When to Align, When to Predict: A Phase Diagram for Multimodal Learning

Ilay Kamai[*]
Technion
ilay.kamai@campus.technion.ac.il

Hugues Van Assel
Genentech

Aviv Regev
Genentech

Hagai B. Perets
Technion

Randall Balestriero
Brown University

Abstract

Cross-modal alignment (CA) and cross-modal prediction (CP) are the dominant paradigms for multimodal representation learning, yet there is no systematic understanding of when each succeeds and when each fails, a gap that leaves practitioners, especially in scientific domains with heterogeneous instruments and multiple levels of measurement, unable to diagnose why standard methods underperform the best single modality. We study both objectives under a spiked signal-plus-noise model with structured cross-modal nuisance correlation, the ingredient that breaks the classical recovery guarantees, and derive separation ratios that expose complementary failure modes: alignment whitens each modality and fails when nuisance is strongly correlated across views; prediction encodes whatever is cross-predictable through a one-sided whitening, with recovery governed by source-modality quality. The resulting phase diagram partitions multimodal problems into four regimes — Both, CA only, CP only, and Neither — refined by a recovery count that separates partial recovery from complete failure. We present a data-driven procedure to locate real-world datasets in this diagram using a small labeled subsample, identifying the preferred objective and prediction direction before any cross-modal training, and identifying when no objective in the CA/CP family can improve on the stronger modality alone. Experiments on synthetic data, stereo-vision benchmarks, image–caption pairs, and two real scientific domains, astronomy and single-cell multi-omics, validate the predictions in the nonlinear regime, including both faces of the Neither regime. Code to reproduce the results is available at `https://github.com/IlayMalinyak/mm_align_vs_pred`.

1 Introduction

Multimodal representation learning aims to extract a shared latent structure from paired observations across different modalities, such as images and captions, audio and video, molecular cell profiles and tissue images, or different telescopes observing the same object. Multimodal learning is crucial when a single modality alone is insufficient to fully describe a phenomenon of interest or when the information in a single modality is degenerate or noisy. Moreover, combining multiple modalities of the same object is an important building block for foundational models. Multimodal learning has achieved many successes across domains and scales (e.g., [14, 36, 1, 9]). However, the field is mostly empirical, and theoretical studies are relatively sparse, though phenomena like the modality gap in contrastive multimodal models [26] have started attracting principled analysis (e.g., [43]).

[*]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Here, we focus on the interplay between the two leading multimodal learning paradigms - *Cross-modal alignment* and *Cross-modal prediction*. *Cross-modal alignment* (CA) projects paired samples into a common embedding space, encouraging matched pairs to be close; CLIP [37], ImageBind [18], and VICReg [8] are prominent examples. *Cross-modal prediction* (CP) reconstructs one modality from the other through a bottleneck, so that the learned representation retains whatever is useful for prediction; masked autoencoders [20], data2vec [5], and the decoder side of encoder-decoder models follow this approach. Both paradigms are widely used, yet they are typically studied in isolation and selected by practitioners based on empirical performance or architectural convenience rather than a principled understanding of their relative strengths and suitability to the problem at hand. We shed light on fundamental characteristics in multimodal learning, when implemented with CA or CP, and provide practical guidelines for success and failure modes of the two. We derive exact solutions and recovery conditions in the linear case, verify the results for the non-linear case using various experiments with deep neural networks, and provide a data-driven method for the analysis of multimodality problems. To the best of our knowledge, this is the first work to systematically compare CA and CP under a multimodal spiked model with structured cross-modal nuisance correlation, and to translate the resulting recovery conditions into a practical diagnostic procedure that applies to real paired datasets. Our main contributions are:

- **Unified linear analysis of CA and CP.** Using the known equivalences of CA with Canonical Correlation Analysis (CCA) and CP with truncated reduced-rank regression (RRR) as a starting point, we derive closed-form solutions for both objectives and analyze them under a spiked signal-plus-noise model with cross-modal nuisance correlation. We derive separation ratios Δ_{CA} and Δ_{CP} that determine when each method recovers the shared signal subspace, and expose complementary failure modes.
- **Phase diagram with four recovery regimes.** The separation ratios partition the space of multimodal problems into four regions — CA only, CP only, Both, and Neither — visualized as a phase diagram in signal-noise space. We identify the Neither regime as the natural habitat of complementary scientific modalities and an important open problem for multimodal representation learning.
- **Data-driven recovery regime estimation.** We propose an algorithm that predicts the separation ratios for any paired dataset, based on a small labeled subsample, and before any cross-modal training. Beyond this regime prediction, the per-modality noise estimates identify which modality is stronger and in which direction CP should be applied, a non-trivial question to address in practice.
- **Experimental validation across scales.** Experiments on synthetic data, controlled stereovision benchmarks, and image–caption pairs confirm that the failure modes identified in the linear theory persist with deep networks. On real scientific data, spanning astronomy and single-cell multi-omics, we confirm the predicted regime shift and observe both faces of the Neither regime: complete failure and partial recovery.

