COORDINATES OR EQUIVALENCE? ALIGNMENT REALLOCATES SENSITIVITY ACROSS INPUT VARIATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive image-text training changes how an encoder responds to variations that its captions may leave unspecified. We measure this change using distributions induced by a fixed panel of input transformations. Each family's within-input covariance is compared with the covariance of the complete mixture, yielding response shares and cross-family covariance similarities that are invariant to invertible affine reparameterisations. An exact covariance budget separates within-component variation from variation of component means. A conditional linear-model analysis further expresses response changes through the encoder's input-space readout operator. On Stanford Cars, disrupting image-caption correspondence changes the mean per-seed correlation between a frozen description-based predictor and response-share changes from $+0.833$ to -0.443. The five-dose series changes sign between corruption fractions 0.25 and 0.50. Pairing contrasts also appear on CUB and Flowers and under a second public initialisation, with a gate-limited result on ConvNeXt. A separate analysis relates geometric reallocation to perturbation-specific accuracy changes, whose associations vary across datasets. Attributing the direction to description content needs separable caption sets; a pre-training screen passes one of six caption pairs across five datasets, and the content comparison on that pair is not significant. Together, these results provide an affine-invariant diagnostic of pairing-dependent response allocation, leaving open whether description content determines its direction.

1 INTRODUCTION

A caption describes selected properties of an image. Training an image encoder against that caption may change its responses to colour, position, visible extent and local detail in different ways. Classification accuracy summarises task performance, but does not reveal how those responses are redistributed. We ask which declared input variations account for more of the response after image-text alignment, and how that allocation depends on the pairing used in training.

The measurement must distinguish encoder change from a global change of embedding coordinates. If $f_t(x) = Af_0(x) + b$ with invertible A, affine prediction heads can be transported by the inverse map even though Euclidean distances and anisotropy change. We therefore study an explicitly affine equivalence class. A nonzero population-level change in our invariants rules out this affine null on the declared panel; it does not establish information loss under arbitrary nonlinear invertible maps.

Our construction fixes anchor images, transformation kernels and mixture weights. An anchor and a transformation family induce a distribution of embeddings, whose covariance records the response to that variation. Comparing each within-component covariance with the complete mixture covariance yields affine-invariant readouts without fitting a map between checkpoints. Normalising across families gives response shares. The covariance budget includes both within-component variation and the spread of component means, so it describes relative allocation rather than conservation of information.

For each transformation and strength, a description predictor measures how much of the reference response lies in directions that the training captions can express, and we correlate it with the share change. We test this measurement by changing image-caption correspondence while holding the predictor and evaluation panel fixed. On Cars the correlation is positive for all three class-caption seeds and negative for all four seeds trained on permuted captions. The graded intervention samples

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.

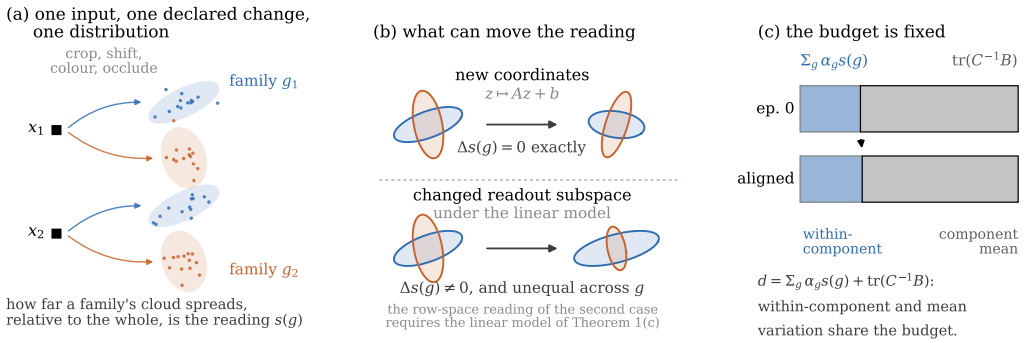

Figure 1: The setting. (a) A declared input transformation induces an embedding distribution; $s(g)$ measures its covariance relative to the mixture. (b) A common invertible affine map preserves the readouts. A row-space interpretation additionally uses the linear model of Theorem 1(c). (c) Within-component readings and component-mean variation sum to d. The diagram uses the reported budget proportions.

how this correlation changes with corruption, while separate analyses examine cross-family geometry and associations with accuracy.

The paper contributes a fixed-panel covariance decomposition with affine-invariant response shares and cross-family geometry, a linear readout analysis with explicit directional conditions, and pairing interventions across datasets and public initialisations. The pairing contrast is empirical and does not identify a unique training mechanism. Behavioural associations (Section 6.6) and the two-description comparison (Section 6.5) test further consequences. Figure 1 summarises the measurement; Appendix J examines the conditional readout interpretation.

2 Related Work

Contrastive alignment. Contrastive objectives have been analysed through mutual-information bounds, an alignment–uniformity split and guarantees under data assumptions (van den Oord et al., 2018; Poole et al., 2019; Wang & Isola, 2020; Saunshi et al., 2019; 2022; HaoChen et al., 2021; Zhang et al., 2023), and image–text training through the modality gap and through what alignment retains or discards (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Liang et al., 2022; Schrodi et al., 2025; Mistretta et al., 2025; Dufumier et al., 2025; Thoreau et al., 2025; Tjandrasuwita et al., 2025; Fang et al., 2026; Chaudhuri et al., 2025; Billa, 2026; Xiu et al., 2026). We propose no objective and use a conditional directional model to interpret its effects (Section 4).

Distributional representational similarity. The nearest construction compares response distributions conditioned on stimuli, modulo a transformation group (Williams et al., 2021; Duong et al., 2023), using whitening, Procrustes and canonical-correlation methods (Umeyama, 1991; Hotelling, 1936; Raghu et al., 2017; Kornblith et al., 2019); see Klabunde et al. (2025). Proposition 1 places our distributional affine null in the same whitened orthogonal framework. Our components are anchors crossed with declared variations, and their covariance decomposition yields response shares, a closed budget and cross-family overlap. No transformation is fitted between checkpoints. Under the per-cell adaptations in Appendix C, CKA, Procrustes and the frozen-subspace SVCCA variant do not reproduce the sign reversal; a stochastic shape metric does ($+0.831/-0.271$). The latter result makes the comparison substantive: the proposed readouts provide an explicit decomposition of the measured variation, while a fitted distributional distance quantifies overall discrepancy.

Invariance, and representations as distributions. Responses to input transformations and their relation to training augmentations have been studied by Bouchacourt et al. (2021) and Wang et al. (2022). Relative to per-transformation sensitivity scores, the decomposition here adds a closed covariance budget and cross-family geometry. Probabilistic and stochastic embeddings (Venkataramanan

et al., 2025; Nguyen et al., 2025; Popescu, 2026) model induced distributions to read uncertainty. Here the declared transformations specify the distributions without a parametric model. Separating a measurement from claims about what a representation contains is also central to probing (Xu et al., 2020; Hewitt & Liang, 2019).

The target axis. A concurrent submission concerns usable information about specified targets (Anonymous, 2027). Its target-conditioned scatter corresponds to our between-component term B_t only when the partitions and weights match.

3 Statistical Tokens and Pencil Invariants

3.1 The object

Let $\mathcal{X}$ be the input space and $\mathcal{Z} = \mathbb{R}^d$ the representation space, fixed at the pre-normalisation image embedding. Declare a finite set of *variations* $\mathcal{G} = \{g_1, \dots, g_G\}$, each a Markov kernel $K_g(dx' \mid x)$ whose generating rule is frozen across compared checkpoints. A member of $\mathcal{G}$, called a cell, is one of five transformation families at one of three calibrated strengths, so $G = 15$; g indexes cells throughout. We call what an input becomes once such a change is allowed, a distribution over representations rather than a vector, a *statistical token*.

Definition 1 (Statistical token)**.** *At checkpoint t with encoder f_t, the token of input x under cell g is the pushforward $P^g_{t,x} := (f_t)_{\#}K_g(\cdot \mid x) \in \mathcal{P}(\mathcal{Z})$. Fixing anchors $x_1, \dots, x_N$ with weights w_i and cell weights α_g, the composite index is $j = (g, i)$ with $\lambda_j = \alpha_g w_i$.*

Drawing M augmented views samples from $P^g_{t,x}$ without a parametric model. Distributions are not needed to identify an affine map, since $d + 1$ anchors in general position determine it, and finite point matches say little about unrestricted diffeomorphisms (Appendix A). We use them because they record within-component rearrangement and supply the second moments used below. With $\mathrm{Aff} = \{z \mapsto Az + b : A \in GL(d)\}$, the null is

$$H^w(\mathrm{Aff}): \quad \exists\, \kappa \in \mathrm{Aff},\ \ \kappa_{\#}P^g_{0,x_i} = P^g_{t,x_i}\ \ \forall (g, i). \tag{1}$$

The pointwise version $f_t = \kappa \circ f_0$ implies it but not conversely, since matching distributions permits rearrangement within a component; the readouts test second-moment consequences of the distributional version.

Proposition 1 (Affine gauge)**.** *If $\bar{P}_0 = \sum_j \lambda_j P_{0,j}$ and $\bar{P}_t$ have finite second moments and non-singular covariances C_0, C_t, then $H^w(\mathrm{Aff})$ holds if and only if there is a single $Q \in O(d)$ with $\widetilde{P}_{t,j} = Q_{\#}\widetilde{P}_{0,j}$ for all j, where $\widetilde{\ }$ denotes whitening by the respective mixture mean and covariance.*

Proposition 1 follows by centring and whitening each mixture (Appendix A). The affine class is motivated by transport of affine prediction heads. Our readouts remove its effect algebraically rather than fitting a cross-checkpoint map. Appendix B compares fitted alternatives on the same anchors.

3.2 The readouts

Define, with no per-sample whitening,

$$\Sigma^g_t = \sum_i w_i \,\mathrm{Cov}\big(P^g_{t,x_i}\big), \quad B_t = \mathrm{Cov}_{j\sim\lambda}(m_{t,j}), \quad C_t = \sum_g \alpha_g \Sigma^g_t + B_t, \tag{2}$$

with $m_{t,j}$ the mean of $P_{t,j}$ and $H^g_t = C_t^{-1/2}\Sigma^g_t C_t^{-1/2}$, and read

$$s_t(g) = \mathrm{tr}\big(C_t^{-1}\Sigma^g_t\big), \quad \pi_t(g) = \frac{\alpha_g s_t(g)}{\sum_h \alpha_h s_t(h)}, \quad \Gamma_t(g,h) = \frac{\mathrm{tr}(H^g_t H^h_t)}{\sqrt{\mathrm{tr}[(H^g_t)^2]\,\mathrm{tr}[(H^h_t)^2]}}. \tag{3}$$

Weights α are uniform, and the 5×5 display of Γ is the same cosine read at each family's middle calibrated strength (Section 5). The headline quantity is the share $\pi_t(g)$, a ratio of two invariants and

free of overall scale, which matters because the compared fine-tuning runs deform the representation by very different amounts. With $T_t = \sum_h \alpha_h s_t(h)$,

$$\Delta\pi(g) = \frac{\alpha_g \Delta s(g) - \pi_0(g)\Delta T}{T_t}. \tag{4}$$

A family can gain share while its own $s(g)$ falls. Section 4 gives directional statements about the strength-normalised change $\Delta s(g)/\tau_g$ in an input-space model. The empirical statistic instead uses $\Delta\pi$ and a reference-embedding predictor. The raw readout $s(g)$ serves as a check: it keeps the pairing signs, although the two predictors of Appendix J are ordered differently under it. Since $H_t^g \succeq 0$, $\Gamma_t(g,h) \in [0,1]$ wherever defined.

Theorem 1 (Pencil invariants). *Assume finite second moments and $C_t \succ 0$.* (a) *The readouts $s(g)$, $\pi(g)$, the spectrum of $C^{-1}\Sigma^g$, and $\Gamma(g,h)$ are unchanged by $z \mapsto Az + b$, $A \in GL(d)$.* (b) $\sum_g \alpha_g s_t(g) + \mathrm{tr}(C_t^{-1} B_t) = d$. (c) *Suppose the encoder acts as $x \mapsto Wx + c$ on the component supports, with W of full row rank written as $W = AP$ for an orthonormal row basis P. Let μ_j and Ξ_j be the input-space component means and covariances, and $\Xi_g = \sum_i w_i \Xi_{(g,i)}$. With $B_X = \mathrm{Cov}_{j\sim\lambda}(\mu_j)$ and $\Lambda = B_X + \sum_h \alpha_h \Xi_h \succ 0$,*

$$s(g) = \mathrm{tr}\left[(P\Lambda P^\top)^{-1} P \Xi_g P^\top\right], \tag{5}$$

which depends on W only through its row space.

