_j v_j v_j^\top$ with $\lambda_1 \geq \cdots \geq \lambda_d$ and let P_k project onto* $\mathrm{span}(v_1, \ldots, v_k)$*. Then*

$$\frac{\mathbb{E}\|P_k(z-\mu)\|^2}{\mathbb{E}\|z-\mu\|^2} = \frac{\sum_{j\leq k}\lambda_j}{\sum_j \lambda_j}, \qquad \frac{\mathbb{E}\|P_k w\|^2}{\mathbb{E}\|w\|^2} = \frac{k}{d}, \tag{8}$$

and both hold as exact sample identities under in-sample whitening. Suppose further that the tail is conditionally centred *Gaussian given the input, $P_k^\perp w \mid x \sim \mathcal{N}(0, D_x)$, with $\mathbb{E}D_x = \mathrm{I}_m$ and $m = d - k$. Then the tail energy $T = \|P_k^\perp w\|^2$ has*

$$\begin{aligned} \mathbb{E}T = m, \qquad \mathrm{Var}T = \mathrm{Var}(\mathrm{tr}D_x) + 2\,\mathbb{E}\,\mathrm{tr}(D_x^2), \\ \mathrm{Cov}(T_1, T_2) = \mathrm{Cov}\big(\mathrm{tr}D_x^{(1)}, \mathrm{tr}D_x^{(2)}\big) + 2\,\mathbb{E}\,\mathrm{tr}\big(D_x^{(12)} D_x^{(21)}\big) \end{aligned} \tag{9}$$

for the energies T_1, T_2 of any two disjoint blocks of tail directions, where $D_x^{(a)}$ denotes the diagonal block of D_x on block a and $D_x^{(12)}, D_x^{(21)}$ its off-diagonal blocks; each tail coordinate j has excess kurtosis $3\mathrm{Var}\big((D_x)_{jj}\big)$. In particular, if $D_x = s(x)\mathrm{I}_m$ with $\mathbb{E}s = 1$ and $\mathrm{Var}s = V$, then relative to a Gaussian clone, and for two halves of the tail,

$$\frac{\mathrm{Var}T}{\mathrm{Var}_{\mathcal{N}}T} = 1 + V\Big(1 + \frac{m}{2}\Big), \qquad \mathrm{corr}(T_1, T_2) = \frac{V}{V + \frac{4}{m}(1+V)}, \qquad \gamma_2(w_j) = 3V, \tag{10}$$

while if the two halves carry scales s_1, s_2 with $\mathrm{corr}(s_1, s_2) = \rho_s$ the second expression is multiplied by ρ_s.

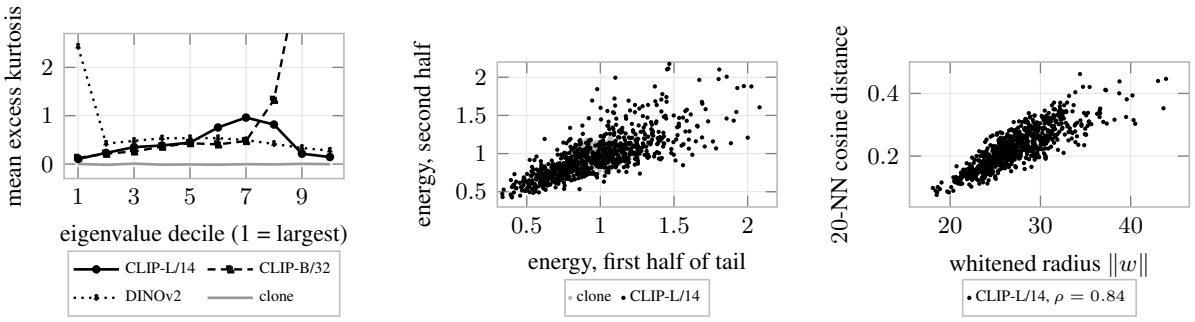
The proof is in Appendix A. The first display is the inversion: for CLIP ViT-L/14 the head share of the norm falls from 66% to 12%, and the $d-k$ directions the encoder treats as noise, amplified to unit variance each, come to dominate the norm by count. The second display says what the norm then measures, and the third identifies it. Because $m \approx 768 - 95$, the factor $m/4$ multiplies any per-input variation of the residual scale by more than a hundred, which is why a scale with a coefficient of variation of a third produces a radius over-dispersed sixfold while the coordinates, whose excess kurtosis is only $3V$, stay close to normal. The first two displayed moments are two equations in the two unknowns (V, ρ_s), so the measured over-dispersion and the measured halves correlation determine both, with no free parameter left over. Solving them jointly matters: reading V off the over-dispersion alone, as the single-scale formula invites, biases it low by the factor $(1+\rho_s)/2$, which is negligible at ρ_s near one but a factor 1.40 for CLIP ViT-B/32. On units: eq. (10) is a ratio of variances of the tail energy and Table 4 a ratio of radial standard deviations, which agree to first order by the delta method, $\mathrm{sd}(R) \approx \mathrm{sd}(T)/(2\sqrt{\mathbb{E}T})$. Table 5 reports the fit for five encoder-modality pairs.

