ATTENUATION OR REMOVAL? CALIBRATED READING OF USABLE CONDITIONAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal alignment trains a representation to match a description that is almost never complete. Does the alignment remove content the description omits, or merely make it harder to access? A finite probe cannot tell from a single operating point: its reading depends both on representation change and on where the target sits along its resolution curve. We separate representation-level conditional information from finite-resolution accessibility, and read usable information across Gaussian-noise resolution σ. Similarity transforms translate the resulting curve along $\log \sigma$, which is why the raw magnitude of a fixed-resolution diagnostic does not identify mechanism. Calibrating the probe with known invertible transforms then defines information-preserving nulls with an explicit floor and power region: a trajectory landing on such a null is not thereby shown invertible; it belongs to the instrument's observational equivalence class of information-preserving change. On fine-tuned CLIP checkpoints the Stanford Cars category channel contracts by 0.60 nats of log-resolution while a construction-independent control does not, yet the trajectory remains indistinguishable from information-preserving attenuation; on CUB both targets expand instead, so we claim no universal sign. A local model of contrastive fine-tuning explains how this arises: alignment leaves the covariance-private block unchanged whereas the negative-sampling term contracts it, with an exactly preserving scalar regime and a perturbation bound around it. Synthetic ground truth separates attenuation from genuine removal, temperature interventions move contraction as predicted, and the calibrated pattern extends to DeepFashion, RSICD and COCO.

1 INTRODUCTION

Learning a representation by matching it against a description in another modality is now a standard recipe, and the description is almost never a complete account of the input: captions name some of what an image shows, clinical notes some of what a scan records, metadata some of what a signal carries. Whenever an objective rewards agreement with such a partial reference, the same question arises—what happens to the content the reference never specifies? Improved alignment could coexist with weaker access to it (Figure 1), and neither representation distance nor downstream accuracy settles the matter: invertible changes move distances without moving information, and language priors support high multimodal accuracy with little dependence on the other modality at all (Asadi et al., 2026). We study the question in its most widely deployed instance, CLIP-style contrastive image–text training (Radford et al., 2021), though the obstacle is a measurement one and not specific to that objective.

What a probe estimates from finite data and a fixed model class is not a Shannon quantity. It is a *usable* analogue (Xu et al., 2020), read at a resolution fixed by adding Gaussian noise of scale σ to the embedding. Our starting observation is simple: σ is not a nuisance parameter to be fixed and apologised for. It is an external ruler. Sweeping it turns a single number into a curve, and that curve's shape carries identification a single reading cannot.

Four levels separate what changes in a representation from what a finite probe can establish, and the paper follows them in order. At the *representation* level (Section 3) we distinguish preservation from removal using $I(Y; Z \mid G)$; at the *accessibility* level, finite-resolution readout may weaken even when that quantity is unchanged; at the *geometric* level (Section 4), sweeping

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.

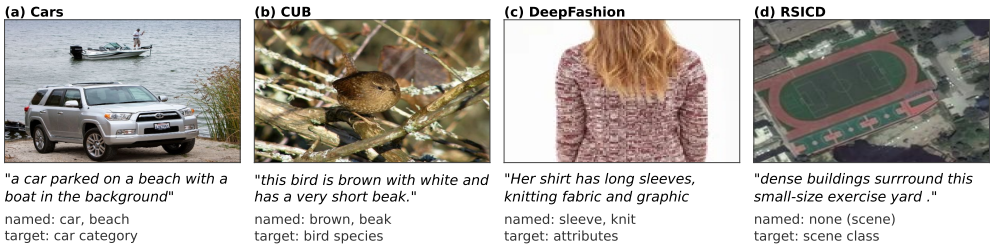

Figure 1: Descriptions specify part of what an input shows. Four domains, each with its caption reproduced verbatim, the properties that caption names, and a target present in the image but not named. When alignment changes access to such a target, has its conditional information been removed, or only its finite-resolution accessibility? Pairs are illustrative; every curve in this paper is a dataset-level probe estimate.

the resolution turns such changes into translations and deformations of target-indexed curves, so a single-resolution slope does not identify mechanism and spectrum-only capacity measures are blind to directional reallocation; and *calibration* against known invertible transforms (Section 5) identifies an observational equivalence class rather than the underlying checkpoint map. Section 6 then gives a local model of contrastive fine-tuning in which the alignment source vanishes on the covariance-private subspace while the negative-sampling term contracts it, and Section 7 validates the instrument on synthetic ground truth, measures target-specific trajectories on real checkpoints, and reads six corpora with it.

2 Related Work

Contrastive objectives have been characterised through mutual-information bounds, an alignment–uniformity decomposition 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), and a large empirical literature asks whether alignment preserves modality-specific information and how the modality gap behaves (Thoreau et al., 2025; Dufumier et al., 2025; Schrodi et al., 2025; Liang et al., 2022; Mistretta et al., 2025; Fang et al., 2026a; Tjandrasuwita et al., 2025; Xiu et al., 2026; Billa, 2026); these speak to the pretext task or to downstream accessibility, whereas Section 6 asks what the objective's two terms do to information the reference never mentions and our resolution axis makes representation-level accessibility a measured coordinate. Our estimate is a usable analogue for a fixed predictor family (Xu et al., 2020), and our calibration lifts control tasks (Hewitt & Liang, 2019; Pimentel et al., 2020) from random *labels* to known *transforms*: it measures what a probe reports for a change known to destroy nothing. A concurrent anonymous submission asks the dual question on the input side, reading which changes to an input an aligned representation still tells apart (Anonymous, 2027). Appendix A gives the fuller discussion.

3 Representation Information and Finite-Resolution Accessibility

Let (X, T, Y) be jointly distributed with Y finite, and at checkpoint t let $Z_t = f_t(X) \in \mathbb{R}^d$. For $N \sim \mathcal{N}(0, I_d)$ independent of everything, write $\widetilde{Z}_t(\sigma) = Z_t + \sigma N$ and $u = \log \sigma$. A probe conditions not on raw T but on $G = g(T)$ from a frozen sentence encoder, so the reference does not drift when a checkpoint's own text tower moves; since G is a function of T,

$$I(Y; \widetilde{Z}_t \mid T) - I(Y; \widetilde{Z}_t \mid G) = I(Y; T \mid \widetilde{Z}_t, G) - I(Y; T \mid G), \tag{1}$$

which has no fixed sign in general and is non-negative when G is sufficient for Y given T.

Definition 1 (Preservation, removal, attenuation)**.** *For representations Z_0, Z_t of the same input and fixed reference G, write $I_t^Y := I(Y; Z_t \mid G)$. The change $Z_0 \mapsto Z_t$ is* information-preserving for Y *if $I_t^Y = I_0^Y$ and exhibits* removal *if $I_t^Y < I_0^Y$. Writing $\mathcal{I}_\mathcal{Q}$ for the measurement procedure fixed*

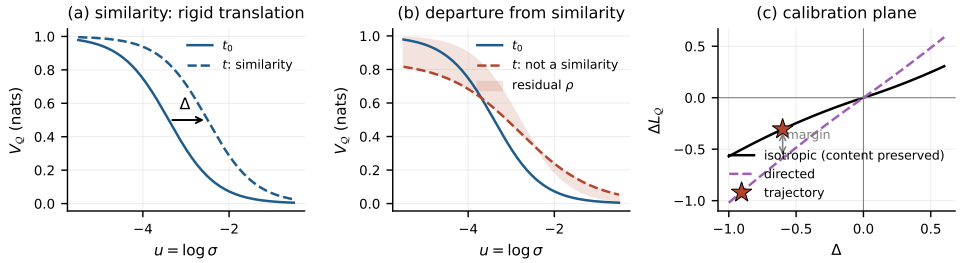

Figure 2: From curve geometry to calibrated equivalence. **(a)** A similarity translates the whole curve along $\log \sigma$ (Proposition 1). **(b)** A non-similarity change may deform it, leaving a shape residual ρ. **(c)** A finite probe can report an apparent plateau loss even under a known information-preserving transform. Calibration therefore defines a *similarity null*: landing on it denotes instrumental equivalence to the tested family, not identification of the underlying map.

by the probe family, the resolution grid and the fitting rule (Definition 2), an information-preserving change exhibits attenuation under $\mathcal{I}_\mathcal{Q}$ *if additionally* $V_\mathcal{Q}(t,u) < V_\mathcal{Q}(0,u)$ *on a non-empty resolution interval: reduced accessibility without representation-level loss. Removal is therefore a representation-level statement, whereas attenuation additionally refers to finite-resolution accessibility under a stated instrument.*

Since Z_t need not be a function of Z_0, the data-processing inequality does not apply and $I_t^Y > I_0^Y$ is possible in principle; the expansions reported on CUB and RSICD below are movements of accessibility, not claims about that case. Proposition 1 shows the two notions are distinct: an invertible contraction leaves $I(Y; Z \mid G)$ exactly unchanged while lowering $V_\mathcal{Q}$ throughout the live band.