The paper is organized as follows: in Section 3, we present the methods and construct a multimodal spiked model with both modality-specific and cross-modal correlated noise features for the linear case. We then derive signal recovery conditions and phase diagrams. In Section 4, we provide experimental results that support the theory and an algorithm for estimating recovery regimes. Conclusions are provided in Section 5.

2 Related Work

Theory of multimodal learning. A small but growing body of work studies multimodal learning theoretically. One study [22] analyzed when using more modalities reduces population risk, showing through generalization bounds that the benefit depends on the gap in representation quality between modality subsets. Their linear analysis assumes orthonormal projections and full-rank covariance. Another work [28] proves that multimodal learning can achieve lower sample complexity than unimodal learning by decoupling the complexity of the connection function from the predictor. However, their analysis does not capture the effects of modality-specific noise, nuisance features, or rank deficiency that arise in practice. Both works establish that multimodality *can* help under specific assumptions; our work characterizes when it works and when it *can also hurt*, and shows that the specific approach (*i.e.* CA or CP) used to combine modalities matters. In a complementary

direction, [39] studies spike detection in a multimodal spiked covariance model, comparing self-covariance, cross-covariance, and joint-covariance decompositions under finite-sample (Wishart) noise. They derive Baik–Ben Arous–Péché (BBP) phase transitions for each matrix and produce phase diagrams showing which method detects the signal first as a function of signal strength and sampling ratio. Concurrently, [33] establish BBP-type thresholds for partial least squares (PLS) and CCA under finite-sample Wishart noise, and [40] extend this line to a multimodal spiked matrix–tensor model, showing that joint maximum-likelihood optimization is strictly worse than a sequential strategy. Both perform finite-sample BBP-type analyses of a single signal model. Our contribution is complementary, comparing CA against CP at the population level under structured cross-modal nuisance covariance η_j in a matrix–matrix setting. This parameterization is what produces phenomena that prior spiked-model analyses did not capture, like source–target asymmetry of CP, and the *Neither* regime in which neither paradigm fully recovers the signal.

Theory of self-supervised and contrastive learning. Our work also connects to a rich literature on the theory of unimodal self-supervised learning (SSL), where "views" are generated by data augmentation rather than being structurally distinct modalities. Several works derive closed-form solutions for linear SSL models and analyze the role of augmentations in shaping learned representations (e.g., [11, 6]). Most closely related is [42], which compares joint-embedding and reconstruction-based SSL in a unified linear framework, the unimodal counterpart of our analysis. They show that joint-embedding methods impose a strictly weaker alignment condition on augmentations than reconstruction methods when irrelevant features have large magnitude, providing provable guidelines for choosing between the two paradigms. Our work extends this line of inquiry to the genuinely multimodal setting, where the two "views" are not designed augmentations of the same input but fixed, structurally distinct modalities with inherent asymmetries in quality and noise. This introduces new phenomena, such as modality bottlenecks, cross-modal nuisance correlation, and asymmetry of cross-prediction, that do not arise in the unimodal SSL framework. Other theoretical works on contrastive and non-contrastive learning (e.g., [3, 19, 41]) study downstream task performance as a function of augmentation design, but do not compare between alignment and prediction or address multimodal noise structure.

Beyond theory, our analysis also speaks to a core architectural choice: whether to add a predictor head on top of a joint embedding like in JEPA ([25]) architectures (*e.g.,* [4, 7]) or align embeddings directly (*e.g.,* SimSiam [13], VICReg [8]). This choice has typically been motivated by collapse prevention or by the presence of conditional context. Our analysis adds a complementary axis to this design choice in the multimodal case: even without conditional information or asymmetric features, the right choice between aligning and predicting depends on the noise structure of the modality pair.