Table 5: The inversion and the scale model of Proposition 4, in-sample PCA whitening. "head share" is the share of the norm carried by the top-k directions before and after whitening. "over" is the tail-energy variance relative to a Gaussian clone and "halves" the correlation between the energies of the two halves of the tail; these two measurements determine (V, ρ_s) jointly through eq. (10). "V, one scale" is what the over-dispersion alone would give under $\rho_s = 1$, shown to expose the bias. "tail γ_2" is the tail coordinate kurtosis, as implied by the fitted scale ($3V$) and as measured; the excess above $3V$ is scale variation within a half, which ρ_s does not capture. "tail var" is the share of $\mathrm{Var}(\|w\|^2)$ carried by the tail.

	k	head share		over	halves	V	ρ_s	V, one	tail γ_2		tail
		raw	whitened					scale	$3V$	measured	var
CLIP-L/14 images	95	66%	12%	42.5	+0.76	0.137	0.79	0.123	0.41	0.49	96%
CLIP-L/14 captions	67	69%	9%	73.9	+0.85	0.221	0.88	0.207	0.66	0.70	101%
CLIP-B/32 images	68	67%	13%	50.7	+0.41	0.312	0.42	0.223	0.94	1.37	99%
SigLIP images	73	66%	10%	64.7	+0.71	0.211	0.73	0.183	0.63	0.77	99%
DINOv2 images	205	69%	27%	46.4	+0.91	0.164	0.96	0.161	0.49	0.45	82%
Gaussian clone	94	n/a	n/a	1	-0.08	0	n/a	0	0	-0.01	91%

The fit is coherent in every row, and the two parameters separate cleanly. The identified V runs 0.14 to 0.31, a coefficient of variation of the per-input scale between a third and a half, and ρ_s runs 0.42 to 0.96: in every encoder most of the scale is common to the whole tail, and what is left is idiosyncratic to a half. The two appear separately in the coordinate kurtosis, which exceeds the $3V$ a common scale implies by the idiosyncratic part, so that across the five rows the ordering of the excess is the ordering of $1-\rho_s$. At ρ_s near one the measured tail kurtosis is close to $3V$ (DINOv2, $\rho_s = 0.96$, kurtosis 0.45 against 0.49), and at ρ_s small it is well above (CLIP ViT-B/32, $\rho_s = 0.42$, kurtosis 1.37 against 0.94). The tail carries 82-101% of $\mathrm{Var}(\|w\|^2)$, the head contributing about $2k$ as a Gaussian would, except for DINOv2, whose leading decile is itself heavy-tailed. Figure 2 shows the three facts that matter most: the non-Gaussianity lives where the eigenvalues are smallest and is absent from the clone; two disjoint halves of the tail move together at $+0.76$ against -0.08 for the clone; and the scale that moves them is semantic atypicality, the subject of the next section. Coordinates all but normal in one privileged basis, a radius over-dispersed sixfold, and dependence living between the squares is the sign-type law of Section 3, introduced there only to show that coordinatewise evidence cannot exclude non-sphericity; eq. (7) says why encoders produce it.

Equation (10) also says where the magnitude of the effect should vary: the factor $m/4$ grows with



(a) The tail, not the head. **(b)** One scale per input. **(c)** What the radius measures.