The joint probe family $\mathcal{Q}_{ZG}$ has *optional ignorance* of Z relative to $\mathcal{Q}_G$ if every $q_G \in \mathcal{Q}_G$ has a counterpart $q_{ZG} \in \mathcal{Q}_{ZG}$ with $q_{ZG}(y \mid z, g) = q_G(y \mid g)$ everywhere. That is a property of the classes as sets, and an unconstrained fit is in fact systematically negative at coarse resolution by up to a nat, so we enforce it at fit time: initialise the joint probe at the text-only solution and retain its fit only if it wins on validation. We report

$$V_\mathcal{Q}(t,u) = \widehat{R}_{\text{test}}(\widehat{q}_G) - \widehat{R}_{\text{test}}(\widehat{q}_{ZG}) \tag{2}$$

where $\widehat{R}_{\text{test}}$ is held-out log-loss in nats, so $V_\mathcal{Q}(t,u)$ counts the nats the visual embedding adds to a text-only predictor at resolution u. It is reported with the per-resolution fallback rate, the operational statement that a resolution is exhausted for that target; the difference is exactly zero on test whenever the fallback fires, and is reported unclamped so that the reporting rule does not alter the estimator's distribution there (Appendix B). We call the interval of resolutions on which a target is not yet exhausted its *live band*, and read every coordinate below on that band. Gaussian maximal entropy motivates a capacity *scale*, not a certificate,

$$C_G(t,u) = \tfrac{1}{2} \log \det\big(I_d + \Sigma_V(t)/\sigma^2\big)\,, \tag{3}$$

with $\Sigma_V(t)$ the covariance of Z_t, computable from embeddings with no probe. At the population level $I(Y; \widetilde{Z}_t \mid G)$ is bounded by $H(Y \mid G)$ and non-increasing in u, so that curve saturates rather than growing as a Gaussian gain spectrum; the restricted, finite-sample $V_\mathcal{Q}$ inherits neither property as a theorem, so monotonicity is reported as observed (Table 1).

4 Resolution Curves and What They Identify

4.1 Similarities translate the curve

Write $\mathcal{G}_{\text{sim}} = \{z \mapsto cQz : c > 0,\ Q \in O(d)\}$.

Proposition 1 (Translation). *For any* $c > 0$, $Q \in O(d)$: $I(Y; cQZ + \sigma N \mid G) = I(Y; Z + (\sigma/c)N \mid G)$ *and* $C_G(cQZ, \sigma) = C_G(Z, \sigma/c)$.

Proof. $Q^\top N \overset{d}{=} N$, so $cQZ + \sigma N \overset{d}{=} cQ(Z + (\sigma/c)N)$, and mutual information is invariant under a fixed bijection of one argument; the second claim is $\frac{1}{2}\sum_i \log(1 + c^2\lambda_i/\sigma^2)$. □

The noise is a ruler, not a degradation (Figure 2a). It is the right ruler precisely because of this proposition: a similarity acts on it as a pure translation, so scale is separated from shape by construction. The same holds for the restricted estimate provided $\mathcal{Q}_{ZG}$ is closed under $z \mapsto cQz$, which an MLP with an unconstrained linear first layer satisfies as a class. The *estimator* need not be, since a penalty $\lambda\|W_1\|_F^2$ is orthogonally but not scale invariant; making the fit depend on the input only through $\widetilde{Z}/s(Z)$, for a scale-equivariant statistic s of the *clean* training embeddings, restores $\widehat{V}(cZ, \sigma) = \widehat{V}(Z, \sigma/c)$ for any penalty and optimiser (Proposition 5, Appendix C).

4.2 Three coordinates

Write V_Y for $V_{\mathcal{Q}}$ read on target Y. Fix a reference checkpoint t_0 and a weight w supported on the live band B, and define

$$(\Delta_Y, \rho_Y) = \Big(\arg\min_{\Delta},\ \min_{\Delta}\Big)\ \big\|V_Y(t,\cdot) - V_Y(t_0, \cdot - \Delta)\big\|_w, \qquad L_Y(t) = V_Y(t, u_{\min}), \tag{4}$$

with Δ_C, ρ_C defined analogously from C_G and $u_{\min}$ the finest grid resolution, a finite probe delivering no asymptote. Writing $\mathcal{G}_{\mathcal{Q}}$ for the stabiliser $\{\phi : \mathcal{Q}_{ZG} \circ \phi = \mathcal{Q}_{ZG}\} \supseteq \mathcal{G}_{\text{sim}}$, a similarity gives $\rho_Y = 0$ with every target reporting the same shift $\Delta_Y = \log c$, whereas L_Y is invariant under $\mathcal{G}_{\mathcal{Q}}$ only. But a single target sees the representation only through its Y-conditional channel: with $Z = (Z_1, Z_2)$, $Z_2 \perp (Y, Z_1, G)$ and $\phi(Z_1, Z_2) = (Z_1, AZ_2)$ for invertible $A \neq cQ$, we get $\Delta_Y = \rho_Y = 0$ although $\phi \notin \mathcal{G}_{\text{sim}}$. So $\rho_Y > 0$ certifies departure from similarity while $\rho_Y = 0$ does not certify it: the map from mechanisms to curve signatures is non-injective, which Section 5 formalises.

Δ is dimensionless, additive under composition and independent of d and of the operating resolution, whereas C_G at fixed σ is none of these, so cross-corpus statements are made in Δ; Δ_C, ρ_C are read separately on unnormalised embeddings.

4.3 Why single-resolution diagnostics are confounded, and capacity is blind

A single-resolution analysis summarises a run by the slope of usable information against a capacity proxy at one fixed σ, $\kappa := \partial_t V(t,u)/\partial_t C(t,u)$, and reads it as the usable content given up per unit of capacity. Under translation that ratio factorises, and the factors are not all mechanistic.

Proposition 2 (Slope factorisation)**.** *If both curves move by translation, $V(t,u) = V_0(u - \Delta_Y(t))$ and $C(t,u) = C_0(u - \Delta_C(t))$, then* $\kappa = \dfrac{\dot\Delta_Y}{\dot\Delta_C} \cdot \dfrac{V_0'(u - \Delta_Y)}{C_0'(u - \Delta_C)}$.

Both curves fall monotonically in u, so κ preserves the sign of the mechanistic ratio, but its raw magnitude does not identify that ratio without correcting for local curve geometry: a slope comparison at fixed σ also compares how steep each target's curve happens to be there; Appendix C gives two further consequences. Table 1 gives the practical cost: at the conventional operating point both targets are exhausted on both corpora.

Proposition 3. *$C_G(t,\cdot)$ is an exact translate of $C_G(t_0,\cdot)$ iff the non-zero spectra satisfy $\{\lambda_i'\} = \{c^2\lambda_i\}$ as multisets. Consequently any change $\Sigma_V \mapsto R\Sigma_V R^\top$, $R \in O(d)$, leaves Δ_C and ρ_C unchanged whatever it does to the eigenbasis relative to G or to a target; and since C_G depends on Σ_V only through its spectrum, the same holds for every spectrum-only capacity measure.*

Proofs are in Appendix C, with a table of the forward signature each mechanism predicts in the coordinates of equation 4. The mechanism a directional account posits is exactly of this congruence type, so a proxy of this family is not a conservative reference system for it but a blind one; directional change is recoverable only through target-indexed curves.

5 CALIBRATION: TURNING THE PROBE INTO AN INSTRUMENT

Testing whether a plateau change ΔL_Y reflects a real loss of content by checking $\Delta L_Y = 0$ cannot work: a probe with finite capacity reports a plateau change even under a transform we *know* preserves content exactly, because the transform pushes part of the signal below the probe's own resolution. So $\Delta L_Y \neq 0$ is by itself evidence of nothing. The fix is to measure how L_Y *responds* to a change whose effect on content we already know: take a family $\mathcal{T}$ of known invertible transforms and read the resulting plateau with the same probe. This traces a relation $\Delta L_{\mathcal{T}}(\Delta)$, the apparent loss produced by a transform that shifts the curve by Δ, purely as an artefact of the instrument. What the trajectory is then asked is not whether its plateau moved, but whether it moved by more than a transform that destroys nothing would have moved it.

Definition 2 (Instrumental equivalence)**.** *Fix a target Y, probe family $\mathcal{Q}$, resolution band B, and measurement procedure $\mathcal{I}_{\mathcal{Q}}$. Changes ϕ_1, ϕ_2 are ε-equivalent under $\mathcal{I}_{\mathcal{Q}}$ if the distance $d_{\mathcal{I}}$ between their measured curve coordinates is at most the instrument floor ε, and a family $\mathcal{T}$ has instrumental equivalence class $[\mathcal{T}]_{\mathcal{I},\varepsilon} := \{\phi : \inf_{\psi\in\mathcal{T}} d_{\mathcal{I}}(\phi,\psi) \leq \varepsilon\}$, with $d_{\mathcal{I}}$ the placement residual at matched Δ. The relation is reflexive and symmetric but not transitive; we use it only as ε-indistinguishability from a stated family.*

We use two families, chosen so that the null and its alternative differ in geometry alone. *Isotropic* transforms $z \mapsto cz$ are similarities, so by Proposition 1 they slide a target's curve rigidly and change no shape. *Directed* transforms scale only the leading r directions of a target's own between-class scatter $\Sigma_B^{(Y)}$, leaving the complement fixed, so they contract part of that target's channel and deform the curve rather than only sliding it. Both are invertible, so both conserve $I(Y; Z \mid G)$ exactly, and an expansion ($c > 1$) is as legitimate a calibration point as a contraction: what separates the two families is where the contraction falls, never whether content is destroyed. The directed family is thus the closest invertible surrogate for the mechanism a removal account posits—a contraction concentrated inside the target's own channel, row three of Table 4—which is what makes landing on the isotropic relation informative rather than true by construction.

Proposition 4 (Soundness, equivalence, and power)**.** *Let $\mathcal{C}_{\mathrm{iso}}$ be the calibration relation generated by $\mathcal{G}_{\mathrm{sim}}$.* (i) *Every $\phi \in \mathcal{G}_{\mathrm{sim}}$ lies on $\mathcal{C}_{\mathrm{iso}}$.* (ii) *If an observed trajectory lies within floor ε of $\mathcal{C}_{\mathrm{iso}}$ then $\phi_{\mathrm{obs}} \in [\mathcal{G}_{\mathrm{sim}}]_{\mathcal{I},\varepsilon}$, which does not imply $\phi_{\mathrm{obs}} \in \mathcal{G}_{\mathrm{sim}}$.* (iii) *The isotropic and directed alternatives are distinguishable at shift Δ only if $|\Delta L_{\mathrm{iso}}(\Delta) - \Delta L_{\mathrm{dir}}(\Delta)| > \varepsilon$; since both relations meet at the origin, power vanishes as $\Delta \to 0$.*

Each calibration family is traced on the same arm's cached reference embeddings and read with the identical probe pipeline. Placement is evaluated at matched Δ_{obs}, and each residual is compared with a seed-derived instrument floor. Calibration is rebuilt per corpus and training arm; it therefore supplies a measured null with an explicit power region, not a transported fitted baseline or an inverse map from observed coordinates to mechanisms.