Shares require a positive total within-component reading, and Γ nonzero covariance patterns. Parts (a) and (b) are exact for any encoder; part (c) assumes one affine map on the component supports (Assumption 5), and B_X generally differs from the clean-anchor covariance.

Proposition 2 (Rank floor). *Since $\hat{B}$ is built from NG empirical means,* $\mathrm{rank}\hat{B} \le NG - 1$ *and* $\sum_g \alpha_g \hat{s}(g) \ge d - \mathrm{rank}\hat{B} \ge d - (NG-1)$, *independently of the data.*

The bound follows from $\widehat{C} \succeq \widehat{B}$ and constrains only the between-component term; it decreases with the number of anchors and does not involve the number of views, and at the operating budget it equals $d - \mathrm{rank}\widehat{B} = 1$ (Section 6.1). Comparisons use a fixed $(\mathcal{G}, \alpha, w)$.

4 What Can Move the Reading

4.1 Strength and shape

The linear model expresses the response as a functional of the input covariance. It separates strength from shape without requiring a shared eigenbasis.

Theorem 2 (Strength and shape separate). *Under Theorem 1(c), keep the component distributions, kernels and weights fixed and set $R_t = W_t^\top (W_t \Lambda W_t^\top)^{-1} W_t$. For $\Xi_g = \tau_g U_g$ with $\tau_g = \mathrm{tr}\,\Xi_g > 0$ and $\mathrm{tr}\,U_g = 1$,*

$$s_t(g) = \mathrm{tr}(R_t \Xi_g), \qquad \Delta s(g) = \tau_g \varphi(U_g), \qquad \varphi(U) = \mathrm{tr}[(R_t - R_0)U]. \tag{6}$$

Both operators act on the input space; fixed input distributions, rather than equal traces alone, keep Λ fixed.

Corollary 1 (Strength normalisation). *Where $\Delta s(g) \neq 0$, $\log|\Delta s(g)| - \log\tau_g = \log|\varphi(U_g)|$. If $\varphi(U_g) = \bar{\varphi} + \epsilon_g$ with $\bar{\varphi} \neq 0$ and $\max_g |\epsilon_g| \to 0$, then, for a nonconstant predictor p and nonconstant strengths τ, $\mathrm{corr}(p, \Delta s) \to \mathrm{sign}(\bar{\varphi})\,\mathrm{corr}(p, \tau)$.*

Corollary 2 (A sufficient directional condition). *Suppose $a, b \ge 0$ with $a + b > 0$ and an input-space subspace $\mathcal{S}_x$ satisfy, on the measured covariances,*

$$\varphi(U_g) = (a+b)q_g - b, \qquad q_g = \mathrm{tr}(\Pi_{\mathcal{S}_x} U_g). \tag{7}$$

Then $\mathrm{sign}\,\Delta s(g) = \mathrm{sign}[(a+b)q_g - b]$, and for nonconstant q, $\mathrm{corr}(q, \Delta s/\tau) = 1$. The same correlation conclusion holds for unnormalised Δs when all τ_g are equal.

The experiments replace the exact normalisation of Corollary 1 by a reference-side proxy (Section 5, Assumption 4). Because $\mathrm{tr}(\Lambda\Delta R) = 0$ and $\mathrm{rank}\Delta R \le 2d$ (Appendix A), ΔR is neither positive in every direction nor of full input-space rank, so Corollary 2 restricts only its action on the panel. The input-space q_g differs from the reference-embedding predictor $p_{\mathrm{sh}}^g = \mathrm{tr}(\Pi_{\mathcal{S}}\Sigma_0^g)/\,\mathrm{tr}\,\Sigma_0^g$. Thus the pairing reversal tests an empirical predictor-response relation; interpreting it through Corollary 2 requires an additional link between these quantities.

4.2 TWO SOURCES, AND WHAT FOLLOWS FROM ASSUMING THEM

We use a conditional two-source model to organise possible directional effects, separately from the exact measurement identities.

Assumption 1 (Fixed sources)**.** *Over the interval being compared,* $\Delta R_f = a(f)\Delta R_{\mathrm{al}} + b(f)\Delta R_{\mathrm{bg}}$*, where the source operators are fixed,* $a \geq 0$*,* $b > 0$*,* a *is non-increasing,* b *is non-decreasing and* $a(1) = 0$*. The alignment source is associated with description-expressible directions, and the background source summarises the remaining dynamics.*

For normalised image and text embeddings, with text vectors e_j, temperature γ and softmax weights ω_{ij}, the image-side InfoNCE gradient is $\gamma^{-1}(\sum_j \omega_{ij} e_j - e_i)$. With fixed embeddings and a fixed candidate text set, permuting which text is positive for which image leaves the denominator terms unchanged and alters only the positive term. During training the embeddings, softmax weights, normalisation and encoder Jacobians all enter the update, and the negative term depends on the text vectors rather than on image covariance alone.

The source model is therefore a conditional interpretation, not a decomposition derived from this gradient. Under Theorem 2 and Assumption 1, $\Delta s_f(g)/\tau_g = a(f)\,\mathrm{tr}(\Delta R_{\mathrm{al}} U_g) + b(f)\,\mathrm{tr}(\Delta R_{\mathrm{bg}} U_g)$, so the centred family vector of strength-normalised changes is $y_f = a(f)v_{\mathrm{al}} + b(f)v_{\mathrm{bg}}$ with fixed centred source vectors $v_{\mathrm{al}}, v_{\mathrm{bg}}$. Proposition 3 gives its fixed-vector consequence. The real-data dose contrast is evaluated directly (Section 6.4), and the controlled system of Appendix L is not used as evidence for Proposition 3.

Proposition 3 (Single crossing)**.** *Let* p*,* v_{al} *and* v_{bg} *be fixed centred vectors with* $p \neq 0$ *and* $\langle p, v_{\mathrm{al}}\rangle\langle p, v_{\mathrm{bg}}\rangle < 0$*, and let* $y_f = a(f)v_{\mathrm{al}} + b(f)v_{\mathrm{bg}}$ *under Assumption 1. Where* $y_f \neq 0$*, the sign of* $\mathrm{corr}(p, y_f)$ *is the sign of* $\theta\langle p, v_{\mathrm{al}}\rangle + \langle p, v_{\mathrm{bg}}\rangle$*, where* $\theta = a/b$ *is non-increasing in* f*. It changes sign at most once, at* $\theta_* = -\langle p, v_{\mathrm{bg}}\rangle/\langle p, v_{\mathrm{al}}\rangle > 0$*, and only when the sampled doses straddle that root.*

Proposition 3 (proved in Appendix A) does not guarantee positive correlation at intact pairing unless the initial ratio lies on the corresponding side of the root, and it does not constrain the share-based residual rank statistic. We therefore report the observed intact/permuted sign contrast and the five-dose crossing separately from this model. The endpoint interpolation tested in Section 6.4 is one specific weighting choice. Table 10 distinguishes exact null consequences from conditional interpretations.

5 ESTIMATION

The anchor budget, calibration targets, canonical-correlation grid, primary statistic and primary nulls were fixed before the corresponding readings; the family-block nulls were added later (Appendices D, G and H).

Budget and calibration. Cars uses $N = 4000$ anchors and $M = 4$ views per cell; both checkpoints see identical perturbed inputs. Strength is calibrated on the reference trace to targets $\tau^\star/2, \tau^\star, 2\tau^\star$, with $\tau^\star = 8.53$. A cell is accepted when $|\log(\mathrm{tr}\,\Sigma_0^g/\tau_\ell)| \leq 0.1$ at its target τ_ℓ; the weakest cropping cell is retained with a flag. Fixing this design keeps the input mixture fixed; equal reference traces imply equal input strengths only under the conditions of Assumption 4.

The description subspace, and what clearing its gate licenses. $\mathcal{S}$ is estimated at the reference checkpoint by cross-fitted regularised canonical correlation between the clean visual embeddings of the anchor pool and that arm's captions under a frozen text encoder, using a gate in held-out canonical correlation above 0.1. This declared screening rule selects a reference predictor; it does not establish description-content attribution. In the pairing interventions, the class predictor is frozen across the class, graded and permuted Cars arms, and the CUB attribute predictor is frozen across its paired and permuted arms. Other comparisons use their stated arm-specific predictors.

Statistic and controls. Let $\mathrm{resid}(y \mid x)$ be the OLS residual on an intercept and x. The primary statistic is

$$\rho_{\mathrm{part}} = \mathrm{Spearman}(\mathrm{resid}(p_{\mathrm{sh}}^g \mid \log \mathrm{tr}\,\Sigma_0^g),\ \mathrm{resid}(\Delta\pi(g) \mid \log \mathrm{tr}\,\Sigma_0^g))\,. \tag{8}$$

Approximately 97% of the strength covariate's variance lies between levels; continuous residualisation is not an exact within-level comparison. Controls replace $\mathcal{S}$ by a caption-permuted subspace, a mid-spectrum variance-matched subspace, or V_0, the leading 64 directions of $\mathrm{Cov}(Z_{\mathrm{clean}}) + G^{-1}\sum_g \Sigma_0^g$ (Appendix K).

Uncertainty and nulls. Seed intervals are Student t intervals for means of per-seed correlations and describe training variability conditional on the fixed panel. Cell resampling and family-block permutation address other sources of uncertainty; anchors are not resampled. The two-text arm-label null enumerates 56 assignments. Behavioural cell-label and family-block tests are reported together in Section 6.6 and Appendix H. The pairing contrast is assessed through per-seed signs, sampled doses and the reported replications.

Conditioning. The covariance is ill-conditioned at this budget. Disjoint-half estimation, frozen-reference truncation, ridge variation and the finite-view correction assess different sensitivities (Appendices E and K). They agree in sign with the main readout, and re-estimating $\widehat{\Sigma}^g$ and $\widehat{C}$ after a random A exchanges no residual rank and leaves ρ_{part} unchanged on the class arm (Appendix A).

6 Experiments

Setup. All arms fine-tune OpenCLIP ViT-B/32 from a LAION-2B initialisation (Ilharco et al., 2021; Schuhmann et al., 2022) on Stanford Cars (Krause et al., 2013) for 30 epochs with AdamW (Loshchilov & Hutter, 2019) under one schedule, batch size and augmentation pipeline, differing only in the captions supplied: class names (3 seeds), colour-only captions (5), class names under a fixed permutation against images (4), and a graded series replacing a fraction $f \in \{0.25, 0.5, 0.75\}$ of pairings by class names drawn from a fixed permutation (3 each), with $f = 0$ and $f = 1$ being the class and permuted arms. Each arm changed its own zero-shot accuracy, and no reading is gated on these checks (Appendix E). Because all arms share one augmentation pipeline, interactions between augmentation and supervision are not tested (Table 10).

6.1 The instrument

At the operating budget $\mathrm{rank}\widehat{B} = 511$, so the floor of Proposition 2 is 1, and the empirical budget closes to 512 within 10^{-8}. Closure checks the implementation of the identity. With four views, sample-mean scatter has a finite-view bias (Appendix E). Its plug-in size in the estimated mixture metric is $\beta\sum_g \alpha_g \widehat{s}(g)$, where $\beta = M^{-1}(1 - 1/(NG))$: approximately 35.43 at the reference and 36.48 on the class arm. These are plug-in estimates rather than population errors of the nonlinear readout. Table 7 reports stable correlation signs after correcting the denominator. Under the uncorrected empirical convention, the within-component reading rises from 141.74 to 145.93 and its complementary mean-scatter term falls by the same amount (Figure 3). The bias term thus grows by 1.05, about a quarter of the 4.19 transfer. Subtracting the plug-in bias from each mean-scatter term, while still in the $\widehat{C}$ metric, gives a change of -5.24, with the same sign; this is not a readout under C_{bc}^{-1}. The absolute reading has not converged in N (181.66 at $N = 500$, 141.74 at $N = 4000$; Table 8), so we interpret directions and orderings of change rather than magnitudes.

6.2 A family-specific reallocation

At the population level, nonzero invariant changes exclude a common affine reparameterisation, and a common scaling of all $s(g)$ would leave shares unchanged. On the class arm, however, the seed-mean $\Delta\pi$ changes sign across families (Table 9), and in 11 of 15 cells $|\overline{\Delta\pi}|$ exceeds that cell's seed scatter. Per cell, the change under fine-tuning has a median ratio of 8.11 to the change between seeds (range 0.55 to 97.5); this single-seed ratio describes spread across cells, not a multi-seed effect size.

Cross-family geometry also changes: mean absolute off-diagonal differences are 0.090 for class and 0.104 for colour, compared with at most 0.010 between seeds (Figure 4). After an in-sample least-squares regression, with intercept, of the ten off-diagonal entries on the two families' share changes, the mean residual magnitude is 0.046 versus 0.005 between seeds. The geometry therefore changes beyond what the share changes explain, at least in this in-sample fit.