Figure 2: The mechanism in three panels. (a) Whitening amplifies the encoder's least-informative directions, and that is where the non-Gaussianity is; the Gaussian clone is flat at zero. (b) Two disjoint halves of the tail have energies that move together for real embeddings and not for the clone: one per-input scale multiplies the whole tail. (c) That scale is semantic atypicality.

the tail, so the lower the effective rank of an encoder's representation of a corpus, the more of the norm one scale multiplies. Betser et al. (2025) already report that the corpus used to estimate the whitening transform matters, their cross-dataset table showing the average Anderson-Darling statistic rising from 0.489 to 0.641 on images when a Flickr8k test set is whitened with a COCO transform. We encoded six corpora, COCO scenes, CIFAR-100, Oxford pets, Flowers-102, DTD textures and SVHN digits, at $2{,}500$ images per block in two disjoint blocks per corpus, so that the within-corpus spread provides a measured null; Oxford pets admits one block rather than two, giving 22 blocks in all.

Table 6: Over-dispersion by corpus, two encoders. The within-corpus figure is the standard deviation of a single block, estimated from disjoint pairs of the same distribution.

	COCO	CIFAR-100	pets	flowers	DTD	SVHN	within / between
CLIP ViT-B/32	5.54×	5.75×	6.72×	6.47×	6.02×	8.11×	0.105 / 0.929
DINOv2 ViT-B/14	6.18×	4.18×	6.08×	6.13×	7.82×	7.62×	0.149 / 1.312

Between-corpus spread is $8.8\times$ the within-corpus null for both encoders ($F = 156$ and 154 on $(5, 5)$ degrees of freedom, $p = 1.7 \times 10^{-5}$), yet the two encoders order the corpora differently ($\rho = +0.14$, $p = 0.79$): what matters is not the corpus but how low-rank that encoder's representation of that corpus is. Across all 22 encoder-corpus blocks, over-dispersion is predicted by the effective rank ($\rho = -0.60$, $p = 3.4 \times 10^{-3}$), by the strength of the shared scale ($+0.68$, $p = 5.6 \times 10^{-4}$) and by the dispersion of tail energy ($+0.50$, $p = 1.8 \times 10^{-2}$). Within CLIP alone, effective rank gives -0.92 and tail dispersion $+0.93$. Homogeneous corpora are the more over-dispersed, not the less, because a tight cluster has low effective rank: SVHN under CLIP has effective rank 9 of 512, so 98% of the whitened norm is tail and one scale multiplies nearly all of it.

5 What the norm measures

The mean cosine distance to an input's 20 nearest neighbours is a nonparametric density estimate that knows nothing about Gaussians. Against it, the whitened radius agrees at $\rho = 0.70$ to 0.93 within an encoder, 0.80 on average, and the consensus radius over six encoders agrees with the consensus distance

at 0.83. The raw norm runs the other way, -0.66 to -0.09: prototypical inputs have large raw norms, the familiar "norm as confidence" effect, and whitening inverts it. The quantity is a property of the input rather than of the encoder: six encoders rank the same $5{,}000$ images by whitened radius with mean cross-encoder agreement 0.62 (0.39-0.87), including 0.39 between CLIP ViT-L/14 and DINOv2, which share no objective, no training data and no text supervision, against 0.18 for the raw norm. Whitening strips what is encoder-specific and leaves what is input-specific.

Table 7: AUROC on a corpus-shift benchmark, averaged over five out-distributions, with the in-distribution fitted on one block of $2{,}500$ images and evaluated on a disjoint block. "tail" and "head" are the parts of $\|w\|^2$ beyond and within the effective rank. Mahalanobis^{++} is Mahalanobis after L^2 normalisation.

score	CLIP ViT-B/32		DINOv2 ViT-B/14	
	COCO in	CIFAR-100 in	COCO in	CIFAR-100 in
Mahalanobis, $\|w\|^2$	0.895	0.947	0.974	0.907
tail term only	0.896	0.949	0.973	0.916
head term only	0.414	0.801	0.167	0.212
Mahalanobis^{++}	0.853	0.944	0.973	0.954
k-NN cosine distance	0.688	0.895	0.957	0.939
raw norm	0.159	0.769	0.367	0.508