6 A MECHANISTIC ACCOUNT OF CONTRASTIVE FINE-TUNING

For linear heads $Z_V = WX$, $Z_T = MT$ with $C_V = W\Sigma_X W^\top$, $C_T = M\Sigma_T M^\top$, and in-batch negatives from the marginal, the Gaussian moment generating function gives $\mathbb{E}_{z'} \exp(z \cdot z'/\tau) \approx \exp(z^\top C_T z / 2\tau^2)$, so

$$\mathcal{L}(W, M) \approx -\frac{1}{\tau}\operatorname{tr}\big(W\Sigma_{XT}M^\top\big) + \frac{1}{2\tau^2}\operatorname{tr}(C_V C_T) + \text{const}, \qquad \dot{W} = \frac{M\Sigma_{TX}}{\tau} - \frac{C_T W \Sigma_X}{\tau^2}. \tag{5}$$

The first term is alignment, the second the negative-sampling term. We treat fine-tuning as a local perturbation of a pretrained point; equation 5 is a mechanistic model with testable predictions, not an identification of the full nonlinear CLIP dynamics.

6.1 ALIGNMENT LEAVES PRIVATE GAIN FIXED; UNIFORMITY ATTENUATES IT

Let $P = \ker \Sigma_{TX}$ with projection Π_P, and $W_P = W\Pi_P$. Under the alignment term alone $\dot{W}\Pi_P = \tau^{-1} M\Sigma_{TX}\Pi_P = 0$: the paired-alignment source vanishes on the covariance-private subspace. Covariance privacy concerns $\ker \Sigma_{TX}$ and is not the same thing as literal omission from

the description (Section 6.2). We assume throughout that P is Σ_X-invariant, $\Sigma_X \Pi_P = \Pi_P \Sigma_X \Pi_P$, which holds in the orthogonal shared–private model of Appendix E and trivially under whitening.

Theorem 1 (Private-block flow)**.** *Under that invariance the private block obeys* $\dot{W}_P = -\tau^{-2} C_T W_P \Sigma_X$. *Writing* $A = C_T/\tau^2$ *and* $B = \Sigma_X|_P$, *the solution is*

$$\operatorname{vec}\big(W_P(t)\big) = \exp\big(-t\,(B \otimes A)\big)\ \operatorname{vec}\big(W_P(0)\big), \tag{6}$$

and in the eigenbases of A *and* B, *with* $\widetilde{W} = U^\top W_P V$, *the entries decouple as* $\widetilde{W}_{ij}(t) = e^{-\alpha_i \beta_j t}\, \widetilde{W}_{ij}(0)$ *for* $\alpha_i \in \operatorname{spec}(A)$, $\beta_j \in \operatorname{spec}(B)$.

Proof. $\operatorname{vec}(A W_P B) = (B^\top \otimes A) \operatorname{vec}(W_P)$ and B is symmetric, giving equation 6; conjugating by the eigenbases diagonalises the generator. □

Two features are easy to get wrong: the generator acts on W_P from both sides at once, so the flow is *not* $e^{-At} W_P(0) e^{-Bt}$, and each entry decays at the *product* rate $\alpha_i \beta_j$, not the sum.

Corollary 1 (What sets the rate)**.** *Decay rates* $\alpha_i \beta_j$, *with* $\alpha_i \in \operatorname{spec}(C_T)/\tau^2$ *and* $\beta_j \in \operatorname{spec}(\Sigma_X|_P)$, *are set in the population flow by temperature, the text-tower covariance spectrum, and private-input covariance. Batch size does not appear in equation 6; it can affect the finite-batch approximation to the negative distribution, but not this population rate formula. Because the rate is a product, components of* W_P *aligned with the nullspace of* C_T *are frozen, while contraction strength elsewhere depends on their overlap with occupied text-tower eigendirections.*

Corollary 2 (Exact scalar regime)**.** *If* A *or* B *acts as a scalar on* P, *the solution reduces to* $W_P(0) e^{-\alpha B t}$ *or* $e^{-\beta A t} W_P(0)$, *each* $W_P(0)$ *composed with an invertible map, so rank is exactly preserved at every finite* t. *If* $W_P(0)$ *is injective on the target-bearing private coordinates and that private channel is recoverable from* Z_t *given* G *(as in the orthogonal shared–private model of Appendix E), then its contribution to* $I(Y; Z_t \mid G)$ *is exactly conserved. Whitening is an exact scalar case for* B.

The flow is invertible on the ambient matrix space at every finite t, but this alone does not preserve the rank of $W_P(t)$; outside the scalar regime, rank can be lost at isolated times (Appendix G).

Theorem 2 (Near-isotropic stability)**.** *Suppose on the private block* $A = \alpha I + E_A$ *and* $B = \beta I + E_B$, *with* $\alpha, \beta > 0$, $\|E_A\|_2 \le \varepsilon_A$, *and* $\|E_B\|_2 \le \varepsilon_B$. *Let* $\delta := \beta\varepsilon_A + \alpha\varepsilon_B + \varepsilon_A\varepsilon_B$. *Then, for every finite* t,

$$\|W_P(t) - e^{-\alpha\beta t} W_P(0)\|_F \le e^{-\alpha\beta t}(e^{\delta t} - 1)\|W_P(0)\|_F.$$

Consequently, rank is preserved on any horizon for which $(e^{\delta t} - 1)\|W_P(0)\|_F < s_{\min}(W_P(0))$. *Under the same private-channel recoverability condition as Corollary 2, this also preserves the target-bearing private contribution to conditional information. The proof is in Appendix G.*

The guarantee is finite-horizon and conservative. That $C_G(t)$ can dip before rising follows from the balanced flow in Appendix F; this is a model prediction rather than a claim that every empirical capacity trajectory must be non-monotone.

6.2 The discriminant is shared projection, not description-visibility

Theorem 3. *The decay term in equation 5 carries no dependence on whether a private coordinate encodes real semantic content or an injected pattern. Two target directions lying in* P *therefore obey the same homogeneous private-block operator up to their initial gains. A direction can receive an alignment source only if it has non-zero projection onto the shared cross-covariance subspace; conversely, the source is non-zero whenever* M *is injective on the corresponding non-zero component of* $\operatorname{range}(\Sigma_{TX})$.

This contradicts the intuitive account on which a coordinate is exposed precisely when the description fails to mention it: a real but caption-poor pixel attribute and a constructed marker should behave alike. The claim reads off the operator $W_P \mapsto -\tau^{-2} C_T W_P \Sigma_X$, in which no factor sees a coordinate's semantic status, together with the support condition for the alignment source, so the semantic-status conclusion does not depend on the form of the solution or on which regime the flow is in. It is well-posed for targets with low-dimensional label geometry: a K-way label spans a

Table 1: Usable information V_Q / fallback rate at epoch 30, standardised, five seeds (V dispersion 0.03–0.08, rising with σ). A fallback rate of 1.0 means the retain rule never accepted a joint fit: the resolution is exhausted for that target.

Corpus	Target	0.0125	0.025	0.05	0.075	0.1	0.15	0.2
CUB	category	3.31/0	3.15/0	2.40/0	1.55/0	0.64/0	0.12/0.4	0/1.0
CUB	injected	1.61/0	1.48/0	0.95/0	0.35/0.2	$\approx$ 0/0.8	0/1.0	0/1.0
Cars	category	2.95/0	2.41/0	1.55/0	0.55/0.2	0/1.0	0/1.0	0/1.0
Cars	injected	1.30/0	1.15/0	0.60/0.2	0.15/0.6	0/1.0	0/1.0	0/1.0

between-class scatter of dimension up to $K - 1$, so for a 196-way target the prediction is a matter of degree across a subspace rather than a dichotomy; we do not extend it to a cross-corpus sign prediction, and a high-cardinality target's trajectory is a mixture over canonical modes with differing correlations ρ_k, which are not the shape residuals ρ_Y, ρ_C of equation 4.

7 Evaluation

The main results fine-tune OpenCLIP ViT-B/32 from a LAION-2B initialisation for 30 epochs on CUB-200-2011 (Wah et al., 2011) and Stanford Cars (Krause et al., 2013), reading checkpoints at epochs $\{0, 1, 3, 5, 10, 20, 30\}$, with a texture target on Oxford Flowers (Nilsback & Zisserman, 2008); three further corpora enter below (Appendix L). Two targets carry the main comparison: *category*, the class label, and *injected*, a marker patch written into each preprocessed image whose k-ary label is drawn independently of the caption, so that $H(Y \mid G) = \log k$ holds by construction. The injected target is therefore exposed to the same training signal as the category target while sharing nothing with the reference. All corpora are read on one fixed distribution of (X, T, Y) with $G = g(T)$ frozen, under the splits, retain rule and standardisation of Appendix B, with shifts fitted on each curve's live band and the band pre-declared on a control target (instrument checks in Appendix K). Floors come from five probe seeds: 0.031 nats for L, 0.038 for Δ, and 0.041 for the placement residual on average (0.038–0.042 per cell), the last by paired bootstrap.

Validating the instrument: known attenuation versus known removal. We build a synthetic representation with a known shared block S, on which G depends, and a known private, target-bearing block P, on which it does not, and pass it through the same pipeline used throughout. Three known transforms are applied: (A) a similarity cZ; (B) a directed but still invertible rescaling of part of P; and (C) a non-invertible removal projecting P to zero. (A) and (B) conserve every bit of $I(Y; Z \mid G)$ and both land on the isotropic relation within floor; (C) does not, its residual being -2.06 nats (Figure 4b). (B) was built where the relations have not yet separated, so it tests soundness, not discrimination (Appendix L.3).

