RECOVERING CLEAN REPRESENTATION SIMILARITY WITH INDEPENDENT REPLICAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling debiasing does not necessarily remove the effect of observation noise on centered kernel alignment (CKA). Even when noise leaves the cross-model HSIC numerator unbiased, it can inflate the two self factors and distort both similarity scores and layer rankings. We introduce Independent-Replica CKA (IR-CKA), which pairs Gram matrices from two conditionally independent observations inside each minibatch HSIC_1 factor and averages the factors before normalization. Under the stated mean-preserving noise assumptions, all three factors are unbiased for their clean finite-pool counterparts. Comparisons with feature averaging and m-replica cross-Grams identify the corrections retained when independence is used elsewhere in the construction. We establish concentration under random partitions of a fixed input pool and $O(n^{-1})$ mean-square error for PSD-stabilized IR with ridge $\lambda_n = O(n^{-1/2})$, under moment and positive self-factor margin conditions. Experiments with frozen ViT and ResNet representations show that IR and its variants recover clean similarity matrices and bidirectional layer relationships more accurately than shared-replica, feature-averaging, and cross-Gram baselines under moderate feature dropout. Under stronger corruption, PSD stabilization restores complete estimates and improves several retrieval measures, although accurate recovery is not uniform across layers or metrics. These results demonstrate the practical value of placing replica independence within all three HSIC factors for recovering clean representation similarity from noisy observations.

1 INTRODUCTION

Centered kernel alignment (CKA) compares neural representations across layers and architectures (Kornblith et al., 2019; Nguyen et al., 2021). Its interpretation depends on how features are observed: stochastic corruption or input augmentation can change measured similarity even when model weights remain fixed. When the intended reference is the representation before corruption, a stable score is insufficient. The estimator must recover that specified clean similarity, including the layer relationships it induces.

Sampling bias and observation noise are different obstacles. The unbiased HSIC estimator of Song et al. (2012) addresses the former, and minibatch CKA aggregates its factors for scalable comparison (Nguyen et al., 2021). Applied to noisy features, however, a sampling-unbiased estimator need not recover clean similarity. Our sample-size controls illustrate this distinction: under independent feature dropout, debiased CKA varies little near the end of the observed range while remaining below its clean score. Can repeated noisy observations recover this score without access to clean features at estimation time?

CKA's normalization is central to this question. Under independent noise across models, the cross-model numerator can retain its clean expectation while both self factors acquire positive noise corrections. Layer-dependent corrections can alter score magnitudes and layer rankings. Independent measurements already support noise correction in representational analysis (Walther et al., 2016; Diedrichsen et al., 2021; Pospisil and Bair, 2022), but their availability does not determine an estimator: averaging features, forming an unbiased cross-Gram, and pairing independent Grams inside HSIC yield different self-factor means.

We introduce Independent-Replica CKA (IR-CKA), which pairs ordinary Gram matrices from two conditionally independent observations inside the numerator and both self factors of HSIC_1, then

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.

pools factors before normalization. Under the stated row-wise independent, mean-preserving noise assumptions, each factor has the clean full-pool expectation. Proposition 1 compares IR with shared-replica (SR) CKA, pooled SR, feature averaging, and m-replica cross-Grams. It identifies where residual corrections remain despite entrywise Gram unbiasedness. Our target is clean finite-pool HSIC_1-normalized linear CKA; unbiased factors do not imply an unbiased ratio.

Theorems 2–3 connect these means to normalized accuracy, accounting for dependence between disjoint minibatches of a random partition. Under moment and positive self-factor margin conditions, PSD projection with a positive, vanishing ridge gives a signed score in $[-1, 1]$ with controlled risk and vanishing bias. Matched-observation experiments then evaluate full cross-layer matrices and bidirectional layer retrieval. IR improves recovery under moderate feature dropout, while strong dropout exposes layer-dependent failures that stabilization cannot fully resolve. The contribution is a statistical account of replica placement, partitioning, and normalization, supported by evaluations separating score recovery, ranking, and availability. General image augmentation need not preserve mean features and falls outside the clean-recovery guarantee.

2 Related work

Measures of representational similarity. Representation comparisons differ in the geometry and invariances they emphasize. Canonical correlation analysis (CCA) identifies maximally correlated linear projections; SVCCA combines dimensionality reduction with CCA (Raghu et al., 2017), whereas projection-weighted CCA weights canonical correlations by their contribution to the original representation (Morcos et al., 2018). Representational similarity analysis (RSA) compares pairwise dissimilarities across shared stimuli (Kriegeskorte et al., 2008; Nili et al., 2014). Orthogonal Procrustes comparisons align feature spaces under orthogonal transformations, and generalized shape metrics formalize related comparisons with explicit invariances (Williams et al., 2021). CKA compares centered Gram matrices and has been widely adopted for neural representation analysis following Kornblith et al. (2019). Linear CKA measures agreement in inner-product geometry, whereas nonlinear kernels, such as the RBF kernel, define alternative sample-similarity structures.

Biases, limitations, and variants of CKA. Several developments address distinct limitations of CKA. The unbiased HSIC estimator of Song et al. (2012) underlies commonly used debiased CKA formulations, although unbiased factors do not imply an unbiased normalized ratio. Murphy et al. (2024) demonstrate that conventional CKA can assign high similarity to independent random representations when feature dimensionality is large relative to sample size, compromising comparisons across representations with different feature-to-sample ratios. Chun et al. (2025) further analyze feature-sampling bias and develop corrections for both input and feature sampling. Beyond sampling bias, Cui et al. (2022) show that input similarity structure can produce high CKA even for random networks, and propose deconfounded CKA using covariate adjustment. Functional interpretation presents another challenge: similarity measures can respond differently to functionally relevant and irrelevant changes (Ding et al., 2021), and CKA can change substantially under transformations that largely preserve model behavior (Davari et al., 2023). For computational scalability, Nguyen et al. (2021) aggregate minibatch HSIC_1 factors before normalization, reducing the memory required for full Gram matrices. Our sample-size experiments show that debiased CKA can change little with the number of inputs yet remain far below clean similarity under independently applied feature dropout (Section 5.1). In this setting, correcting sampling bias does not recover the clean score. This observation motivates using independent noisy replicas to estimate clean similarity, as evaluated in the recovery experiments below.

Independent measurements for similarity estimation and learning. Independent measurements have long supported noise correction in representational analysis. Crossvalidated distances, including crossnobis, multiply pattern differences from independent measurement partitions to remove additive noise bias under appropriate assumptions (Walther et al., 2016; Diedrichsen and Kriegeskorte, 2017). Diedrichsen et al. (2021) introduce whitened unbiased-distance-matrix similarity (WUC) and relate it to linear CKA, including a cross-partition second-moment formulation under i.i.d. measurement noise. Our MCG baseline adapts the cross-partition second-moment construction to minibatches, and MCG1 substitutes HSIC_1 for ordinary HSIC; neither implements general covariance-weighted WUC. Pospisil and Bair (2022) correct the numerator and denominator of squared correlation between

expected neural responses using repeated measurements. Their target is closely related to ordinary centered linear CKA, which equals squared Pearson correlation for scalar representations; this identity does not generally extend to finite-sample HSIC_1-normalized CKA.

Related principles also appear in learning and statistical estimation. Noise2Noise learns restoration functions from independently corrupted pairs under a suitable conditional-mean condition (Lehtinen et al., 2018). Noise2Self uses conditional noise independence across coordinate groups and restrictions on the predictor (Batson and Royer, 2019). Double/debiased machine learning combines cross-fitting, which separates nuisance fitting from estimating-equation evaluation, with Neyman-orthogonal scores (Chernozhukov et al., 2018). These methods use different sources of independence for different targets. We study its placement within all three minibatch HSIC_1 factors: our mean calculations distinguish IR from shared-replica and cross-Gram constructions, and our analysis establishes concentration under random partitioning and risk guarantees for PSD stabilization. The target is clean finite-pool HSIC_1-normalized linear CKA under the stated noise assumptions.

3 Observations, targets, and estimators

Fix clean row features $x_i \in \mathbb{R}^{d_A}$ and $y_i \in \mathbb{R}^{d_B}$ for $i \in [n]$, collected in X, Y. Observe $\widetilde{X}^{(s)} = X + E^{(s)}$ and $\widetilde{Y}^{(s)} = Y + F^{(s)}$, $s = 1, \dots, m$, with zero-mean noise conditional on the clean pool. The $m \geq 2$ complete replicas are conditionally independent and identically distributed. The joint noise rows $(e_i^{(s)}, f_i^{(s)})$ are independent across inputs; within each row, dependence between feature coordinates and between the two models is allowed. Write $C_i = \mathbb{E} e_i^{(s)} f_i^{(s)\top}$, $\Sigma_{A,i} = \mathbb{E} e_i^{(s)} e_i^{(s)\top}$, and $\Sigma_{B,i} = \mathbb{E} f_i^{(s)} f_i^{(s)\top}$. Let $A = XX^\top$, $B = YY^\top$, and $K_A^{(s)} = \widetilde{X}^{(s)} \widetilde{X}^{(s)\top}$, with $K_B^{(s)}$ analogous. The mean identities require finite noise second moments. Concentration additionally requires uniform moment control: uniformly bounded noise fourth moments, bounded clean rows, and fixed dimensions suffice. We do not require positive noise variance for the minibatch results. Expectations in Sections 3–4 condition on X, Y; the experiments use $m = 2$.

For symmetric $b \times b$ matrices P, Q, $b \geq 4$, let P°, Q° be their zero-diagonal versions. With $r_i^P = \sum_{j \neq i} P_{ij}$ and $s_P = \sum_i r_i^P$, define the estimator of Song et al. (2012):

$$\mathrm{HSIC}_1(P, Q) = \frac{1}{b(b-3)} \left[\sum_{i \neq j} P_{ij} Q_{ij} + \frac{s_P s_Q}{(b-1)(b-2)} - \frac{2}{b-2} \sum_i r_i^P r_i^Q \right]. \tag{1}$$

This positive-semidefinite bilinear form is invariant to arbitrary changes on either diagonal. Lemma A1 gives its U-centered inner-product and four-sample U-statistic representations.

For comparison define ordinary HSIC by $h_0(P, Q) = \mathrm{tr}(PH_bQH_b)/(b-1)^2$, $H_b = I_b - \mathbf{1}\mathbf{1}^\top/b$, and write $h_1 = \mathrm{HSIC}_1$. Subscripts r restrict a Gram to batch $\mathcal{B}_r$, of size $b \geq 4$. All methods average factors before forming their ratio:

$$\bar{Z}_q = \beta^{-1} \sum_{r=1}^{\beta} Z_{q,r}, \quad q \in \{N, A, B\}, \qquad \widehat{R} = \bar{Z}_N / \sqrt{\bar{Z}_A \bar{Z}_B}. \tag{2}$$

A score is undefined if a required factor is nonfinite or either self-factor is nonpositive. A finite score outside $[-1, 1]$ is not counted as undefined.

SR, pooled SR, and IR. These estimators use the first two replicas. Put $T_r^{st} = h_1(K_{A,r}^{(s)}, K_{B,r}^{(t)})$ and $S_{A,r}^{st} = h_1(K_{A,r}^{(s)}, K_{A,r}^{(t)})$, with S_B analogous. Define

	$Z_{N,r}$	$Z_{A,r}$	$Z_{B,r}$
SR	T_r^{11}	$S_{A,r}^{11}$	$S_{B,r}^{11}$
pooled SR	$(T_r^{11} + T_r^{22})/2$	$(S_{A,r}^{11} + S_{A,r}^{22})/2$	$(S_{B,r}^{11} + S_{B,r}^{22})/2$
IR	$(T_r^{12} + T_r^{21})/2$	$S_{A,r}^{12}$	$S_{B,r}^{12}$.

(3)

SR has the pooled-factor form of Nguyen et al. (2021). Pooled SR uses the same two observations as IR, but pairs same-index Grams in each factor. Shared replica indices do not require shared dropout masks across models. The three IR factors need not be independent.

MCG and MCG1. For $m \geq 2$, both average over all ordered pairs of distinct replicas:

$$\check{A}_{m,r} = \frac{\sum_{s \neq t} \widetilde{X}_r^{(s)} \widetilde{X}_r^{(t)\top}}{m(m-1)}, \qquad \check{B}_{m,r} = \frac{\sum_{s \neq t} \widetilde{Y}_r^{(s)} \widetilde{Y}_r^{(t)\top}}{m(m-1)}. \tag{4}$$

The sums run over $s, t \in [m]$. These cross-Grams are symmetric and unbiased, including on the diagonal. Their factors are $(h_a(\check{A}_{m,r}, \check{B}_{m,r}), h_a(\check{A}_{m,r}, \check{A}_{m,r}), h_a(\check{B}_{m,r}, \check{B}_{m,r}))$, with $a = 0$ for MCG and $a = 1$ for MCG1. Thus MCG1 uses HSIC_1 in *all* factors. At $m = 2$, Eq. (4) reduces to the symmetrized two-replica cross-Gram used in our experiments. Both h_a are positive-semidefinite bilinear forms on symmetric matrices, so the pooled ratios are bounded by one in absolute value whenever defined.

IR-clip, IR-PSD. IR-clip clips a defined IR score to $[-1, 1]$ and preserves undefined entries. IR-PSD projects the *pooled* factor matrix:

$$\widehat{\Gamma} = \begin{pmatrix} \bar{Z}_A & \bar{Z}_N \\ \bar{Z}_N & \bar{Z}_B \end{pmatrix}, \quad \widehat{\Gamma}_{+,\lambda_N} = \Pi_+(\widehat{\Gamma}) + \lambda_N I_2, \quad \widehat{R}^{+}_{\mathrm{IR}}(\lambda_N) = \frac{(\widehat{\Gamma}_{+,\lambda_N})_{12}}{\sqrt{(\widehat{\Gamma}_{+,\lambda_N})_{11}(\widehat{\Gamma}_{+,\lambda_N})_{22}}}. \tag{5}$$

Here Π_+ clips negative eigenvalues of a symmetric 2×2 matrix and $\lambda_N > 0$ is deterministic. The score is defined for finite factors and belongs to $[-1, 1]$. Projection is per layer pair, not on the entire layer-similarity matrix.

Feature mean. This method first averages the m observations and then forms ordinary Grams:

$$\overline{X}_m = \frac{1}{m} \sum_{s=1}^{m} \widetilde{X}^{(s)}, \quad \overline{Y}_m = \frac{1}{m} \sum_{s=1}^{m} \widetilde{Y}^{(s)}, \quad K_A^{\mathrm{F}(m)} = \overline{X}_m \overline{X}_m^\top, \quad K_B^{\mathrm{F}(m)} = \overline{Y}_m \overline{Y}_m^\top. \tag{6}$$

Its factors are defined as for MCG1, replacing the cross-Grams by $K_A^{\mathrm{F}(m)}, K_B^{\mathrm{F}(m)}$, and are pooled as in Eq. (2). The definition also applies at $m = 1$, where it reduces to SR. The clean-matrix and layer-retrieval comparisons use all eight methods with the same observations and partitions; every method using repeated observations uses $m = 2$.

4 Factor means and concentration

We first compute expectations over observation noise for a fixed batch, then relate uniform minibatches to the full pool. Concentration and normalization are treated separately below.

Clean factors and noise corrections. For a batch $\mathcal{B}$ of size b, define

$$\Xi_{AB}^{(\mathcal{B})} = h_1(A_\mathcal{B}, B_\mathcal{B}), \qquad \Xi_A^{(\mathcal{B})} = h_1(A_\mathcal{B}, A_\mathcal{B}), \qquad \Xi_B^{(\mathcal{B})} = h_1(B_\mathcal{B}, B_\mathcal{B}).$$

The ordinary-HSIC targets $\Xi_{q,0}^{(\mathcal{B})}$ replace h_1 by h_0. Throughout, a numerator subscript N is interchangeable with AB. Separate the SR noise correction into terms linear and quadratic in the row noise covariances:

$$\begin{aligned} \mathcal{L}_{AB}^{(\mathcal{B})} &= \frac{\sum_{i \neq j} (x_j^\top C_i y_j + x_i^\top C_j y_i)}{b(b-1)} - \frac{2 \sum_{i,j,k \text{ distinct}} x_j^\top C_i y_k}{b(b-1)(b-2)}, \\ \mathcal{Q}_{AB}^{(\mathcal{B})} &= \frac{\sum_{i \neq j} \langle C_i, C_j \rangle_F}{b(b-1)}, \qquad \Psi_q^{(\mathcal{B})} = \mathcal{L}_q^{(\mathcal{B})} + \mathcal{Q}_q^{(\mathcal{B})}. \end{aligned} \tag{7}$$

All sums are within $\mathcal{B}$. Self terms substitute $(Y, C_i) = (X, \Sigma_{A,i})$ on side A, and $(X, C_i) = (Y, \Sigma_{B,i})$ on side B. For ordinary-HSIC cross-Grams, let $\Delta_{A,m} = \check{A}_{m,\mathcal{B}} - A_\mathcal{B}$ and $\Delta_{B,m} = \check{B}_{m,\mathcal{B}} - B_\mathcal{B}$, and define

$$\Omega_{AB,0}^{(m;\mathcal{B})} = \frac{\mathbb{E}\langle H_b \Delta_{A,m} H_b, H_b \Delta_{B,m} H_b \rangle_F}{(b-1)^2}. \tag{8}$$

Self terms use the same model in both slots. This correction includes diagonal entries; Eq. (28) gives its covariance expansion, with coefficients $2/m$ and $1/[m(m-1)]$.