Table 1: Per-arm Cars readouts on $\Delta\pi$. The class, graded and permuted arms use the frozen class predictor; colour uses its own. The ρ_{part} column reports means of per-seed correlations. The seed-mean class value, 0.846, is given in Appendix E. Cell bootstrap intervals resample the fifteen cells on the seed-mean $\Delta\pi$ vector. The seed-pair row is ρ_{part} with one class seed as reference and another as the trained checkpoint, not a correlation between two $\Delta\pi$ vectors. $|\Delta\Gamma|$ is the reported mean off-diagonal change; ZS is the task-specific text-classifier accuracy of the first seed. NR denotes an unreported quantity.

Arm	seeds	ρ_{part}	seed 95% CI	cell bootstrap	$\lvert\Delta\Gamma\rvert$	ZS
class	3	$+0.833$	$[+0.729, +0.938]$	$[+0.47, +0.96]$	0.090	0.914
colour	5	$+0.406$	$[+0.360, +0.452]$	crosses zero	0.104	0.850
f=0.25	3	$+0.208$	NR	NR	0.113	0.803
f=0.50	3	-0.167	NR	NR	0.151	0.594
f=0.75	3	-0.262	NR	NR	0.169	0.319
permuted (f=1)	4	-0.443	$[-0.515, -0.371]$	crosses zero	0.159	0.018
seed pairs	3 pairs	-0.011 / 0.264 / 0.232	NR	NR	≤ 0.010	NR

6.3 The arm difference, and where it replicates

The fully permuted arm changes image-caption correspondence while retaining the image pool, caption multiset and training recipe. Its initial visual statistics and frozen predictor match the class arm's. Mean per-seed correlations are $+0.833$ and -0.443, with all three class seeds positive and all four permuted seeds negative (Table 1). Since the two arms start from the same checkpoint and predictor, the contrast shows that the reading is sensitive to the pairing itself. The contrast also survives the checks of Appendices E, G and K. Dropping any single cell leaves the class value between $+0.736$ and $+0.895$ and the permuted value between -0.495 and -0.169; the finite-view-corrected denominator gives $+0.833$ and -0.425; and estimating $\widehat{C}$ and Σ^g on disjoint anchor halves keeps both signs in either direction of the split.

A second corpus, where the captions are not the labels. On Cars the captions *are* the class names, so the reversal could be read as a fact about label supervision. On CUB (Wah et al., 2011) the attribute captions of Reed et al. (2016) describe plumage and beak shape and never name the species; keeping them and shuffling only which text sits on which image gives $+0.71$ paired against -0.34 permuted over the same fifteen cells, both seed intervals off zero, $[+0.54, +0.88]$ and $[-0.51, -0.17]$. Their subspace clears the gate only just (held-out canonical correlation 0.116), so we read the sign and not the magnitude; the species-caption arm on the same images gives $+0.82$ (Appendix I).

A third image domain. Oxford Flowers (Nilsback & Zisserman, 2008) keeps class-name captions and the recipe, shuffles only the pairing, and has its calibration rebuilt on its own reference checkpoint. The arms read $+0.393$ and -0.286, every seed carrying that sign and both seed intervals off zero, $[+0.304, +0.482]$ and $[-0.374, -0.197]$. The mean absolute share change, 0.007, is comparable to Cars class (0.0062 in Table 9) and smaller than Cars colour (0.019). Because Flowers has its own calibration and receives 240 optimisation steps against 1920 on Cars under the shared schedule (Appendix E), we compare signs rather than magnitudes across corpora.

A second public initialisation. The same ViT-B/32 from OpenAI CLIP rather than LAION-2B repeats the two arms with its calibration rebuilt on its own reference ($\tau^\star = 5.22$). Its description subspace clears the gate at the declared pair, with held-out canonical correlation 0.147. The class and permuted arms give $+0.730$ and -0.300, with uniform per-seed signs.

A second architecture family, with a caveat. Repeating both arms on `convnext_base_w` (LAION-2B) changes the architecture with the pretraining corpus fixed and calibration rebuilt. With embedding dimension 640, only signs and orderings transfer; the arms read $+0.91$ and -0.41 with uniform per-seed signs. At the declared $(\lambda, k_{\mathcal{S}}) = (0.1, 64)$, however, its description subspace reaches held-out canonical correlation 0.073, below the 0.1 gate. It clears the gate only at $\lambda = 0.01$ (0.230), outside the declared grid, so we report it as a sensitivity check rather than a replication. Each setting answers a different objection. Caption and label coincide on Cars but not on CUB; the image domain changes on Flowers; the public initialisation changes with OpenAI CLIP; and ConvNeXt adds a gate-limited check on architecture. Each is read for its sign (Figure 2b).

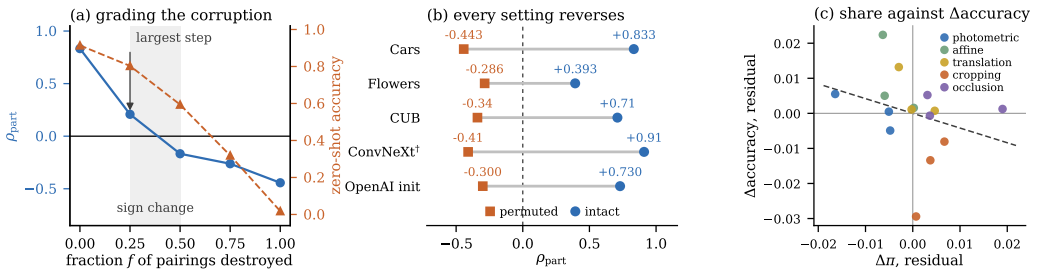

Figure 2: (a) Five sampled corruption doses, using means of per-seed correlations. The observed crossing lies between $f = 0.25$ and 0.50; zero-shot accuracy (right axis; first seed, as in Table 1) is 0.803 at $f = 0.25$; the two axes use different scales and are not compared in magnitude. (b) Intact and permuted pairing in five settings; † identifies the gate-limited ConvNeXt comparison. (c) Cars class share and accuracy changes after the residualisation in Eq. (8), using seed-mean vectors ($\rho_{\text{part}} = -0.514$). The mean of per-seed correlations is -0.456. Accuracy exceeds the reference on all fifteen cells; family-block inference is reported in Section 6.6.

Table 2: Whether two caption sets occupy separable subspaces, on the reference checkpoint. The criterion, fixed in advance, is both held-out canonical correlations above 0.1 and $\text{corr} \leq 0.3$. A / B follows the order of the caption-pair column. All six corr entries are family-level correlations of the two description shares: Pearson over twelve families for Stanford Cars class vs. colour and for CUB (Appendix I), and rank correlation over the five main families for the other four rows.

Corpus	caption pair	held-out CCA, A / B	$\text{corr}(p_A, p_B)$	go?
Stanford Cars	class name vs. colour	0.520 / 0.164	-0.639	yes
CUB	attribute vs. species	0.116 / 0.402	$+0.913$	no
DeepFashion-MM	attributes vs. category	0.105 / 0.226	$+1.00$	no
Oxford Flowers	colour vs. species	0.322 / 0.502	$+1.00$	no
COCO	object presence vs. style	0.452 / 0.524	$+0.90$	no
Stanford Cars	class name vs. body type	0.520 / 0.345	$+1.00$	no

6.4 Grading the intervention

The graded series permutes a prefix of $f \times 8144$ training rows as at $f = 1$, so unlike the full permutation it also shifts class-name counts (Appendix E). Figure 2a reports per-seed means $+0.833$, $+0.208$, -0.167, -0.262, -0.443, with one observed crossing, located only between the sampled values $f = 0.25$ and 0.50. Every seed carries the reported sign at each dose; the intermediate values have no reported intervals and are of the same order as the seed-pair readings 0.264 and 0.232. Meanwhile mean $|\Delta\Gamma|$ rises from 0.090 at $f = 0$ to 0.169 at $f = 0.75$, and is 0.159 at $f = 1$, so correlation decline and geometric deformation describe different aspects of the intervention.

If the change at intermediate f were the mixture $(1 - f)\Delta\pi_{f=0} + f\Delta\pi_{f=1}$ of the three-seed class mean and the four-seed permuted mean, the doses would be a zero-parameter prediction; it gives $+0.650$, $+0.271$ and -0.225 against the measured $+0.208$, -0.167 and -0.262, so it does not reproduce the measured values at $f = 0.25$ and 0.50. The endpoint directions are substantially opposed (cosine -0.592), so a linear mixture cancels mid-range.

Section 6.6 compares ρ_{part} between arms at similar accuracy.

6.5 Whether two descriptions can be told apart

The interventions above vary how much of the description survives, not what it says. Varying the latter needs two caption sets on the same images whose description subspaces are separable. This can be checked before training. We require both held-out canonical correlations to exceed 0.1 and $\text{corr}(p_A, p_B) \leq 0.3$, evaluated on the reference checkpoint with one forward pass per caption set.

Screening six pairs across five datasets (Table 2) passes one, Cars class name against colour, whose arms are trained above. The five failing pairs limit content attribution without showing that their description subspaces coincide. On Cars the separation is visible in Table 9: the class-name score is highest on cropping and occlusion, the colour score on photometric jitter.

On the separable pair the two-text statistic, a rank correlation between the predictor difference and the arms' share-change difference (Appendix G), gives $+0.761$, in the direction expected under the conditional model of Section 4, but it passes neither null: the arm-label null enumerates 56 assignments, returning $p = 0.179$; the family-block null returns $p = 0.067$. Neither establishes the content contrast at 0.05. On CUB, where the attribute captions never name the species, the two predictors correlate at $+0.913$ and the exploratory comparison gives $+0.18$ ($p = 0.21$; Appendix I).

6.6 What the reading predicts about behaviour

A cell whose share rises takes a larger part of the within-component response, so the representation moves relatively more under that perturbation. We therefore test the prespecified hypothesis that larger share increases accompany smaller accuracy gains. The statistic is the residual rank correlation between $\Delta\pi(g)$ and $\Delta\mathrm{acc}(g)$ on identical perturbations; the covariance definitions do not imply this association (Appendix H).

On Cars class, the mean per-seed correlation is -0.456, negative in all three seeds. Accuracy exceeds the reference on every cell of this arm, so the association concerns where the gain is smallest. The cell-label and family-block tests give $p = 0.026$ and $p = 0.092$, so the association holds on this panel with limited cross-family evidence. On CUB, species captions give -0.486 and attribute captions -0.393, with cell-label $p = 0.018$ and 0.024 and family-block $p = 0.058$ and 0.075. The attribute arm falls from 0.652 to 0.413 in species accuracy, so there the association concerns larger losses. The two CUB tests share images and task and are not independent. Cars colour gives -0.758 over five seeds (family-block $p = 0.025$) but is corroboration only, because its largest share increase, photometric, is the property its captions describe. Flowers and ConvNeXt show no significant association, which delimits its scope, and the permuted arm is excluded by a rule fixed in advance because its accuracy is nearly constant across cells (Appendix H). On Cars class and the two CUB arms, shares rise most on occlusion or cropping, which none of their captions describes.

At epoch 15, reconstructed class and $f = 0.25$ runs differ in correlation ($+0.882$ and $+0.186$) at similar three-seed mean accuracies (0.913 and 0.901); the epoch-30 means are 0.913 and 0.803. The reconstructed runs match the epoch-15 accuracies recorded for the main runs within their spread but are separate runs, and the class arm is near its accuracy ceiling (Appendix H). This two-arm comparison at sampled checkpoints motivates, but does not evaluate, a prospective detection rule, and it does not require Theorem 1(c).

7 Scope and Conclusion

Scope. Calling a contraction a gain in *invariance* requires the family to preserve the identity of the input, which cropping and occlusion do not; for those we claim only a change in sensitivity. The exact identities hold for the fixed panel, whereas the row-space reading requires Theorem 1(c); its exploratory use in Appendix J is not required by the comparisons above. Coverage is two backbones and two public initialisations under one recipe, with ConvNeXt gate-limited. The covariance estimate is ill-conditioned; the diagnostics of Appendices E and K bound but do not remove this concern.

Conclusion. In the tested settings, image-text fine-tuning changes affine-invariant response allocation, visible in both the covariance budget and the cross-family geometry. With the reference predictor fixed, intact and disrupted pairing give opposite orderings on Cars, CUB, Flowers and a second initialisation, and a gate-limited ConvNeXt result has the same signs. On Cars the dose series changes sign once, and linear interpolation between its endpoints does not reproduce the doses at $f = 0.25$ and 0.50. Share increases accompany smaller accuracy gains or larger losses on Cars and CUB, but not on Flowers or ConvNeXt. Whether description content sets the direction remains open: one of six screened caption pairs is separable, and its content comparison is not significant.

REFERENCES

Anonymous. Attenuation or removal? calibrated reading of usable conditional information. In *Submitted to The Fifteenth International Conference on Learning Representations*, 2027. URL https://openreview.net/forum?id=aXeWS1nXYM.