A fixed resolution is not a common operating point. Table 1 gives $V_Q(\sigma)$ at epoch 30. The conventional operating point $\sigma = 0.2$ is exhausted for both targets on both corpora while the same targets are live at finer resolutions, and the usable bands of the two corpora differ by roughly a factor of two in σ. The finest-resolution column is the plateau L of equation 4. A cross-corpus dissociation read at one fixed σ therefore cannot be separated from reading two corpora at a resolution informative for only one, and Proposition 2 applies to any slope fitted there. Appendix H plots the full sweeps.

A target-specific effect, and where it places. On Stanford Cars the category target contracts by $\Delta = -0.600$ from epoch 0 to 30 while the construction-independent control does not, shifting by $+0.052$ in the opposite direction; the separation of 0.652 nats of log-resolution is seventeen times the shift floor. On CUB both targets expand together ($+0.325$, $+0.400$), a separation barely twice the floor. The dissociation is thus target-specific, and it belongs to the training run rather than to one optimisation path: three independent fine-tuning seeds give $\Delta_{\text{cat}} = -0.603 \pm 0.025$ against $\Delta_{\text{inj}} = +0.023 \pm 0.031$ (Appendix L.2). The capacity curves themselves collapse almost exactly under translation ($\rho_C = 0.005$ on CUB, 0.002 on Cars, with $\Delta_C = +0.50$ and -0.10), so by Proposition 3 the capacity change is very nearly a pure spectral rescaling on both corpora—and could not have revealed the Cars effect at all.

Table 2: Trajectory placement, epoch $0 \to 30$. Residuals are to each relation; the margin $|\Delta L_{\text{iso}} - \Delta L_{\text{dir}}|$ at the same abscissa determines whether the placement discriminates (Proposition 4(iii)).

Corpus	Target	Δ_{obs}	ΔL_{obs}	ΔL_{iso}	ΔL_{dir}	res. iso	res. dir	margin	floor
Cars	category	-0.600	-0.472	-0.448	-0.560	-0.024	$+0.088$	0.112	0.041
Cars	injected	$+0.052$	$+0.061$	$+0.020$	$+0.062$	$+0.041$	-0.001	0.042	0.038
CUB	category	$+0.325$	$+0.090$	$+0.066$	$+0.154$	$+0.024$	-0.064	0.088	0.039
CUB	injected	$+0.400$	$+0.117$	$+0.088$	$+0.195$	$+0.029$	-0.078	0.107	0.042

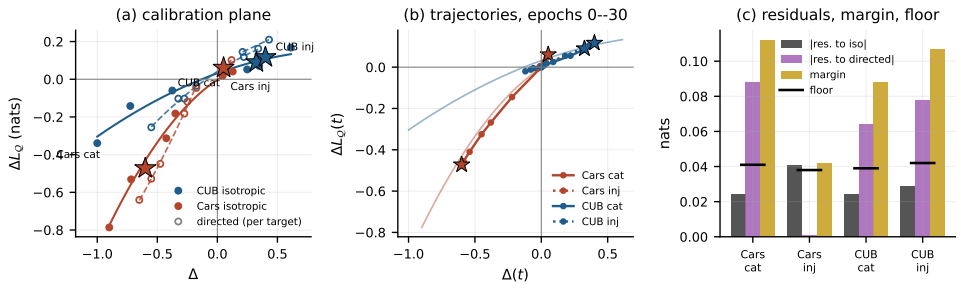

Figure 3: **(a)** The calibration plane. Filled markers and solid lines are isotropic transforms; open markers and dashed lines are directed transforms built per target. Stars are measured trajectories, epoch $0 \to 30$. **(b)** The same trajectories per checkpoint: all four track the isotropic relation, and Cars category contracts along it without crossing to the directed limb. **(c)** Residual to each relation, with margin and floor.

Placed against the two calibration relations, which bracket every observed shift so that placement never extrapolates (Appendix D), the Cars category point lands on the isotropic relation—residual -0.024 nats against a floor of 0.041—and well off the directed relation, margin 0.112, large enough to be informative (Proposition 4(iii)). At the stated floor, the decline is indistinguishable from the instrument response to a provably information-preserving transform. The CUB trajectories land on the isotropic relation too, on the expansion side; the exception is the Cars injected control, which sits near the origin where the relations meet: its margin (0.042) is essentially at the floor, so Proposition 4(iii) leaves the cell without power and we do not read it either way. Figure 3(b) and Table 2 show this is not an endpoint artefact (per-checkpoint coordinates in Appendix I): all four trajectories track the isotropic relation at every checkpoint, and the Cars category path contracts smoothly along it without crossing onto the directed branch.

A direct intervention: temperature. Section 6 makes one prediction we can test directly: lowering τ should raise the contraction rate, roughly as τ^{-2} (Corollary 1). We fine-tune three further Cars arms with the logit scale fixed at $\tau_0/2$, τ_0 and $2\tau_0$, rebuilding calibration on each arm's own embeddings. The category shift moves in the predicted direction by a large margin: $|\Delta_{\text{cat}}| = 0.975$, 0.625 and 0.525 nats as τ^{-2} falls from 39,944 to 2497 (Figure 4a), exceeding the injected control's shift in every arm. Placement is less clean: only the lowest-temperature arm lands within floor of either relation, and on the *directed* one (residual -0.129). We therefore read this as directional confirmation of the mechanism rather than a second calibrated placement (Appendix L.1).

Which targets move, and how far the reading travels. Theorem 3 predicts *which* targets should move: not those a description happens to omit, but those whose directions project onto the shared subspace. DeepFashion-MultiModal (Jiang et al., 2022) tests this directly, recording many discrete garment attributes on the same images; across a pre-registered scan of 18 attributes, a target's overlap with the shared subspace (p_{shared}) predicts its shift far better than literal caption mention—Spearman $\rho = -0.62$ ($p < 0.01$) against $\rho = -0.21$—with every target on the isotropic relation (Appendix L.4). The same holds on the main corpora for two low-cardinality targets captions largely do not name, colour and a texture statistic on Oxford Flowers: both stay inside the plateau floor while only the category targets move (Appendix J), where a visibility account predicts they should move like the category target. On RSICD (Lu et al., 2018), whose satellite imagery has very different

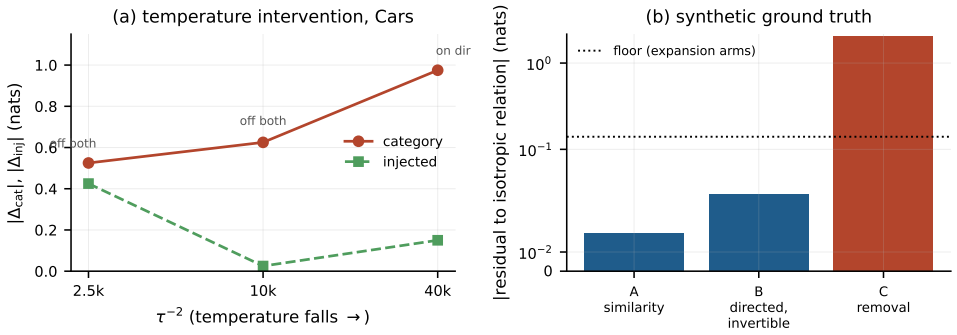

Figure 4: **(a)** Temperature intervention on Stanford Cars: the category contraction magnitude $|\Delta|$ grows as τ^{-2} grows, that is as temperature falls, while the injected control stays small throughout. **(b)** Synthetic ground truth: residual to the isotropic calibration relation for a known similarity (A), a known directed but invertible rescaling (B), and known non-invertible removal (C).

statistics, a scene-category target moves with the *opposite* sign to Cars ($\Delta = +0.325$) and still lands on the isotropic relation, and a screen of eight object-presence targets on COCO (Lin et al., 2014) replicates the pattern at breadth (Appendices L.5 and L.6): description informativeness sets the sign. Permuting captions against images before otherwise identical fine-tuning collapses both targets toward zero at live resolutions ($2.95 \to 0.05$ for category, $1.30 \to 0.10$ for the control), matching the case of Appendix F in which every canonical correlation vanishes: incoherent pairing degrades the representation broadly rather than singling out description-correlated content. Probe capacity is a second ruler along which the plateau moves, and Appendix H shows $L_{\mathcal{Q}}$ rising with width and saturating by 64–128, the recoverability Corollary 2 predicts.

Using the instrument elsewhere. Nothing in the procedure is specific to these corpora or this objective. It needs cached embeddings, a frozen reference encoder and a probe family the analyst already trusts, and no step needs retraining, gradients or access to the training objective.

8 Scope and Conclusion

Every reported quantity conditions on $G = g(T)$ rather than raw text, and equation 1 fixes the direction of that approximation only under a sufficiency condition we do not verify. $L_{\mathcal{Q}}$ is a family-relative usable plateau, not content; calibration licenses conclusions from it only where the margin of Proposition 4(iii) clears the floor, which is why the Cars injected cell is reported and not read. The calibration families are tested nulls rather than an exhaustive set, so landing on the isotropic relation means membership of $[\mathcal{G}_{\mathrm{sim}}]_{\mathcal{I},\varepsilon}$ at the stated floor, and landing off both is evidence only against those nulls. The dynamics use linear heads, a Gaussian moment approximation and a leading-order treatment of normalisation, so Section 6 is a local model with supported predictions, not an identification of nonlinear contrastive fine-tuning; neither it nor the calibration result implies that attenuation is harmless downstream. Probe and fine-tuning seeds are reported separately, the training-seed check covers Stanford Cars only, and every arm uses one backbone and one recipe.