3 Cross-Alignment and Cross Prediction Approaches

We study two objectives for multimodal representation learning. The first is *cross-alignment* (CA), which aligns paired samples in a shared latent space. The second is *cross-prediction* (CP), which predicts one modality from the other through an encoder–decoder factorization. Both are formalized below and analyzed throughout the paper. More details and proofs are deferred to Appendix A.

3.1 Objectives

Let $f_{\mathcal{X}} : \mathbb{R}^{d_x} \to \mathbb{R}^k$ and $f_{\mathcal{Y}} : \mathbb{R}^{d_y} \to \mathbb{R}^k$ be encoders producing latent codes $\mathbf{z}_x^{(i)} := f_{\mathcal{X}}(\mathbf{x}_i)$ and $\mathbf{z}_y^{(i)} := f_{\mathcal{Y}}(\mathbf{y}_i)$ in a shared latent space of dimension k, and let $\bar{\mathbf{z}}_x^{(i)} := \mathbf{z}_x^{(i)} - \frac{1}{n}\sum_j \mathbf{z}_x^{(j)}$ and $\bar{\mathbf{z}}_y^{(i)} := \mathbf{z}_y^{(i)} - \frac{1}{n}\sum_j \mathbf{z}_y^{(j)}$ denote the centered codes. Let $f_{\mathcal{D}} : \mathbb{R}^k \to \mathbb{R}^{d_y}$ be a decoder. The two objectives are

$$\text{(CA)} \quad \min_{f_{\mathcal{X}}, f_{\mathcal{Y}}} \ \frac{1}{n}\sum_i \|f_{\mathcal{X}}(\mathbf{x}_i) - f_{\mathcal{Y}}(\mathbf{y}_i)\|_2^2 \quad \text{s.t.} \quad \frac{1}{n}\sum_i \bar{\mathbf{z}}_x^{(i)}\bar{\mathbf{z}}_x^{(i)\top} = \frac{1}{n}\sum_i \bar{\mathbf{z}}_y^{(i)}\bar{\mathbf{z}}_y^{(i)\top} = \mathbf{I}_k, \tag{1}$$

$$\text{(CP)} \quad \min_{f_{\mathcal{X}}, f_{\mathcal{D}}} \ \frac{1}{n}\sum_i \|\mathbf{y}_i - f_{\mathcal{D}}(f_{\mathcal{X}}(\mathbf{x}_i))\|_2^2. \tag{2}$$

The CA constraint fits the cross-modal similarity matrix to the identity: matched pairs are pulled together while each view is separately whitened, so that no coordinate collapses and no two coordinates are redundant. Whitening each view rather than their cross-moment is what makes equation 1

reduce exactly to CCA (Theorem A.1); it is also the form implemented softly by the variance and covariance terms of VICReg [8], which we use as the CA objective in all non-linear experiments. Figure 1 illustrates the two paradigms. To understand when each objective succeeds or fails, we restrict to the linear case, where equation 1 and equation 2 admit closed-form solutions and the population geometry is fully tractable. This linear analysis captures the core mechanism and provides our main analysis tool, and its predictions transfer to the nonlinear regime in our experiments (Section 4).

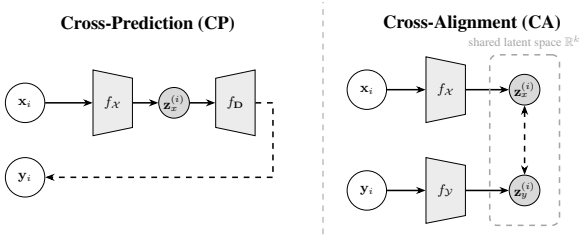


Figure 1: Two multimodal learning paradigms studied in this work. *(Left)* **Cross-prediction** (CP), Equation (2): an encoder $f_{\mathcal{X}}$ maps modality $\mathbf{x}$ to a latent code $\mathbf{z}$, and a decoder $f_{\mathbf{D}}$ reconstructs the paired target $\mathbf{y}$. *(Right)* **Cross-alignment** (CA), Equation (1): encoders $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$ project paired samples $(\mathbf{x}_i, \mathbf{y}_i)$ into a shared latent space where matched pairs are pulled together.