Jayadev Billa. Modality collapse as mismatched decoding: Information-theoretic limits of multimodal llms, 2026.

Diane Bouchacourt, Mark Ibrahim, and Ari Morcos. Grounding inductive biases in natural images: invariance stems from variations in data. *Advances in Neural Information Processing Systems*, 34: 19566–19579, 2021.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse, 2025.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Lyndon R. Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H. Williams. Representational dissimilarity metric spaces for stochastic neural networks. In *International Conference on Learning Representations*, 2023.

Wanlong Fang, Tianle Zhang, and Alvin Chan. To align or not to align: Strategic multimodal representation alignment for optimal performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21056–21064, 2026. doi: 10.1609/aaai.v40i25.39248.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019.

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

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D. Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *International Conference on Learning Representations*, 2025.

Anthony Nguyen, Emanuel Aldea, Sylvie Le Hégarat-Mascle, and Renaud Lustrat. Stochastic embeddings: A probabilistic and geometric analysis of out-of-distribution behavior. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, pp. 3200–3220, 2025.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180, 2019.

Sebastian Popescu. Distributional deep gaussian processes. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, pp. 5504–5540, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pp. 5628–5637, 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286, 2022.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *International Conference on Learning Representations*, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Romain Thoreau, Jessie Levillain, and Dawa Derksen. Can multimodal representation learning by alignment preserve modality-specific information?, 2025.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment, 2025.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Aishwarya Venkataramanan, Paul Bodesheim, and Joachim Denzler. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with gaussian process latent variable models. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, pp. 4309–4328, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6WsBGk4Iag.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *International Conference on Machine Learning*, pp. 41677–41693, 2023.

Reproducibility Statement

Anchor pools, augmentation draw rules, calibrated family strengths, protocol files, judgement tables and cached readouts are fixed across every compared checkpoint and will be released with the code, together with scripts that recompute the reported quantities. The readouts are keyed by corpus, checkpoint, seed, family and level, and contain s, π, p^g_{sh}, $\Delta\mathrm{acc}$ and the two filtering masks. Both null constructions for the two-text comparison are documented (Appendix G).

AI Use Statement

Generative AI tools were used to assist with language editing and presentation. The authors reviewed and verified all AI-assisted edits and take full responsibility for the content of the paper.

A Proofs

A.1 Why the object has to be a distribution

Lemma 1 (Limits of point observations)**.** *Let $\mathcal{M}$ be a connected smooth manifold without boundary of dimension at least two.* (a) *For two ordered lists of distinct points $z_1, \dots, z_n$ and $y_1, \dots, y_n$, there is a compactly supported diffeomorphism ϕ with $\phi(z_\ell) = y_\ell$ for every ℓ.* (b) *If κ is a diffeomorphism and $F \subsetneq \mathcal{M}$ is closed, infinitely many diffeomorphisms agree with κ on F.*

Proof. (a) Induction on m, starting from the identity. Suppose a diffeomorphism ϕ already maps z_j to y_j for $j < m$. Its current image $\phi(z_m)$ and y_m lie outside the already placed target points $y_1, \dots, y_{m-1}$. Removing finitely many points leaves a connected manifold in dimension at least two, so a smooth embedded path in that complement joins these two points. A compactly supported isotopy along a thin neighbourhood of the path moves $\phi(z_m)$ to y_m while fixing the earlier targets; composing it with ϕ completes the step. If the two points coincide, no move is needed. (b) Since $\kappa(F)$ is closed and proper, choose a coordinate ball in its complement. Flows of smooth vector fields supported in that ball give infinitely many distinct maps $\psi \circ \kappa$ agreeing with κ on F. □

The lemma concerns an unrestricted diffeomorphism class. It does not preclude identification in larger but structured classes than affine maps. Affine maps are determined by sufficiently many anchors in general position. Distributional observations additionally allow within-component rearrangements and provide the second moments used by the present readouts.

A.2 Proof of Proposition 1

Write $\mu_t = \mathbb{E}_{\bar{P}_t}[Z]$, $C_t = \mathrm{Cov}_{\bar{P}_t}(Z)$ and $\mathcal{W}_t(z) = C_t^{-1/2}(z - \mu_t)$, with $\widetilde{P}_{t,j} = (\mathcal{W}_t)_{\#} P_{t,j}$.

Proof. ($\Rightarrow$) Suppose $\kappa(z) = Az + b$ satisfies $\kappa_{\#} P_{0,j} = P_{t,j}$ for all j. Taking the mixture gives $\mu_t = A\mu_0 + b$ and $C_t = AC_0A^\top$. Put $Q = C_t^{-1/2} A C_0^{1/2}$. Then $QQ^\top = C_t^{-1/2} A C_0 A^\top C_t^{-1/2} = C_t^{-1/2} C_t C_t^{-1/2} = I$, so $Q \in O(d)$, and for any z, $\mathcal{W}_t(\kappa(z)) = C_t^{-1/2}(Az + b - \mu_t) = C_t^{-1/2} A(z - \mu_0) = Q\,\mathcal{W}_0(z)$. Pushing forward, $\widetilde{P}_{t,j} = (\mathcal{W}_t)_{\#}\kappa_{\#} P_{0,j} = Q_{\#}\widetilde{P}_{0,j}$. ($\Leftarrow$) Given such a Q, set $A = C_t^{1/2} Q C_0^{-1/2}$ and $b = \mu_t - A\mu_0$; then $\mathcal{W}_t \circ \kappa = Q \circ \mathcal{W}_0$ as above, and since $\mathcal{W}_t$ is a bijection, $\kappa_{\#} P_{0,j} = P_{t,j}$. □

Two remarks. The equivalence is what makes Aff a gauge: quotienting it is a change of description rather than a discarded alternative hypothesis. And the second-moment assumption is the only one used: no Gaussian or unimodality assumption enters, and the Q of Proposition 1 is never estimated by the readouts.

A.3 Proof of Theorem 1

Proof. (a) Under $z \mapsto Az + b$ with $A \in GL(d)$, every covariance transforms as $\Sigma \mapsto A\Sigma A^\top$ and every mean by the affine map, so $\Sigma^g \mapsto A\Sigma^g A^\top$, $B \mapsto ABA^\top$ and hence $C \mapsto ACA^\top$. Therefore

$$C^{-1}\Sigma^g \;\longmapsto\; (ACA^\top)^{-1} A\Sigma^g A^\top = A^{-\top} C^{-1} \Sigma^g A^\top,$$

a similarity transformation. Similarity preserves traces, spectra and traces of products. Since $\mathrm{tr}(H_g H_h) = \mathrm{tr}(C^{-1}\Sigma_g C^{-1}\Sigma_h)$, this proves the invariance of Γ wherever its denominator is nonzero. Shares are ratios of invariant traces. Part (b) follows by applying $\mathrm{tr}(C^{-1}\cdot)$ to Eq. (2).

For part (c), the common affine encoder gives $\Sigma_g = W\Xi_g W^\top$, $B = W B_X W^\top$ and $C = W\Lambda W^\top$, where $\Lambda = B_X + \sum_h \alpha_h \Xi_h \succ 0$. Writing $W = AP$ yields

$$C^{-1}\Sigma_g = A^{-\top}(P\Lambda P^\top)^{-1}(P\Xi_g P^\top)A^\top.$$

Taking the trace proves Eq. (5). Full row rank and $\Lambda \succ 0$ ensure invertibility. Two orthonormal bases of the same row space differ by an orthogonal factor that cancels. □

Numerical checks. Applying a random $A \in GL(d)$ to cached tokens changes $\max_g |s(g) - s_{\mathrm{aff}}(g)|$ by 2.2×10^{-16}, and comparing $W = P$ with $W = AP$ in the linear model of part (c) changes $s(g)$ by 6.2×10^{-15}. The 2.2×10^{-16} entries conjugate cached covariances. We also re-estimate $\widehat{\Sigma}^g$ and $\widehat{C}$ from transformed tokens at $\mathrm{cond}(A) = 10, 100, 1000$ and recompute shares, residualised $\Delta\pi$ and ρ_{part} on the class arm, with p^g_{sh} and $\log \mathrm{tr}\,\Sigma^g_0$ frozen in the original coordinates (the trace is not affine invariant). Residual ranks and ρ_{part} are unchanged. The median $\max|\Delta s|$ values, 1.5×10^{-8}, 5.4×10^{-7} and 1.9×10^{-5}, are on the scale of s; the largest resulting change in a residualised $\Delta\pi$ is 1.1×10^{-9}, against a smallest spacing of 0.012 between the ranked residuals of the seed-mean vector.

A.4 Proof of Proposition 2

Proof. $\hat{B}$ is the covariance of NG empirical component means, hence $\mathrm{rank}\hat{B} \le NG-1$. Since $\hat{C} = \sum_g \alpha_g \hat{\Sigma}^g + \hat{B} \succeq \hat{B}$, every generalised eigenvalue of the pair $(\hat{B}, \hat{C})$ lies in $[0,1]$, so $\mathrm{tr}(\hat{C}^{-1}\hat{B}) \le \mathrm{rank}\hat{B}$. Substituting into Theorem 1(b) gives $\sum_g \alpha_g \hat{s}(g) = d - \mathrm{tr}(\hat{C}^{-1}\hat{B}) \ge d - \mathrm{rank}\hat{B}$. On $\ker \hat{B}$ we have $\hat{C} = \sum_g \alpha_g \hat{\Sigma}^g$, so the corresponding generalised eigenvalues equal one exactly and contribute $d - \mathrm{rank}\hat{B}$ to the reading regardless of the data. □

The bound constrains the between-component contribution and nothing else. It does not say that $\hat{C}$ is singular whenever $NG \le d$: with one component and enough views spanning $\mathbb{R}^d$ the within-component covariance can be full rank while $\hat{B} = 0$. For empirical view clouds with M views the total mixture scatter has rank at most $NGM - 1$, an upper bound rather than a guarantee, so the split between anchors and views is a variance question and not a rank question.

A.5 Proof of Theorem 2 and its corollaries

Proof. With $z = W_t x$ the response and total covariances are $\Sigma_t^g = W_t \Xi_g W_t^\top$ and $C_t = W_t \Lambda W_t^\top$, so cyclicity gives

$$s_t(g) = \mathrm{tr}\big(W_t^\top (W_t \Lambda W_t^\top)^{-1} W_t \Xi_g\big) = \mathrm{tr}(R_t \Xi_g).$$

Subtracting at fixed Λ and writing $\Xi_g = \tau_g U_g$ gives $\Delta s(g) = \tau_g \,\mathrm{tr}(\Delta R\, U_g)$. Writing $W_t = A_t P_t$ with P_t orthonormal rows gives $R_t = P_t^\top (P_t \Lambda P_t^\top)^{-1} P_t$, so invertible output coordinates cancel. □

Two constraints bound the directional model. The matrix $\Pi_t = \Lambda^{1/2} W_t^\top (W_t \Lambda W_t^\top)^{-1} W_t \Lambda^{1/2}$ is the orthogonal projector onto $\mathrm{row}(W_t \Lambda^{1/2})$, giving

$$\mathrm{tr}(\Lambda R_t) = d, \qquad \mathrm{tr}(\Lambda \Delta R) = 0, \qquad \mathrm{rank}\Delta R \le 2d.$$

Thus ΔR cannot be positive everywhere. For $a, b > 0$, $a\Pi_{\mathcal{S}_x} - b\Pi_{\mathcal{S}_x^\perp}$ has full input-space rank; Corollary 2 restricts its action on the panel's covariances, rather than requiring this equality globally. Its input-space overlap q_g also differs from the embedding-space predictor p_{sh}^g.

Proof of Corollaries 1 and 2. For Corollary 1, divide Eq. (6) by τ_g. Taking logarithms of absolute values where nonzero subtracts $\log \tau_g$ exactly. For Corollary 2, Eq. (7) makes $\Delta s(g)/\tau_g$ a positive-slope affine function of q_g. Pearson correlation is therefore 1 when q_g is nonconstant. Unequal τ_g can change the ordering of raw Δs; the experimental proxy and share normalisation introduce further distinctions. The limit statement of Corollary 1 is Corollary 3 below. □

Corollary 3 (A common shape component)**.** *For fixed nonconstant p and τ, let $\varphi(U_g) = \bar{\varphi} + \epsilon_g$ with $\bar{\varphi} \ne 0$ and $\max_g |\epsilon_g| \to 0$. Then $\Delta s \to \bar{\varphi}\tau$ and* $\mathrm{corr}(p, \Delta s) \to \mathrm{sign}(\bar{\varphi})\,\mathrm{corr}(p, \tau)$.

Proof. By Eq. (6), $\Delta s(g) = \tau_g(\bar{\varphi} + \epsilon_g) \to \bar{\varphi}\tau_g$. Pearson correlation is continuous at nonconstant vectors, unchanged by positive scaling and reversed by negative scaling, which gives the limit. □

Fixed kernels, component distributions and weights keep Λ fixed; equal traces alone do not. Reference-side calibration is a separate proxy described in Assumption 4.