Table 7 is the mechanism on a benchmark the detection literature recognises. The Mahalanobis score is its tail term, agreeing to three decimals in all four settings, while the head term is at chance or, for DINOv2, strongly anti-informative, out-of-distribution inputs projecting less onto the in-distribution semantic directions. The whitened norm beats k-NN distance in three of the four settings, the parametric tail estimate being sharper than the local one, and its calibration on the same task is poor at the same time: at AUROC 0.974, a threshold set at the 1% in-distribution tail flags only 61% of out-of-distribution inputs, and 13% of them pass at 95% true-positive rate. L^2 normalisation before whitening, the Mahalanobis^{++} variant, does not act on the per-input scale: on the in-distribution test set the normalised and unnormalised scores agree at $\rho = 0.99$ and the tail-halves correlation is unchanged ($0.88 \to 0.88$), because the shared scale lives in direction space, which normalisation preserves. Where normalisation helps substantially (DINOv2 with CIFAR-100 in-distribution, $0.907 \to 0.954$) it is because the out-distribution's raw norms differ from the in-distribution's and corrupt the covariance fit; where there is no such shift it slightly hurts ($0.895 \to 0.853$). Its benefit is between distributions, not per input.

Two natural readings of the scale fail. The radius is uncorrelated with pixel-level complexity (-0.09 to -0.04 against JPEG bytes per pixel, permutation standard deviation 0.014), and against caption-based scene complexity it is positive but small ($+0.04$ to $+0.13$, real at 9 standard deviations, under 2% of variance). Atypicality here is semantic and relational, isolation on the representation manifold, not a property of the pixels.

6 The radial law is free, and the repair

Theorem 5 (the radial law is free)**.** *Let $\mathcal{L}$ be an InfoNCE loss built on cosine similarity between encoders f and g. (a) For any measurable $\lambda : \mathcal{X} \to (0,\infty)$ and $\kappa : \mathcal{Y} \to (0,\infty)$, $\mathcal{L}(\lambda f, \kappa g) = \mathcal{L}(f,g)$ exactly, for every batch and temperature. (b) If the law of x is atomless conditionally on $f(x)/\|f(x)\|$, then for every probability measure ν on $(0,\infty)$ there is a measurable λ such that λf attains the same loss as f on every batch and has radial law ν, independent of direction. (c) Let ρ be the law of the embedding with polar*

marginals $(\rho_U, \rho_{R|u})$, let σ be uniform on $\mathbb{S}^{d-1}$ and ν any law on $(0,\infty)$, write $\mathcal{L}(\rho)$ for the population loss, which by (a) depends on ρ only through ρ_U, and suppose a minimiser exists. Then for every $\beta > 0$ every minimiser of $\mathcal{L}(\rho) + \beta \mathrm{KL}(\rho \,\|\, \sigma \otimes \nu)$ has $R \sim \nu$ independent of U, whatever ρ_U is. The radial law is thus inherited from the reference measure rather than from $\mathcal{L}$: $\nu = \chi_d$ makes the radius χ_d, which is Gaussian exactly when the directional factor is uniform, and any other ν gives a non-Gaussian law with the same directional behaviour and the same value of the contrastive term.

Proof. (a) Each similarity depends on $f(x_i)$ only through $f(x_i)/\|f(x_i)\|$, unchanged by multiplication by $\lambda(x_i) > 0$, and the loss is a function of the similarity matrix alone. (b) Take $\mathcal{X}$ standard Borel and $f \neq 0$ a.s., and condition on $u = f(x)/\|f(x)\|$. Fix a Borel isomorphism $g : \mathcal{X} \to [0,1]$ and let $F_u(t) = \mathbb{P}(g(x) \leq t \mid u)$ be a regular conditional distribution function, jointly measurable in (u,t) by construction. Put $T_u(x) = F_u(g(x))$: the conditional law is atomless, so $T_u(x) \sim \mathrm{Unif}(0,1)$ given every u, and $(u,x) \mapsto T_u(x)$ is measurable. Set $\lambda(x) = F_\nu^{-1}(T_u(x))/\|f(x)\|$. Then $\|\lambda(x) f(x)\| \sim \nu$ conditionally on every u, hence unconditionally and independently of u, and equality of losses is (a). (c) By (a), $\mathcal{L}(\rho)$ depends on ρ only through ρ_U, and $\mathrm{KL}(\rho \,\|\, \sigma \otimes \nu) = \mathrm{KL}(\rho_U \| \sigma) + \mathbb{E}_{u\sim\rho_U}[\mathrm{KL}(\rho_{R|u} \| \nu)]$. The first term and $\mathcal{L}$ depend only on ρ_U; the second is non-negative and vanishes iff $\rho_{R|u} = \nu$ for ρ_U-a.e. u. □