3.2 Linear analysis under a spiked model

With linear encoders $f_X(\mathbf{x}) = \mathbf{W}\mathbf{x}$, $f_Y(\mathbf{y}) = \mathbf{V}\mathbf{y}$ for CA, and linear encoder $f_X(\mathbf{x}) = \mathbf{E}\mathbf{x}$ and decoder $f_D(\mathbf{z}) = \mathbf{D}\mathbf{z}$ for CP, both objectives admit closed-form solutions expressible through the SVDs of two modality-coupling matrices:

$$\mathbf{C} := \mathbf{\Sigma}_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1/2}, \qquad \mathbf{A} := \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1/2}, \tag{3}$$

where $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{yy}$, $\mathbf{\Sigma}_{xy}$ denote the population (cross-)covariances; throughout, $\mathbf{x}$ and $\mathbf{y}$ are zero-mean, so second moments coincide with covariances, and $\hat{\mathbf{\Sigma}}$ denotes sample estimates (Algorithm 1). CA projects onto the leading k singular directions of $\mathbf{C}$ (symmetric whitening, equivalent to CCA [21, 2]); CP projects onto the leading k singular directions of $\mathbf{A}$ (source-side whitening only, equivalent to truncated reduced-rank regression (RRR) [23, 17]). Although CCA and RRR are classically connected (e.g., [16]), the two paradigms diverge once structured cross-modal nuisance is present, which is the regime our analysis characterizes.

Spiked model. To analyze recovery, we posit a signal-plus-noise model in which each modality decomposes into k shared signal coordinates and $d - k$ modality-specific nuisance coordinates. In suitable orthogonal bases, the covariances are block diagonal:

$$\mathbf{\Sigma}_{xx} = \operatorname{diag}\big(\mathbf{K}^2 + \mathbf{\Gamma}_x^{(s)},\, \mathbf{\Gamma}_x^{(n)}\big), \quad \mathbf{\Sigma}_{yy} = \operatorname{diag}\big(\mathbf{K}^2 + \mathbf{\Gamma}_y^{(s)},\, \mathbf{\Gamma}_y^{(n)}\big), \quad \mathbf{\Sigma}_{xy} = \operatorname{diag}\big(\mathbf{K}^2,\, \mathbf{\Gamma}_{xy}\big), \tag{4}$$

where $\mathbf{K} = \operatorname{diag}(\kappa_1, \ldots, \kappa_k)$ collects the cross-modal signal strengths, $\mathbf{\Gamma}_x^{(s)}, \mathbf{\Gamma}_y^{(s)}$ are the view-specific noise variances on the shared coordinates, $\mathbf{\Gamma}_x^{(n)}, \mathbf{\Gamma}_y^{(n)}$ are the view-specific noise variances on the nuisance coordinates, and $\mathbf{\Gamma}_{xy} = \operatorname{diag}(\eta_1, \ldots, \eta_{d-k})$ encodes cross-modal nuisance correlation, with $0 \le \eta_j \le \sqrt{\tilde{\gamma}_j^x \tilde{\gamma}_j^y}$. Full parameterization is given in Section A.

Under this model, the singular values of $\mathbf{C}$ and $\mathbf{A}$ decompose cleanly into signal and nuisance contributions, yielding recovery conditions for each objective.

Proposition 3.1 (CA vs. CP separation)**.** *Under the spiked model equation 4, the singular values of* $\mathbf{C}$ *and* $\mathbf{A}$ *split into signal values* $\{\rho_i, \tau_i\}_{i \in [\![k]\!]}$ *and nuisance values* $\{\nu_j, \xi_j\}_{j \in [\![d-k]\!]}$*, given by*

$$\rho_i = \frac{\kappa_i^2}{\sqrt{(\kappa_i^2 + \gamma_i^x)(\kappa_i^2 + \gamma_i^y)}}, \quad \tau_i = \frac{\kappa_i^2}{\sqrt{\kappa_i^2 + \gamma_i^x}}, \quad \nu_j = \frac{\eta_j}{\sqrt{\tilde{\gamma}_j^x \tilde{\gamma}_j^y}}, \quad \xi_j = \frac{\eta_j}{\sqrt{\tilde{\gamma}_j^x}}. \tag{5}$$