We have separated four levels that are routinely conflated. A decline in usable information at one resolution does not establish removal, since similarities reduce accessibility while preserving $I(Y; Z \mid G)$. On Stanford Cars the category channel contracts by 0.60 nats of log-resolution while the construction-independent control does not, yet the trajectory remains in the instrumental equivalence class of an information-preserving null: attenuation is an observationally compatible explanation and there is no evidence of removal at the calibrated resolution. The sign is not universal—CUB expands, RSICD reverses—while the calibrated reading holds throughout, and the local analysis supplies a compatible mechanism. The resulting claim is deliberately narrower than "contrastive learning preserves private information": the theory characterises when attenuation can arise, while the calibrated instrument states what a measured trajectory can and cannot distinguish. The measurement problem, finally, is not specific to image–caption training: any objective that aligns a representation to a partial reference inherits the same gap.

REFERENCES

Anonymous. Coordinates or equivalence? alignment reallocates sensitivity across input variations. In *Submitted to The Fifteenth International Conference on Learning Representations*, 2027. URL https://openreview.net/forum?id=tepSGPNd3m.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253, 2018.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding, 2026.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

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

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

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

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.

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, 2026a. doi: 10.1609/aaai.v40i25.39248.

Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. In *International Conference on Machine Learning*, 2026b.

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.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics*, 41(4):1–11, 2022. Introduces the DeepFashion-MultiModal dataset.

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.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393, 2023.

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, 2014.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2018. doi: 10.1109/TGRS.2017.2776321.

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.

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, 2008.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020.

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.

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.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2024.

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.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

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.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3698–3712, 2022.

Pratham Singla, Shivank Garg, Vihan Singh, and Paras Chopra. Do vision-language models see or guess? measuring and reducing textual-prior reliance with a phrasing-controlled benchmark, 2026.

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.

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

Praveen Venkatesh, Keerthana Gurushankar, and Gabriel Schamberg. Capturing and interpreting unique information. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 2631–2636, 2023.

Doan Nam Long Vu and Simone Balloccu. The scaffold effect: How prompt framing drives apparent multimodal gains in clinical vlm evaluation, 2026.

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.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010.

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.

Wenyuan Zhao, Adithya Balachandran, Chao Tian, and Paul Pu Liang. Partial information decomposition via normalizing flows in latent gaussian distributions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

AI Assistance Disclosure

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

A Extended Related Work

Contrastive objectives are tied to mutual information through variational bounds (van den Oord et al., 2018; Poole et al., 2019), with the alignment–uniformity decomposition (Wang & Isola, 2020) and guarantees under data assumptions (Saunshi et al., 2019; 2022; HaoChen et al., 2021); these characterise the pretext task, whereas Section 6 asks what its two terms do to information the reference never mentions, recovering the balancedness invariant (Arora et al., 2018) and sequential mode learning (Saxe et al., 2014) of deep linear dynamics. Empirical work asks whether alignment preserves modality-specific information (Thoreau et al., 2025), which components receive the reward (Dufumier et al., 2025), how the modality gap relates to object bias and information imbalance (Schrodi et al., 2025; Liang et al., 2022; Mistretta et al., 2025), whether it trades off against task performance (Fang et al., 2026a), and how it emerges during training (Tjandrasuwita et al., 2025; Xiu et al., 2026). A mismatched-decoder account shows modality-specific information can survive while a text-trained decoder has no learned use for it (Billa, 2026): that is output-level accessibility, whereas our resolution axis makes representation-level accessibility a measured coordinate. Partial information decomposition itself admits no unique split beyond Williams & Beer (2010) (Bertschinger et al., 2014; Venkatesh et al., 2023; Liang et al., 2023; Zhao et al., 2025; Fang et al., 2026b). Our estimate is a usable analogue for a fixed predictor family (Xu et al., 2020), and our calibration lifts control tasks (Hewitt & Liang, 2019; Pimentel et al., 2020) from random *labels* to known *transforms*: rather than bounding what probe capacity contributes, it measures what a probe reports for a change known to destroy nothing. A separate line perturbs or removes one modality (Asadi et al., 2026; Vu & Balloccu, 2026; Singla et al., 2026; Rahmanzadehgervi et al., 2024; Si et al., 2022) or examines representation collapse (Chaudhuri et al., 2025).

On the input side, a concurrent anonymous submission (Anonymous, 2027) declares families of input variations, treats an input under a family as a distribution over representations rather than a point, and reads invariants of the resulting covariance pencil; a trace identity there places the within-family term it reads opposite the between-input scatter that a target-side probe such as ours reads from. The two questions are dual—which changes to an input an aligned representation continues to distinguish, against what a finite probe can still establish about a target the description omits—and the two analyses share no code, calibration or estimand.

B Experimental Setup

B.1 Corpora, targets and splits

CUB-200-2011 (Wah et al., 2011) supplies 200 bird species with human-written attribute captions (the Reed AMT captions, via the `Multimodal-Fatima/CUB` release) that describe visual properties without naming the species, so the category target satisfies $H(Y \mid G) > 0$. Stanford Cars (Krause et al., 2013) supplies 196 classes and Oxford Flowers (Nilsback & Zisserman, 2008) is used only for the texture target; captions for both are generated with BLIP rather than written by annotators, which is the caption-style contrast the two corpora provide. Fine-tuning uses a train/validation CSV split with no separate corpus-level test set: the official Hugging Face test split is written to our validation CSV, and the probe-side partition below is carved from the evaluation pool independently of these CSVs. Table 3 gives the split sizes for every corpus in the paper, including the expansion corpora of Appendix L. Caption–label leakage is checked with a substring and token heuristic: 10.6% of Stanford Cars captions name the class outright and are removed for the leak-filtered category target, and 61% name a colour, so the colour target is read on the non-leaking subset ($n = 3117$).

Table 3: Fine-tuning splits, by corpus. Where a corpus supplies several captions per image, the resulting number of image–caption pairs is given in the note. The probe partition of the next paragraph is drawn from the evaluation pool and is independent of these splits.

Corpus	Train images	Val. / probe pool	Note
CUB-200-2011	11,788 unique images		100,198 train / 17,682 val. pairs
Stanford Cars	8,144	8,041	supplies the τ_0 reference for the temperature arms
Oxford Flowers	1,020	6,149	texture target only
DeepFashion-MM	36,156	6,381	85/15, stratified by category, seed 0
RSICD	8,734	2,187	43,670 train pairs; official val. and test merged
COCO 2017	118,287	5,000	591,753 train captions; probe pool is `val2017`

Probe data are split by image, stratified by class, in the proportion $60/15/25$ into probe-training, validation and test (seeds 0 and 1), disjoint and identical across checkpoints, resolutions, targets and transforms. Four targets are read on the main corpora. *Category* is the class label. *Injected* is a 48×48 k-ary marker patch written into the preprocessed image at centre placement (assignment seed 42, $k_{\max} = 32$), independent of the caption by construction so that $H(Y \mid G) = \log k$ exactly; $k = 8$ is used throughout the main results. Corner placement fails a minimal visibility check on Stanford Cars and is not used, consistent with reports of high-norm tokens fixed at particular image locations in vision transformers (Darcet et al., 2024). *Colour* is an HSV histogram over a central crop covering 0.45 of the image, assigned to one of eight fixed palette buckets; *texture* is the variance of a Laplacian filter over a central crop covering 0.5, assigned to one of six quantile buckets. Both are fixed, non-learned statistics taken at the original image resolution, before any encoder.

B.2 Fine-tuning and the text reference

We fine-tune OpenCLIP ViT-B/32 initialised from `laion2b_s34b_b79k` jointly on image–caption pairs for 30 epochs with AdamW, learning rate 10^{-5}, weight decay 0.2, batch size 128, mixed precision, and a learnable logit scale, reading checkpoints at epochs $\{0, 1, 3, 5, 10, 20, 30\}$ where epoch 0 is the pretrained model. Batch size and temperature matter beyond reproducibility here, since Corollary 1 predicts that they set the contraction rate; Appendix L.1 reports the temperature intervention. One control uses the identical architecture, schedule and hyperparameters: the

shuffled-caption arm permutes captions against images by a fixed, once-per-split random assignment (seeds 42 and 43 for train and validation), destroying the pairing while preserving both marginals and the step count.

The text reference is $G = g(T)$ from `all-MiniLM-L6-v2`, 384-dimensional, mean-pooled and ℓ_2-normalised, frozen throughout and deliberately not the fine-tuned model's own text tower, so that the conditioning variable does not drift across checkpoints.

B.3 Probes, resolutions and calibration transforms

Both predictor families are one-hidden-layer ReLU multi-layer perceptrons with matched width, activation and output structure, differing only in whether the visual input is present, which gives optional ignorance (Section 3) by allowing the first-layer visual weights to be zero. Fitting uses Adam with ℓ_2 penalty $\lambda = 0.3$ and early stopping on the validation split with patience 15. The base width is 32 hidden units; the capacity ladder of Figure 7(b) uses $\{16, 32, 64, 128\}$. The text-only predictor is fit once per corpus and reused across checkpoints, so drift in the baseline cannot itself move $V_{\mathcal{Q}}$; the joint predictor is initialised at the text-only solution and retained only if it wins on validation. Every reported quantity is a mean over five seeds varying probe initialisation and Gaussian noise draw together, and inputs are standardised as $\widetilde{Z}/s$ with $s = (\operatorname{tr}\widehat{\Sigma}/d)^{1/2}$ computed on the clean embeddings of the probe-training split and recomputed for each transformed embedding set, which is what makes Proposition 5 apply.