A.6 Proof of Proposition 3

Proof. Use the fixed centred predictor p and source vectors $v_{\mathrm{al}}, v_{\mathrm{bg}}$. Since $b(f) > 0$, the scale of $y_f = b(f)(\theta v_{\mathrm{al}} + v_{\mathrm{bg}})$ cancels from Pearson correlation. At defined points,

$$\mathrm{corr}(p, \theta v_{\mathrm{al}} + v_{\mathrm{bg}}) = \frac{\theta \langle p, v_{\mathrm{al}} \rangle + \langle p, v_{\mathrm{bg}} \rangle}{\|p\|\, \|\theta v_{\mathrm{al}} + v_{\mathrm{bg}}\|}.$$

The numerator has one positive root because its coefficients have opposite signs. Along a monotone θ sweep, defined nonzero signs can change at most once; a change occurs only if the sampled domain straddles the root. If the response vector vanishes there, correlation is undefined and the one-sided signs must be used. □

The proposition does not assert monotonic correlation magnitude or fix an initial sign without an endpoint condition. It also does not constrain the residual rank correlation of separately normalised endpoint shares. The real-data interpolation $(1-f)\Delta\pi_0 + f\Delta\pi_1$ is a particular empirical model, whose failure to reproduce the doses does not test the general monotone fixed-source statement.

B Why the Affine Group Is Worth Quotienting

Section 3 argues for the affine quotient on the grounds that a downstream linear head could undo any such map. This appendix gives an empirical counterpart and compares the closed-form quotient with maps fitted on the same anchors.

We fit a general linear map $\hat{A}$ between checkpoints by cross-fitted least squares on the same anchors and folds, and record how much of the paired residual it absorbs. This is a descriptive decomposition: $\hat{A}$ is estimated from data, so a small residual is *not* evidence for $H^w(\mathrm{Aff})$ and we do not report it as such.

Table 3: A fitted affine map, applied to three pairs of checkpoints under one pipeline. "Explained" is held-out Euclidean energy, $1 - \|Z_t - \widehat{\kappa}(Z_0)\|_F^2 / \|Z_t - Z_0\|_F^2$, on a five-fold split of the same anchors, with $\widehat{\kappa}$ either Umeyama's similarity or a general linear map (ridge $\lambda = 0.01$ when the unregularised condition number exceeds 10^6); "in-fold gap" is that quantity in-fold minus held-out for the linear map; "residual ratio" is $\widehat{R}(\widehat{A}Z_0, Z_t)/\widehat{R}(Z_0, Z_t)$. Here $\widehat{R}(Z, Z')$ is the mean squared gap between isotropic-Gaussian log ratio fields over the N anchors, $\frac{1}{NM}\sum_{i,m} \|R(z_{i,m}) - R'(z'_{i,m})\|^2$, on the same paired draws; it is not the Euclidean energy in the explained column. The last row shuffles which anchors are paired when the map is fitted. Under this implementation the general linear map absorbs most of the reference-to-aligned change; on seed pairs it explains less held-out energy than the similarity, and its residual ratio exceeds one.

pair	explained, similarity	explained, general linear	in-fold gap	residual ratio
reference → aligned	0.477	0.729	0.169	0.140
seed ↔ seed (3 pairs)	0.371–0.385	0.322–0.347	0.345–0.354	1.50–1.63
shuffled anchor pairing	0.518	0.254	NR	1.665

Table 3 shows that fitted similarity and general linear maps explain different amounts of held-out Euclidean change under this implementation. The general linear map absorbs more reference-to-trained change, while its performance is worse on the seed-pair comparison. The separately normalised ratio-field residual increases on those seed pairs. We therefore base the analysis on the invariant readouts and use fitted maps only as a descriptive comparison.

C Comparison with Representational Similarity Measures

Appendix B fits a global map; this appendix adapts similarity measures to individual variation cells. CKA, Procrustes and a frozen-subspace SVCCA variant do not reproduce the reported sign reversal, while a stochastic shape metric does. The SVCCA variant freezes the pre-transformation 99%-variance subspaces, of ranks 221 for clean embeddings and 208 to 236 for the cells. We label this variant explicitly; it does not test standard SVCCA with its SVD step recomputed after each transformation.

The measures need row-aligned embeddings, which the cached second moments do not provide, so embeddings were re-extracted from the same checkpoints under the draw rule of Appendix E, in float64, without normalisation. The cached readouts were recomputed from them before anything new was measured: $+0.846$ and $+0.833$ on the class arm, -0.443 on the permuted arm, a median fine-tuning-to-seed ratio of 8.11, and every entry of Table 3, all identical to the values reported above; two independent extractions agree on Σ_0^g to the last bit.

The per-family adaptation is specific to this comparison. Linear CKA, median-bandwidth RBF CKA and the frozen-subspace SVCCA variant enter through

$$\tilde{s}(g,\ell) = 1 - \mathrm{sim}(Z_{\mathrm{clean}}, Z_{g,\ell}), \qquad \tilde{\pi} = \tilde{s} \Big/ \textstyle\sum_{h,\ell'} \tilde{s}(h,\ell'), \tag{9}$$

while the Procrustes residual is already a dissimilarity and enters with that orientation; $\Delta\tilde{\pi}$ is read across checkpoints with the statistic of Eq. (8), on the same fifteen cells and conditioning on the same $\log \operatorname{tr} \Sigma_0^g$. A measure reproduces the contrast when ρ_{part} is positive on the class arm, negative on the permuted arm, and of one sign across the seeds of each; this criterion was fixed before the readings. The stochastic shape metric of Duong et al. (2023), added afterwards under the same criterion, whitens each side, fits an orthogonal map and a translation, and takes $\widetilde{s}(g, \ell)$ to be the mean per-anchor squared Wasserstein distance between the two Gaussians.

Invariance. Table 4 applies one $A = U\Lambda V^\top$ to the clean and perturbed embeddings of the reference checkpoint, at three condition numbers with five draws each, and repeats the exercise with A restricted to $O(d)$.

Table 4: Relative change of each measure under a change of coordinates, median over five draws with the range in brackets. The shape-metric ranges are 0.0009–0.0011, 0.0015–0.0015 and 0.0017–0.0018. The last column is the orthogonal control. For $s(g)$, $\pi(g)$ and Γ the entries are the cached-covariance conjugation check; full re-estimation errors are given in Appendix A. SVCCA† is the frozen-subspace variant described above. The frozen-subspace row transports $U \mapsto A^{-\top}U$ and is not standard SVCCA; recomputing the SVD after A yields relative changes 2.1×10^{-4}, 1.8×10^{-3} and 9.4×10^{-3} at the three condition numbers.

measure	$\operatorname{cond}(A) = 10^1$	10^2	10^3	$Q \in O(d)$
$s(g)$, $\pi(g)$, Γ	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
linear CKA	0.062 (0.038–0.091)	0.241 (0.188–0.312)	0.553 (0.481–0.640)	3×10^{-15}
RBF CKA	0.138 (0.094–0.186)	0.447 (0.390–0.521)	0.781 (0.712–0.844)	6×10^{-15}
Procrustes residual	0.105 (0.071–0.152)	0.219 (0.174–0.268)	0.394 (0.331–0.462)	4×10^{-15}
SVCCA†	6×10^{-15}	6×10^{-15}	6×10^{-15}	6×10^{-15}
shape metric	0.0010	0.0015	0.0017	5×10^{-14}

The Gram-based measures and Procrustes residual change under the tested nonorthogonal maps. The frozen-subspace SVCCA row concerns its stated variant, not standard recomputed SVCCA. The shape metric changes by about 10^{-3} in this implementation. For a fitted distance, exact invariance of the population quantity is distinct from numerical error introduced by regularisation or by solving the fit.

The arm contrast. Table 5 gives the per-family reading.

Table 5: Mean per-seed residual rank correlations of adapted responses. † denotes frozen-subspace SVCCA, as in Table 4. There are three class and four permuted seeds.

measure	class arm	permuted arm	reverses	one sign within each arm
π	$+0.833$	-0.443	yes	yes
linear CKA	$+0.429$	$+0.232$	no	yes
RBF CKA	$+0.214$	$+0.188$	no	yes
Procrustes	$+0.071$	$+0.161$	no	no
SVCCA†	$+0.286$	$+0.143$	no	yes
shape metric	$+0.831$	-0.271	yes	yes

The four adapted comparison baselines have positive means on both arms; the shape metric reverses. Per-seed values are $+0.482$, $+0.375$, $+0.429$ and $+0.268$, $+0.196$, $+0.250$, $+0.214$ for linear CKA, $+0.268$, $+0.161$, $+0.214$ and $+0.232$, $+0.143$, $+0.196$, $+0.179$ for RBF CKA, $+0.196$, -0.089, $+0.107$ and $+0.125$, $+0.214$, $+0.161$, $+0.143$ for Procrustes, and $+0.321$, $+0.250$, $+0.286$ and $+0.179$, $+0.107$, $+0.161$, $+0.125$ for SVCCA; the paper's readout gives $+0.882$, $+0.811$, $+0.807$ and -0.375, -0.464, -0.471, -0.461. The shape metric reads $+0.818$, $+0.818$, $+0.857$ and -0.232, -0.286, -0.271, -0.296; its class value $+0.831$ belongs to the shape metric and is distinct from the $+0.833$ of π in the first row.

Scale of the change. Each measure does register fine-tuning. Against the paper's median ratio of 8.11, the fine-tuning-to-seed ratio is 3.4 for $1-$CKA, 4.5 for $1-$RBF CKA, 4.1 for the Procrustes

residual and 2.5 for $1-$SVCCA, on clean-to-perturbed distances of 0.072, 0.183, 0.358 and 0.047 on the class arm against seed-pair values of 0.021, 0.041, 0.088 and 0.019.

SVCCA at this sample size. Estimating the canonical directions on one half of the anchors and evaluating on the other gives held-out mean canonical correlations of 0.548, 0.511 and 0.562 across the three seed pairs, against 0.492 for a shuffled-anchor control; the in-fold values are 0.961, 0.948 and 0.972. At $N = 4000$ and $d = 512$ the fitted directions carry only weakly to held-out anchors (0.51–0.56 against 0.49 for shuffled pairing).

D CALIBRATION

The anchor and view budget, the augmentation draw rule, the family list and levels, the calibration target, the canonical-correlation grid, the primary statistic, the primary null constructions and a judgement table naming what each outcome would be written as were fixed before the corresponding readings and will be released with the code; the family-block nulls of Appendices G and H were added later. Anchor pools are frozen per corpus, and each corpus or backbone added later carries its own calibration target and anchor pool rather than Cars' knobs.

On Cars the calibration target is $\tau^\star = 8.5255$, the geometric mean of the five families' traces at their median parameter setting on the reference checkpoint. Table 6 gives the solved strength parameters. One cell is out of tolerance: cropping displaces the embedding by at least 6.26 at the weakest calibrated level, so that cell cannot reach $\tau^\star/2 = 4.26$ and is carried at $|\log(\operatorname{tr}\Sigma_0^g/(\tau^\star/2))| = 0.38$ with a flag. Appendix G reports what dropping it does to the headline statistic.

Table 6: Family strength parameters after equal-trace calibration on the reference checkpoint. The flagged cell (†) cannot reach its target because of a floor intrinsic to the family. Each entry is a single scalar strength for one family. Photometric entries scale brightness, contrast and saturation jitter of amplitude 0.2 times the entry and hue jitter of 0.05 times the entry (capped at 0.5), applied in that order. Affine entries scale one draw each of rotation, shift and isotropic scale, and translation entries scale shift only. Cropping entries scale the random crop window, which is resized back, and occlusion entries scale the area of one grey patch. Each of the four views draws its amplitudes independently and uniformly inside these ranges, including the sign.

family	$\tau^\star/2$	$\tau^\star$	$2\tau^\star$
photometric	1.3689	2.5768	4.6533
affine warp	0.2852	0.9971	3.8073
translation	0.2090	0.9920	2.4166
cropping	$0.0100^\dagger$	0.4412	1.2966
occlusion	0.1550	0.3897	1.0635

E EXPERIMENTAL DETAILS

Backbone and schedule. All arms start from OpenCLIP ViT-B/32 with the LAION-2B checkpoint `laion2b_s34b_b79k` (Ilharco et al., 2021; Schuhmann et al., 2022) and fine-tune both towers and the logit scale jointly. Zero-shot tasks use each checkpoint's own text tower; geometry uses pre-normalisation image embeddings. Training uses AdamW (Loshchilov & Hutter, 2019), momentum coefficients $(0.9, 0.98)$, numerical constant 10^{-6}, weight decay 0.2, batch size 128, cosine schedule with 100 warm-up steps. The schedule has two phases: epochs 0–15 at learning rate 10^{-5}, then a resume from the epoch-15 checkpoint for epochs 15–30 at 10^{-6}. Every arm uses this schedule without exception, so within each comparison the arms differ only in the captions supplied and the seed. Readings are taken at epoch 30 against the pretrained model as reference. At this batch size the shared schedule is 1920 optimisation steps on Cars, whose training split is 8144 images, and 240 on Flowers, whose training split is 1020 images, so the 100 warm-up steps are a tenth of the first phase on Cars and most of it on Flowers.