Table 1: Factor means conditional on X, Y and a fixed batch. Batch superscripts are suppressed, and $\Psi_q = \mathcal{L}_q + \mathcal{Q}_q$. Feature mean, MCG, and MCG1 use m replicas; all experiments use $m = 2$. Ordinary-HSIC terms generally retain batch-size dependence after subset averaging. IR-clip and IR-PSD share the unmodified IR input factors.

Method	$\mathbb{E}Z_N$	$\mathbb{E}Z_A$	$\mathbb{E}Z_B$
SR	$\Xi_{AB} + \Psi_{AB}$	$\Xi_A + \Psi_A$	$\Xi_B + \Psi_B$
Pooled SR	$\Xi_{AB} + \Psi_{AB}$	$\Xi_A + \Psi_A$	$\Xi_B + \Psi_B$
Feature mean (m)	$\Xi_{AB} + \frac{\mathcal{L}_{AB}}{m} + \frac{\mathcal{Q}_{AB}}{m^2}$	$\Xi_A + \frac{\mathcal{L}_A}{m} + \frac{\mathcal{Q}_A}{m^2}$	$\Xi_B + \frac{\mathcal{L}_B}{m} + \frac{\mathcal{Q}_B}{m^2}$
MCG (m)	$\Xi_{AB,0} + \Omega^{(m)}_{AB,0}$	$\Xi_{A,0} + \Omega^{(m)}_{A,0}$	$\Xi_{B,0} + \Omega^{(m)}_{B,0}$
MCG1 (m)	$\Xi_{AB} + \frac{\mathcal{L}_{AB}}{m} + \frac{\mathcal{Q}_{AB}}{m(m-1)}$	$\Xi_A + \frac{\mathcal{L}_A}{m} + \frac{\mathcal{Q}_A}{m(m-1)}$	$\Xi_B + \frac{\mathcal{L}_B}{m} + \frac{\mathcal{Q}_B}{m(m-1)}$
IR	Ξ_{AB}	Ξ_A	Ξ_B

Proposition 1 (Expectations of minibatch HSIC factors)**.** *Fix X, Y and $\mathcal{B}$, and assume the observation model in Section 3. For $q \in \{N, A, B\}$, SR and pooled SR satisfy $\mathbb{E}Z_q = \Xi_q^{(\mathcal{B})} + \Psi_q^{(\mathcal{B})}$, whereas IR satisfies $\mathbb{E}Z_q^{\mathrm{IR}} = \Xi_q^{(\mathcal{B})}$. The m-replica constructions have means*

$$\begin{aligned} \mathbb{E}Z_q^{\mathrm{F}(m)} &= \Xi_q^{(\mathcal{B})} + \frac{\mathcal{L}_q^{(\mathcal{B})}}{m} + \frac{\mathcal{Q}_q^{(\mathcal{B})}}{m^2}, \\ \mathbb{E}Z_q^{\mathrm{C1}(m)} &= \Xi_q^{(\mathcal{B})} + \frac{\mathcal{L}_q^{(\mathcal{B})}}{m} + \frac{\mathcal{Q}_q^{(\mathcal{B})}}{m(m-1)}, \\ \mathbb{E}Z_q^{\mathrm{C0}(m)} &= \Xi_{q,0}^{(\mathcal{B})} + \Omega_{q,0}^{(m;\mathcal{B})}. \end{aligned} \tag{9}$$

Here $\mathbb{E}\check{A}_{m,\mathcal{B}} = A_{\mathcal{B}}$ and $\mathbb{E}\check{B}_{m,\mathcal{B}} = B_{\mathcal{B}}$. At $m = 2$, the MCG1 correction is $\Psi_q^{(\mathcal{B})}/2$.

For every h_1 factor construction in Table 1, a uniform b-subset $\mathcal{B}$, independent of the observations, satisfies

$$\mathbb{E}_{\mathcal{B}}[Z_{q,\mathcal{B}} \mid \mathcal{W}] = Z_{q,[n]}, \tag{10}$$

where $\mathcal{W}$ denotes the complete observed replica pool and $Z_{q,[n]}$ applies the same factor construction to all n rows. Consequently, taking noise expectations replaces $(\mathcal{B})$ by $[n]$ in all h_1 rows of Table 1; pooled factor means are independent of b, also under a uniform random partition. This conclusion does not generally hold for ordinary-HSIC MCG, and none of these identities asserts unbiasedness after normalization, clipping, or PSD projection.

The proof is in Appendix A.2. If the models' noises are independent within a replica, $C_i = 0$ and the numerator corrections vanish, but nonnegative self-factor corrections generally remain. The corrections for feature mean and cross-Grams vanish as $m \to \infty$ for a fixed clean batch and noise distribution, but generally remain at finite m. IR removes them by placing independence between the arguments of each HSIC_1 factor, already with two replicas.

4.1 Random partitions and normalized risk

The factor means specify the targets; concentration determines whether a finite noisy pool estimates them reliably. Scheme (P) uniformly permutes $n = \beta b$ identities and splits them into β blocks. Joint noisy rows are independent across inputs and independent of the partition. Scheme (I), included for comparison, samples independent uniform b-subsets and regenerates fresh noise for every occurrence. Set $N = n_{\mathrm{eff}} = n$ in (P) and $N = \beta b$ in (I). With $m_q = \mathbb{E}\bar{Z}_q$, define

$$m_A, m_B \geq m_* > 0, \qquad r = \frac{m_N}{\sqrt{m_A m_B}}, \qquad r_{\mathrm{IR}} = R_0 = \frac{\Xi_{AB}^{[n]}}{\sqrt{\Xi_A^{[n]} \Xi_B^{[n]}}}. \tag{11}$$

For SR, $r_{\mathrm{SR}} = (\Xi_{AB}^{[n]} + \Psi_{AB}^{[n]}) / \sqrt{(\Xi_A^{[n]} + \Psi_A^{[n]})(\Xi_B^{[n]} + \Psi_B^{[n]})}$. This is its factor-mean center, not necessarily $\mathbb{E}\widehat{R}_{\mathrm{SR}}$.

Theorem 2 (Factor and ratio concentration). *For SR and IR, suppose the four-input factor kernels in Eqs.* (19) *and* (20) *have second moments bounded by M, and the uniform margin in Eq.* (11) *holds. Under (I) or (P),*

$$\mathrm{Var}(\bar{Z}_q) \le C_1 M/N, \qquad C_1 = 80 \text{ in (I)}, \quad C_1 = 272 \text{ in (P)}. \tag{12}$$

For the theoretical safeguard $\widehat{R}_\lambda = \bar{Z}_N / \sqrt{(\bar{Z}_A \vee m_A/2)(\bar{Z}_B \vee m_B/2)}$,

$$\mathbb{E}|\widehat{R}_\lambda - r|^2 \le C/N, \qquad \mathbb{E}\widehat{R}_\lambda - r = O(N^{-1}), \qquad \mathbb{P}(|\widehat{R}_\lambda - r| > \varepsilon) \le C/(N\varepsilon^2). \tag{13}$$

Any measurable completion agreeing with the raw ratio on $\{\bar{Z}_A \ge m_A/2, \bar{Z}_B \ge m_B/2\}$ obeys

$$\mathbb{P}(|\widehat{R}^{\mathrm{comp}} - r| > \varepsilon) \le C'/N + C''/(N\varepsilon^2). \tag{14}$$

Constants depend on M, m_*. Uniform fourth moments and bounded clean rows imply the kernel condition. If the absolute kernels are bounded by L, then $\mathbb{P}(\max_q |\bar{Z}_q - m_q| > u) \le 6\exp[-Nu^2/(640L^2)]$.

Proofs, including the dependence between partition blocks, are in Appendix B. The safeguard uses unknown population thresholds; its expectation bound is not a claim about an unprotected ratio near zero denominators. An implementable safeguard with a known margin is analyzed in Corollary A7. Increasing the number of partitions of the same noisy pool does not multiply N: it reduces only partition variance, not full-pool noise variance (Eq. (38)).

4.2 PSD STABILIZATION

For IR, $\Gamma_0 = \mathbb{E}\widehat{\Gamma} = \left(\begin{smallmatrix} m_A & m_N \\ m_N & m_B \end{smallmatrix}\right) \succeq 0$ by U-centered Gram geometry. Projection need not preserve unbiased factors, but controls normalized risk.

Theorem 3 (IR-PSD concentration and vanishing bias). *Under the IR conditions of Theorem 2, put $a_N = 4m_*^{-1}(2\sqrt{C_1 M/N} + \sqrt{2}\lambda_N)$. Then*

$$\mathbb{E}|\widehat{R}^+_{\mathrm{IR}}(\lambda_N) - R_0|^2 \le a_N^2, \qquad |\mathbb{E}\widehat{R}^+_{\mathrm{IR}}(\lambda_N) - R_0| \le a_N. \tag{15}$$

For every $0 < \alpha < 1$, with probability at least $1 - \alpha$,

$$|\widehat{R}^+_{\mathrm{IR}}(\lambda_N) - R_0| \le \frac{4}{m_*}\left(2\sqrt{\frac{C_1 M}{N\alpha}} + \sqrt{2}\lambda_N\right). \tag{16}$$

Under bounded kernels, the same probability assertion holds with

$$|\widehat{R}^+_{\mathrm{IR}}(\lambda_N) - R_0| \le \frac{4}{m_*}\left(2L\sqrt{\frac{640\log(6/\alpha)}{N}} + \sqrt{2}\lambda_N\right). \tag{17}$$

For $\lambda_N = O(N^{-1/2})$, mean-square error is $O(N^{-1})$ and absolute bias is $O(N^{-1/2})$.

Appendix B.4 derives these bounds from Theorem 2 and the nonexpansiveness of PSD projection. The result allows boundary targets $|R_0| = 1$ but does not promise $O(N^{-1})$ bias there. A small self-factor margin can still make a bounded score inaccurate. An absolute ridge and unweighted projection also need not preserve finite-sample invariance to separate rescaling of the models.

Our clean-target recovery guarantees apply under independent replicas and mean-preserving feature noise. The inverted feature dropout used in our experiments satisfies the mean-preservation condition, whereas general image augmentation need not.

5 EXPERIMENTS

We first compare sampling debiasing under clean and corrupted observations across ViT-L/16–ViT-B/16, ViT-B/16–ResNet-50, and ResNet-50–ResNet-34. We then evaluate clean-matrix recovery and bidirectional layer retrieval primarily on ViT-B/16–ResNet-50 under independent feature dropout. All networks are pretrained and frozen. Basic crop/flip and RandAugment appear in the sample-size controls; the recovery experiments use mean-preserving feature corruption.

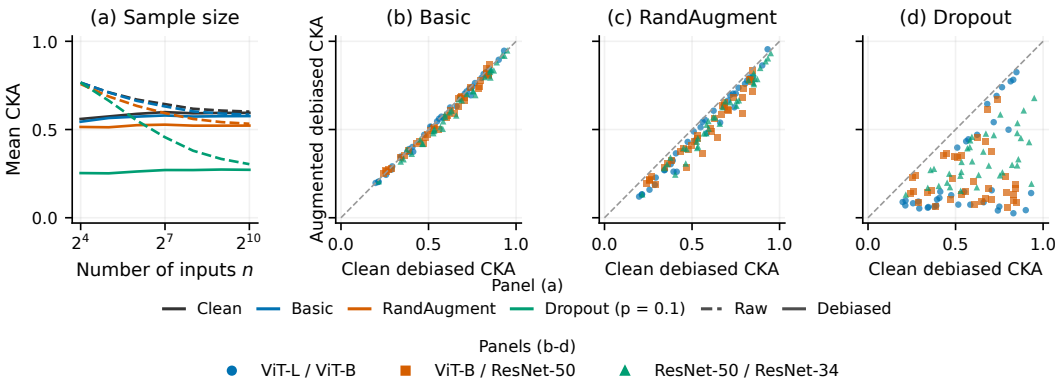


Figure 1: Sampling debiasing and clean-score recovery. (a) Raw (dashed) and debiased (solid) CKA, equally averaged over all 108 layer pairs. (b–d) Clean versus augmented debiased CKA at $n = 1024$; each point retains one layer pair, with model pair indicated by color and marker. Points below the diagonal have lower augmented scores. All panels show recorded means without confidence intervals.

5.1 Sampling debiasing does not ensure clean-score recovery

We compare full-sample raw and debiased CKA, $R_a(K, L) = h_a(K, L)/\sqrt{h_a(K, K)h_a(L, L)}$ for $a = 0, 1$, on CIFAR-10 test representations from ViT-L/ViT-B, ViT-B/ResNet-50, and ResNet-50/ResNet-34. Six readouts per model give 108 layer pairs. We vary $n \in \{16, 32, 64, 128, 256, 512, 1024\}$ under clean observations, basic crop/flip, RandAugment, and feature dropout ($p = 0.1$). Image augmentations are shared across the two models: basic uses crop padding 4 and flip probability 0.5, and RandAugment additionally uses magnitude 9. Feature-dropout masks are independent across models.

We first examine whether sampling debiasing is sufficient to preserve clean representation similarity under corruption. Debiased CKA shows less dependence on sample size than raw CKA (Fig. 1a), consistent with its correction of finite-sample HSIC bias. Its agreement with the clean scores, however, depends strongly on the corruption considered. At $n = 1024$, the mean augmented-minus-clean differences are -0.0186 under basic augmentation and -0.0725 under RandAugment, compared with a substantially larger decrease of -0.3230 under feature dropout (Fig. 1b–d). For dropout, the scores are lower than their clean counterparts for all 108 evaluated layer pairs.

These results highlight a limitation of sampling debiasing as a means of recovering clean similarity. Although the average discrepancies under basic augmentation and RandAugment are smaller, debiased CKA does not preserve the clean scores under independent feature dropout. Inverted feature dropout preserves each feature in conditional expectation, making it a controlled setting in which to distinguish sampling-bias correction from noise correction. The following experiments therefore evaluate whether IR and its variants recover the clean similarity matrices and layer relationships in this setting.

5.2 Recovering the clean similarity matrix

Frozen ImageNet-pretrained ViT-B/16 and ResNet-50 provide 13 and 15 readouts on all 10,000 CIFAR-10 test images, excluding classification heads. We apply independent inverted feature dropout at $p \in \{0, 0.1, 0.3, 0.5, 0.7\}$. Each of five paired trials pools two partitions of the same noisy pool with $b = 1{,}000$. All eight methods share observations and partitions; SR uses the first observation. The reference is full-pool clean HSIC_1-normalized CKA, also for MCG, whose finite-sample ordinary-HSIC target differs. Matrix RMSE, whole-matrix Spearman correlation, and row-wise maximizing-set Jaccard are computed within trials, then summarized by trial means and SDs. Incomplete-matrix RMSE conditions on finite entries; the other metrics require completeness. Appendix C gives the protocols and definitions.

At $p = 0.1$, the IR variants recover scores with near-zero matrix RMSE, while the other five estimators incur substantial error (Figure 3). Figure 2 shows their different layer-pair patterns: SR and MCG1 scores lose information in early and middle readouts. The accompanying factor

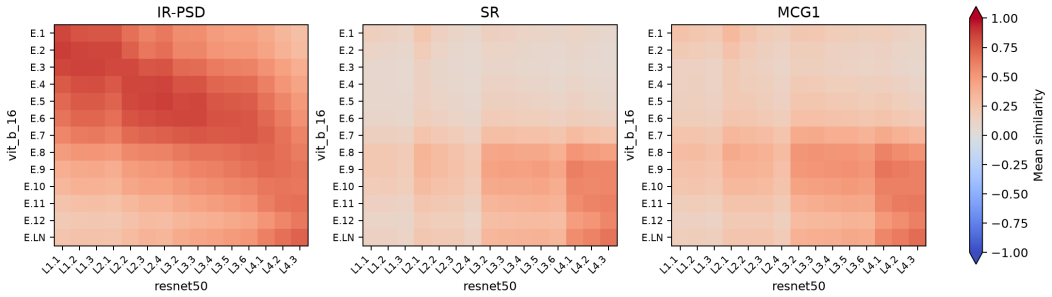


Figure 2: ViT-B/16 versus ResNet-50 at feature dropout $p = 0.1$. Panels show trial-mean IR-PSD, SR, and MCG1 scores, respectively, using a common color scale; rows are ViT readouts and columns are ResNet blocks. SR and MCG1 attenuate many early- and middle-layer scores. Clean-target accuracy is assessed separately in Figure 3; the clean-reference heatmap can be found in Fig. 5.