CA (resp. CP) recovers the shared signal subspace whenever its signal singular values exceed its nuisance singular values. Defining the separation ratios

$$\Delta_{\mathrm{CA}} := \frac{\min_i \rho_i}{\max_j \nu_j}, \qquad \Delta_{\mathrm{CP}} := \frac{\min_i \tau_i}{\max_j \xi_j}, \tag{6}$$

full recovery holds when $\Delta_{\mathrm{CA}} > 1$ (for CA) or $\Delta_{\mathrm{CP}} > 1$ (for CP). In the homogeneous case ($\kappa_i \equiv \kappa$, $\gamma_i^y \equiv \gamma^y$, $\tilde{\gamma}_j^y \equiv \tilde{\gamma}^y$), the two ratios satisfy

$$\Delta_{\mathrm{CA}} \;=\; \Delta_{\mathrm{CP}} \cdot \sqrt{\frac{\tilde{\gamma}^y}{\kappa^2 + \gamma^y}}. \tag{7}$$

More generally, $\Delta_{\mathrm{CA}}/\Delta_{\mathrm{CP}}$ is monotonically non-decreasing in each $\tilde{\gamma}_j^y$.

Proposition 3.2 (Partial recovery)**.** *Under the spiked model equation 4, the top-k singular vectors of $\mathbf{C}$ (resp. $\mathbf{A}$) contain at least*

$$r_{\mathrm{CA}} := \big|\{i : \rho_i > \max_j \nu_j\}\big|, \qquad r_{\mathrm{CP}} := \big|\{i : \tau_i > \max_j \xi_j\}\big| \tag{8}$$

shared signal directions.

Full recovery is $r = k$ (Proposition 3.1); $r = 0$ corresponds to complete failure; and the intermediate regime $0 < r < k$ is *partial recovery*, in which the learned representation is guaranteed to capture at least r signal directions mixed with nuisance. Partial recovery arises only under heterogeneous signal strengths: in the homogeneous case $\kappa_i \equiv \kappa$, $\gamma_i^x \equiv \gamma^x$, $\gamma_i^y \equiv \gamma^y$ (Figure 2) all signal singular values coincide, so $r \in \{0, k\}$ and partial recovery collapses to the binary regime of Proposition 3.1. As κ_i become increasingly heterogeneous, the four-region phase diagram of figure 2 smears into a graded continuum (Figure 7, Section D). In particular, the Neither regime ($\Delta_{\mathrm{CA}}, \Delta_{\mathrm{CP}} \leq 1$) means that full recovery fails for both objectives, not that the signal is lost: it necessarily coincides with complete failure ($r = 0$) only in the homogeneous case.

Failure modes. Equation (5) exposes a fundamental asymmetry. CA's nuisance values ν_j are *cross-modal correlation coefficients* in $[0, 1]$, independent of nuisance variance: when any $\nu_j \to 1$, no signal direction can match it under whitening (since $\rho_i < 1$ whenever any modality-specific noise is present), so $\Delta_{\mathrm{CA}} < 1$ regardless of signal strength. CP's nuisance values $\xi_j = \nu_j\sqrt{\tilde{\gamma}_j^y}$ depend on the target nuisance variance: large $\tilde{\gamma}_j^y$ amplifies even moderate nuisance correlation into a recovery-breaking singular value (here and in Figure 2, we hold ν_j fixed when varying $\tilde{\gamma}_j^y$, consistent with the (κ, ν) axes of the phase diagram). CP is also asymmetric: swapping source and target replaces $\tilde{\gamma}_j^y$ with $\tilde{\gamma}_j^x$, so the direction of prediction matters. A direct corollary, demonstrated in Figure 9, is that CP can achieve lower reconstruction MSE than CA while recovering the wrong subspace: the MSE objective itself does not distinguish signal from nuisance. The relation of equation 7 makes the complementarity explicit: the ratio $\Delta_{\mathrm{CA}}/\Delta_{\mathrm{CP}}$ grows with $\tilde{\gamma}_j^y$, so a large target-side nuisance favors CA, while a small target-side nuisance favors CP. The same relation predicts when latent-space prediction helps: regressing onto a learned target representation rather than the raw target, as in data2vec or JEPA-style predictors, suppresses target-side nuisance variance, shrinking $\tilde{\gamma}_j^y$ and enlarging the region where CP recovers. Masked prediction, which changes which target coordinates are scored rather than their variance, is outside the model. In the resulting phase diagram (Figure 2) the (κ, ν) plane partitions into four regions — *Both*, *CA only*, *CP only*, *Neither* — with boundaries set by $\Delta_{\mathrm{CA}} = 1$ and $\Delta_{\mathrm{CP}} = 1$.

Key takeaway. These failure modes suggest that CA is preferable when nuisance correlation is moderate, modality quality is uncertain, and modality-specific noise is large. CP may be preferable, with the correct source-target orientation, when the signal is strong and the target noise is weak.