A derivation concluding that a contrastive objective induces a Gaussian has therefore put the radial law into the reference measure. Betser et al. (2026a) prove Gaussianity of fixed-k marginals, their Corollary 1 giving $\sqrt{d}\, u_k \Rightarrow \mathcal{N}(0, \mathrm{I}_k)$ for every fixed k, a Maxwell-Poincaré statement with which our measurements agree: coordinate excess kurtosis in the encoders' own axes is $+0.022$ (Table 4). But $\|w\|^2$ is a function of all d coordinates, not of a fixed-k marginal, and marginal normality plus dependence does not give χ_d; the dependence of Figure 2(b) is what breaks it. Their Proposition 2, for unnormalised encoders, reaches $\sqrt{d}\, z_k \Rightarrow \mathcal{N}(0, r_0^2 \mathrm{I}_k)$ under an assumption they name thin-shell concentration, $r/r_0 \to 1$: the Gaussian shell property of Section 2, exactly the assumption Theorem 5(b) shows no objective can supply, and exactly what our measurement finds violated, the whitened radius being over-dispersed $6.04\times$ rather than concentrating. Their Theorem 1 takes the regularised route, with $\beta \mathrm{KL}(\rho \,\|\, \gamma_\lambda^B)$ against a Gaussian reference truncated to a ball, $\gamma_\lambda^B \propto e^{-\lambda\|z\|^2} \mathbf{1}_B$; that is Theorem 5(c) with ν the radial law of γ_λ^B, so the radial conclusion is the reference measure's. Their framing is equivalent and worth making explicit, since $\mathrm{KL}(\rho \,\|\, \gamma_\lambda)$ is up to constants a norm penalty minus an entropy: to regularise towards low feature norm and high feature entropy *is* to choose a Gaussian reference. Numerically, on a random batch the loss is 35.766749; after multiplying every embedding by an independent lognormal factor that changes the radial standard deviation from 0.67 to 140.4 it is 35.766749 again, to every printed digit. Nor does the training recipe or the output layer set the radial law empirically. Three ViT-B/32 encoders differing only in recipe have raw-norm dispersion differing by 17 standard errors and recall@1 differing by five points, yet statistically identical whitened over-dispersion ($\chi^2 = 0.96$ on two degrees of freedom, $p = 0.62$), and DINOv2, which ends in a LayerNorm and pins its raw norm more than twice as tightly as any CLIP, is nonetheless mid-range at $6.42\times$. Recall@1 runs 32.3% to 37.6% and the three pool to $6.49\times$; DINOv2's raw norm has $\mathrm{cv} = 0.0199$ against 0.0710 for the widest CLIP. Equation (7) explains the dissociation: whitening removes the global scale and the covariance and leaves a quantity that depends on the shape of the residual, on which a normalisation layer has no purchase.

Proposition 6 (rankings and recalibration)**.** *For a spherically symmetric law with radial density p_R: (a) if $|\frac{\mathrm{d}}{\mathrm{d}r} \log p_R(r)| < (d-1)/r$ on the range of interest, then $-\log f(w)$ is strictly increasing in $\|w\|$, so $\|w\|^2$ and the negative log-likelihood induce the same ordering and every rank-based quantity is identical; (b)*

the only correction to the surrogate is a reparameterisation of the radius,

$$s(w) = -\log \hat{p}_R(\|w\|) + (d-1)\log\|w\| + \textit{const}, \tag{11}$$

with $\hat{p}_R$ a univariate density estimate; if $\hat{p}_R$ satisfies the inequality of (a) then s is increasing in $\|w\|$ and every ranking is unchanged. For a twice-differentiable p_R a kernel estimate attains mean integrated squared error of order $n^{-4/5}$, a rate independent of d (Tsybakov, 2009, §1.2).