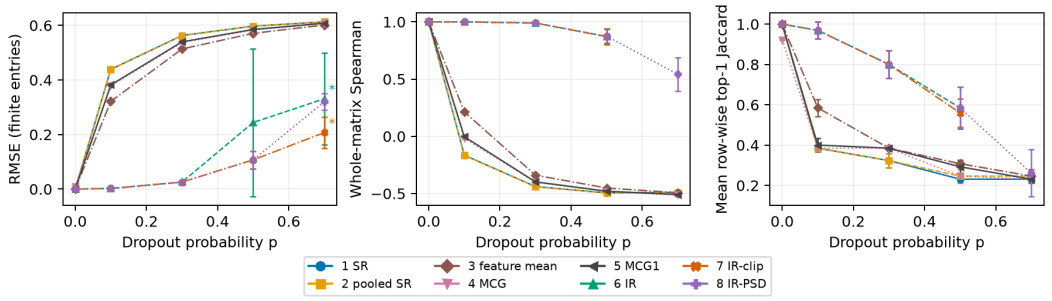


Figure 3: Clean-matrix recovery for ViT-B/16 versus ResNet-50. Left: RMSE over layer-pair entries within each trial; middle: Spearman correlation of vectorized matrices; right: mean row-wise top-1 Jaccard. Points and bars report means and sample SDs across five trials. Asterisks mark conditional RMSE computed on finite entries. At $p = 0.7$, IR and IR-clip have incomplete matrices in every trial, so their Spearman and Jaccard points are unavailable. Legend indices identify overlapping curves.

diagnostics support Proposition 1: median numerator ratios to clean values are approximately one for pooled SR, MCG1, and IR, whereas ViT self-factor ratios are $4.991, 2.994, 1.000$ and ResNet ratios are $1.699, 1.349, 1.000$. Thus MCG1 retains approximately half the excess self-factor contribution despite using independent replicas. At $p = 0.5$, raw IR remains defined but reaches 4.615. IR-clip and IR-PSD reduce matrix RMSE from 0.245 ± 0.271 to 0.107 ± 0.032. Their RMSEs agree to the reported precision; this is a stabilization benefit, not evidence of a PSD-specific advantage. At $p = 0.7$, 251 of 975 layer-pair–trial IR scores are undefined because a self factor is negative, and 53 defined scores exceed one. Failures concentrate in early ViT layers. IR-PSD produces complete matrices but projects 304 factor matrices to scores with absolute values above 0.999; 23 scores remain negative. Appendix C.1 traces these failures to self-factor fluctuations and noisy numerators.

IR-PSD has lower complete-matrix RMSE than pooled SR in every trial at $p = 0.7$, and lower trial-averaged squared error on 177 of 195 layer pairs. Nevertheless, their Jaccard values are similar (0.262 ± 0.117 versus 0.246 ± 0.034). Accurate score reconstruction and agreement on maximizing layers are therefore distinct outcomes. The next experiment evaluates retrieval with explicit clean-score tolerances, tie handling, and missingness penalties.

5.3 Recovering Clean Layer Relationships

Layer matching retrieves candidate layers using noisy similarities and evaluates their clean scores. We test ViT-to-ResNet and reverse retrieval between six readouts per model on 4,096 CIFAR-10 test images. Rows and columns define different anchor sets, and matching is not one-to-one. Three seeds provide five trials each, with two partitions per trial and $b = 128$. Uniform dropout uses $p = 0.625s$; heterogeneous rates range from $0.25s$ to s.

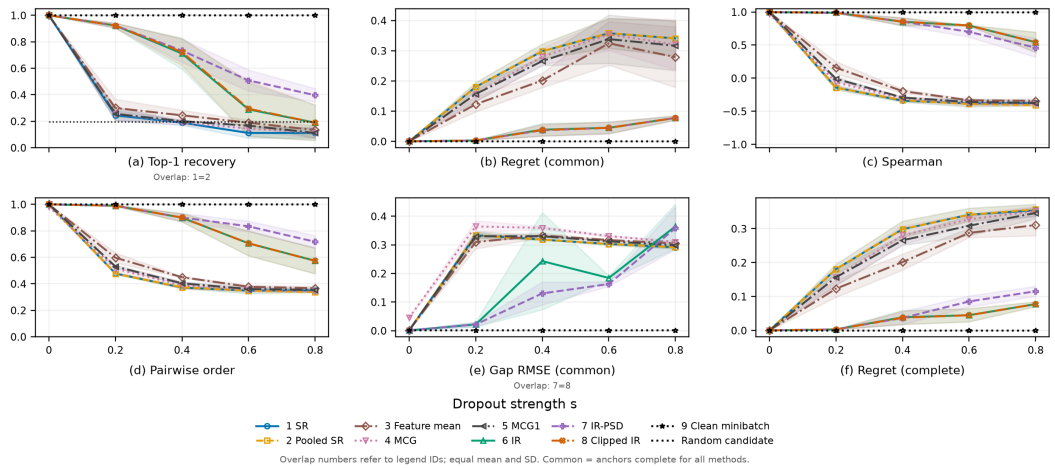


Figure 4: ResNet-to-ViT retrieval under heterogeneous dropout: (a) top-1 recovery, (b) common-anchor clean regret, (c) Spearman correlation, (d) pairwise order accuracy, (e) common-anchor gap RMSE, and (f) method-specific complete-anchor regret. Curves aggregate six anchors and six candidates each; shading is the SD of three seed means. Panels (b,e) require common complete anchors; (c,f) use method-specific valid anchors. Complete results and readouts appear in Appendix D.

Top-1 recovery measures the fraction of noisy maximizers within 0.005 of the clean maximum, assigning zero to incomplete anchors. Clean regret, anchor-wise Spearman, pairwise order accuracy, and pairwise-gap RMSE assess selection, ranking, and score spacing (Appendix D.2). This Spearman compares candidates within anchors, unlike the whole-matrix metric above. Uncertainty is the SD of three seed means after averaging trials.

At uniform $s = 0.2$, IR variants recover $92.2\%/97.8\%$ of forward/reverse matches, versus $50.0\%/16.7\%$ for MCG1 and pooled SR. IR gap RMSE is $0.0077/0.0146$, versus MCG1's $0.1553/0.3287$. Heterogeneous dropout shows the same low-noise advantage (Table 2); all methods are complete, excluding missingness as its explanation (see Fig. 7,Fig. 8). At heterogeneous $s = 0.8$ (Fig. 4), reverse-direction recovery is 39.6% for IR-PSD, 18.9% for IR, and 11.1% for MCG1. IR-PSD remains complete but has a worse common-anchor gap RMSE than MCG1 (0.358 versus 0.298), despite better order accuracy (71.7% versus 35.3%). The common-anchor gap RMSE uses only six of fifteen trials; order accuracy counts missing comparisons as failures across all trials. Thus availability and ranking gains do not guarantee recovery of score spacing.

6 Conclusion

We introduced IR-CKA for recovering clean representation similarity from repeated noisy observations. Under the stated noise assumptions, placing replica independence within all three minibatch HSIC_1 factors makes their expectations equal to the corresponding clean finite-pool quantities. Our comparisons with shared-replica, feature-averaging, and cross-Gram estimators clarify why the placement of independence matters. We further establish concentration under random minibatch partitioning and risk guarantees for PSD stabilization.

Experiments with frozen ViT and ResNet representations show that IR and its variants improve clean-similarity recovery and bidirectional layer retrieval relative to the compared baselines under moderate feature dropout. Stronger noise can produce negative self factors and undefined or extreme raw IR scores. IR-PSD addresses these numerical failures: PSD projection with a positive ridge yields defined, bounded scores for finite input factors. At high noise levels, however, estimates can saturate near ± 1, distorting similarity magnitudes and layer relationships. IR-PSD therefore provides a practical stabilization of IR, while accurate recovery still depends on the noise level and the available observations.

REPRODUCIBILITY STATEMENT

Estimator definitions and sampling schemes are given in Sections 3–4.1. Appendices provide proofs, metric definitions, fixed-checkpoint protocols, noise settings, and complete layer-matching results. Reported variation is over the specified noise and partition draws, not retraining. The source package includes the sample-size summary data and a script reproducing its descriptive statistics and figures. These summaries contain three records per clean condition and nine per augmented condition, but no seed/trial identifiers; Section 5.1 therefore reports descriptive comparisons without trial-level confidence intervals or significance tests. The two $p = 0.7$ trial CSV files supporting the strong-dropout diagnostics are also included. Other matrix-recovery conditions and layer-retrieval runs are documented through their figures, protocols, and reported summaries; their full trial records are not included in this package.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, mathematical exposition, editing, and preparation of the experimental analysis. Responsibility for the correctness of the arguments, empirical claims, and final submission rests with the authors.

REFERENCES

J. Batson and L. Royer. Noise2Self: Blind denoising by self-supervision. *Proceedings of ICML*, PMLR 97:524–533, 2019. https://proceedings.mlr.press/v97/batson19a.html.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. https://doi.org/10.1111/ectj.12097.

C. Chun, A. Canatar, S. Chung, and D. D. Lee. Estimating neural representation alignment from sparsely sampled inputs and features. *arXiv preprint arXiv:2502.15104*, 2025. https://arxiv.org/abs/2502.15104.

T. Cui, Y. Kumar, P. Marttinen, and S. Kaski. Deconfounded representation similarity for comparison of neural networks. *Advances in Neural Information Processing Systems*, 35:19138–19151, 2022. https://doi.org/10.52202/068431-1391.

M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky. Reliability of CKA as a similarity measure in deep learning. *International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=8HRvyxc606.

J. Diedrichsen and N. Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017. https://doi.org/10.1371/journal.pcbi.1005508.

J. Diedrichsen, E. Berlot, M. Mur, H. H. Schütt, M. Shahbazi, and N. Kriegeskorte. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons, Behavior, Data Analysis, and Theory*, 5(3):1–31, 2021. https://doi.org/10.51628/001c.27664.

F. Ding, J.-S. Denain, and J. Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021. https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *Proceedings of ICML*, PMLR 97:3519–3529, 2019. https://proceedings.mlr.press/v97/kornblith19a.html.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. https://doi.org/10.3389/neuro.06.004.2008.

J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning image restoration without clean data. *Proceedings of ICML*, PMLR 80:2965–2974, 2018. https://proceedings.mlr.press/v80/lehtinen18a.html.

A. Morcos, M. Raghu, and S. Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018. https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

A. Murphy, J. Zylberberg, and A. Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. *ICLR Workshop on Representational Alignment (Re-Align)*, 2024. https://arxiv.org/abs/2405.01012.

T. Nguyen, M. Raghu, and S. Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=KJNcAkY8tY4.

H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10(4):e1003553, 2014. https://doi.org/10.1371/journal.pcbi.1003553.

D. A. Pospisil and W. Bair. Accounting for bias in the estimation of r^2 between two sets of noisy neural responses. *Journal of Neuroscience*, 42(50):9343–9355, 2022. https://doi.org/10.1523/JNEUROSCI.0198-22.2022.

M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012. https://jmlr.org/papers/v13/song12a.html.

G. J. Székely and M. L. Rizzo. Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42(6):2382–2412, 2014. https://doi.org/10.1214/14-AOS1255.

A. Walther, H. Nili, N. Ejaz, A. Alink, N. Kriegeskorte, and J. Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200, 2016. https://doi.org/10.1016/j.neuroimage.2015.12.012.

A. H. Williams, E. Kunz, S. Kornblith, and S. Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021. https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

A MINIBATCH HSIC FACTOR IDENTITIES

A.1 THE DIAGONAL-FREE HSIC ESTIMATOR

Lemma A1. *For symmetric $b \times b$ matrices P, Q, $b \geq 4$, define*

$$(\mathcal{U}_b P)_{ij} = P_{ij} - \frac{r_i^P + r_j^P}{b-2} + \frac{s_P}{(b-1)(b-2)} \quad (i \neq j), \qquad (\mathcal{U}_b P)_{ii} = 0,$$

where $r_i^P = \sum_{j \neq i} P_{ij}$ and $s_P = \sum_i r_i^P$. Then

$$\mathrm{HSIC}_1(P, Q) = \frac{\langle \mathcal{U}_b P, \mathcal{U}_b Q \rangle_F}{b(b-3)}, \qquad \mathrm{HSIC}_1(P, P) \geq 0. \tag{18}$$

If $\mathcal{D}$ is the collection of four-element subsets of the batch, then

$$\begin{aligned} \mathrm{HSIC}_1(P, Q) &= \binom{b}{4}^{-1} \sum_{D \in \mathcal{D}} h_D(P, Q), \\ h_D &= \frac{1}{24} \sum_{(i,j,k,l) \text{ permutes } D} (P_{ij} Q_{ij} - 2 P_{ij} Q_{ik} + P_{ij} Q_{kl}). \end{aligned} \tag{19}$$

Proof. Expanding the standard U-centering operator (Székely and Rizzo, 2014, Definition 2) gives (18); nonnegativity follows by setting $P = Q$. Equation (19) is the order-four HSIC U-statistic representation of Song et al. (2012, Theorem 3), grouped by unordered four-element subsets. These algebraic identities hold for arbitrary symmetric P, Q and do not require positive semidefiniteness. □

SR self-factors are squared norms in (18); IR self-factors are inner products between replicas. The denominator margin below is needed in both cases. For every four-element subset D, define the symmetrized IR numerator kernel by

$$h_{N,D}^{\mathrm{IR}} = \tfrac{1}{2} \left[h_D(K_A^{(1)}, K_B^{(2)}) + h_D(K_A^{(2)}, K_B^{(1)}) \right]. \tag{20}$$

Together with the denominator kernels from (19), this gives $Z_q = \binom{b}{4}^{-1} \sum_{|D|=4} h_{q,D}$ for every factor in (3). The two summands in (20) need not be independent; all variance bounds below treat their average as one kernel.

A.2 PROOF OF PROPOSITION 1: FACTOR EXPECTATIONS

Proof. All expectations in Steps 1–5 are conditional on the clean pool and a fixed batch $\mathcal{B}$ of size $b \geq 4$. We suppress the batch restrictions and the superscript $(\mathcal{B})$. The numerator calculation uses $C_i = \mathbb{E} e_i^{(s)} f_i^{(s)\top}$; each self-factor follows by replacing (Y, F, C_i) with $(X, E, \Sigma_{A,i})$, or by the corresponding side-B substitution. Set $\Xi_N = \Xi_{AB}$ and use the same convention for Ψ, $\mathcal{L}$, $\mathcal{Q}$, and Ω.

Step 1: SR and its covariance corrections. Put $K = K_A^{(1)}$ and $L = K_B^{(1)}$. Centering and row independence give $\mathbb{E}K = A + D_A$ and $\mathbb{E}L = B + D_B$, where D_A, D_B are diagonal. For distinct i, j, k, expansion of the off-diagonal entries gives

$$C_{ij}^{(1)} := \mathrm{Cov}(K_{ij}, L_{ij}) = x_j^\top C_i y_j + x_i^\top C_j y_i + \langle C_i, C_j \rangle_F, \tag{21}$$

$$C_{ijk}^{(2)} := \mathrm{Cov}(K_{ij}, L_{ik}) = x_j^\top C_i y_k. \tag{22}$$

Entries on disjoint pairs of rows have zero covariance. Thus the expectation of the order-four kernel in Eq. (19) differs from its clean value only through the shared-edge and shared-vertex terms. A fixed ordered pair has $(b-2)(b-3)$ completions to an ordered quadruple, and a fixed ordered triple has $b-3$. Consequently,

$$\mathbb{E} h_1(K, L) - h_1(A, B) = \frac{\sum_{i \neq j} C_{ij}^{(1)}}{b(b-1)} - \frac{2 \sum_{i,j,k \text{ distinct}} C_{ijk}^{(2)}}{b(b-1)(b-2)} = \Psi_{AB} = \mathcal{L}_{AB} + \mathcal{Q}_{AB}. \tag{23}$$

This proves the SR means, including both self-factors by substitution. The calculation permits full marginal covariances; it does not require independence between feature coordinates.

Step 2: pooled SR and IR. Identical replica distributions and linearity give, for every $q \in \{N, A, B\}$,

$$\mathbb{E}Z_q^{\mathrm{pool}} = \tfrac{1}{2}\big(\mathbb{E}Z_q^{\mathrm{SR},1} + \mathbb{E}Z_q^{\mathrm{SR},2}\big) = \Xi_q + \Psi_q.$$

For IR, independence of the complete replicas, bilinearity, and diagonal invariance instead imply

$$\mathbb{E}h_1(K_A^{(s)}, K_B^{(t)}) = h_1(A + D_A, B + D_B) = h_1(A, B), \qquad s \neq t.$$

Averaging the two numerator pairings and making the self substitutions proves $\mathbb{E}Z_q^{\mathrm{IR}} = \Xi_q$. Neither averaging argument requires independence between the three factors.