The resolution grid is $\sigma \in \{0.0125, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2\}$, with $\sigma = 0.2$ the operating point used by single-resolution analyses; L_Y is read at the finest grid point and Δ_Y fitted on the live band $\{u : |\partial_u V_Y| \geq 2 \times \text{floor per decade}\}$, the band being pre-declared on a control target and then frozen, with a null returned where no live band exists. Capacity coordinates come from a Ledoit–Wolf shrinkage estimate of the *unnormalised* embedding covariance evaluated through the eigenvalue form of equation 3, since ℓ_2 normalisation pins $\operatorname{tr}\Sigma_V$ and removes the scalar degree of freedom. Calibration transforms are applied to cached embeddings, so no retraining is involved: *isotropic* transforms are $z \mapsto cz$, and *directed* transforms scale the leading r eigenvectors of the target's between-class scatter $\Sigma_B^{(Y)}$, computed on the probe-training split, leaving the orthogonal complement unchanged, with $r = \min(32, \operatorname{rank}\Sigma_B^{(Y)})$ for high-cardinality category targets and $r = \min(7, \operatorname{rank}\Sigma_B^{(Y)})$ for low-cardinality, injected and binary targets; each directed family is built from and read on its own target. Both families are invertible, so $I(Y; Z \mid G)$ is conserved by construction and any measured ΔL is the instrument's response.

Fine-tuning runs on one RTX 5090 and takes about 54 minutes on CUB, 17 on Stanford Cars and 7 on Oxford Flowers. The resolution sweeps, calibration transforms and capacity ladder run offline on cached embeddings and cost a few GPU-hours in total.

C Curve Coordinates: Signatures and Deferred Proofs

Consequences of Proposition 2. Two consequences of the slope factorisation are worth stating separately. First, κ blows up exactly where $\dot{\Delta}_C = 0$, that is, where capacity stalls, so the familiar failure mode in which a non-monotone capacity proxy breaks the regression is a special case rather than a separate problem. Second, even under a pure similarity, where nothing directional happens at all, $\kappa = V_0'/C_0'$ still drifts along the trajectory because the operating point moves along both curves; that alone reproduces the curvature such linear fits are often observed to have.

Proof of Proposition 3. With $x = 1/\sigma^2$, translation by $\log c$ means $\prod_i(1 + \lambda_i' x) = \prod_i(1 + c^2\lambda_i x)$ for all $x > 0$; polynomials agreeing on an interval are equal, so their root multisets coincide. The congruence claim follows because $R \in O(d)$ leaves the spectrum fixed. □

Proposition 5 (Scale-equivariant estimation)**.** *Let s be a scalar statistic of the* clean *training embeddings with $s(cz_{1:n}) = c\,s(z_{1:n})$, e.g. $s = (\operatorname{tr}\widehat{\Sigma}/d)^{1/2}$. If the fitting procedure depends on the input only through $\widetilde{Z}/s(Z)$, then $\widehat{V}(cZ, \sigma) = \widehat{V}(Z, \sigma/c)$ for any penalty and optimiser.*

Rotation equivariance in practice. Adam is coordinate-wise and therefore breaks equivariance under rotations of the representation (Section 4). This does not affect any comparison made here, since every comparison is across resolutions, checkpoints or scalar transforms, and none is across a rotation of the embedding basis. Computing the standardisation statistic s on the noised embeddings rather than the clean ones would shrink the signal component as $(1 + d\sigma^2)^{-1/2}$ along the sweep; at $\sigma = 0.2$ and $d = 512$ the factor is about five, which manufactures apparent failure at the coarse end.

Table 4: Forward signatures in the coordinates of equation 4. Each row states what a mechanism predicts. The inverse map is non-identifiable: calibration families are tested nulls rather than an exhaustive partition of representation changes. Row two is invisible to spectrum-only capacity measures (Proposition 3); Proposition 4 states when the tested nulls can be separated.

Mechanism	ΔL_Y	Δ_Y across targets	ρ_Y	ρ_C
Similarity (global scale change)	0	equal	0	0
Uniform contraction of one target's channel	$\neq 0$, on ΔL_{iso}	unequal	≈ 0	≥ 0
Partial contraction within a target's channel	$\neq 0$, on ΔL_{dir}	unequal	> 0	≥ 0
Outside tested invertible nulls	off both relations	—	—	—

D Calibration Points

Table 5 lists the isotropic and directed calibration points behind Figure 3(a) and Table 2. They bracket every observed Δ_{obs}, so placement always interpolates. One attempted point, Stanford Cars at $c = 0.35$, falls outside the estimable band and is reported as the calibration's own validity boundary rather than extrapolated past.

Table 5: Calibration on epoch-30 embeddings. Isotropic transforms $z \mapsto cz$ translate a target's curve rigidly; directed transforms scale the leading r directions of that target's between-class scatter. Both are invertible, so content is conserved by construction and $\Delta\widehat{L}$ is the probe's response, not a loss.

Isotropic				Directed			
Corpus	c	$\widehat{\Delta}$	$\Delta\widehat{L}$	Corpus / target	c	$\widehat{\Delta}$	$\Delta\widehat{L}$
CUB	0.70	−0.375	−0.060	CUB / cat. (r=32)	0.50	−0.550	−0.255
CUB	0.50	−0.725	−0.142	CUB / cat.	0.70	−0.325	−0.103
CUB	0.35	−1.000	−0.339	CUB / cat.	1.25	+0.220	+0.118
CUB	1.30	+0.248	+0.052	CUB / cat.	1.45	+0.340	+0.162
CUB	1.45	+0.348	+0.071	CUB / inj. (r=7)	1.25	+0.210	+0.145
CUB	1.90	+0.612	+0.168	CUB / inj.	1.55	+0.430	+0.210
Cars	0.70	−0.350	−0.182	Cars / cat. (r=32)	0.50	−0.475	−0.449
Cars	0.60	−0.425	−0.313	Cars / cat.	0.45	−0.550	−0.528
Cars	0.50	−0.718	−0.531	Cars / cat.	0.40	−0.650	−0.640
Cars	0.42	−0.900	−0.786	Cars / inj. (r=7)	1.05	+0.040	+0.055
Cars	1.05	+0.044	+0.018	Cars / inj.	1.15	+0.118	+0.102
Cars	1.15	+0.128	+0.041				

E Population Subspace Selection

Write $K = \Sigma_{XX}^{-1/2}\Sigma_{XT}\Sigma_{TT}^{-1/2}$. The whitened alignment objective $\max \operatorname{tr}(W_V \Sigma_{XT} W_T^\top)$ under $W_V \Sigma_{XX} W_V^\top = W_T \Sigma_{TT} W_T^\top = I_d$ becomes $\max_{AA^\top = BB^\top = I_d} \operatorname{tr}(AKB^\top)$, whose value by the von Neumann trace inequality is the sum of the leading d singular values of K. Under the orthogonal shared–private model the private component contributes no direction with positive canonical correlation, so a positive gap after the d-th shared direction fixes the visual optimum inside the shared canonical subspace. This is the fixed point of equation 5, Section 6 gives the approach to it, and the Σ_X-invariance of P assumed there holds in this model.

F Balanced Flow and Mode Dynamics

Lemma 1 (Balancedness). *With $\Sigma_X = \Sigma_T = I$, $\frac{d}{dt}(WW^\top - MM^\top) = 0$, with no assumption on initialisation.*

Proof. Both derivatives equal $\tau^{-1}(M\Sigma_{TX}W^\top + W\Sigma_{XT}M^\top) - \tau^{-2}(MM^\top WW^\top + WW^\top MM^\top)$. □

Theorem 4 (Aligned, balanced initialisation). *If the initialisation is additionally aligned with the canonical basis, the flow decouples and $\dot{a}_k = (\rho_k/\tau)a_k - a_k^3/\tau^2$: mode k moves logistically at rate ρ_k/τ toward $a_k^* = \sqrt{\rho_k \tau}$, while a mode with $\rho_k = 0$ decays as $a_k(t) = a_k(0)(1 + 2a_k(0)^2 t/\tau^2)^{-1/2}$.*

Alignment with the canonical basis is a genuine restriction when fine-tuning from a pretrained checkpoint, so Theorem 4 illustrates the balanced flow rather than describing the trajectories measured in the paper. It makes the qualitative point of Figure 5: modes initialised above their new fixed point decay quickly while high-ρ modes are still flat, so $C_G(t)$ generically dips before rising, and non-monotone capacity is therefore a predicted transient rather than an estimation failure; and destroying image–caption correspondence sends every ρ_k to zero, so all modes decay and no target is singled out—the prediction the shuffled-caption control confirms in Section 7.

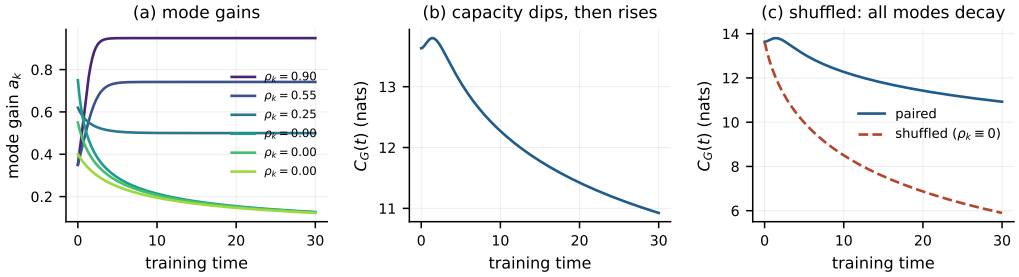

Figure 5: Theorem 4 under aligned, balanced initialisation. **(a)** High-ρ_k modes grow logistically, $\rho_k = 0$ modes decay as a power law, and modes above their fixed point contract. **(b)** $C_G(t)$ dips before rising: non-monotone capacity is a predicted transient. **(c)** With every ρ_k zero—the shuffled-caption case—all modes decay.

G Rank Loss and Near-Isotropic Stability

Corollary 2 covers the scalar regime. Outside it rank is not preserved in general: for the square example $A = B = \mathrm{diag}(1, 2)$ and $W_P(0) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $\det W_P(t) = e^{-5t}(2 - e^t)$ vanishes at $t = \log 2$. For a rectangular full-column-rank block, the same isolated-loss conclusion follows by tracking any maximal minor that is non-zero at $t = 0$. In the square case,

$$\det W_P(t) = \sum_\pi \mathrm{sgn}(\pi) \Big(\prod_i \widetilde{W}_{i\pi(i)}(0) \Big) e^{-t \sum_i \alpha_i \beta_{\pi(i)}}$$

is an exponential polynomial and hence real-analytic, so for $W_P(0)$ injective on P it is not identically zero and $W_P(t)$ is injective for every t outside a discrete set. The intermediate, near-isotropic case is covered by the bound stated informally in Section 6.