Proof. Differentiate eq. (3): $\frac{\mathrm{d}}{\mathrm{d}r}(-\log f) = -\frac{\mathrm{d}}{\mathrm{d}r}\log p_R + (d-1)/r > 0$ under (a), and rank statistics are invariant under strictly increasing transformations. For (b), eq. (3) shows that within the spherical family the density is determined by p_R, so the surrogate is wrong exactly by the univariate factor $p_R(r)/\chi_d(r)$, and a univariate kernel estimate attains the stated rate. □

The margin in (a) is a factor of 37 at the measured parameters, so within the spherical family nothing can reorder the scores; the family is wrong, so the question is quantitative. Conditioning the radial law on a single direction, the mildest available correction, moves the ranking to Spearman 0.836 (images) and 0.947 (captions), with 60% and 78% of the flagged top 1% surviving: aggregate rank-based results are not in danger. Thresholds are another matter.

Table 8: Actual exceedance of a threshold set from the Gaussian surrogate, and from the one-dimensional recalibration eq. (11), when the truth is the measured radial law.

nominal	images		captions	
	Gaussian	recalibrated	Gaussian	recalibrated
10%	38.3%	10.1%	38.4%	10.0%
5%	35.8%	5.0%	36.5%	5.1%
1%	31.2%	1.0%	33.0%	1.1%
0.1%	26.5%	0.1%	29.3%	0.1%

A nominal 1% tail carries 31% of the image mass and 33% of the caption mass; at the 99th percentile of the image radius the χ_{768} law is wrong by a factor 10^{31}, and for captions 10^{50}. Cross-modal comparison inherits the same error, the surrogate reporting gaps of -10, $+42$, $+98$ nats at equal within-modality percentiles against -10, $+30$, $+53$ correct, and orientation contaminates the threshold, a single nominal 1% cut firing on 3.65% of image embeddings pointing away from the leading principal axis and 0.00% of those aligned with it. Table 8 shows eq. (11) restoring every nominal level. Equation (11) corrects the radial part, which is where the calibration error lives; the residual departure from sphericity is low-dimensional, and conditioning on a few directional coordinates handles it. Either way the argument of Proposition 6(b) holds: no high-dimensional density estimate is needed, which matters because estimating one in $\mathbb{R}^{768}$, even a log-concave one, is impractical at this sample size (Cule et al., 2010; Kim and Samworth, 2016). The practical recommendation is simpler still: calibrate on the empirical law of the tail energy, or use the k-NN density the score approximates.

7 Conclusion

Whitening a foundation-model embedding does something more interesting than Gaussianise it. It inverts the encoder's hierarchy, so what the squared whitened norm measures is the energy an input places outside the encoder's semantic subspace: a Mahalanobis estimate of atypicality that agrees with nonparametric

density at $\rho \approx 0.80$, that six encoders compute alike, and that carries the whole of the Mahalanobis out-of-distribution signal while the semantic part carries none. Its variance is set by a per-input scale, identified with its spread from two measured moments, and the factor $m/4$ multiplying that scale is why a coefficient of variation of a third produces a radius over-dispersed sixfold while the coordinates keep an excess kurtosis of only $3V$.

That account keeps what worked and explains what did not. Detection works because the score is a density ranking. Thresholds fail by 10^{31} because the radial law is the corpus's atypicality distribution, not χ_d. Coordinatewise tests passed because eq. (6) predicts they will at the group size used, and because in-sample whitening removes the one statistic that would have shown otherwise. And a different choice of objective would not have changed the radial law, because a cosine-similarity contrastive loss is blind to it. What sets $\sigma(x)$ at training time, whether eq. (11) conditioned on a few directions is the right estimator, and whether the tail structure can be exploited rather than corrected, are open. The methodological point is about nulls: a distributional claim about a learned representation is testable only against a null that shares the representation's covariance, and without one an identity can pass for a prediction and a factor-of-six departure can read as agreement.

References

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952. doi: 10.1214/aoms/1177729437.

T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49 (268):765–769, 1954. doi: 10.1080/01621459.1954.10501232.

R. Balestriero and Y. LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *NeurIPS*, 2022.

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. arXiv:2603.15026.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened CLIP as a likelihood surrogate of images and captions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. arXiv:2505.06934.

Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces Gaussian distribution. In *International Conference on Learning Representations*, 2026a. Oral presentation. arXiv:2602.24012.