Step 3: feature-mean factors. For $m \geq 1$, write $\bar{e}_i = m^{-1}\sum_{s=1}^m e_i^{(s)}$ and $\bar{f}_i = m^{-1}\sum_{s=1}^m f_i^{(s)}$. The averaged noise remains centered and row-wise independent, with

$$\mathbb{E}\bar{e}_i\bar{f}_i^\top = C_i/m, \qquad \mathbb{E}\bar{e}_i\bar{e}_i^\top = \Sigma_{A,i}/m, \qquad \mathbb{E}\bar{f}_i\bar{f}_i^\top = \Sigma_{B,i}/m.$$

Apply Step 1 with these covariances. The terms defining $\mathcal{L}_q$ are linear in the covariances, whereas those defining $\mathcal{Q}_q$ are quadratic. Hence

$$\mathbb{E}Z_q^{\mathrm{F}(m)} = \Xi_q + \frac{\mathcal{L}_q}{m} + \frac{\mathcal{Q}_q}{m^2}.$$

Step 4: cross-Grams and MCG1 with m replicas. Let $m \geq 2$ and put $\Delta_{A,m} = \check{A}_m - A$ and $\Delta_{B,m} = \check{B}_m - B$. The cross-Gram remainder decomposes as

$$\begin{aligned}\Delta_{A,m,ij} &= L_{A,ij} + V_{A,ij},\\ L_{A,ij} &= x_i^\top \bar{e}_j + \bar{e}_i^\top x_j, \qquad V_{A,ij} = \frac{1}{m(m-1)}\sum_{s\neq t} e_i^{(s)\top} e_j^{(t)},\end{aligned} \tag{24}$$

and similarly for side B. These identities include $i = j$. Every term has mean zero, so $\mathbb{E}\check{A}_m = A$ and $\mathbb{E}\check{B}_m = B$, including on the diagonal. A product between an L term and a V term contains a centered, independent replica appearing only once and therefore has zero mean. Writing $\delta_{ij} = \mathbf{1}\{i = j\}$, the linear terms satisfy

$$\mathbb{E}L_{A,ij}L_{B,kl} = \frac{1}{m}\big[\delta_{jl}x_i^\top C_j y_k + \delta_{jk}x_i^\top C_j y_l + \delta_{il}x_j^\top C_i y_k + \delta_{ik}x_j^\top C_i y_l\big].$$

For the quadratic terms, independence and centering leave only matching or reversed ordered replica pairs. Row independence then gives

$$\mathbb{E}V_{A,ij}V_{B,kl} = \frac{\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{m(m-1)}\langle C_i, C_j\rangle_F.$$

Thus the complete entry covariance, including diagonal entries, is

$$\begin{aligned}\mathbb{E}\Delta_{A,m,ij}\Delta_{B,m,kl} = \frac{1}{m}\big[&\delta_{jl}x_i^\top C_j y_k + \delta_{jk}x_i^\top C_j y_l\\ &+ \delta_{il}x_j^\top C_i y_k + \delta_{ik}x_j^\top C_i y_l\big]\\ &+ \frac{\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{m(m-1)}\langle C_i, C_j\rangle_F.\end{aligned} \tag{25}$$

For off-diagonal entries, this yields the contractions

$$\begin{aligned}\mathrm{Cov}(\check{A}_{m,ij}, \check{B}_{m,ij}) &= \frac{x_j^\top C_i y_j + x_i^\top C_j y_i}{m} + \frac{\langle C_i, C_j\rangle_F}{m(m-1)},\\ \mathrm{Cov}(\check{A}_{m,ij}, \check{B}_{m,ik}) &= \frac{x_j^\top C_i y_k}{m}, \qquad i, j, k \text{ distinct},\end{aligned}$$

while the disjoint-row covariance is zero. Substitution into the same kernel contraction as in Eq. (23) therefore gives

$$\mathbb{E}Z_q^{\mathrm{MCG1}(m)} = \Xi_q + \frac{\mathcal{L}_q}{m} + \frac{\mathcal{Q}_q}{m(m-1)}.$$

At $m = 2$ this reduces to $\Xi_q + \Psi_q/2$; for general m the linear and quadratic corrections have different coefficients.

Step 5: ordinary-HSIC cross-Grams. Set $H = H_b$, $x_i^c = (HX)_i$, and $y_i^c = (HY)_i$, with row features written as columns. Since the cross-Gram remainders have mean zero, bilinearity gives

$$\mathbb{E}Z_q^{\mathrm{C0}(m)} = \Xi_{q,0} + \Omega_{q,0}^{(m)}, \qquad \Omega_{AB,0}^{(m)} = \mathbb{E}h_0(\Delta_{A,m}, \Delta_{B,m}). \tag{26}$$

Ordinary HSIC retains diagonal entries, so this correction requires the full covariance in Eq. (25). Expanding its trace gives

$$(b-1)^2\Omega_{AB,0}^{(m)} = \sum_{i,j,k,l} H_{jk}H_{li}\mathbb{E}\Delta_{A,m,ij}\Delta_{B,m,kl}. \tag{27}$$

The first and fourth linear contractions each give $\sum_i (x_i^c)^\top C_i y_i^c/m$; the other two each give $\sum_i H_{ii}\langle C_i, X^\top HY\rangle_F/m$. The two quadratic contractions give the terms containing H_{ij}^2 and $H_{ii}H_{jj}$, respectively. Consequently,

$$\begin{aligned}\Omega_{AB,0}^{(m)} = \frac{1}{(b-1)^2}\Bigg[\frac{2}{m}\sum_i \left\{(x_i^c)^\top C_i y_i^c + H_{ii}\langle C_i, X^\top HY\rangle_F\right\} \\ + \frac{1}{m(m-1)}\sum_{i,j}(H_{ij}^2 + H_{ii}H_{jj})\langle C_i, C_j\rangle_F\Bigg].\end{aligned} \tag{28}$$

The sums include $i = j$. The self substitutions give $\Omega_{A,0}^{(m)}$ and $\Omega_{B,0}^{(m)}$, and in particular

$$\Omega_{A,0}^{(m)} = \frac{\mathbb{E}\|H\Delta_{A,m}H\|_F^2}{(b-1)^2} \geq 0, \qquad \Omega_{B,0}^{(m)} = \frac{\mathbb{E}\|H\Delta_{B,m}H\|_F^2}{(b-1)^2} \geq 0.$$

These calculations use only finite second moments, centering, and the stated row and replica independence. No Gaussian assumption is needed.

Step 6: one uniform-subset identity for all HSIC$_1$ methods. Condition now on all realized noisy replicas over the entire pool. For any fixed symmetric $n \times n$ matrices P, Q and a uniform b-subset $\mathcal{B}$ independent of these observations, Eq. (19) gives

$$\begin{aligned}\mathbb{E}_{\mathcal{B}}h_{1,b}(P_{\mathcal{B}}, Q_{\mathcal{B}}) &= \binom{b}{4}^{-1}\sum_{\substack{D\subset[n]\\|D|=4}}\mathbb{P}(D\subset\mathcal{B})\, h_D(P,Q) \\ &= \binom{n}{4}^{-1}\sum_{\substack{D\subset[n]\\|D|=4}} h_D(P,Q) = h_{1,n}(P,Q),\end{aligned} \tag{29}$$

because $\mathbb{P}(D \subset \mathcal{B}) = \binom{b}{4}/\binom{n}{4}$. The subscripts on h_1 only specify its matrix size. Every Gram used by SR, pooled SR, feature mean, MCG1, and IR is the principal restriction of the corresponding full-pool matrix: feature averaging and cross-replica multiplication commute with restriction to the selected rows. Each factor is therefore an instance of Eq. (29), or a fixed linear combination of such instances. Taking the remaining expectation over observation noise proves the common replacement of $(\mathcal{B})$ by $[n]$ in all their factor means. Every block of a uniform random partition has the same uniform-subset marginal, so linearity also proves the pooled-factor claim without assuming independence between blocks. This identity concerns unnormalized factors, not their ratio. It does not apply to MCG's ordinary h_0: averaging its batchwise means averages both $\Xi_{q,0}^{(\mathcal{B})}$ and $\Omega_{q,0}^{(m;\mathcal{B})}$, which generally depend on b and cannot be replaced by their full-pool counterparts. □

B MINIBATCH CONCENTRATION FOR TWO SAMPLING SCHEMES

Lemma A2 (Uniform kernel moments)**.** *Suppose the joint observation rows (including both replicas when used) are independent across inputs, the clean rows are uniformly bounded, and the dimensions are fixed. Assume the uniform fourth-moment bound*

$$\sup_{n,i,s}\max\{\mathbb{E}\|e_i^{(s)}\|^4, \mathbb{E}\|f_i^{(s)}\|^4\} \leq M_4 < \infty.$$

Then there is $M < \infty$, independent of n, b, β, such that

$$\sup_{q,D} \mathbb{E}h_{q,D}^2 \le M, \qquad J := \sup_{q,D} |\mathbb{E}h_{q,D}| \le \sqrt{M}.$$

Here $h_{q,D}$ is the factor kernel in (19) *or* (20)*, evaluated on four distinct pool identities.*

Proof. Each kernel term is a product of two inner products. For four distinct indices, its total degree in the joint noise vector of any single sample is at most two: even $P_{ij}Q_{ij}$ uses only two factors from row i and two from row j. After squaring, every row has degree at most four. Expand coordinates and deterministic shifts. Within a row, generalized Hölder bounds each resulting mixed monomial by the marginal fourth moments and the bounded signal coefficients. Across distinct rows the expectation factors by row independence. There are finitely many monomials, uniformly in n because dimensions and the 24 permutations are fixed. Applying $(\sum_{j=1}^k u_j)^2 \le k \sum_j u_j^2$ proves a common bound M for every individual pairing. For the symmetrized numerator,

$$\mathbb{E}(h_{N,D}^{\mathrm{IR}})^2 \le \tfrac{1}{2}\mathbb{E}h_D(K_A^{(1)}, K_B^{(2)})^2 + \tfrac{1}{2}\mathbb{E}h_D(K_A^{(2)}, K_B^{(1)})^2 \le M.$$

Jensen also gives $J \le \sqrt{M}$. Only independence across input rows is used in this moment bound; dependence within a row is allowed. □

Lemma A3 (Noise variance at fixed batch)**.** *For every fixed batch $\mathcal{B}$ of size $b \ge 4$ and every factor q,*

$$\mathrm{Var}(Z_q \mid \mathcal{B}) \le 16M/b. \tag{30}$$

Proof. Write $Z_q = \binom{b}{4}^{-1} \sum_{D \subset \mathcal{B}, |D|=4} h_{q,D}$. A row includes both networks and, for IR, both replicas. Conditional on the identities, disjoint D, D' use independent rows and have zero covariance. All other covariances have absolute value at most M by Cauchy–Schwarz. For fixed D and uniform D',

$$\mathbb{P}(D \cap D' \ne \varnothing) \le \sum_{i \in D} \mathbb{P}(i \in D') = 4 \cdot \frac{4}{b} = \frac{16}{b}.$$

Thus

$$\mathrm{Var}(Z_q \mid \mathcal{B}) = \binom{b}{4}^{-2} \sum_{D,D'} \mathrm{Cov}(h_{q,D}, h_{q,D'} \mid \mathcal{B}) \le M\frac{16}{b}.$$

□

Lemma A4 (Permutation variance bound)**.** *Let π be uniform on permutations of $[n]$. Suppose exchanging any two positions changes $F(\pi)$ by at most L. If F depends only on the first k positions, its variance is at most kL^2; in general it is at most nL^2. The same assertion holds if the function only depends on the unordered set of the first k positions.*

Proof. Reveal the permutation sequentially and define $M_j = \mathbb{E}[F \mid \pi_1, \dots, \pi_j]$, $M_0 = \mathbb{E}F$. Conditional on a prefix, consider two possible next labels a, b. Exchanging labels a, b in the remaining permutation is a bijection between the two sets of completions, preserves uniform measure, and changes F by at most L. Therefore the two conditional values of M_j differ by at most L. Since M_{j-1} is their weighted average over all next labels, $|M_j - M_{j-1}| \le L$. For $i < j$, the martingale property gives

$$\mathbb{E}[(M_i - M_{i-1})(M_j - M_{j-1})] = \mathbb{E}[(M_i - M_{i-1})\mathbb{E}(M_j - M_{j-1} \mid \pi_1, \dots, \pi_{j-1})] = 0.$$

If k positions determine F, then $M_k = F$, and hence $\mathrm{Var}\, F = \sum_{j=1}^k \mathbb{E}(M_j - M_{j-1})^2 \le kL^2$. Take $k = n$ for the general assertion. □

Theorem A5 (Pooled variance under (I) and (P))**.** *Suppose the joint observation rows (including both replicas when used) are independent across inputs and the noise is independent of the sampling. Under (I), use independent uniform b-subsets and fresh independent noise for every occurrence; under (P), use a uniform partition of $n = \beta b$ identities. Assume*

$$\sup_{q,D} \mathbb{E}h_{q,D}^2 \le M < \infty, \qquad J := \sup_{q,D} |\mathbb{E}h_{q,D}| \le \sqrt{M}.$$

Then every factor satisfies

$$\text{(I)} \qquad \operatorname{Var} \bar{Z}_q \le 80M/(\beta b), \tag{31}$$

$$\text{(P)} \qquad \operatorname{Var} \bar{Z}_q \le 272M/n = 272M/(\beta b). \tag{32}$$

All covariances between the three pooled factors are bounded in absolute value by the corresponding common upper bound.

Proof. Write $m_q(\mathcal{B}) = \mathbb{E}[Z_q \mid \mathcal{B}] = \binom{b}{4}^{-1} \sum_{D \subset \mathcal{B}, |D|=4} \mathbb{E} h_{q,D}$. Replacing one member of $\mathcal{B}$ affects $\binom{b-1}{3}$ terms, each by at most $2J$. Consequently

$$|m_q(\mathcal{B}') - m_q(\mathcal{B})| \le 2J \frac{\binom{b-1}{3}}{\binom{b}{4}} = 8J/b. \tag{33}$$

A uniform subset is the set of the first b positions of a uniform permutation. A transposition either does not change that set or replaces one member. Lemma A4 therefore gives $\operatorname{Var}_{\mathcal{B}} m_q(\mathcal{B}) \le b(8J/b)^2 = 64J^2/b$. The law of total variance and (30) yield

$$\operatorname{Var} Z_q = \mathbb{E} \operatorname{Var}(Z_q \mid \mathcal{B}) + \operatorname{Var} m_q(\mathcal{B}) \le (16M + 64J^2)/b \le 80M/b.$$

Under (I), both selection and noise are independent between occurrences, so

$$\operatorname{Var} \bar{Z}_q = \beta^{-2} \sum_r \operatorname{Var} Z_{q,r} \le 80M/(\beta b).$$

For (P), condition on the entire partition $\Pi = (\mathcal{B}_1, \dots, \mathcal{B}_\beta)$. Disjoint batches use independent sample noise vectors, so

$$\mathbb{E} \operatorname{Var}(\bar{Z}_q \mid \Pi) = \mathbb{E}\left[\frac{1}{\beta^2} \sum_r \operatorname{Var}(Z_{q,r} \mid \Pi)\right] \le \frac{16M}{\beta b} = \frac{16M}{n}.$$

Let $F_q(\Pi) = \mathbb{E}[\bar{Z}_q \mid \Pi] = \beta^{-1} \sum_r m_q(\mathcal{B}_r)$. Exchanging two permutation positions either stays within a block or changes one member in each of two blocks. By (33),

$$|F_q(\Pi') - F_q(\Pi)| \le \frac{1}{\beta} 2 \frac{8J}{b} = \frac{16J}{n}.$$

Lemma A4 gives $\operatorname{Var} F_q \le n(16J/n)^2 = 256J^2/n$. A second application of total variance proves (32). Finally $|\operatorname{Cov}(\bar{Z}_q, \bar{Z}_{q'})| \le \sqrt{\operatorname{Var} \bar{Z}_q \operatorname{Var} \bar{Z}_{q'}}$ proves the covariance bound. □

B.1 Normalization stability

Lemma A6. *Let U, V, W have finite second moments, means u, v, w, with $v, w > 0$. Define*

$$\Delta = \sqrt{vw}, \quad r = u/\Delta, \quad |r| \le c_0, \quad \delta^2 = \mathbb{E}a^2 + \mathbb{E}s^2 + \mathbb{E}t^2,$$
$$a = (U-u)/\Delta, \qquad s = (V-v)/v, \qquad t = (W-w)/w.$$

Use fixed proportional thresholds $v/2, w/2$ and write

$$R_\lambda = \frac{U}{\sqrt{(V \vee (v/2))(W \vee (w/2))}}.$$

Then

$$\mathbb{E}|R_\lambda - r|^2 \le C(1 + c_0^2)\delta^2, \qquad |\mathbb{E}R_\lambda - r| \le C(1 + c_0)\delta^2. \tag{34}$$

On $\mathcal{G} = \{V \ge v/2, W \ge w/2\}$ the raw and truncated ratios agree, and $\mathbb{P}(\mathcal{G}^c) \le 4\delta^2$. Any measurable completion of the raw ratio outside $\mathcal{G}$ inherits the probability bound

$$\mathbb{P}(|R - r| > \varepsilon) \le 4\delta^2 + C(1 + c_0^2)\delta^2/\varepsilon^2. \tag{35}$$

If the completed raw ratio is bounded by a fixed constant, it also satisfies $\mathbb{E}R = r + O((1 + c_0)\delta^2)$.