Redundant vs. Complementary Modalities. The failure modes of CA and CP can be understood through a distinction between two regimes of multimodal data. *Redundant* modalities, such as image–caption pairs, where captions are written to describe image content, share dominant structure across views, corresponding to large κ_i and small ν_j in our model, or the lower right corner in Figure 2. In this regime, the redundancy assumption underlying standard multimodal SSL holds. *Complementary* modalities, by contrast, arise when each view provides a structurally distinct perspective on the same object, as in multi-sensor measurements in astrophysics and earth science, or multi-omics or multi-scale profiling in biology. Here, view-specific structure dominates the variance of each modality while cross-modal nuisance correlation is non-negligible, pushing toward

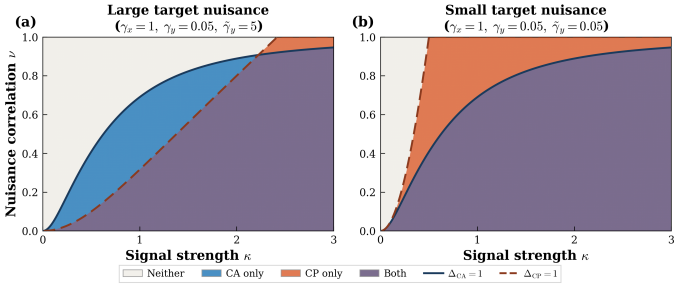


Figure 2: Phase diagram for signal recovery in (κ, ν) space under the homogeneous model (all signal and noise components are equal). Solid and dashed lines respectively show the $\Delta_{\mathrm{CA}} = 1$ and $\Delta_{\mathrm{CP}} = 1$ boundaries from Proposition 3.1. (a) Large target nuisance ($\tilde{\gamma}^y \gg \gamma^y$). (b) Small target noise ($\tilde{\gamma}^y \sim \gamma^y$). Phase diagrams for the non-homogeneous case with partial recoveries are shown in Figure 7.

large ν_j and small effective κ_i, precisely the Neither region (upper left corner) of Figure 2 where both paradigms fail. The intermediate region, which is represented by the center area in Figure 2 is noise dependent - CA is preferred when target noise is large, and CP is preferred when target noise is weak.

4 Experiments

To test the validity of our theory in real-world problems, we conducted experiments across varying levels of complexity and controllability. The first is a linear experiment that directly implements the spiked model, followed by non-linear experiments with synthetic vision datasets, simulating multi-modality by projecting the scene onto two virtual cameras. Next, we use MS-COCO, a real image-caption multi-modality dataset. Finally, we verify our predictions on two real scientific problems: an astrophysics multimodal experiment, and a single-cell multi-omics pair. In all non-linear experiments, we use the VICReg approach [8] as the CA objective, and MSE reconstruction as CP. VICReg's variance and covariance terms are the soft form of the constraint in equation 1: the variance term keeps each coordinate of a view from collapsing, and the covariance term decorrelates coordinates within a view, together penalizing departures from $\frac{1}{n}\sum_i \bar{\mathbf{z}}^{(i)}\bar{\mathbf{z}}^{(i)\top} = \mathbf{I}_k$, while the invariance term is the alignment loss itself. VICReg was shown to be a formulation of DeepCCA [2], the non-linear version of CCA [see e.g., 12], and we compare the two directly in our two synthetic experiments (see Section D). We provide implementation details of all experiments in Section B.

4.1 Spiked Synthetic Data

We first verify the theoretical predictions of Section 3 using closed-form solvers on finite-sample, synthetic data drawn from the spiked covariance model. This confirmed that cross-modal noise correlation breaks CP well before CA (Figure 6). The empirical subspace distance shows that CP fails to recover the signal (dist $\to 1$) at $\nu \approx 0.15$, while CA maintains near-perfect recovery until $\nu \approx 0.75$ (Figure 6, left). Based on the theoretical separation ratios (Figure 6, right), Δ_{CP} crosses the recovery threshold $\Delta = 1$ at a noise correlation roughly $5\times$ lower than Δ_{CA}. The wide gap between the two thresholds validates the complementary failure modes identified in Section 3. Replacing the CA solver with an InfoNCE-trained linear encoder, as in CLIP-style training, does not change this picture: InfoNCE fails only at similarly high nuisance correlation, far from CP (Section C.1), since negatives break pairing across samples while nuisance correlation is a within-pair property. We provide additional linear experiments in Section D.