Roy Betser, Omri Hofman, Roman Vainshtein, and Guy Gilboa. General and domain-specific zero-shot detection of generated images via conditional likelihood. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026b. arXiv:2512.05590.

Włodzimierz Bryc. *The Normal Distribution: Characterizations with Applications*, volume 100 of *Lecture Notes in Statistics*. Springer, New York, 1995. ISBN 978-0-387-97990-8. doi: 10.1007/978-1-4612-2560-7.

Madeleine Cule, Richard Samworth, and Michael Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B*, 72(5):545–607, 2010. doi: 10.1111/j.1467-9868.2010.00753.x.

Ralph D'Agostino and E. S. Pearson. Tests for departure from normality. *Biometrika*, 60(3):613–622, 1973.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3): 793–815, 1984. doi: 10.1214/aos/1176346703.

K. Ethayarajh. How contextual are contextualized word representations? *EMNLP*, 2019.

Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Number 36 in Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1990. ISBN 0-412-31430-4.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *NeurIPS*, 2021.

G. Ilharco et al. OpenCLIP. Zenodo, doi:10.5281/zenodo.5143773, 2021.

Arlene K. H. Kim and Richard J. Samworth. Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, 44(6):2756–2779, 2016. doi: 10.1214/16-AOS1480.

B. Klartag. A central limit theorem for convex sets. *Inventiones Mathematicae*, 168(1):91–131, 2007. doi: 10.1007/s00222-006-0028-8.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31*, pages 7167–7177, 2018.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. arXiv:2411.14517.

B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li. On the sentence embeddings from pre-trained language models. *EMNLP*, 2020.

W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou. Mind the gap: the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 2022.

T.-Y. Lin et al. Microsoft COCO: common objects in context. *ECCV*, 2014.

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

M. A. Stephens. Tests based on EDF statistics. In R. B. D'Agostino and M. A. Stephens, editors, *Goodness-of-Fit Techniques*, chapter 4. Marcel Dekker, New York, 1986. ISBN 0-8247-7487-6.

J. Su, J. Cao, W. Liu, and Y. Ou. Whitening sentence representations for better semantics and faster retrieval. arXiv:2103.15316, 2021.

V. N. Sudakov. Typical distributions of linear functionals in finite-dimensional spaces of high dimension. *Soviet Mathematics Doklady*, 19:1578–1582, 1978.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 2022.

Chen Tasker, Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. The universal normal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. arXiv:2603.21786.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/978-0-387-79052-7.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *ICML*, 2020.

A Proofs

Proof of Proposition 2. Write $h_x(s) = \Phi(xs^{-1/2})$, so that $F(x) = \mathbb{E}\, h_x(S)$. Differentiating twice,

$$h_x'(s) = -\tfrac{1}{2}\, x\, s^{-3/2}\phi(xs^{-1/2}), \qquad h_x''(1) = \tfrac{1}{4}\, x^2\phi'(x) + \tfrac{3}{4}\, x\phi(x) = \tfrac{1}{4}\, x(3-x^2)\phi(x).$$

Since $\mathbb{E}(S-1) = 0$, $\mathbb{E}(S-1)^2 = V$ and $\mathbb{E}|S-1|^3 = o(V)$, a second-order expansion about $s = 1$ gives

$$F(x) - \Phi(x) = \tfrac{V}{8}\, x(3-x^2)\phi(x) + o(V),$$

where the remainder is controlled as follows. The Anderson-Darling weight $\phi/(\Phi(1-\Phi))$ grows like $|x|$ and is not integrable, so a uniform bound on $F-\Phi$ does not suffice and the expansion must be truncated on the mixing variable. Fix $\delta \in (0, 1/2)$ and split on $A_\delta = \{|S-1| \le \delta\}$. On A_δ, h_x''' is bounded uniformly in x by a constant c_δ, so Taylor's theorem with remainder gives $|\,\mathbb{E}[h_x(S)\mathbf{1}_{A_\delta}] - \Phi(x) - \frac{V}{8}x(3-x^2)\phi(x)\,| \le c_\delta\,\mathbb{E}|S-1|^3 + \Phi(x)\,\mathbb{P}(A_\delta^c)$, and $\mathbb{P}(A_\delta^c) \le \mathbb{E}|S-1|^3/\delta^3$ by Markov. Off A_δ the contribution to $F-\Phi$ at x is at most $\mathbb{P}(S > M) + \sup_{s \le M} |\Phi(xs^{-1/2}) - \Phi(x)|$ with $\mathbb{P}(S > M) \le \mathbb{E}|S-1|^3/(M-1)^3$, which decays in x fast enough to be integrable against the weight. Since $\mathbb{E}|S-1|^3 = o(V)$, every remainder term is $o(V^2)$ after weighting, leaving $\Delta^2 = \kappa V^2 + o(V^2)$ with κ as in eq. (5). That integral converges because its integrand decays as $|x|^7\phi(x)^2$, and its value is by quadrature. Finally $\mathbb{E}X^4 - 3 = 3\,\mathbb{E}S^2 - 3 = 3V$. □