Proof. Put $s_* = s \vee (-1/2)$ and $t_* = t \vee (-1/2)$. Then $|s_*| \le |s|$, $|t_*| \le |t|$, and for $f(s,t) = ((1+s)(1+t))^{-1/2}$ on $s, t \ge -1/2$,

$$|f| \le 2, \quad |\partial_s f|, |\partial_t f| \le 2, \quad |\partial_{ss} f|, |\partial_{tt} f| \le 6, \quad |\partial_{st} f| \le 2.$$

These derivative bounds follow by differentiating each power. Thus

$$|R_\lambda - r| = |(r+a) f(s_*, t_*) - r| \le 2|a| + 2c_0(|s| + |t|).$$

Squaring and using $(x+y+z)^2 \le 3(x^2+y^2+z^2)$ proves the first bound. For the bias, centering and $\{|s| > 1/2\}$ imply

$$0 \le \mathbb{E}s_* = \mathbb{E}(s_* - s) \le \mathbb{E}[|s| \mathbf{1}_{s<-1/2}] \le 2\mathbb{E}s^2, \qquad 0 \le \mathbb{E}t_* \le 2\mathbb{E}t^2.$$

Taylor's formula with the bounded Hessian gives

$$f(s_*, t_*) = 1 - \tfrac{1}{2}(s_* + t_*) + R_2, \qquad |R_2| \le C(s_*^2 + t_*^2).$$

Hence $|\mathbb{E}(f-1)| \le C\delta^2$. Since $\mathbb{E}a = 0$,

$$|\mathbb{E}[af]| = |\mathbb{E}[a(f-1)]| \le 2\sqrt{\mathbb{E}a^2}\sqrt{\mathbb{E}(|s|+|t|)^2} \le C\delta^2.$$

Combining $\mathbb{E}(R_\lambda - r) = r\mathbb{E}(f-1) + \mathbb{E}[af]$ proves the bias bound. Chebyshev yields $\mathbb{P}(\mathcal{G}^c) \le 4(\mathbb{E}s^2 + \mathbb{E}t^2)$. Equation (35) follows by splitting the event and applying Markov to the squared error of R_λ. For a bounded completion,

$$\mathbb{E}[|R_\lambda| \mathbf{1}_{\mathcal{G}^c}] \le 2|r|\mathbb{P}(\mathcal{G}^c) + 2\sqrt{\mathbb{E}a^2 \mathbb{P}(\mathcal{G}^c)} \le C(1+c_0)\delta^2;$$

the completed ratio contributes only its bound times $\mathbb{P}(\mathcal{G}^c)$. This proves the last assertion. □

B.2 Proof of Theorem 2

Proof. Proposition 1 proves the stated means and centers. Both centers have absolute value at most one. For IR this follows from (18) applied to the clean full-pool Grams. For SR the pointwise inequality $|Z_N| \le \sqrt{Z_A Z_B}$ follows from the same identity; after expectation, $|m_N| \le \mathbb{E}\sqrt{Z_A Z_B} \le \sqrt{m_A m_B}$. Theorem A5 proves (12) and, with (11), gives

$$\delta^2 = \frac{\operatorname{Var} \bar{Z}_N}{m_A m_B} + \frac{\operatorname{Var} \bar{Z}_A}{m_A^2} + \frac{\operatorname{Var} \bar{Z}_B}{m_B^2} \le \frac{3C_1 M}{n_{\text{eff}} m_*^2}, \tag{36}$$

where $C_1 = 80$ or 272. Apply Lemma A6 with $c_0 = 1$. Its squared-error bound gives the first assertion of (13), Markov's inequality gives the tail assertion, and its bias bound gives the middle assertion. Finally

$$\mathbb{P}(\mathcal{G}^c) \le 4\frac{\operatorname{Var} \bar{Z}_A}{m_A^2} + 4\frac{\operatorname{Var} \bar{Z}_B}{m_B^2} \le \frac{8C_1 M}{n_{\text{eff}} m_*^2},$$

and the raw/truncated ratios coincide on $\mathcal{G}$. Splitting the deviation event over $\mathcal{G}$ and its complement proves (14). □

Corollary A7 (Truncation at a known denominator margin)**.** *Under the assumptions of Theorem 2, suppose the lower bound m_* is known. Set $\lambda_* = m_*/2$, $N = \beta b$, and*

$$\widehat{R}_* = \frac{\bar{Z}_N}{\sqrt{(\bar{Z}_A \vee \lambda_*)(\bar{Z}_B \vee \lambda_*)}}. \tag{37}$$

Then for both SR and IR, under either sampling scheme,

$$\mathbb{E}|\widehat{R}_* - r|^2 \le \frac{C}{N}, \qquad |\mathbb{E}\widehat{R}_* - r| \le \frac{C}{N}, \qquad \mathbb{P}(|\widehat{R}_* - r| > \varepsilon) \le \frac{C}{N\varepsilon^2},$$

where C depends only on M and m_*. This estimator does not require knowledge of the individual means m_A, m_B.

Proof. Write $U = \bar{Z}_N$, $V = \bar{Z}_A$, $W = \bar{Z}_B$, with means u, v, w, and put $V_* = V \vee \lambda_*$, $W_* = W \vee \lambda_*$. Since $v, w \geq m_* = 2\lambda_*$,

$$|V_* - v| \leq |V - v|, \qquad 0 \leq \mathbb{E}(V_* - v) = \mathbb{E}[(\lambda_* - V)\mathbf{1}_{V<\lambda_*}] \leq \frac{\operatorname{Var} V}{v - \lambda_*} \leq \frac{2 \operatorname{Var} V}{m_*},$$

and the same bounds hold for W_*. For $g(x, y) = (xy)^{-1/2}$ on $[\lambda_*, \infty)^2$, the function and its first two derivatives are bounded by constants depending only on m_*. Taylor expansion at (v, w) therefore gives, with $S = \operatorname{Var} V + \operatorname{Var} W$,

$$|\mathbb{E}[g(V_*, W_*) - g(v, w)]| \leq CS, \qquad \mathbb{E}|g(V_*, W_*) - g(v, w)|^2 \leq CS.$$

Every factor mean is an average of kernel means, so $|u| \leq \sqrt{M}$. Using $\mathbb{E}(U - u) = 0$ and Cauchy–Schwarz,

$$|\mathbb{E}[Ug(V_*, W_*)] - ug(v, w)| \leq C|u|S + C\sqrt{\operatorname{Var} U\, S},$$
$$\mathbb{E}|Ug(V_*, W_*) - ug(v, w)|^2 \leq C \operatorname{Var} U + Cu^2 S.$$

Theorem A5 bounds each factor variance by $C_1 M/N$. Substitution proves the first two assertions, and Markov's inequality proves the tail bound. □

Weak signals and thresholds. Equation (36) makes the dependence on denominator strength explicit. Without a fixed margin, consistency still follows if the three normalized variances in (36) tend to zero; absolute variances tending to zero alone are insufficient. Corollary A7 makes the truncation implementable when a positive lower bound is known. A data-dependent threshold requires control of both its location and its exceptional events; convergence below the target alone does not establish the $O(N^{-1})$ bias bound.

Repeated partitions of one noisy pool. To separate repeated partitioning from fresh measurements, let $\mathcal{F}$ denote the observed noisy pool, and draw T independent uniform partitions conditional on $\mathcal{F}$. Their pooled factors $Z^{(t)}$ are conditionally independent. Every four-element subset has probability $\beta\binom{b}{4}/\binom{n}{4}$ to lie within one block. Therefore

$$U_{\mathcal{F}} := \mathbb{E}[Z^{(t)} \mid \mathcal{F}] = \binom{n}{4}^{-1} \sum_{|D|=4} h_D(\mathcal{F}),$$

which is the corresponding full-pool factor, using (20) for the IR numerator. Total variance gives

$$\operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^{T} Z^{(t)}\right) = \frac{1}{T}\mathbb{E} \operatorname{Var}(Z^{(1)} \mid \mathcal{F}) + \operatorname{Var}(U_{\mathcal{F}}). \tag{38}$$

The first term decreases with T; the second, generally nonzero, does not. This identity concerns unnormalized factors. A ratio of pooled factors requires the additional ratio-stability step, as above.

B.3 Exponential concentration for bounded perturbations

Corollary A8 (Bounded observations)**.** *Use (I) or (P), with independent sample noise vectors and the specified replica law. Suppose every observed row, in every required replica, satisfies $\|\widetilde{x}_i^{(s)}\| \leq R_A$ and $\|\widetilde{y}_i^{(s)}\| \leq R_B$ almost surely, uniformly over the pool. Set*

$$L = 4 \max\{R_A^4, R_B^4, R_A^2 R_B^2\} > 0, \qquad N = \beta b.$$

Bounded clean rows and almost surely bounded noise norms imply this condition by the triangle inequality. For SR and IR, and all $u > 0$,

$$\mathbb{P}\left(\max_{q \in \{N, A, B\}} |\bar{Z}_q - m_q| > u\right) \leq 6 \exp\left(-\frac{Nu^2}{640L^2}\right). \tag{39}$$

For IR, assume centered noise independent across inputs and independent replicas, and let $m_A, m_B \geq m_* > 0$. Define

$$R_0 = \frac{m_N}{\sqrt{m_A m_B}}, \qquad t_\delta = \frac{L}{m_*}\sqrt{\frac{640 \log(6/\delta)}{N}}, \quad 0 < \delta < 1.$$

If $t_\delta < 1$, then with probability at least $1-\delta$ the untruncated ratio is well defined and

$$|\widehat{R}_{\mathrm{IR}} - R_0| \le \frac{2t_\delta}{1-t_\delta}. \tag{40}$$

In addition,

$$\mathbb{P}(\bar{Z}_A \le 0 \text{ or } \bar{Z}_B \le 0) \le 4\exp\left(-\frac{Nm_*^2}{640L^2}\right). \tag{41}$$

These assertions do not require independence of feature coordinates.

Proof. Cauchy–Schwarz bounds a Gram entry by R_A^2 or R_B^2. The absolute coefficients of each summand in (19) sum to four; hence $|h_{q,D}| \le L$, also for the average in (20). The remaining concentration argument uses only this kernel bound and the sampling assumptions, and thus also proves the bounded-kernel assertion of Theorem 2. Replacing a complete observation row, including both replicas when required, affects only four-tuples containing that row. Consequently the change of one batch factor is at most

$$2L\frac{\binom{b-1}{3}}{\binom{b}{4}} = \frac{8L}{b},$$

and the change of a pooled factor is at most $8L/N$.

We use the following elementary martingale bound. If a centered martingale increment D_j obeys $|D_j| \le c_j$, convexity of the exponential gives

$$\mathbb{E}[e^{\lambda D_j} \mid \mathcal{F}_{j-1}] \le \cosh(\lambda c_j) \le e^{\lambda^2 c_j^2/2}.$$

Iteration therefore bounds the moment generating function of the sum by $\exp(\lambda^2 \sum_j c_j^2/2)$.

Write $\mathcal{S}$ for all batch identities and $F_q = \mathbb{E}[\bar{Z}_q \mid \mathcal{S}]$. Conditional on $\mathcal{S}$, reveal one joint observation row at a time. There are N independent rows under either scheme (occurrences in (I), distinct pool rows in (P)). Coupling the unrevealed rows identically bounds each Doob increment by $8L/N$, so

$$\mathbb{E}[e^{\lambda(\bar{Z}_q - F_q)} \mid \mathcal{S}] \le \exp\left(\frac{32\lambda^2 L^2}{N}\right). \tag{42}$$

Under (I), reveal the b sampled identities sequentially in each of the β independent subsets. The permutation-completion coupling in Lemma A4 changes at most one selected identity and thus changes F_q by at most $8L/N$. There are N increments, giving

$$\mathbb{E}e^{\lambda(F_q - m_q)} \le e^{32\lambda^2 L^2/N}.$$

Under (P), a transposition changes at most two blocks and therefore changes F_q by at most $16L/N$. Revealing the $n = N$ permutation positions gives

$$\mathbb{E}e^{\lambda(F_q - m_q)} \le e^{128\lambda^2 L^2/N}.$$

Combining either bound with (42) by conditional expectation yields the common bound

$$\mathbb{E}e^{\lambda(\bar{Z}_q - m_q)} \le e^{160\lambda^2 L^2/N}.$$

Markov's inequality, minimized at $\lambda = Nu/(320L^2)$, and the same argument for the lower tail imply

$$\mathbb{P}(|\bar{Z}_q - m_q| \ge u) \le 2e^{-Nu^2/(640L^2)}.$$

A union bound proves (39) and, applied to deviations of size m_*, proves (41).

For IR, diagonal invariance and replica independence give the clean factor means as in Proposition 1; this step uses no independence between coordinates. Set

$$u_A = \frac{\bar{Z}_A - m_A}{m_A}, \quad u_B = \frac{\bar{Z}_B - m_B}{m_B}, \quad w = \frac{\bar{Z}_N - m_N}{\sqrt{m_A m_B}}.$$

On the event in (39) with $u = L\sqrt{640\log(6/\delta)/N}$, all three absolute values are at most t_δ. Thus $d = \sqrt{(1+u_A)(1+u_B)}$ lies in $[1-t_\delta, 1+t_\delta]$, and both self-factors are positive. Finally,

$$\left|\frac{R_0 + w}{d} - R_0\right| \le \frac{|w| + |R_0||1-d|}{d} \le \frac{(1+|R_0|)t_\delta}{1-t_\delta} \le \frac{2t_\delta}{1-t_\delta}.$$

Here $|R_0| \le 1$ follows from Cauchy–Schwarz for the clean U-centered Grams. This proves (40). □

For example, $0 < \varepsilon \le 1$ and

$$N \ge \frac{640L^2}{m_*^2}\frac{(2+\varepsilon)^2}{\varepsilon^2}\log\frac{6}{\delta}$$

suffice for error at most ε with probability at least $1-\delta$. Under (P), $N = n$ independently of the number of blocks. These common upper bounds do not order the actual SR and IR variances.

B.4 Proof and implementation scope of the PSD correction

We first give the two deterministic facts used in Theorem 3. They do not assume an invertible target factor matrix; positive target diagonal entries suffice.

Lemma A9 (Projection and normalization)**.** *Let* $\Gamma_0 = \left(\begin{smallmatrix} v & u \\ u & w \end{smallmatrix}\right) \succeq 0$ *with* $v, w \ge m_* > 0$*, and let* H *be positive semidefinite with strictly positive diagonal entries. Define* $r = u/\sqrt{vw}$ *and* $r(H) = H_{12}/\sqrt{H_{11}H_{22}}$*. For any symmetric* 2×2 *matrix* D*,*

$$\|\Pi_+(D) - \Gamma_0\|_F \le \|D - \Gamma_0\|_F, \tag{43}$$

$$|r(H) - r| \le \frac{4}{m_*}\|H - \Gamma_0\|_F. \tag{44}$$

Proof. Write $P = \Pi_+(D)$, obtained by setting the negative eigenvalues of D to zero. Then $D - P \preceq 0$ and $(D-P)P = 0$. Since $\Gamma_0 \succeq 0$,

$$\langle D - P, \Gamma_0 - P\rangle_F = \operatorname{tr}((D-P)\Gamma_0) \le 0.$$

Expanding $D - \Gamma_0 = (D-P) + (P-\Gamma_0)$ gives $\|D-\Gamma_0\|_F^2 \ge \|P-\Gamma_0\|_F^2$, proving (43).

Put $e = \|H - \Gamma_0\|_F$. Positive semidefiniteness gives $|r(H)|, |r| \le 1$. If $e \ge m_*/2$, then $|r(H) - r| \le 2 \le 4e/m_*$. If $e < m_*/2$, set $t = e/m_* < 1/2$ and

$$s = \frac{H_{11} - v}{v}, \quad z = \frac{H_{22} - w}{w}, \quad a = \frac{H_{12} - u}{\sqrt{vw}}.$$

Each absolute value is at most t. Thus $d = \sqrt{(1+s)(1+z)} \in [1-t, 1+t]$ and

$$|r(H) - r| = \left|\frac{r+a}{d} - r\right| \le \frac{|a| + |r||1-d|}{d} \le \frac{2t}{1-t} \le 4t.$$

This proves (44) on both events. □

Proof of Theorem 3. The U-centered clean full-pool matrices have inner product m_N and squared norms m_A, m_B, with the common factor $1/[n(n-3)]$. Consequently $\Gamma_0 \succeq 0$, including when $|R_0| = 1$. Proposition 1 gives $\mathbb{E}\widehat{\Gamma} = \Gamma_0$. Under its kernel second-moment assumption, Theorem 2 gives $\operatorname{Var}\bar{Z}_q \le C_1 M/N$. Lemma A2 supplies a sufficient fourth-moment condition for this kernel bound. It follows that

$$\begin{aligned}\mathbb{E}\|\widehat{\Gamma} - \Gamma_0\|_F^2 &= \operatorname{Var}\bar{Z}_A + 2\operatorname{Var}\bar{Z}_N + \operatorname{Var}\bar{Z}_B \\ &\le \frac{4C_1 M}{N}.\end{aligned} \tag{45}$$

No covariance terms enter this squared Frobenius norm, so independence between the three factors is unnecessary. Markov's inequality proves

$$\mathbb{P}\left(\|\widehat{\Gamma} - \Gamma_0\|_F > 2\sqrt{\frac{C_1 M}{N\alpha}}\right) \le \alpha. \tag{46}$$

This is the required matrix concentration statement, derived from the observation model rather than postulated as a separate hypothesis.