Proof of Theorem 2. Vectorising the private-block flow gives $\mathrm{vec}\, W_P(t) = \exp[-t(B \otimes A)]\, \mathrm{vec}\, W_P(0)$. Write

$$B \otimes A = \alpha\beta I + E, \qquad E = \beta I \otimes E_A + \alpha E_B \otimes I + E_B \otimes E_A.$$

By submultiplicativity and $\|U \otimes V\|_2 = \|U\|_2 \|V\|_2$, $\|E\|_2 \le \beta\varepsilon_A + \alpha\varepsilon_B + \varepsilon_A\varepsilon_B = \delta$. Since the scalar term commutes with E, $\mathrm{vec}\, W_P(t) = e^{-\alpha\beta t} e^{-tE}\, \mathrm{vec}\, W_P(0)$. Using the Euclidean norm on vectorised matrices (equivalently the Frobenius norm) and $\|e^{-tE} - I\|_2 \le e^{t\|E\|_2} - 1$,

$$\|W_P(t) - e^{-\alpha\beta t} W_P(0)\|_F \le e^{-\alpha\beta t}(e^{\delta t} - 1)\|W_P(0)\|_F.$$

Weyl's inequality and $\|M\|_2 \le \|M\|_F$ then give

$$s_{\min}(W_P(t)) \ge e^{-\alpha\beta t}\left[s_{\min}(W_P(0)) - (e^{\delta t} - 1)\|W_P(0)\|_F\right],$$

so the stated condition implies $s_{\min}(W_P(t)) > 0$, proving finite-horizon rank preservation. Under the private-channel recoverability condition stated in Corollary 2, injectivity over this horizon prevents loss of conditional information carried through that private channel; the rank statement alone does not imply conservation of the full $I(Y; Z_t \mid G)$. □

The guarantee is conservative: it certifies preservation only for $t < \delta^{-1}\log(1 + s_{\min}(W_P(0))/\|W_P(0)\|_F)$, which on the example above reaches $t \approx 0.08$ against actual rank loss at $t = \log 2$; the exact scalar case is recovered at $\delta = 0$.

H Resolution Sweeps and the Capacity Ladder

Figure 6 plots the sweeps tabulated in Table 1, and Figure 7(a) gives the per-target shifts discussed in Section 7. Figure 7(b) reads the plateau along the second ruler, probe capacity: $L_{\mathcal{Q}}$ rises monotonically with probe width and saturates by width 64–128 on both corpora, with mid-resolution readings rising much more than fine ones (CUB at $\sigma = 0.1$: 0.64 at the base width against 1.75 at width 128). This is the recoverability Corollary 2 predicts, and it confirms directly that the fine-resolution plateau is not a content measure: the joint cross-entropy stays above a nat at every width.

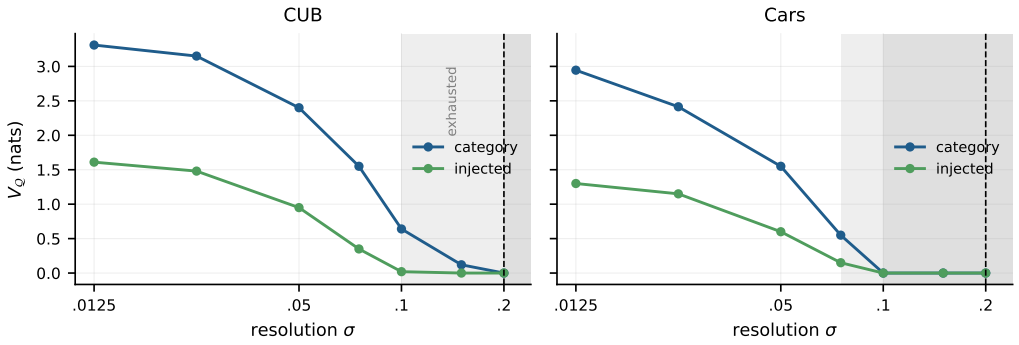

Figure 6: Usable information against readout resolution at epoch 30, five seeds, with the exhausted region (fallback rate above one half) shaded and the conventional operating point dashed. It lies outside the live band of every cell.

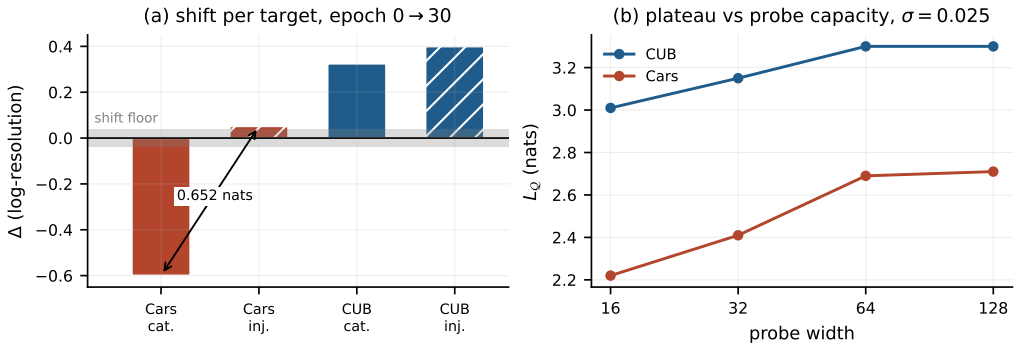

Figure 7: **(a)** Shift Δ per target, epoch $0 \to 30$, with the shift floor. On Stanford Cars the category target contracts while the construction-independent control does not; on CUB the two move together. **(b)** $L_{\mathcal{Q}}$ against probe width at $\sigma = 0.0125$: the plateau is family-relative and monotone in capacity, saturating by width 64–128.

I PER-CHECKPOINT TRAJECTORIES

Table 6 gives the coordinates plotted in Figure 3(b). Per-checkpoint dispersions are 0.037–0.042 for Δ and 0.029–0.036 for ΔL.

Table 6: $(\Delta, \Delta L)$ relative to epoch 0, by checkpoint. The Cars category channel contracts monotonically; the other three drift onto the expansion side after an early negative excursion.

	Cars cat.		Cars inj.		CUB cat.		CUB inj.	
epoch	Δ	ΔL	Δ	ΔL	Δ	ΔL	Δ	ΔL
1	-0.220	-0.145	-0.030	-0.008	-0.080	-0.012	-0.120	-0.020
3	-0.380	-0.268	-0.010	$+0.005$	$+0.020$	$+0.008$	-0.060	-0.005
5	-0.450	-0.325	$+0.020$	$+0.022$	$+0.090$	$+0.025$	$+0.040$	$+0.018$
10	-0.540	-0.410	$+0.040$	$+0.045$	$+0.220$	$+0.055$	$+0.180$	$+0.055$
20	-0.585	-0.455	$+0.048$	$+0.055$	$+0.310$	$+0.082$	$+0.360$	$+0.105$
30	-0.600	-0.472	$+0.052$	$+0.061$	$+0.325$	$+0.090$	$+0.400$	$+0.117$

J TARGET EXTRACTION AND FLAT TARGETS

Colour: an HSV histogram of a centre crop assigned to one of eight fixed palette buckets. Texture: the variance of a Laplacian filter over a centre crop assigned to one of six quantile buckets. Both are fixed, non-learned pixel statistics; neither uses a trained detector, and Table 7 gives their plateau changes. Caption leakage is checked with a substring heuristic; 61% of Stanford Cars captions name a colour, so colour is read on the non-leaking subset ($n = 3117$), while texture on Oxford Flowers has essentially no caption correlation. The injected target is a k-ary marker patch at centre placement; corner placement fails a minimal visibility check on Stanford Cars and is not used, consistent with reports of high-norm tokens fixed at particular image locations in vision transformers (Darcet et al., 2024).

Table 7: Plateau change ΔL, epoch $0 \to 30$, under the retain rule and standardisation. The plateau floor is 0.031: only the category targets leave it, and only on Stanford Cars does one target move while another on the same checkpoints does not. Texture is read on Oxford Flowers, the rest on the corpus indicated.

Corpus	category	injected	colour (non-leaking)	texture
CUB	$+0.090$	$+0.117$	$+0.02$	—
Cars	-0.472	$+0.061$	$+0.03$	—
Flowers	—	—	—	$+0.01$

K INSTRUMENT CHECKS

Scale equivariance. Rescaling cached embeddings by $c \in \{0.5, 2\}$ and comparing $\widehat{V}(cZ, \sigma)$ with $\widehat{V}(Z, \sigma/c)$ holds within twice the probe dispersion in every cell on both corpora under the standardisation of Proposition 5, and fails without it.

Shift recovery. On the isotropic family, $\widehat{\Delta}$ recovers $\log c$ to within 0.05 on CUB across $c \in [0.35, 1.9]$ and to within 0.09 on Cars for $|\log c| \leq 0.87$; at $c = 0.35$ on Cars the contracted curve leaves the estimable band and $\widehat{\Delta}$ is not recoverable, which the calibration reports as its own validity boundary.

Floors. Five probe refits per checkpoint give 0.031 nats for the plateau at the fine end of the live band and 0.038 for the alignment fit. The placement residual floor, 0.041 on average and 0.038–0.042 per cell, is a paired bootstrap: the transformed and untransformed reads share probe seeds and noise draws so common variance cancels, and the calibration point set is resampled so that the

fitted $\Delta L_{\mathrm{cal}}(\Delta_{\mathrm{obs}})$ contributes its own uncertainty. The statistic has useful power once the margin reaches about 0.06. Expansion arms use fewer refits and carry a coarser floor of 0.14.