Proof of Proposition 4. $\mathbb{E}\|P_k(z-\mu)\|^2 = \mathrm{tr}(P_k\Sigma) = \sum_{j \le k}\lambda_j$ and $\mathbb{E}\|P_k w\|^2 = \mathrm{tr}(P_k\,\mathbb{E}ww^\top) = \mathrm{tr}P_k = k$, and under in-sample whitening the second identity of eq. (2) makes both exact sample averages. For the moments, given x the tail is $\mathcal{N}(0, D_x)$, so $\mathbb{E}[T \mid x] = \mathrm{tr}D_x$ and $\mathrm{Var}[T \mid x] = 2\mathrm{tr}(D_x^2)$, and the law of total variance gives $\mathrm{Var}T$; for two blocks, the conditional covariance of two quadratic forms of a Gaussian vector is $2\mathrm{tr}(P_1D_xP_2D_x) = 2\mathrm{tr}(D_x^{(12)}D_x^{(21)})$, and adding $\mathrm{Cov}(\mathbb{E}[T_1 \mid x], \mathbb{E}[T_2 \mid x])$ gives the third expression; a coordinate has $\mathbb{E}[w_j^4 \mid x] = 3(D_x)_{jj}^2$, whence its excess kurtosis. For the two-scale specialisation put $\nu = m/2$ per half, so $\mathrm{tr}D_x = \nu(s_1+s_2)$ and $\mathrm{tr}D_x^2 = \nu(s_1^2+s_2^2)$. Then $\mathrm{Var}(\mathrm{tr}D_x) = \nu^2\big(2V + 2\rho_sV\big) = \frac{m^2}{2}V(1+\rho_s)$ and $\mathbb{E}\mathrm{tr}D_x^2 = m(1+V)$, so $\mathrm{Var}T = \frac{m^2}{2}V(1+\rho_s) + 2m(1+V)$ against $2m$ for the clone, giving the first expression of eq. (10). The off-diagonal blocks of D_x vanish, so $\mathrm{Cov}(T_1, T_2) = \mathrm{Cov}(\nu s_1, \nu s_2) = \nu^2\rho_sV$, while $\mathrm{Var}T_i = \nu^2V + 2\nu(1+V)$ using $\mathbb{E}s_i = 1$ and $\mathrm{Var}s_i = V$; their ratio is the second expression. Setting $\rho_s = 1$ recovers the single-scale case $\mathrm{Var}T/\mathrm{Var}_{\mathcal{N}}T = 1 + V(1+m/2)$. □

B Reproduction

Encoding. This is the only stage that needs a GPU and a network. It produces CLIP ViT-L/14 image and caption embeddings for $5{,}000$ MS-COCO val2017 pairs; the six-encoder sweep over the same images; and the corpus sweep, six corpora by two encoders in disjoint blocks of $2{,}500$ images.

Analysis. CPU only, no network, every step seeded. One step reproduces the published protocol and builds the reference ensemble of Table 1; one recomputes the arithmetic behind the published moment table; one evaluates κ and $n^\star$ by quadrature and predicts the pass rates of Proposition 2; one measures the inversion and identifies the two-scale model of Table 5 and Figure 2; one builds the covariance-matched clones of Table 3; and the rest produce Table 7, Table 8, Table 6, the retrieval cost of the radial coordinate, and what the radius tracks in Section 5.
The one instruction that matters: extract embeddings before L^2 normalisation.