For every realization, $\widehat{\Gamma}_{+,\lambda_N} \succeq \lambda_N I_2$. Its diagonal entries are positive, its determinant is nonnegative, and therefore $|\widehat{R}^+_{\mathrm{IR}}(\lambda_N)| \le 1$. By Lemma A9 and the triangle inequality,

$$\begin{aligned}\left|\widehat{R}^+_{\mathrm{IR}}(\lambda_N) - R_0\right| &\le \frac{4}{m_*}\|\widehat{\Gamma}_{+,\lambda_N} - \Gamma_0\|_F \\ &\le \frac{4}{m_*}\left(\|\widehat{\Gamma} - \Gamma_0\|_F + \sqrt{2}\lambda_N\right).\end{aligned} \tag{47}$$

The L^2 triangle inequality and (45) give

$$\left(\mathbb{E}\left|\widehat{R}^{+}_{\mathrm{IR}}(\lambda_N) - R_0\right|^2\right)^{1/2} \le \frac{4}{m_*}\left(2\sqrt{\frac{C_1 M}{N}} + \sqrt{2}\lambda_N\right) = a_N.$$

The absolute bias is at most the L^1 error, which is at most this L^2 error. This proves (15). Combining (46) and (47) proves (16).

Under bounded kernels, Theorem 2 gives

$$\max_q |\bar{Z}_q - m_q| \le L\sqrt{\frac{640\log(6/\alpha)}{N}}$$

with probability at least $1-\alpha$. On this event the Frobenius error in (47) is at most twice the displayed bound, which proves (17).

Finally $a_N \to 0$ under the stated uniform margin and moment bounds with $\lambda_N \to 0$. The mean-square bound proves consistency, and the bias bound proves $\mathbb{E}\widehat{R}^{+}_{\mathrm{IR}}(\lambda_N) - R_0 \to 0$. No limit of the pool-dependent R_0 itself is required. Taking $\lambda_N = O(N^{-1/2})$ gives the stated rates. □

C CLEAN-MATRIX RECOVERY: PROTOCOL AND DIAGNOSTICS

This section provides the implementation details of Section 5.2. The main experiment uses all 10,000 CIFAR-10 test images and frozen ImageNet-pretrained ViT-B/16 and ResNet-50 from torchvision, with both models evaluated in inference mode and $b = 1{,}000$. Section C.2 separately reports a ViT-L/16–ViT-B/16 strong-dropout illustration.

We use a maximum of fifteen readouts per model for layer pairing. For ViT-B/16, this retains all twelve transformer encoder outputs, E.1–E.12, together with the output of the final encoder LayerNorm, E.LN. We extract the CLS token vector at each readout. E.LN (Short for Layer Normalization) is a separate normalized readout of the final block output, rather than an additional transformer block.

For ResNet-50, the implementation selects fifteen of its sixteen residual blocks by rounding evenly spaced indices in forward order. The selected blocks are L1.1–L1.3, L2.1–L2.4, L3.2–L3.6, and L4.1–L4.3, where L$x.y$ denotes block y of stage x. For one image, the selected block output collapses to a vector after global pooling over the two spatial dimensions. Classification heads are excluded from feature extraction.

These representations define a 13×15 similarity matrix containing all 195 selected layer pairs. Five trials give $13 \times 15 \times 5 = 975$ layer-pair–trial records at each dropout probability.

Denote $|A|$ and $|B|$ as the number of layers of model A and model B.

Feature Dropout. For either observation under consideration, inverted feature dropout gives

$$\widetilde{x}^{(s)}_{ij} = x_{ij}\frac{M^{(s)}_{ij}}{1-p}, \qquad M^{(s)}_{ij} \sim \mathrm{Bernoulli}(1-p), \quad s \in \{1,2\}. \tag{48}$$

Here i indexes inputs and j indexes feature coordinates. Masks are independent across samples, coordinates, models, and replicas conditional on the clean features. Corruption is applied after feature extraction, without retraining. At each $p \in \{0, 0.1, 0.3, 0.5, 0.7\}$, five trials independently generate noisy observations. Each trial uses two random partitions of the same observations into batches of $b = 1{,}000$. Every method averages its three factors over both partitions before normalization; clipping or PSD stabilization is then applied to the pooled estimate. The IR-PSD ridge is $10^{-8}/\sqrt{n}$. Repeating partitions does not produce additional independent replicas.

Metrics and uncertainty. Let $R^0 \in \mathbb{R}^{|A|\times|B|}$ be full-pool clean HSIC_1-normalized CKA, and let $\widehat{R}$ be a trial estimate. For its finite-entry set $\mathcal{F}$, the reported reconstruction error is

$$\mathrm{RMSE}(\widehat{R}, R^0) = \left[\frac{1}{|\mathcal{F}|}\sum_{(i,j)\in\mathcal{F}}(\widehat{R}_{ij} - R^0_{ij})^2\right]^{1/2}. \tag{49}$$

Incomplete-matrix RMSE is explicitly marked as conditional. Whole-matrix Spearman correlation ranks the vectorized matrices and requires all 195 scores to be finite and both rank vectors to be non-constant. For exact maximizing sets $\mathcal{T}_i = \arg\max_j \widehat{R}_{ij}$ and $\mathcal{C}_i = \arg\max_j R^0_{ij}$, Jaccard similarity is

$$J = \frac{1}{|A|} \sum_{i=1}^{|A|} \frac{|\mathcal{T}_i \cap \mathcal{C}_i|}{|\mathcal{T}_i \cup \mathcal{C}_i|}. \tag{50}$$

It is reported only for complete matrices. Unlike the tolerance-aware retrieval metric of Section 5.3, this diagnostic uses exact ties and no clean-score tolerance. It tests agreement with clean CKA choices, not independently established functional layer equivalence. We compute all metrics within trials, then average the five values and report their sample SD. Layers and partitions are not uncertainty replicates; the images and checkpoints are fixed.

Additional trial summaries. IR-PSD's matrix RMSE at $p = 0.3$ is 0.026 ± 0.005. At $p = 0.5$, raw IR has RMSE 0.245 ± 0.271, compared with 0.107 ± 0.032 for both IR-clip and IR-PSD. IR-PSD's Spearman correlations at $p = 0.1, 0.3, 0.5$ are $1.000, 1.000, 0.875$ respectively, and its Jaccard coefficients are $0.969, 0.800, 0.585$, respectively. At $p = 0.7$, its RMSE is 0.320 ± 0.030.

Factor diagnostics. Each noisy factor is divided by its clean counterpart (factor at $p = 0$) using the same input partition; the following summaries are medians across layer pairs and trials. At $p = 0.1$, numerator ratios for pooled SR, MCG1, and IR are 1.000. ViT self-factor ratios are 4.991, 2.994, and 1.000; ResNet self-factor ratios are 1.699, 1.349, and 1.000. The relation $2.994 \approx 1 + (4.991 - 1)/2$, also observed on the ResNet side, agrees with the half-correction in Proposition 1.

Outliers, ties, and layer dependence. At $p = 0.5$, one IR trial has RMSE 0.729 and the maximum score is 4.615, although every score is finite. Clipping and PSD stabilization have equal matrix RMSE to the displayed precision. Their scores differ by at most 4.55×10^{-9}, but their exact maximizing sets can differ because scalar clipping creates exact ties. Such Jaccard differences do not establish a substantive PSD-specific ranking benefit.

At $p = 0.7$, raw and clipped IR have missing scores in every trial. IR-PSD has lower complete-matrix RMSE than pooled SR in each of the five trials, but the layer-pair advantage is not uniform: its trial-averaged squared error is lower on 177 of 195 pairs. Errors concentrate in earlier ViT readouts; mean row RMSE is about 0.566 at E.3, versus 0.057–0.063 at E.10–E.12 and E.LN. This heterogeneity explains why overall error alone cannot characterize recovery at every layer.

C.1 Strong-dropout factor failures: ViT-B/16 versus ResNet-50

All counts in this subsection concern $p = 0.7$ in the main experiment: $n = 10{,}000$, $b = 1{,}000$, two partitions per trial, and five trials. Write the pooled numerator and self factors as (u, v, w), so that

$$\widehat{R}_{\mathrm{IR}} = \frac{u}{\sqrt{v}\sqrt{w}}, \qquad v > 0, \quad w > 0.$$

Of the 975 layer-pair–trial records, 251 (25.74%) are undefined, all because at least one self factor is negative. Negative ViT self factors affect 240 records, negative ResNet self factors affect 13, and two records have both factors negative. No undefined scores arise from missing or nonfinite input factors. Among the remaining 724 scores, 53 exceed 1, none fall below -1, and the maximum is 3.1234.

The denominator failures occur predominantly in earlier ViT layers. Across trials, 16 distinct ViT layer–trial self-factor estimates are negative, all within `E.1`–`E.6`. For ResNet-50, only `L2.4` in trial 1 has a negative self factor. Each such failure affects an entire row or column of the similarity matrix, so the affected entries are not independent failures. By contrast, the final-layer pair, `E.LN` versus `L4.3`, gives IR scores between 0.7354 and 0.7483, close to its clean reference of 0.7379. Final-layer accuracy therefore does not imply reliable recovery throughout the networks.

For `E.7` versus `L1.3`, the numerator increases from 4.0719 in trial 0 to 4.8381 in trial 2, whereas the denominator decreases from 4.9665 to 1.5502. IR consequently increases from 0.8199 to 3.1210, against a clean reference of 0.6416. The numerator increases by 18.8%, while the denominator contracts to 31.2% of its earlier value. This example shows denominator-driven amplification, but

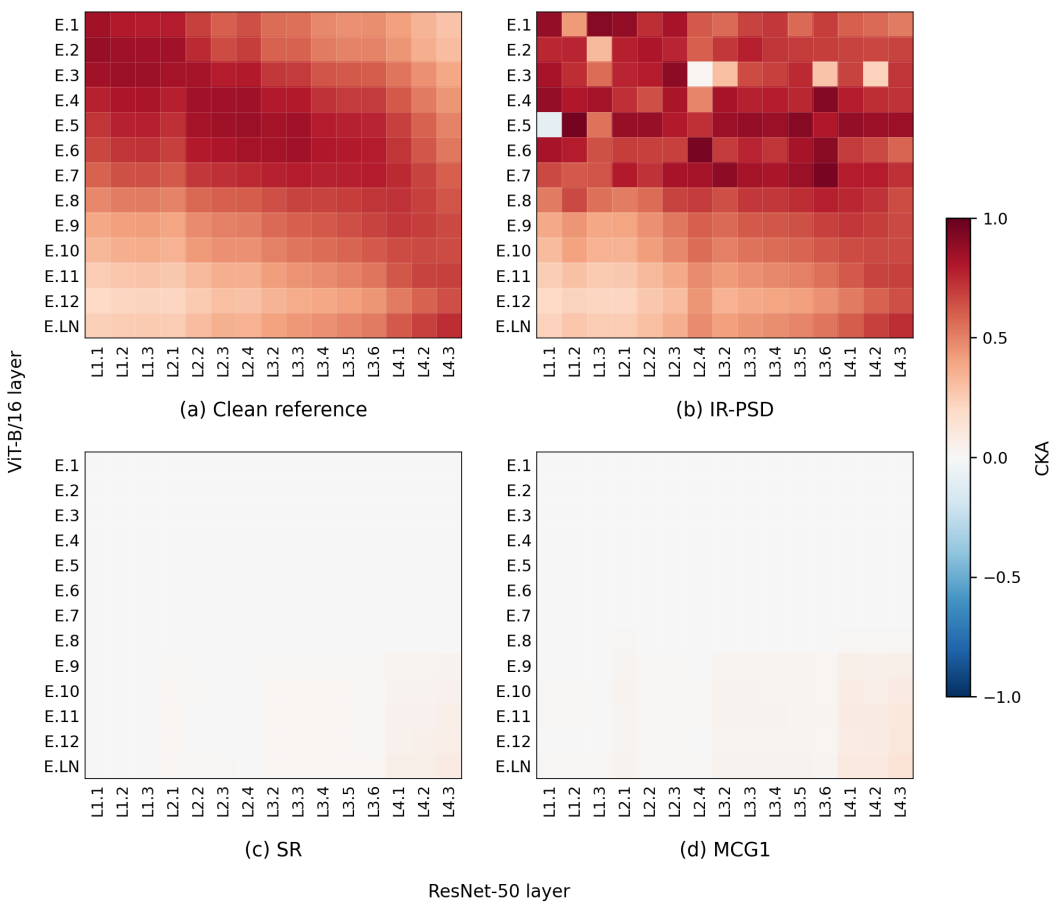


Figure 5: Similarity Heatmap for ViT-B/16 and ResNet-50 on CIFAR-10 under feature dropout with $p = 0.7$. Panel (a) shows the similarity score of the clean reference. Panels (b)–(d) show that of IR-PSD, SR, and MCG1, respectively. SR and MCG1 substantially underestimate the clean scores. IR-PSD better preserves their magnitude and the late-layer pattern, but early-layer distortions remain, including a negative mean score at E.5-L1.1.

numerator fluctuations also matter: 23 records have negative numerators despite positive clean reference scores, including 11 with valid denominators. The saved $p = 0.7$ trial file contains normalized clean reference scores, rather than all clean factors; that file alone therefore does not support a complete decomposition into numerator and denominator estimation errors.

IR-PSD projects 304 factor matrices: the 251 associated with undefined IR scores and the 53 associated with scores above 1. All resulting scores are finite and bounded, but 23 remain negative. Under the configured stabilization, all 304 projected scores satisfy $|\widehat{R}_{\text{IR-PSD}}| > 0.999$, indicating frequent saturation near the boundaries. After averaging across the five trials, one of the 195 layer-pair entries remains negative: `E.5` versus `L1.1` has mean IR-PSD -0.0995, compared with a clean reference of 0.7135.

The unnormalized HSIC_1 is not bounded by one. Its normalized same-replica CKA construction lies in $[-1, 1]$ whenever defined, and IR-PSD enforces the same interval for finite input factors. Neither bound requires a nonnegative score: PSD projection allows negative off-diagonal entries. A negative score is therefore not, by itself, an invalid output. In these records, however, negative estimates of positive clean scores are recovery errors. Boundedness and numerical availability do not establish statistical accuracy.

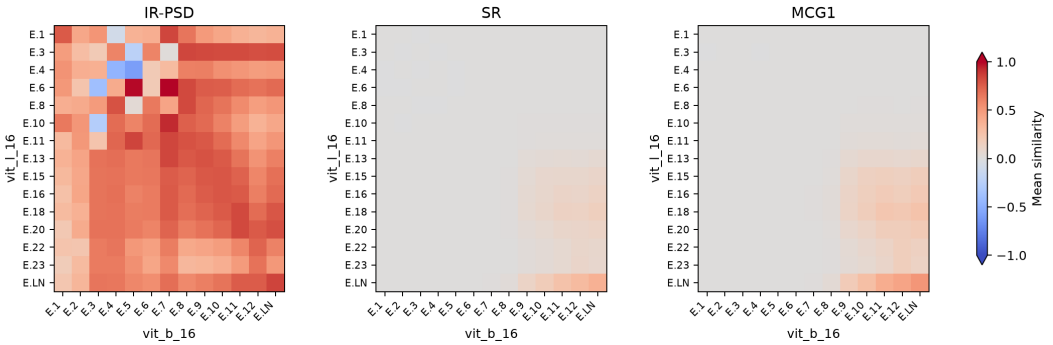


Figure 6: Trial-mean similarity matrices for ViT-L/16 (rows) and ViT-B/16 (columns) at feature dropout probability $p = 0.7$. Left: IR-PSD; middle: SR; right: MCG1. SR and MCG1 assign small scores to most early-layer pairs while retaining some positive late-layer scores. IR-PSD produces a more varied matrix, including negative entries. The panels use a common color scale and show noisy estimates, not the clean reference. This separate model pair is not the ViT-B/16–ResNet-50 pair analyzed in Section C.1.

C.2 A separate strong-dropout illustration: ViT-L/16 versus ViT-B/16

Figure 6 concerns a different model pair from the preceding diagnostics: ViT-L/16 versus ViT-B/16 on CIFAR-10 at $p = 0.7$. This run uses $n = 10{,}000$, $b = 1{,}000$, two partitions per trial, and five trials. Its 15×13 selected layer pairs again give 975 layer-pair–trial records. The 15 ViT-L/16 readouts shown on the vertical axis are E.1, E.3, E.4, E.6, E.8, E.10, E.11, E.13, E.15, E.16, E.18, E.20, E.22, E.23, and E.LN; ViT-B/16 uses E.1–E.12 and E.LN. These counts and axes must not be confused with the ViT-B/16–ResNet-50 results above.