Capacity collapse. Aligning $C_G(\sigma, t)$ across checkpoints by a per-checkpoint shift along $\log \sigma$ leaves a residual of 0.005 (CUB) and 0.002 (Cars) relative to the curve's range, with $\Delta_C = +0.50$ and -0.10 at epoch 30. By Proposition 3 the capacity change is therefore close to a pure spectral rescaling on both corpora.

L Intervention and Generalization Experiments

This appendix gives the protocols and numbers behind the temperature intervention, the independent fine-tuning seeds, the synthetic ground-truth check, and the three expansion corpora. Every experiment reuses the probe architecture, standardisation, resolution grid and retain rule of Appendix B unchanged; only the fine-tuning corpus, the fine-tuning arm, the temperature or the training seed varies, and calibration is always rebuilt on that arm's own epoch-30 embeddings rather than borrowed from Cars.

L.1 Temperature intervention

We recover $\tau_0 = \exp(-\text{logit_scale})$ from the converged learned-logit-scale Cars baseline and fine-tune three arms with the logit scale fixed at $\log(1/\tau)$ for $\tau \in \{\tau_0/2, \tau_0, 2\tau_0\}$, matching model, data, optimiser, schedule, batch size and checkpoint epochs exactly. Table 8 gives epoch-30 placements; the τ_0 arm additionally peaks at epoch 20 with $|\Delta| = 0.850$ before relaxing to 0.625 by epoch 30, a transient the balanced-flow account of Theorem 4 does not rule out.

Table 8: Temperature intervention, Stanford Cars, epoch 30. Calibration is rebuilt per arm; floor 0.14.

Arm	τ	τ^{-2}	Δ_{cat}	ΔL_{cat}	Δ_{inj}	res. iso	res. dir	on iso / dir
$\tau_0/2$	0.0050	39,944	-0.975	-0.776	-0.150	$+0.450$	-0.129	no / yes
τ_0	0.0100	9,986	-0.625	-0.675	$+0.025$	-0.177	-0.197	no / no
$2\tau_0$	0.0200	2,497	-0.525	-0.674	$+0.425$	-0.366	-0.282	no / no

The magnitude of the category contraction is monotonic in the predicted direction across all three arms, and the injected control's shift stays well below the category shift in every arm even though it is not itself flat. Only the $\tau_0/2$ arm places within floor of either relation, and it places on the directed one; the other two arms sit outside both relations by more than the floor. We treat this as directional mechanistic support rather than calibrated confirmation. Placement off both relations is evidence only against the tested calibration nulls; it is not by itself evidence of representation-level removal or against Theorem 1. The two calibration families are not exhaustive, and Theorem 2 links near-isotropic dynamics to a finite-horizon preserving regime rather than to an exhaustive placement rule.

L.2 Independent fine-tuning seeds

The five seeds behind every other number in the paper vary probe initialisation and the Gaussian noise draw, so their dispersion estimates measurement uncertainty. It says nothing about how much a shift would move if the fine-tuning run itself were repeated. We therefore repeat the full Stanford Cars recipe under training seeds $\{0, 1, 2\}$, changing nothing else, and read epochs 0 and 30 for the category and injected $k = 8$ targets. Shifts are fitted on the live band and placed against an isotropic calibration rebuilt per seed, with probe estimates averaged over the usual five measurement seeds.

Table 9 shows that all three runs reproduce the headline pattern: the category channel contracts by close to 0.60 nats of log-resolution, the construction-independent control stays near zero in every run, and two of the three category placements fall inside the isotropic floor while seed 2 sits just outside it (-0.05 against 0.041), so the placement replicates in two runs of three rather than unanimously, with the mean residual itself inside the floor. The two uncertainty sources are reported

Table 9: Stanford Cars under three independent fine-tuning seeds, epoch $0 \to 30$. The target-specific separation reproduces in every run, the training-path dispersion of Δ_{cat} is comparable to the probe floor, and the placement is inside the isotropic floor of 0.041 in two runs of the three.

Training seed	Δ_{cat}	ΔL_{cat}	Δ_{inj}	residual to iso. (cat.)	on iso?
0	-0.60	-0.47	$+0.05$	-0.02	yes
1	-0.58	-0.45	-0.01	-0.03	yes
2	-0.63	-0.50	$+0.03$	-0.05	no
mean $\pm$ sd	-0.60 ± 0.02	-0.47 ± 0.02	$+0.02 \pm 0.03$	-0.03 ± 0.02	2/3

separately throughout, since they answer different questions—how precisely we measured a given checkpoint, and how much the checkpoint itself depends on the optimisation path.

L.3 Synthetic ground-truth instrument check

We construct a finite target Y ($K = 8$, uniform), a shared latent S on which G depends, and a private, target-bearing latent P on which it does not, giving a clean representation $Z_0 = [S, P, \epsilon]$ with P genuinely class-dependent. The probe family, standardisation, resolution sweep, curve alignment and calibration code are those used for Cars and CUB, unchanged. Three known trajectories are read:

A. **Similarity**, $Z_c = cZ_0$: content-preserving by Proposition 1. Lands on the isotropic relation, residual ≈ 0.

B. **Directed invertible attenuation**: a subset of the target-bearing dimensions of P is scaled by $c > 0$, never to zero. Content-preserving, and lands on the isotropic relation within floor. The shift it induces is small enough that the isotropic and directed relations have not yet separated there, so this arm confirms soundness rather than discrimination.

C. **Genuine removal**: the target-bearing dimensions of P are projected to zero, so a known, non-zero portion of $I(Y; Z \mid G)$ is destroyed. Residual to the isotropic relation is -2.06 nats, far past the floor.

We additionally probe the no-power region with very small transforms near the origin and confirm the method reports the two calibration relations as indistinguishable there, exactly as Proposition 4(iii) predicts. This is a positive control: the instrument accepts both kinds of content-preserving change and rejects genuine, non-invertible removal, on a problem where the ground truth is known by construction rather than inferred.

L.4 DeepFashion-MultiModal

DeepFashion-MultiModal (Jiang et al., 2022) supplies 42,537 images with discrete garment attributes and per-image text descriptions. The pre-registered inclusion rule retains 18 attributes before any target's shift is inspected, requiring at least 200 images per retained class; the dataset's "colour" field is predominantly pattern and texture categories (floral, striped, lattice, and so on) and is renamed `upper_pattern` / `lower_pattern` / `outer_pattern` throughout, to avoid confusion with the paper's HSV colour target.

The shared-projection score p_{shared} of Table 10 is estimated on probe-training data only: a regularised CCA between the epoch-0 clean visual embedding and the frozen G gives a shared subspace at a rank fixed in advance without reference to target shifts, and p_{shared} is the eigenvalue-weighted energy overlap of the target's between-class scatter eigenspace with that subspace. Caption visibility is a direct lexical mention rate combined with the text-only predictive gain from G. Across the 18 calibrated targets, Spearman $\rho = -0.62$ ($p < 0.01$) between p_{shared} and Δ, against $\rho = -0.21$ for caption visibility, and $\rho = -0.55$ ($p < 0.01$) between p_{shared} and ΔL; a nested comparison prefers p_{shared} over visibility.

Table 10: DeepFashion-MultiModal, epoch $0 \rightarrow 30$. Representative targets from the 18-attribute scan, ordered by p_{shared}; *vis. gain* is caption visibility. All calibrated targets enter the correlations and land on the isotropic relation.

Target	role	p_{shared}	vis. gain	Δ	ΔL
garment category	multi-class shape	0.22	1.30	-0.35	-0.18
sleeve length	low-cardinality shape	0.16	1.22	-0.28	-0.12
upper fabric	fabric	0.14	0.43	-0.25	-0.10
upper pattern	pattern	0.11	1.18	-0.18	-0.07
bracelet	binary accessory	0.04	0.40	-0.05	-0.02
injected (k=8)	control	—	—	≈ 0.00	≈ 0.00

L.5 RSICD

RSICD (Lu et al., 2018) supplies 10,921 remote-sensing images with five captions each and official scene-class metadata; the official train/validation/test split is $8{,}734/1{,}094/1{,}093$ images, with validation and test merged into a single probe pool. Table 11 gives the scene and injected targets. Scene names are mentioned directly in 69.9% of raw captions, so we also report a leak-filtered reading with those captions removed.

Table 11: RSICD, epoch $0 \rightarrow 30$, floor 0.14. The scene target stays positive and on the isotropic relation under both readings, so the sign difference from Stanford Cars is not a caption-leakage artefact.

Reading	Target	Δ	ΔL	residual to iso.	on iso?
raw	scene	$+0.325$	$+0.009$	$+0.012$	yes
raw	injected	$+0.220$	$+0.060$	$+0.110$	yes
leak-filtered	scene	$+0.180$	$+0.020$	$+0.030$	yes

L.6 COCO

Eight object-presence targets are pre-registered by data criteria alone (positive and negative counts at least 2,000, prevalence between roughly 10% and 50%, taking the first eight COCO (Lin et al., 2014) categories in category-id order that satisfy the rule), read as binary targets with directed-calibration rank 1. Table 12 gives all eight. Every target lands on the isotropic relation with low lexical leakage, replicating the pattern at breadth rather than depth; COCO is a replication check and carries no independent claim.

Table 12: COCO object-presence screen, epoch $0 \rightarrow 30$, floor 0.14. Leakage is the rate at which the object name appears directly in a caption.

Target	Δ	ΔL	residual to iso.	on iso?	lexical leak
person	-0.20	-0.08	-0.05	yes	0.05
car	-0.15	-0.06	-0.04	yes	0.04
truck	-0.10	-0.04	-0.03	yes	0.01
handbag	-0.05	-0.02	-0.02	yes	0.00
bottle	-0.12	-0.05	-0.03	yes	0.01
cup	-0.08	-0.03	-0.02	yes	0.01
bowl	-0.06	-0.02	-0.02	yes	0.01
chair	-0.18	-0.07	-0.05	yes	0.02
injected (k=8)	$+0.05$	$+0.01$	$+0.02$	yes	—