Arms. The class arm uses class-name captions (3 seeds); the colour arm uses "a photo of a $\{$colour$\}$ car" over eight colours, read on the colour evaluation pool of 3117 images (5 seeds); the permuted

arm uses class-name captions under one fixed permutation against images, fixed once per split (4 seeds, of which analyses marked as two-seed use the first two); the permutation acts on images rather than on classes, so the images of one model receive many different wrong class names rather than one consistent relabelling; and the graded series replaces a fraction $f \in \{0.25, 0.5, 0.75\}$ of pairings by captions drawn from a fixed permutation (3 seeds each), so that $f = 0$ and $f = 1$ coincide with the class and permuted arms. At intermediate f the training list copies the frozen $f = 1$ image-row permutation on a prefix of length $f \times 8144$ and keeps the unpermuted captions on the remaining training rows; the evaluation split is left unshuffled, and the three dose seeds share one training list. This prefix changes class-name counts (180 of 196 classes change at $f = 0.25$), whereas the $f = 1$ arm keeps the training caption multiset. On CUB the attribute arm uses the human-written visual descriptions of Reed et al. (2016), which do not name the species, and the species arm uses "a photo of a {species}" (3 seeds each).

Reading protocol. Cars uses $N = 4000$ fixed anchors from 8041 evaluation images, and $M = 4$ views per cell, with deterministic paired draws and float64 arithmetic. The colour evaluation subset contains 3117 images. CUB uses the standard split of 5994 training and 5794 evaluation images, and all 5794 evaluation images are read.

Let $\widehat{\Sigma}_{\mathrm{w}} = \sum_g \alpha_g \widehat{\Sigma}_g$, where component covariances use denominator $M - 1$, and let $\widehat{B}$ be the weighted scatter of sample means. For independent component draws and uniform weights,

$$\mathbb{E}\widehat{B} = B + \beta \sum_g \alpha_g \Sigma_g, \qquad \beta = M^{-1}(1 - 1/(NG)).$$

The plug-in bias-term readings are 35.43 and 36.48 on Cars. Common random inputs reduce comparison variance but do not remove sampling error. The corrected denominator satisfies

$$C_{\mathrm{bc}} = \widehat{C} - \beta\widehat{\Sigma}_{\mathrm{w}} = (1 - \beta)\widehat{C} + \beta\widehat{B} \succeq (1 - \beta)\widehat{C} \succ 0.$$

It transforms by congruence and therefore preserves affine invariance of its trace readouts. This does not make inverse-based statistics unbiased. A corrected mean-scatter term need not be positive semidefinite, so the nonnegative budget interpretation is not transferred to it. Table 7 lists the pairing correlations under C_{bc}.

Table 7: Pairing correlations after replacing $\widehat{C}$ by $C_{\mathrm{bc}} = \widehat{C} - \beta \sum_g \alpha_g \widehat{\Sigma}_g$. The correction targets the finite-view sample-mean bias; it does not certify the population error of the nonlinear statistic. The first column uses $\widehat{C}$. Corrected-denominator signs remain unchanged in the reported seed summaries.

arm	ρ_{part} with $\widehat{C}$	under C_{bc}
Cars class (per-seed / seed-mean)	$+0.833/+0.846$	$+0.833/+0.846$
Cars permuted	-0.443	-0.425
$f = 0.25, 0.50, 0.75$	$+0.208/-0.167/-0.262$	$+0.208/-0.164/-0.262$
Flowers class / permuted	$+0.393/-0.286$	$+0.364/-0.296$
CUB attribute / permuted	$+0.71/-0.34$	$+0.73/-0.33$

Training gates. Each arm is checked on a zero-shot task before any geometric reading is taken, and no geometric reading is gated on the result: on Cars, class-name accuracy $0.850 \to 0.914/0.912/0.913$ (Table 1 prints the first seed, while 0.913 in Section 6.6 is their mean), colour accuracy $0.175 \to 0.850/0.864/0.867$ (balanced 0.70–0.74 over eight classes) and the permuted arm's class-name accuracy 0.018; on CUB, species accuracy $0.652 \to 0.973/0.973/0.974$ for the species arm and $0.652 \to 0.413$ for the attribute arm, whose captions never name a species; on Flowers, $0.700 \to 0.945$ paired against $0.700 \to 0.105$ permuted; on ConvNeXt, $0.906 \to 0.953$. Held-out canonical correlations for the description subspaces are 0.520 (Cars class), 0.164 (Cars colour), 0.402 (CUB species), 0.116 (CUB attribute) and 0.502 (Flowers species), against a permuted-caption floor of ≈ 0.00; the ConvNeXt subspace reaches only 0.073 at the declared $(\lambda, k_{\mathcal{S}}) = (0.1, 64)$ and 0.230 at $\lambda = 0.01$, and is discussed in Section 6.3. The OpenAI-initialised ViT-B/32 subspace reaches 0.147 at the declared pair. On Cars the OpenAI-initialised arms read class-name accuracy $0.598 \to 0.872/0.868/0.881$ against $0.019/0.016/0.021$ permuted. Equal-trace targets are $\tau^\star = 8.53$ on Cars, 7.42 on CUB, 6.37 on Flowers, 7.49 on ConvNeXt and 5.22 on the OpenAI-initialised ViT-B/32; none is transported between them.

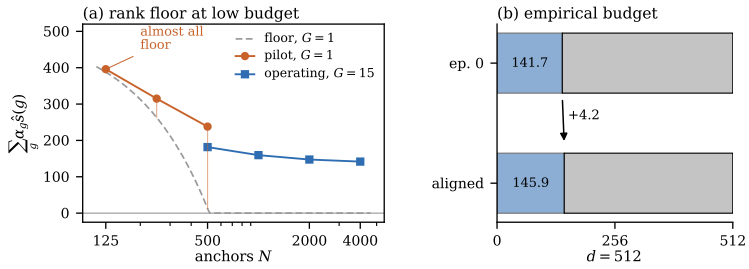

Figure 3: (a) Empirical response reading versus anchor budget, with the rank lower bound of Proposition 2. (b) The two empirical covariance terms sum to d. Their changes describe a redistribution within this estimator; the finite-view bias and its correction are discussed in Appendix E.

Additional arms. A CUB attribute-permuted arm (3 seeds) keeps the attribute captions and the identical recipe and shuffles only which text sits on which image. Two Oxford Flowers arms (class and permuted, 3 seeds each) use `a photo of a {species}` captions and the same schedule on a training set of 1020 images, read on $N = 4000$ anchors from that corpus's evaluation pool, with equal-trace calibration rebuilt on its own reference checkpoint at $\tau^\star = 6.369$. Two ConvNeXt arms (class and permuted, 3 seeds each) use `convnext_base_w` with LAION-2B pretraining at its native 256 resolution, with equal-trace calibration rebuilt on that backbone's own reference checkpoint, so $\tau^\star$ differs from the ViT value and the calibrated knobs are not transported. Two further arms (class and permuted, 3 seeds each) repeat the Cars recipe from the OpenAI CLIP initialisation of the same ViT-B/32, with calibration rebuilt on its own reference at $\tau^\star = 5.22$, where cropping at $\tau^\star/2$ is again clamped at the grid floor.

Caption and aggregation conventions. CUB attribute readings use one caption per image. We distinguish the mean of per-seed correlations from the correlation of seed-mean cell vectors. The Cars class values are $+0.833$ and $+0.846$, respectively. Table 1 uses the former throughout, so its point estimate and seed interval refer to the same summary. At the graded doses the seed-mean-vector values are $+0.189$, -0.168, -0.271, compared with per-seed means $+0.208$, -0.167, -0.262. The two conventions retain the reported signs.

The screen of Section 6.5 additionally evaluates DeepFashion-MultiModal and COCO (Lin et al., 2014); no arm is trained on either, and they enter only Table 2.

F Full Readouts

Table 8: The reading against the anchor budget at the reference checkpoint, averaged over the fifteen cells, with the low-budget pilot at a single family for comparison. The ledger itself closes to 512.00 within 10^{-8} at both checkpoints ($141.74 + 370.26$ at the reference, $145.93 + 366.07$ on the class arm).

N	operating (G=15)	pilot (G=1)
125	NR	396
250	NR	315
500	181.66	238
1000	159.53	NR
2000	147.42	NR
4000	141.74	NR

Table 8 gives the budget sweep behind Figure 3a. At the operating budget, $\mathrm{rank}\hat{B} = 511$ so $d - \mathrm{rank}\hat{B} = 1$. Rank counts the eigenvalues of $\hat{B}$ above $\lambda_{\max}(\hat{B}) \times 10^{-8}$; at the reference checkpoint $\lambda_{\max} = 6.06$, so the threshold is 6.06×10^{-8}. Further, $\mathrm{cond}(\hat{C})$ is 9.4×10^9 at the reference checkpoint and 2.5×10^{10} on the class arm, with $\lambda_{\min}(\hat{C}) \approx 10^{-10}$. The pilot column is

dominated by the floor of Proposition 2 at the two smallest budgets (388 and 263); at $N = 500$ the floor is 13.

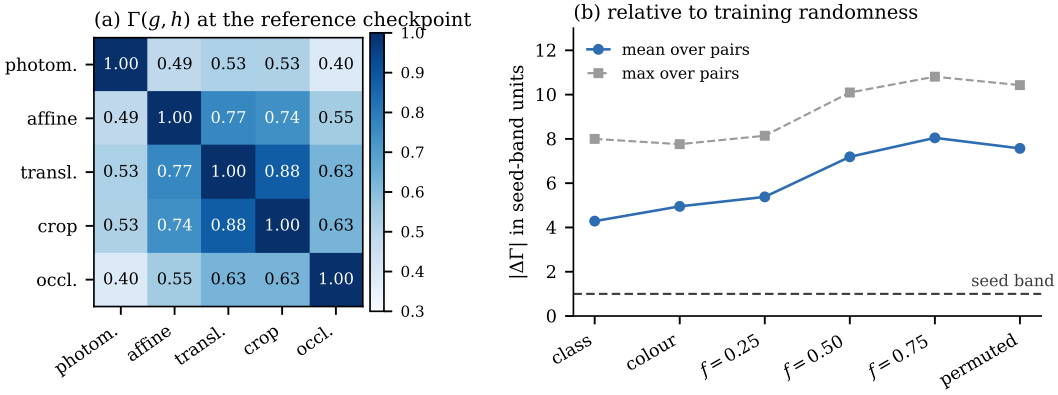

Figure 4: Cross-family geometry. *(a)* $\Gamma(g, h)$ at the reference checkpoint: the declared families are far from orthogonal, and translation and cropping are the most alike pair. *(b)* The mean and maximum absolute change of the off-diagonal entries per arm, in units of the seed band, taken as the largest seed-pair change over family pairs; every arm exceeds it by four to eleven times. The solid and dashed lines are the mean and the maximum over family pairs of the same $\Delta\Gamma$ matrix. The seed-pair value in Section 6.2 is instead a mean over family pairs.