For this pair, raw IR is undefined in 294 of 975 records (30.15%). IR-PSD projects 323 factor matrices: the 294 with undefined raw IR and 29 with finite raw scores outside $[-1, 1]$. All projected outputs are finite and bounded, but 58 of the 975 scores are negative. Six of the 195 layer-pair means remain negative after averaging the five trials. Thus, the varied IR-PSD pattern in the figure coexists with substantial estimation failures; it does not imply reliable recovery at every layer. The figure contains no clean-reference panel and serves as a qualitative comparison of the noisy score matrices.

D Layer matching: protocol, metric definitions, and complete results

D.1 Experimental question and protocol

We evaluate whether representation-similarity estimators preserve the layer relationships obtained from clean observations when only independently corrupted replicas are available. The reference is an empirical clean CKA matrix, rather than a ground-truth correspondence between model functions. For each anchor layer in one model, the candidate set consists of six layers of the other model. Retrieval is evaluated in both directions and allows multiple anchors to select the same candidate; it is not a one-to-one assignment problem.

We compare fixed pretrained ViT-B/16 and ResNet-50 representations on $n = 4096$ saved CIFAR-10 test examples. The ViT layers are E.1, E.3, E.6, E.8, E.11, and E.LN, where E.LN denotes the final encoder LayerNorm output. The ResNet layers are L1.1, L2.1, L2.4, L3.3, L3.6, and L4.3; block indices in these displayed names are one-based. Thus, each observation produces a 6×6 similarity matrix. We denote ViT-to-ResNet retrieval by $A \to B$ and the reverse direction by $B \to A$. Classification heads are excluded.

For each trial, we independently corrupt each coordinate of the extracted feature vectors using inverted dropout:

$$\widetilde{x}_{ij}^{(r)} = x_{ij}\frac{M_{ij}^{(r)}}{1-p_\ell}, \qquad M_{ij}^{(r)} \sim \text{Bernoulli}(1-p_\ell), \qquad r \in \{1,2\}. \tag{51}$$

Masks are independent across replicas and models, conditional on the clean features. In particular, $\mathbb{E}[\widetilde{x}_{ij}^{(r)} \mid x_{ij}] = x_{ij}$. This is post-extraction feature dropout, not training-time network dropout or image augmentation.

We sweep $s \in \{0, 0.2, 0.4, 0.6, 0.8\}$. In the uniform profile, every layer has $p_\ell = 0.625s$. In the heterogeneous profile, each model independently receives a random permutation of

$$(p_1, \ldots, p_6) = s\,(0.25, 0.40, 0.55, 0.70, 0.85, 1.00). \tag{52}$$

The two profiles therefore match mean dropout probability, not noise energy: conditional coordinate variance is $x_{ij}^2 p_\ell/(1-p_\ell)$. At $s = 0.8$, uniform dropout uses $p = 0.5$, whereas heterogeneous dropout ranges from 0.2 to 0.8.

Each trial uses two random partitions of the same examples into non-overlapping batches of size $b = 128$. Each partition contains 32 batches. All 64 batch contributions are pooled at the HSIC-factor level before normalization; they are not 64 independent datasets. Methods share the same replicas and partitions. We use three random seeds and five trials per seed. The seeds vary corruption, heterogeneous allocations, and partitions; they do not represent retrained models or independent sample sets. The underlying random draws are reused across strengths within a trial, yielding paired comparisons.

Estimators. All eight noisy methods are defined in Section 3. IR-PSD uses $\lambda_n = 10^{-8}/\sqrt{n} = 1.5625 \times 10^{-10}$ in HSIC-factor units. Projection acts on each layer pair's factor matrix, not on the 6×6 similarity matrix.

D.2 Clean references and six-panel evaluation

The primary reference is the empirical full-sample clean matrix

$$F_{\ell k} = \frac{h_{1,n}(K_\ell^0, L_k^0)}{\sqrt{h_{1,n}(K_\ell^0, K_\ell^0)h_{1,n}(L_k^0, L_k^0)}}, \tag{53}$$

where $h_{1,n}$ is Eq. (1) with b replaced by n. A second reference, denoted P, pools clean HSIC_1 factors over the same partitions used in the corresponding trial. The "Clean minibatch" curve evaluates P against the chosen reference and is a noise-free computational control, not a noisy estimator. MCG can differ from this HSIC_1 target even without corruption because it uses ordinary HSIC.

For either retrieval direction, let t_{ij} be the reference score for anchor i and candidate j, and let $s_{ij}^{(m)}$ be method m's score. There are $L = q = 6$ anchors and candidates. Define

$$V_i^{(m)} = \mathbf{1}\{s_{ij}^{(m)} \text{ is finite for every } j\}, \tag{54}$$

$$\mathcal{C}_i = \{j : t_{ij} \geq \max_k t_{ik} - \delta\}, \qquad \delta = 0.005, \tag{55}$$

$$\mathcal{T}_i^{(m)} = \{j : s_{ij}^{(m)} \geq \max_k s_{ik}^{(m)} - \epsilon\}, \qquad \epsilon = 10^{-10}, \tag{56}$$

$$\mathcal{J} = \{i : V_i^{(m)} = 1 \text{ for all eight noisy methods } m\}. \tag{57}$$

The predicted set $\mathcal{T}_i^{(m)}$ is used only when $V_i^{(m)} = 1$. A finite out-of-range IR score does not make an anchor incomplete. The clean control is not included when forming $\mathcal{J}$.

(a) Top-1 recovery: higher is better.

$$Q^{(m)} = \frac{1}{L}\sum_{i=1}^{L} q_i^{(m)}, \qquad q_i^{(m)} = \begin{cases} |\mathcal{T}_i^{(m)} \cap \mathcal{C}_i|/|\mathcal{T}_i^{(m)}|, & V_i^{(m)} = 1, \\ 0, & V_i^{(m)} = 0. \end{cases} \tag{58}$$

This is expected retrieval success under uniform tie-breaking among predicted maximizers. It compares candidate identities; an estimated score exceeding the clean maximum is not sufficient for success. The random-candidate baseline is $L^{-1}\sum_i |\mathcal{C}_i|/q$, equal to $1/6$ for $A \to B$ and $7/36$ for $B \to A$ in these data. The reverse direction contains a near-tie within the clean tolerance.

(b) Regret on common complete anchors: lower is better. For a complete anchor define the clean-score loss

$$D_i^{(m)} = \max_j t_{ij} - \frac{1}{|\mathcal{T}_i^{(m)}|} \sum_{j \in \mathcal{T}_i^{(m)}} t_{ij}. \tag{59}$$

Panel (b) reports

$$D_{\text{common}}^{(m)} = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} D_i^{(m)}. \tag{60}$$

All methods are compared on the same anchors within a trial. If $\mathcal{J}$ is empty, the statistic is unavailable rather than zero. The selected subset can nevertheless favor easier anchors.

(c) Spearman correlation: higher is better. Let $\mathcal{I}_\rho^{(m)}$ contain anchors with complete scores and nonconstant reference and predicted rank vectors. The plotted statistic is

$$\rho^{(m)} = \frac{1}{|\mathcal{I}_\rho^{(m)}|} \sum_{i \in \mathcal{I}_\rho^{(m)}} \operatorname{Corr}\big(\operatorname{rank}(t_{i\cdot}), \operatorname{rank}(s_{i\cdot}^{(m)})\big). \tag{61}$$

Ranks average exact ties. This panel uses method-specific complete anchors, not $\mathcal{J}$; it is a conditional ordering diagnostic, not a coverage-adjusted score.

(d) Pairwise order accuracy: higher is better. For anchor i, compare $\mathcal{P}_i = \{(j,k) : j < k,\ |t_{ij} - t_{ik}| > \delta\}$. For each eligible pair, define

$$a_{ijk}^{(m)} = \begin{cases} 0, & s_{ij}^{(m)} \text{ or } s_{ik}^{(m)} \text{ is undefined,} \\ 1/2, & |s_{ij}^{(m)} - s_{ik}^{(m)}| \le \epsilon, \\ \mathbf{1}\{(t_{ij} - t_{ik})(s_{ij}^{(m)} - s_{ik}^{(m)}) > 0\}, & \text{otherwise.} \end{cases} \tag{62}$$

We average $a_{ijk}^{(m)}$ within each anchor and then across anchors with at least one eligible pair. All anchors in this experiment have eligible pairs. Missing comparisons count as failures, while other comparisons in an incomplete anchor still contribute.

(e) Pairwise gap RMSE on common anchors: lower is better.

$$E_i^{(m)} = \left\{ \frac{1}{\binom{q}{2}} \sum_{j<k} \Big[(s_{ij}^{(m)} - s_{ik}^{(m)}) - (t_{ij} - t_{ik}) \Big]^2 \right\}^{1/2}, \tag{63}$$

$$E_{\text{common}}^{(m)} = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} E_i^{(m)}. \tag{64}$$

All candidate pairs are used here, without the clean-gap exclusion of panel (d). With $e_{ij} = s_{ij}^{(m)} - t_{ij}$, the implementation uses the identity

$$(E_i^{(m)})^2 = \frac{2}{q-1} \sum_j (e_{ij} - \bar{e}_i)^2. \tag{65}$$

Consequently, a common additive score shift is ignored, but collapsing distinct clean scores to a common saturated value is penalized. This measures relative score geometry, not merely ordering. An estimator can attain better retrieval with worse gap RMSE.

(f) Regret on method-specific complete anchors: lower is better.

$$D_{\text{complete}}^{(m)} = \frac{\sum_{i : V_i^{(m)} = 1} D_i^{(m)}}{\sum_i V_i^{(m)}}. \tag{66}$$

Unlike panel (b), the evaluated anchors can differ across methods. A smaller value does not establish superiority if difficult anchors are missing. Undefined anchors are excluded from these two conditional metrics.

Aggregation and uncertainty. Each metric is first computed within one trial, including its specified anchor averaging. For seed r, let $\mathcal{U}_r$ be the trials in which that metric is defined. The reported mean and shading are

$$\bar{z}_r = \frac{1}{|\mathcal{U}_r|} \sum_{u \in \mathcal{U}_r} z_{ru}, \qquad \bar{z} = \frac{1}{3} \sum_{r=1}^{3} \bar{z}_r, \qquad \mathrm{SD}(z) = \sqrt{\frac{1}{2} \sum_{r=1}^{3} (\bar{z}_r - \bar{z})^2}. \tag{67}$$

All reported settings retain at least one valid trial in each seed. Conditional metrics can use fewer than five trials per seed; seed means still receive equal weights. Shading is the sample SD across three seed means, not a confidence interval, not the SD across 15 trials, and not a pooled-anchor statistic.

D.3 Results

Clean controls and reference sensitivity. At $s = 0$, every method achieves unit tolerance-aware retrieval in both directions. MCG1 coincides with the minibatch HSIC_1 construction in the absence of corruption. MCG nevertheless has nonzero gap error: approximately 0.0361 for $A \to B$ and 0.0456 for $B \to A$, compared with 0.0009 and 0.0011 for the clean minibatch control. This reflects its different finite-sample HSIC functional and should not be attributed to noise. Across all settings, switching from the full clean reference to the same-partition clean reference leaves top-1 recovery and pairwise order accuracy unchanged. The largest changes in the reported common regret and common gap RMSE are 0.000875 and 0.001213, respectively. The main conclusions are therefore not explained by a full-sample versus minibatch reference mismatch.

What the clean layer relationships encode. The clean $A \to B$ maximizers are E.1↦L1.1, E.3↦L1.1, E.6↦L3.3, E.8↦L3.6, E.11↦L4.3, and E.LN↦L4.3. Thus the task measures recovery of a many-to-one empirical layer relationship rather than diagonal alignment by depth. In the reverse direction, L1.1 has clean scores 0.843168 and 0.842664 for E.1 and E.3, respectively; both are accepted under $\delta = 0.005$. This explains the larger tolerance-aware random baseline in $B \to A$. Directional asymmetry is also amplified by missing-score geometry: one fully missing ViT row removes one $A \to B$ anchor but leaves every $B \to A$ anchor with a missing candidate.

Independent-replica factors improve low-noise matching. Table 2 reports the first nonzero noise level. Under uniform dropout ($s = 0.2$, hence $p = 0.125$), SR, pooled SR, feature averaging, MCG, and MCG1 all achieve 50.0% recovery in $A \to B$ and 16.7% in $B \to A$. IR achieves 92.2% and 97.8%, respectively. All anchors are complete at this noise level, so the comparison is not affected by missing-score selection. IR also reduces common gap RMSE from 0.1553 to 0.0077 relative to MCG1 in $A \to B$, and from 0.3287 to 0.0146 in $B \to A$. These reductions show recovery of relative scores beyond a coincidentally correct maximizer.

Under heterogeneous dropout at $s = 0.2$, IR achieves 88.9% and 92.2% recovery, whereas MCG1 achieves 50.0% and 25.6%. IR's common gap errors are 0.0093 and 0.0217, versus 0.1632 and 0.3297 for MCG1. PSD and clipped IR have essentially identical performance to raw IR at this strength. Thus, the low-noise improvement is attributable to the underlying IR construction rather than stabilization. Because MCG1 also uses two independent observations and HSIC_1 in all three factors, this comparison isolates the importance of how replica information enters the self factors; simply constructing a cross-Gram is insufficient in this experiment.

The advantage persists at intermediate strengths. At $s = 0.4$, raw IR achieves $81.1\%/78.9\%$ recovery in uniform $A \to B/B \to A$, compared with $38.9\%/16.7\%$ for MCG1. Under heterogeneous dropout, the corresponding rates are $80.0\%/71.1\%$ for IR and $43.3\%/20.0\%$ for MCG1. At $s = 0.6$, IR-PSD reaches $70.7\%/59.4\%$ under uniform noise and $67.8\%/50.8\%$ under heterogeneous noise. In the latter reverse-direction setting, raw IR achieves only 28.9%, demonstrating the increasing importance of handling unstable self factors. The six-panel figures below show the full strength sweep; Tables 3–6 detail the strongest-noise results.

Strong noise separates retrieval quality from score geometry. At $s = 0.8$, IR-PSD retains higher recovery and pairwise order accuracy than the non-IR baselines in each profile and direction, but recovery remains incomplete (Table 3). In uniform $A \to B$, recovery is 54.8% for IR-PSD,

Table 2: Top-1 recovery (%) at $s = 0.2$, using the full clean reference. Values are mean ± SD across three seed means. All anchors and all 15 trials are complete for every method.

	Uniform		Heterogeneous	
Method	$A \to B$	$B \to A$	$A \to B$	$B \to A$
SR	50.0 ± 0.0	16.7 ± 0.0	48.9 ± 5.1	24.4 ± 3.8
Pooled SR	50.0 ± 0.0	16.7 ± 0.0	48.9 ± 5.1	24.4 ± 3.8
Feature mean	50.0 ± 0.0	16.7 ± 0.0	53.3 ± 3.3	30.0 ± 6.7
MCG	50.0 ± 0.0	16.7 ± 0.0	50.0 ± 3.3	25.6 ± 5.1
MCG1	50.0 ± 0.0	16.7 ± 0.0	50.0 ± 3.3	25.6 ± 5.1
IR	92.2 ± 3.8	97.8 ± 3.8	88.9 ± 1.9	92.2 ± 1.9
IR-PSD	92.2 ± 3.8	97.8 ± 3.8	88.9 ± 1.9	92.2 ± 1.9
IR-clip	92.2 ± 3.8	97.8 ± 3.8	88.9 ± 1.9	92.2 ± 1.9

Table 3: Uniform dropout, $A \to B$, $s = 0.8$. Full clean reference. Mean ± SD of seed means. Recovery and order accuracy are percentages; other quantities use their native scales. Common and conditional metrics exclude unavailable trials as described in the text.

Method	Recovery	D_{common}	Spearman	Order	E_{common}	D_{complete}
SR	35.6 ± 1.9	0.063 ± 0.014	0.210 ± 0.015	57.7 ± 0.9	0.190 ± 0.004	0.068 ± 0.007
Pooled SR	35.6 ± 1.9	0.062 ± 0.019	0.213 ± 0.025	57.4 ± 0.6	0.190 ± 0.003	0.065 ± 0.013
Feature mean	33.3 ± 0.0	0.058 ± 0.004	0.209 ± 0.003	58.0 ± 0.2	0.176 ± 0.004	0.060 ± 0.005
MCG	33.3 ± 0.0	0.081 ± 0.022	0.170 ± 0.029	56.2 ± 1.0	0.174 ± 0.005	0.084 ± 0.019
MCG1	33.3 ± 0.0	0.066 ± 0.009	0.196 ± 0.014	57.6 ± 0.6	0.181 ± 0.004	0.068 ± 0.010
IR	52.2 ± 9.6	0.045 ± 0.018	0.701 ± 0.064	71.0 ± 8.5	0.355 ± 0.150	0.045 ± 0.018
IR-PSD	54.8 ± 8.8	0.046 ± 0.017	0.683 ± 0.061	80.0 ± 3.3	0.183 ± 0.016	0.046 ± 0.015
IR-clip	49.6 ± 10.7	0.051 ± 0.023	0.692 ± 0.066	69.6 ± 9.0	0.183 ± 0.016	0.051 ± 0.023

52.2% for raw IR, 49.6% for clipped IR, and 33.3% for MCG1. Under heterogeneous $A \to B$, the corresponding values are 44.4%, 42.2%, 41.0%, and 26.7%. These results support improved retrieval, not near-perfect recovery under severe corruption.