Table 9: The fifteen calibrated cells. $\operatorname{tr}\Sigma_0^g$ is the deformation strength after calibration, the two predictors are the shares of that response lying in each caption set's description subspace, and the three $\Delta\pi$ columns are the change in share for the class, colour and permuted arms. The last column is the per-cell ratio of the class arm's raw change $|\Delta s|$ to the mean of the three seed-pair changes for that cell; its median is 8.11 and it spans 0.55 to 97.5. Cells are ordered by family and then by calibrated level. The three $\Delta\pi$ columns are seed-mean. The last column is computed on a single class seed and is not the filter used in Table 11; it spans 0.55 to 97.5 unrounded, and the table prints one decimal, so the minimum appears as 0.6. The $\Delta\pi_{\text{perm}}$ column averages the first two permuted seeds; the four-seed headline is Table 1. The cosine of this column with the class column is -0.584, against -0.592 for the four-seed endpoint used in Section 6.4.

cell	$\operatorname{tr}\Sigma_0^g$	p_{class}	p_{colour}	$\Delta\pi_{\text{class}}$	$\Delta\pi_{\text{colour}}$	$\Delta\pi_{\text{perm}}$	FT/seed (one seed)
photometric, $\tau^\star/2$	4.22	0.232	0.048	−0.0117	+0.0093	+0.0004	97.5
photometric, $\tau^\star$	8.59	0.245	0.059	−0.0165	+0.0263	+0.0039	38.1
photometric, $2\tau^\star$	16.39	0.260	0.070	+0.0011	+0.0973	+0.0127	0.6
affine, $\tau^\star/2$	4.32	0.314	0.040	−0.0062	−0.0085	+0.0085	10.2
affine, $\tau^\star$	8.43	0.333	0.041	−0.0062	−0.0144	+0.0073	4.4
affine, $2\tau^\star$	17.04	0.371	0.048	−0.0001	−0.0188	−0.0116	2.7
translation, $\tau^\star/2$	4.26	0.339	0.039	−0.0019	−0.0042	+0.0130	2.5
translation, $\tau^\star$	8.43	0.365	0.042	−0.0005	−0.0031	+0.0105	2.5
translation, $2\tau^\star$	16.70	0.371	0.047	+0.0031	+0.0096	−0.0096	21.3
cropping, $\tau^\star/2$	6.26	0.400	0.043	+0.0037	−0.0102	+0.0066	6.8
cropping, $\tau^\star$	8.28	0.395	0.044	+0.0033	−0.0157	+0.0014	5.0
cropping, $2\tau^\star$	16.42	0.385	0.050	+0.0066	−0.0253	−0.0157	8.4
occlusion, $\tau^\star/2$	4.19	0.346	0.048	−0.0036	−0.0117	−0.0014	14.5
occlusion, $\tau^\star$	8.55	0.374	0.051	+0.0035	−0.0137	−0.0063	8.1
occlusion, $2\tau^\star$	17.15	0.406	0.054	+0.0253	−0.0168	−0.0198	23.6

In Table 9 photometric jitter carries the least class-expressible and the most colour-expressible response, while cropping and occlusion carry the most class-expressible one; over the twelve candidate families the two shares correlate at -0.639 (Table 2). This separation is what makes the two-text contrast of Section 6.5 possible on this corpus at all. The class share changes have the opposite sign

Table 10: Exact null implications and conditional interpretations. A common affine change forces all readout differences to zero. A directional readout model additionally needs Corollary 2 and a link from the input-space overlap to the measured predictor. Holding training augmentation fixed does not test its interaction with the supervision.

Account	Implication or evidence required
Common affine coordinates	$\Delta s = \Delta\pi = \Delta\Gamma = 0$; correlation with a zero response vector is undefined.
Directional readout model	Positive correlation of q_g with $\Delta s(g)/\tau_g$ under Eq. (7); no unconditional share-sign prediction.
Caption-dependent collapse	Requires a specified covariance model and empirical checks.
Augmentation-dependent response	Requires varying or modelling the augmentation-supervision interaction.
Common scaling of all $s(g)$	Shares remain fixed; their change correlation is undefined.

to the permuted ones in ten of the fifteen cells, and to the colour ones in seven. The cell flagged in Appendix D for being out of trace tolerance is not an outlier in any other column.

G Robustness

Readout and intervals. Shares describe relative allocation, while $s(g)$ is itself affine invariant. The class raw-reading correlation is $+0.861$, compared with $+0.846$ on seed-mean shares. Raw and share versions retain the reported pairing signs, but the ordering of the two mechanistic predictors differs (Appendix J). The cell bootstrap crosses zero on the colour and permuted arms, whereas their individual training seeds carry the reported signs. These summaries address different uncertainty sources.

Which cells carry the statistic. Two objections apply to a rank statistic over fifteen cells: a cell whose fine-tuning change is smaller than its seed-pair change contributes noise, and the reading of Theorem 1(c) is weakest for the families whose deformation is strongest. Table 11 recomputes the statistic on subsets of cells, per seed, with each arm filtered on its *own* signal-to-noise ratio. Rows are to be read against the all-cells value in the same convention, $+0.833$ for the class arm and -0.443 for the permuted arm; the seed-mean convention gives $+0.846$ on the class arm (Appendix E), and the two are aggregations of one quantity rather than two results.

The class arm keeps its sign under every subset ($+0.736$ to $+0.919$). The permuted arm returns -0.443 under both signal-to-noise filters, because those filters remove no cell on that arm rather than because the statistic is insensitive to dropping cells, and it is close to zero on the weak-deformation families alone. That last row is a limit on the granularity at which the difference can be read on this panel: it supports a contrast between the arms, not a per-family account of the permuted arm, and it is not a limit on whether destroying the pairing reverses the sign, which Section 6.3 addresses on a second corpus, a third image domain and a second public initialisation.

Hyper-parameters and the flagged cell. The canonical-correlation grid is $\lambda \in \{0.05, 0.1, 0.3\}$ crossed with rank $k_{\mathcal{S}} \in \{32, 64, 128\}$, with $(\lambda, k_{\mathcal{S}}) = (0.1, 64)$ declared primary in advance; all nine cells are computed and none is selected on the outcome. At the other eight grid points the class arm remains positive and the permuted arm negative. The primary pair $(\lambda, k_{\mathcal{S}}) = (0.1, 64)$ is used throughout. Dropping the single out-of-tolerance cell leaves the class arm at $+0.823$ against $+0.833$ in the per-seed convention of Table 11, and moves the permuted arm to -0.469 against -0.443, so no reported sign depends on it.

Null distributions. The two-text statistic is the Spearman correlation, over the fifteen cells, between $p_{\text{class}} - p_{\text{colour}}$ and the difference of seed-mean share changes, $\overline{\Delta\pi}_{\text{class}} - \overline{\Delta\pi}_{\text{colour}}$ (three class seeds, five colour seeds). The predictors are frozen at the reference checkpoint, the seed means are taken before correlating, and, unlike ρ_{part}, the statistic is not residualised on $\log \operatorname{tr} \Sigma_0^g$. Its null declared in the protocol permutes arm labels over the seed-by-arm pool. With three class and five colour seeds it enumerates all $\binom{8}{3} = 56$ assignments, so the minimum attainable p is 0.018; the observed

Table 11: The primary statistic recomputed on subsets of the fifteen cells. The statistic is per-seed then averaged. Each arm contributes one mask, applied to every seed of that arm, so a single n is well defined. The ratio is $|\,\mathrm{mean}_{\mathrm{seeds}}(s_t - s_0)|$ over the mean of $|s_a - s_b|$ on that arm's seed pairs, not $\mathrm{mean}\,|\Delta s|$ and not the last column of Table 9; the class band uses the three pairwise differences among the same three fine-tuning runs. Neither threshold removes any cell on the permuted arm. "Weak families" keeps photometric, affine and translation, which we expect to be closer to the linear regime (not validated; Assumption 5). "Weak levels" keeps the two lowest calibrated strengths.

subset	cells, class / perm	class arm	permuted arm
all cells	15 / 15	$+0.833$	-0.443
drop cells with ratio below 1	15 / 15	$+0.833$	-0.443
drop cells with ratio below 3	12 / 15	$+0.879$	-0.443
weak-deformation families only	9 / 9	$+0.744$	-0.096
two weakest levels only	10 / 10	$+0.919$	-0.212
drop the out-of-tolerance cell	14 / 14	$+0.823$	-0.469
leave-one-cell-out range	14 / 14	$[+0.736, +0.895]$	$[-0.495, -0.169]$

$+0.761$ gives $p = 10/56 = 0.179$. The protocol also contains a family-block null, added when fewer seeds were available and no arm-label assignment could reach $p < 0.05$; enumerating all 120 family orderings, it gives $p = 8/120 = 0.067$. Neither null passes at 0.05, and the two-text comparison is reported as not established.

H The Behavioural Prediction in Full

The shares need not track changes in cosine or zero-shot predictions. For example, $f_t(x) = a(x)f_0(x)$ for positive input-dependent $a(x)$ leaves normalised embeddings unchanged, while it can change the covariance readouts. The behavioural association below is therefore empirical.

Accuracy change is evaluated on the same calibrated perturbations, using 8041 Cars class images, 3117 colour images, the 5794 CUB evaluation images and 6149 Flowers images. Geometry and accuracy share the transformation panel but may use different image counts. The declared one-sided null permutes accuracy changes over fifteen cells against frozen shares, using 10^4 assignments including the observed one. The family-block alternative preserves the three levels within a family and enumerates $5! = 120$ assignments. On Cars class, the two tests give $p = 0.026$ and $p = 0.092$. Seed agreement describes training repeatability conditional on this panel; it does not replace family-level inference. The family-block result is also reported in Section 6.6.

Two exclusion rules were fixed before the replication arms, and the Cars permuted arm, were read behaviourally. An arm does not carry the claim if its fine-tuned accuracy is near-constant across the fifteen cells (specifically, if its range is under 30% of the reference model's), because $\Delta\mathrm{acc}$ then reduces to the reference accuracy negated; the diagnostic is the correlation of $\Delta\pi$ with $\mathrm{acc}_{\mathrm{FT}}$ alone, reported in the last column of Table 12. And an arm is corroboration rather than an independent test if the family whose share rises most is the property its captions describe. Every arm below except the Cars permuted arm clears the first rule, and the Cars colour arm does not clear the second.

Per-seed values are -0.314, -0.546, -0.507 (Cars class), -0.546, -0.429, -0.482 (CUB species), -0.357, -0.429, -0.393 (CUB attribute) and -0.775, -0.789, -0.743, -0.711, -0.771 (Cars colour); every seed of every arm above the rule carries the sign shown. The two CUB arms share their images, family set and evaluation task, so their two p values are not independent of each other. Computing the statistic on the seed-mean $\Delta\mathrm{acc}$ instead gives -0.514, -0.500, -0.400 and -0.757, so the full-sample reading keeps its sign under either aggregation.

The slices are not. On the Cars class arm the same seed-mean vectors that give -0.514 over fifteen cells give $+0.12$ over the nine and $+0.60$ over the six, against the -0.393 and -0.486 the per-seed convention returns. These slices are sensitive to aggregation and to panel composition, so we draw no robustness conclusion from them; the claim rests on the fifteen-cell statistic reported alongside them.

Settings without a significant behavioural association. Flowers gives $\rho_{\mathrm{part}} = -0.043$ with seeds of both signs and $p = 0.44$; ConvNeXt gives -0.179, seeds of both signs and $p = 0.184$. These

Table 12: The behavioural prediction per arm: share change against accuracy change over the fifteen cells, conditioning on $\log \operatorname{tr} \Sigma_0^g$; p is one-sided against the cell-label null. The two slice columns drop cropping and occlusion, and keep only those two; they are printed for completeness, and the paragraph below this table says why no robustness conclusion is drawn from them. The last column is the exclusion diagnostic. The permuted arms on the replication corpora were not read behaviourally, for the reason given below.

arm	task	seeds	ρ_{part}	p	9 cells	6 cells	$\rho(\Delta\pi, \mathrm{acc}_{\mathrm{FT}})$
Cars, class	class-name	3	-0.456	0.026	-0.393	-0.486	-0.646
CUB, species	species	3	-0.486	0.018	-0.433	-0.371	-0.411
CUB, attribute	species	3	-0.393	0.024	-0.367	-0.314	-0.321
Cars, colour	colour	5	-0.758	0.004	-0.983	-0.657	NR
Flowers, class	species	3	-0.043	0.440	0.000	-0.371	-0.232
ConvNeXt, class	class-name	3	-0.179	0.184	-0.250	$+0.200$	-0.107
Cars, permuted	class-name	4	-0.753	0.0004	-0.567	-0.371	-0.13

settings do not support the association in the reported tests. A smaller absolute share change alone cannot explain a weak rank correlation, which is invariant to positive rescaling. Flowers also has 240 optimisation steps versus Cars' 1920, but attributing the outcome to this difference would require a separate comparison. ConvNeXt's behavioural statistic uses $\Delta\pi$ and $\Delta\mathrm{acc}$, not $\mathcal{S}$, so its CCA gate does not invalidate this behavioural test. Both outcomes delimit the empirical scope of the association.

Why the permuted arm is excluded. Its fine-tuned class-name accuracy is between 0.014 and 0.018 at every cell, so $\Delta\mathrm{acc}$ is almost exactly the reference accuracy negated, and the correlation reproduces $-\rho(\Delta\pi, \mathrm{acc}_{\mathrm{ref}})$ rather than anything about the fine-tuned model: against $\mathrm{acc}_{\mathrm{FT}}$ alone the same statistic is -0.13, against -0.646 for the class arm on the same diagnostic. The number satisfies the prediction and does not mean what the prediction means, so we report it and do not use it, and the permuted arms on the replication corpora were not read at all.

The colour arm. Colour training raises the photometric share most (the strongest photometric cell moves by $+0.0973$), and on that cell the colour task gains least over the reference ($+0.185$ against $+0.53$ to $+0.69$ on affine and translation cells). On the three arms that carry the claim the share rises most on occlusion (Cars class, CUB species) or on cropping (CUB attribute), which no caption set in question describes.

The epoch-15 reading. The epoch-15 checkpoints of the main runs were not retained by the two-phase training script, so the class and $f = 0.25$ arms were retrained through the first phase alone, from the same data and seeds; their epoch-15 class-name accuracies $(0.913, 0.914, 0.914)$ differ from the records of the main runs $(0.911, 0.916, 0.917$, kept to three decimals) by at most 0.004 per seed, within the 0.006 range of those records. At epoch 15 the class arm reads $+0.882$ ($+0.896$, $+0.889$, $+0.861$) and the $f = 0.25$ arm $+0.186$ ($+0.186$, $+0.139$, $+0.232$), against accuracies of 0.913 and 0.901. These are means over the three reconstructed seeds (class 0.913, 0.914, 0.914; $f = 0.25$ 0.895, 0.907, 0.902), not the first-seed convention of Figure 2a. At epoch 30 the readings are $+0.833$ and $+0.208$ and the accuracies 0.913 and 0.803. The comparison is between the two arms at a fixed epoch; the class arm's own accuracy is already at its ceiling by epoch 15, so nothing here is a claim about detecting corruption from a single arm.

I CUB, AND TWO DESCRIPTIONS THAT COULD NOT BE SEPARATED

The two-text comparison contrasts two caption sets on the same images, and requires their description subspaces to be geometrically separable. CUB is the corpus on which it would be most informative, since the attribute captions describe plumage and beak shape and never name the species, so "what the description says" and "what the label is" are genuinely different there. The comparison nonetheless does not meet the separation criterion on the tested families.

Scanning twelve candidate families on the reference checkpoint (the five used throughout plus hue, saturation, greyscale mixing, channel swapping, patch shuffling, small rotations and fine occlusion) gives $\mathrm{corr}(p_A, p_B) = +0.913$ between the attribute and species predictors, the value reported in Table 2, with a maximum per-family separation of 0.128. Restricting to the five main families gives $+0.971$. On the fifteen calibrated cells the two predictors correlate at $+0.934$. These are Pearson correlations over 12 families, 5 families and 15 cells; each family contributes one share, taken at the calibrated $\tau^\star$ level for the five main families and at a single scanned setting for the other seven. At any of these three granularities the CUB pair stays above the 0.3 gate, and the Cars class-colour pair of Table 2 stays below it, so the screening outcome does not depend on which is used. The candidate families were chosen to span colour, texture, position, shape and occlusion precisely because those are the axes an attribute description might load on differently from a species name; none of them meets the separation criterion.

The colour-oriented caption subset (412 images, held-out CCA 0.213) and shape-oriented subset (380 images, 0.057) give predictors correlating at $+0.929$. The shape-oriented subset does not clear the 0.1 gate, so this comparison is descriptive; together with the family scan, it shows limited separation on this panel rather than equality of all appearance-description subspaces.

The exploratory two-text statistic is $+0.18$, with a null 95th percentile of 0.30 and $p = 0.21$. It does not establish a content contrast on the weakly separated predictors. The species arm gives $+0.82$; the attribute and attribute-permuted arms give $+0.71$ and -0.34, or $+0.63$ and -0.29 when residualised against $\log s_0(g)$. Mean absolute share changes are 0.0046, 0.0041 and 0.019, respectively. The five-family aggregation, averaging the three calibrated levels within each family, is $+0.11$ ($p = 0.40$ under the family-block null). It is a resolution check and does not replace the cell-level exploratory value $+0.18$.

J An Exploratory Reading: A Turn of the Read-out Subspace

The conditional model suggests comparing the description predictor $\mathcal{S}$ with the top-variance predictor V_0. On seed-mean shares, their class correlations are $+0.846$ and $+0.668$; the first two permuted seeds give -0.429 and -0.539 in the same convention (Figure 5). We call the pattern in which $\mathcal{S}$ is stronger in the class arm and V_0 at least as negative in the permuted arm the predictor exchange. This is an exploratory predictor comparison, separate from the paired/permuted contrast. Within each family, the correlation between an anchor's own overlap and its own change stays within ± 0.06; because anchor covariances differ, we treat this as a descriptive observation.

Three dependencies keep this comparison exploratory. It rests on the linear model of Theorem 1(c), which we expect to be furthest from the linear regime on cropping and occlusion (not validated; Assumption 5); in the frozen-reference truncation diagnostic the predictor exchange holds only for $k \geq 256$; and an alternate construction of V_0 gives $+0.332$ on the class arm and no exchange. The arm signs do not depend on the second or third; the permuted arm keeps its sign without cropping and occlusion, but its magnitude falls to -0.096 on the weak-deformation families alone (Table 11). The predictor exchange depends on all three (Appendix K).

The predictor exchange depends on the readout. Raw $s(g)$ gives $+0.861$ versus $+0.679$ for class and -0.482 versus -0.386 for permutation; the latter ordering reverses relative to shares. The pairing sign contrast remains. The mid-spectrum control gives $+0.296$ on class, while residualising $\mathcal{S}$ against top-variance directions gives $+0.214$, compared with -0.45 to -0.29 for caption-permuted predictors. These results do not identify a unique semantic mechanism.

The predictor correlation would not have carried the attribution in any case: V_0 overlaps $\mathcal{S}$ by 0.755, and subspaces built the same way from captions permuted across images overlap V_0 by 0.696, 0.721, 0.768, 0.771 and 0.807 over five permutations, a comparable range. Direct overlap between $\mathcal{S}$ and the permuted-caption subspaces is 0.42 on average, so the two are not the same subspace.

K Conditioning of the Estimated Covariance

The readouts contain $\hat{C}^{-1}$ and $\hat{C}$ is ill-conditioned at this budget ($\mathrm{cond} \approx 10^{10}$, $\lambda_{\min} \approx 10^{-10}$), so a reading could in principle be dominated by near-null directions that are themselves poorly estimated. Three analyses address this, all on cached readouts with no retraining.

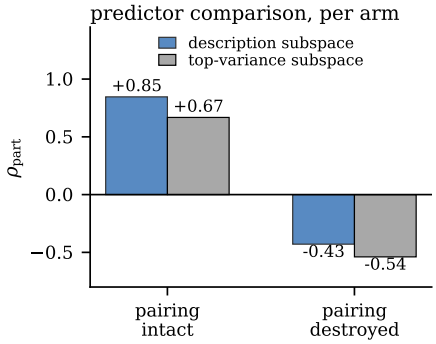

Figure 5: Which predictor explains the reading, per arm, at full spectrum, each arm's $\Delta\pi$ through its own C_t. The top-variance predictor is the leading 64 eigendirections of $\mathrm{Cov}(Z_{\text{clean}}) + \frac{1}{G}\sum_g \Sigma_0^g$, not of the C inverted for $s(g)$; the two arms share the reference checkpoint bitwise. Computed in the seed-mean convention on three class seeds and the two-seed permuted arm, not on the four seeds of Table 1.

Spectral truncation. Table 13 inverts a frozen, truncated reference covariance $C_0^{(k)}$ on its leading k eigendirections, with the predictors held fixed. It is a fixed-reference-metric diagnostic, and what it can be asked is bounded accordingly. The quantity $\mathrm{tr}\left(C_0^{-1}\Sigma_t^g\right)$ is not invariant under $A \in GL(d)$ acting on the target checkpoint, since Σ_t^g transforms by congruence and C_0 does not, so Theorem 1(a) does not cover it; and Table 3 shows that most of what fine-tuning does to these embeddings is absorbed by a fitted global linear map in that implementation. Differences between this table and the main readout therefore mix a change of coordinates with any spectral effect and cannot be attributed to the tail alone.

Within the diagnostic, the arm signs are stable down to $k = 64$, and the two-predictor exchange holds for $k \geq 256$ and fails at 128 and 64. That boundary is a property of this non-invariant readout and is reported as such rather than as a statement about the main statistic. Truncating instead on each checkpoint's own C_t would define a spectrally matched diagnostic, but one that preserves $O(d)$ and not $GL(d)$ invariance, so it would not settle the question either; the diagnostic in this group that keeps the invariance of the main readout intact is the disjoint-half re-estimation below.

Table 13: Fixed-reference-metric diagnostic: spectral truncation of a frozen $C_0^{(k)}$, with the predictors frozen. The readout in this table is not affine-invariant (Appendix K). "Exchange" asks whether the description subspace is strictly stronger in the class arm *and* the top-variance subspace at least as negative in the permuted arm. Throughout this table $\Delta\pi$ is computed through a frozen, truncated reference covariance $C_0^{(k)}$, whereas Figure 5 uses each checkpoint's own C_t; the $k = 512$ row is therefore not the untruncated limit of that figure, and the two need not agree.

k	class, $\mathcal{S}$	class, V_0	permuted, $\mathcal{S}$	permuted, V_0	exchange
64	$+0.929$	$+0.575$	-0.396	-0.243	no
128	$+0.957$	$+0.671$	-0.411	-0.329	no
256	$+0.896$	$+0.761$	-0.446	-0.529	yes
384	$+0.879$	$+0.771$	-0.443	-0.525	yes
512	$+0.864$	$+0.775$	-0.457	-0.536	yes

Disjoint-half re-estimation. Estimating $\hat{C}$ on one half of the anchors and evaluating Σ^g on the other, and then the reverse, gives $+0.86$ and $+0.87$ on the class arm and -0.42 and -0.40 on the permuted arm, so the signs do not depend on estimating both matrices on the same anchors.

Regularisation sweep. Replacing $\widehat{C}$ by $\widehat{C} + \epsilon I$ deliberately breaks the exact invariance of Theorem 1(a). Over $\epsilon/\lambda_{\min}$ from 10^{-3} to 10^{3} the class arm reads $+0.864$ and the permuted arm -0.425 on this diagnostic, while the invariance violation stays below 5.3×10^{-4}; at $\epsilon = 0$ it is the 2.2×10^{-16} conjugation check of Appendix A.

The three dependencies, in full. The linear model of Theorem 1(c) is expected to be least accurate for cropping and occlusion, which is not validated (Assumption 5); these families carry the most description-expressible response, and restricting the statistic to the three weak-deformation families leaves the class arm at $+0.744$ against $+0.833$ on all cells, so the forward reading does not depend on them, while the permuted arm keeps its sign but shrinks to -0.096. In the frozen-reference diagnostic, truncation leaves the arm signs intact down to $k = 64$ but the predictor exchange only for $k \geq 256$, as Table 13 shows. The third is the construction of V_0 itself.

Predictor construction. This alternate predictor is the leading variance of $\widehat{C}$, the inverted mixture covariance, not of $\widehat{B}$, and is not the construction used in Figure 5. It correlates with the primary one at Spearman 0.875, with maximum cell difference 0.048, but gives $+0.332$ on class and -0.436 on permutation, versus $+0.668$ and -0.539 for the primary predictor. At $k = 384$ and 512 it gives -0.407 and -0.404 on the permuted arm, compared with -0.443 and -0.457 for $\mathcal{S}$, and it does not reproduce the exchange at any k.

L Controlled Dynamics

We train a small encoder with an InfoNCE objective against a frozen text reference, with the description-expressible subspace constructed and therefore known. Across twelve paired runs the measured ordering is $\rho = +1$ in every run, and the arm in which the pairing is destroyed shows a collapse specific to the alignment term. Because the subspace is constructed, these runs check the measurement in a known setting; they are not independent evidence for the directional model, and the real-data results do not depend on them.

M Assumptions

Assumption 2 (Fixed declaration)**.** *Kernels, anchors and weights are fixed across compared checkpoints. The readout uses pre-normalisation embeddings and the affine group acts on that space. A decrease under crop or occlusion is described as a response change unless the transformation is independently shown to preserve task identity. Unconverged absolute readings are not interpreted as population magnitudes.*

Assumption 3 (No dimension reduction in the main readout)**.** *The full embedding is retained. Projected diagnostics define different measurements and generally do not preserve the affine null of Eq. (1). The anchor and view choices are design decisions, with their sensitivities bounded by the reported checks.*

Assumption 4 (Calibration proxy)**.** *Reference calibration uses* $\operatorname{tr}\Sigma_0^g = \operatorname{tr}(W_0 \Xi_g W_0^\top)$*, whereas Theorem 2 uses* $\tau_g = \operatorname{tr}\Xi_g$*. These traces coincide up to a common factor only under corresponding isometry conditions. Residualising the share response on the reference trace is therefore an empirical adjustment, not the exact strength subtraction of Corollary 1.*

Assumption 5 (Common affine interpretation)**.** *The exact row-space formula assumes a single affine encoder on the joint component supports. With component means* μ_j*,* $B_X = \operatorname{Cov}_{j\sim\lambda}(\mu_j)$*; it reduces to the clean-anchor covariance when transformations have the corresponding zero conditional mean. A local expansion* $f_t(x_i + \delta) = f_t(x_i) + J_{t,i}\delta + r_{t,i}(\delta)$ *has anchor-dependent Jacobians, and interpreting it through one global row space requires control of their variation and of the remainder, which is not supplied here. Theorem 1(a,b) and the empirical pairing comparison do not require it.*