The reverse direction exposes a stronger reliability limitation. Under uniform and heterogeneous $s = 0.8$, raw IR has complete-anchor fractions of 46.7% and 38.9%, respectively. Its retrieval rates fall to 21.1% and 18.9%, close to the tolerance-aware random-candidate baseline of 19.4%. IR-PSD remains complete and reaches 42.2% and 39.6%. Its pairwise order accuracy is 73.5% and 71.7%, compared with 30.8% and 35.3% for MCG1. Clipping alone preserves raw IR's missing entries and cannot provide the same completeness improvement.

However, the gap metric gives an important counterexample to uniform superiority. Under heterogeneous $s = 0.8$, common gap RMSE is 0.4171 for raw IR, 0.2242 for IR-PSD, and 0.2126 for MCG1 in $A \to B$. In $B \to A$, it is 0.3644, 0.3583, and 0.2981, respectively. Under uniform $B \to A$, raw IR's gap RMSE reaches 0.7974, compared with 0.3435 for IR-PSD and 0.2899 for MCG1. Thus, stabilizing the ratio reduces some extreme errors but does not guarantee more accurate score spacing than the baselines. Ranking, top-1 selection, and numerical gap recovery are distinct outcomes.

Common-support and conditional regret answer different questions. At heterogeneous $s = 0.8$ in $B \to A$, common regret is 0.0775 for raw, clipped, and PSD IR, compared with 0.3164 for MCG1.

Table 4: Uniform dropout, $B \to A$, $s = 0.8$. Full clean reference. Mean ± SD of seed means. Recovery and order accuracy are percentages; other quantities use their native scales. Common and conditional metrics exclude unavailable trials as described in the text.

Method	Recovery	D_{common}	Spearman	Order	E_{common}	D_{complete}
SR	16.7 ± 0.0	0.388 ± 0.001	-0.451 ± 0.013	30.2 ± 0.3	0.283 ± 0.000	0.387 ± 0.000
Pooled SR	16.7 ± 0.0	0.388 ± 0.001	-0.448 ± 0.009	30.0 ± 0.8	0.284 ± 0.000	0.388 ± 0.001
Feature mean	16.7 ± 0.0	0.383 ± 0.001	-0.408 ± 0.016	32.1 ± 0.8	0.299 ± 0.001	0.384 ± 0.001
MCG	16.7 ± 0.0	0.386 ± 0.000	-0.418 ± 0.010	30.8 ± 0.6	0.301 ± 0.000	0.386 ± 0.001
MCG1	16.7 ± 0.0	0.384 ± 0.001	-0.424 ± 0.005	30.8 ± 0.5	0.290 ± 0.000	0.386 ± 0.001
IR	21.1 ± 20.4	0.112 ± 0.029	0.431 ± 0.185	57.9 ± 11.6	0.797 ± 0.637	0.112 ± 0.029
IR-PSD	42.2 ± 8.4	0.107 ± 0.033	0.509 ± 0.082	73.5 ± 3.1	0.344 ± 0.062	0.115 ± 0.011
IR-clip	22.2 ± 21.1	0.108 ± 0.032	0.433 ± 0.186	57.9 ± 11.8	0.344 ± 0.062	0.108 ± 0.032

Table 5: Heterogeneous dropout, $A \to B$, $s = 0.8$. Full clean reference. Mean $\pm$ SD of seed means. Recovery and order accuracy are percentages; other quantities use their native scales. Common and conditional metrics exclude unavailable trials as described in the text.

Method	Recovery	D_{common}	Spearman	Order	E_{common}	D_{complete}
SR	28.9 ± 8.4	0.129 ± 0.036	0.236 ± 0.035	58.8 ± 1.7	0.212 ± 0.009	0.121 ± 0.019
Pooled SR	26.7 ± 8.8	0.134 ± 0.032	0.200 ± 0.041	57.5 ± 2.2	0.212 ± 0.009	0.130 ± 0.021
Feature mean	27.8 ± 7.7	0.132 ± 0.044	0.244 ± 0.037	58.6 ± 1.2	0.213 ± 0.007	0.118 ± 0.026
MCG	26.7 ± 8.8	0.144 ± 0.028	0.233 ± 0.038	58.2 ± 1.2	0.217 ± 0.007	0.136 ± 0.018
MCG1	26.7 ± 8.8	0.140 ± 0.035	0.229 ± 0.030	58.2 ± 1.5	0.213 ± 0.008	0.129 ± 0.020
IR	42.2 ± 8.4	0.043 ± 0.042	0.647 ± 0.065	64.5 ± 5.7	0.417 ± 0.188	0.043 ± 0.042
IR-PSD	44.4 ± 8.4	0.043 ± 0.042	0.575 ± 0.024	74.6 ± 1.2	0.224 ± 0.050	0.072 ± 0.030
IR-clip	41.0 ± 8.8	0.047 ± 0.036	0.644 ± 0.050	64.2 ± 5.5	0.224 ± 0.050	0.047 ± 0.036

Table 6: Heterogeneous dropout, $B \to A$, $s = 0.8$. Full clean reference. Mean $\pm$ SD of seed means. Recovery and order accuracy are percentages; other quantities use their native scales. Common and conditional metrics exclude unavailable trials as described in the text.

Method	Recovery	D_{common}	Spearman	Order	E_{common}	D_{complete}
SR	11.1 ± 3.8	0.341 ± 0.059	-0.377 ± 0.037	35.0 ± 1.3	0.291 ± 0.005	0.354 ± 0.018
Pooled SR	11.1 ± 3.8	0.341 ± 0.059	-0.406 ± 0.021	33.8 ± 0.5	0.291 ± 0.005	0.357 ± 0.014
Feature mean	13.3 ± 3.3	0.279 ± 0.100	-0.342 ± 0.031	36.6 ± 1.5	0.305 ± 0.004	0.310 ± 0.031
MCG	11.1 ± 3.8	0.321 ± 0.084	-0.390 ± 0.027	34.4 ± 1.4	0.310 ± 0.004	0.352 ± 0.015
MCG1	11.1 ± 3.8	0.316 ± 0.082	-0.375 ± 0.027	35.3 ± 1.7	0.298 ± 0.005	0.346 ± 0.017
IR	18.9 ± 13.5	0.078 ± 0.007	0.543 ± 0.153	57.4 ± 9.7	0.364 ± 0.078	0.078 ± 0.007
IR-PSD	39.6 ± 5.0	0.078 ± 0.007	0.461 ± 0.143	71.7 ± 4.7	0.358 ± 0.075	0.115 ± 0.015
IR-clip	18.9 ± 13.5	0.078 ± 0.007	0.543 ± 0.153	57.3 ± 9.7	0.358 ± 0.075	0.078 ± 0.007

Method-specific complete regret is 0.0775 for raw IR but 0.1154 for IR-PSD. The latter difference is not evidence that PSD harms the same anchors: IR-PSD is evaluated on all anchors in panel (f), including cases excluded from raw IR's conditional statistic. In this dataset the common complete set coincides with the raw-IR complete set, because the other baselines and IR-PSD remain complete and clipping preserves raw IR's undefined entries.

The amount of selection is substantial. At $s = 0.8$, only 7 of 15 uniform reverse-direction trials and 6 of 15 heterogeneous reverse-direction trials contain any common complete anchors. For heterogeneous $A \to B$, 12 of 15 trials remain; all 15 remain for uniform $A \to B$. The common anchor fractions are 87.8%, 46.7%, 71.1%, and 38.9% for uniform $A \to B$, uniform $B \to A$, heterogeneous $A \to B$, and heterogeneous $B \to A$, respectively. Table 7 reports the coverage and trial counts omitted from the compact figures. Conditional curves can improve or fluctuate as difficult anchors disappear; they are not unconditional risk estimates.

PSD and clipping have complementary limitations. On common complete anchors, IR-PSD and clipped IR have numerically indistinguishable gap-RMSE curves at the displayed precision. When a finite raw ratio violates the bound and its two self factors are positive, PSD projection can place its factor matrix on a rank-one boundary, producing a score close to the same signed boundary as clipping. Their main empirical difference here is that PSD also yields finite normalized scores when raw self factors fail, whereas clipping does not. PSD saturation can leave tiny score differences that determine a unique maximizer at $\epsilon = 10^{-10}$; neither low regret nor a bounded score alone demonstrates a well-separated recovered ranking.

Table 7: Availability at $s = 0.8$. Mean common-anchor coverage equals raw/clipped IR coverage in these data. IR-PSD and all non-IR baselines have 100% method-specific coverage. The common-metric trial count is out of 15; every seed retains at least one trial.

Profile	Direction	Common coverage (%)	Trials with common anchors
Uniform	$A \to B$	87.8	15/15
Uniform	$B \to A$	46.7	7/15
Heterogeneous	$A \to B$	71.1	12/15
Heterogeneous	$B \to A$	38.9	6/15

Scope of the evidence. Taken together, the results support the IR factor construction as a useful estimator of clean similarity and layer relationships under independent, mean-preserving feature corruption. The MCG1 comparison shows that independent observations alone do not make every self-normalized construction equally effective. The strongest evidence is at low and intermediate noise, where IR improves both retrieval and relative-score error without relying on missing-data selection. At high noise, stabilization improves operational availability and several retrieval measures but does not uniformly improve score geometry. This experiment concerns one fixed model pair, a fixed sample set, and synthetic feature dropout; it does not establish functional layer equivalence, performance under arbitrary image corruptions, or variability across independently trained models. We report descriptive means and seed-level SDs and do not claim statistical significance from three seeds.

D.4 Six-panel figures and captions

The following three full-reference figures complement Figure 4, covering both retrieval directions and both noise profiles. The same-partition-reference counterparts serve as sensitivity checks, not additional independent experiments. In every figure, panels (a)–(f) correspond exactly to the metrics defined in Section D.2. Curves with identical means and SDs are grouped and identified by the legend indices beneath the relevant panel. A missing conditional value is omitted, not replaced by zero.

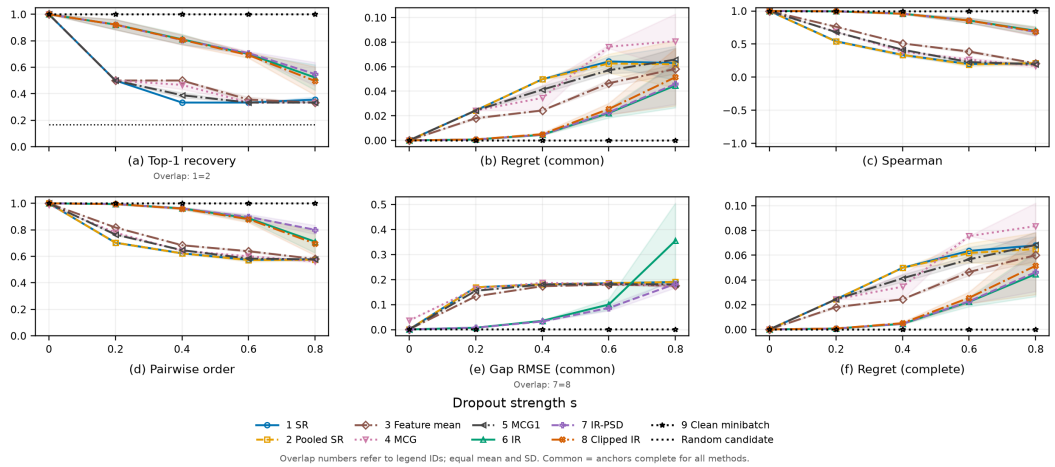


Figure 7: Uniform feature dropout, $A \to B$. (a) Tolerance-aware top-1 recovery; (b) clean regret on anchors complete for all methods; (c) Spearman correlation on method-specific complete, nonconstant rankings; (d) pairwise order accuracy, counting missing comparisons as failures; (e) pairwise gap RMSE on common complete anchors; (f) clean regret on method-specific complete anchors. Higher is better in (a,c,d); lower is better in (b,e,f). Shading is SD across three seed means. The clean tolerance is 0.005, and the predicted tie tolerance is 10^{-10}. The actual dropout probability is $p = 0.625s$. At $s = 0.8$, common coverage is 87.8%, and panels (b,e) use 15/15 trials; their means must not be interpreted as full-coverage risk.

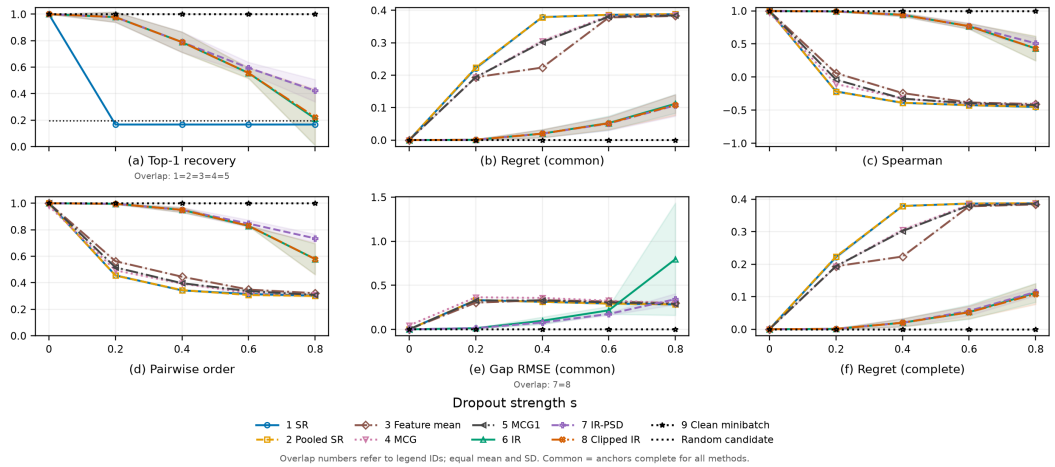


Figure 8: Uniform feature dropout, $B \to A$. (a) Tolerance-aware top-1 recovery; (b) clean regret on anchors complete for all methods; (c) Spearman correlation on method-specific complete, nonconstant rankings; (d) pairwise order accuracy, counting missing comparisons as failures; (e) pairwise gap RMSE on common complete anchors; (f) clean regret on method-specific complete anchors. Higher is better in (a,c,d); lower is better in (b,e,f). Shading is SD across three seed means. The clean tolerance is 0.005, and the predicted tie tolerance is 10^{-10}. The actual dropout probability is $p = 0.625s$. At $s = 0.8$, common coverage is 46.7%, and panels (b,e) use 7/15 trials; their means must not be interpreted as full-coverage risk.

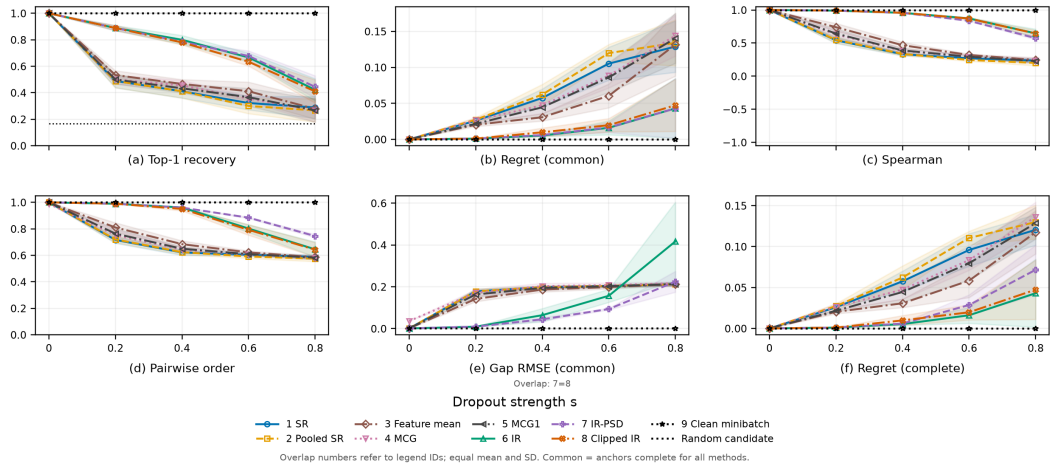


Figure 9: Heterogeneous feature dropout, $A \to B$. (a) Tolerance-aware top-1 recovery; (b) clean regret on anchors complete for all methods; (c) Spearman correlation on method-specific complete, nonconstant rankings; (d) pairwise order accuracy, counting missing comparisons as failures; (e) pairwise gap RMSE on common complete anchors; (f) clean regret on method-specific complete anchors. Higher is better in (a,c,d); lower is better in (b,e,f). Shading is SD across three seed means. The clean tolerance is 0.005, and the predicted tie tolerance is 10^{-10}. Layer probabilities are random permutations of $s(0.25, 0.40, 0.55, 0.70, 0.85, 1)$. At $s = 0.8$, common coverage is 71.1%, and panels (b,e) use 12/15 trials; their means must not be interpreted as full-coverage risk.