4.2 Stereo vision experiments

Stereo-dSprites We next tested whether the complementary failure modes persist in a nonlinear setting using *Stereo-dSprites*, a synthetic stereo-vision benchmark, based on the dSprites dataset ([31]), with controlled nuisance alignment that simulates a multimodal visual setting. Two virtual cameras observe a shared 2D object on a 64×64 image. Here, shape is the signal (low pixel variance but perfectly correlated across views) whereas world position is the nuisance (high pixel variance and highly, but imperfectly, correlated, with the correlation controlled by a camera jitter

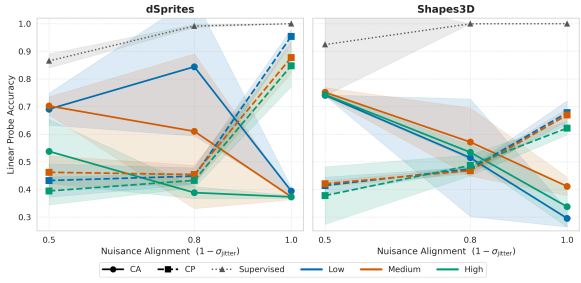


Figure 3: Linear probe accuracy vs. nuisance alignment $\nu_{\max} = 1 - \sigma_{\text{jitter}}$. *(Left)* Stereo-dSprites (3-class, grayscale, 100k samples). *(Right)* Stereo-3DShapes (4-class, RGB, 100k samples). In both settings, color represents weak noise levels. In both panels, the trade-off between the methods is clearly seen. CA (solid, circles) peaks at moderate-to-low alignment and collapses at full alignment; CP (dashed, squares) shows the opposite pattern. The crossover at $\nu_{\max} \approx 0.8$ is consistent across datasets. Lower absolute ceilings in 3DShapes reflect the harder discrimination task.

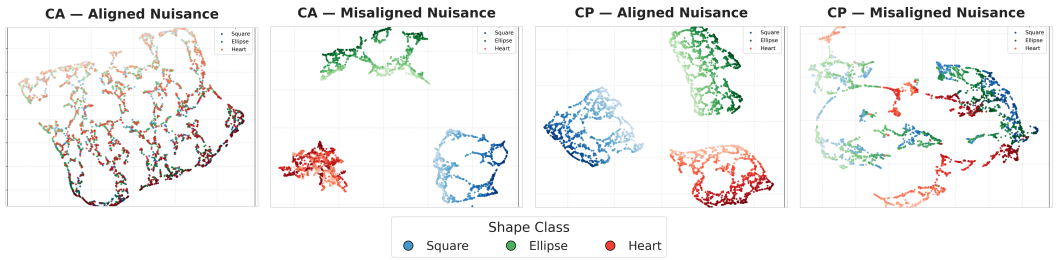


Figure 4: UMAP embeddings of learned representations of the stereo-dSprites experiment (color = shape, intensity = position). From left to right: CA with aligned noise ($\sigma_{\text{jitter}} = 0$), CA with misaligned noise ($\sigma_{\text{jitter}} = 0.5$), CP with aligned noise, and CP with misaligned noise. All experiments have the same modality-specific noise ($\sigma_{\text{noise}} = 0.5$). Each method succeeds exactly where the other fails, and on failures, the models learn the nuisance.

parameter σ_{jitter}). We define nuisance alignment as $\nu_{\max} = 1 - \sigma_{\text{jitter}}$, and evaluate downstream shape classification via linear probe (details in Section B).

Stereo-3DShapes We replicate the Stereo-dSprites protocol on *Stereo-3DShapes* (based on [10]): RGB 64×64 stereo pairs of 3D-rendered objects (Cube, Cylinder, Sphere, Capsule) with controlled position jitter (details in Section B).

Both experiments shows the expected trade-off between objectives - CA fails with perfectly correlated noise and improves as the alignment decreases, while CP shows the opposite behavior (Figure 3). At full nuisance alignment ($\nu = 1$), the cross-modal mapping is deterministic and CP's overcapacity bottleneck encodes both signal and nuisance without compression pressure, circumventing the theory's failure prediction. As soon as jitter breaks determinism ($\nu < 1$), CP is forced to compress, and the separation ratio governs recovery. These overall conclusions are also observed when we examine examples of latent features with different alignment values for CA and CP in the stereo-dSprites experiment with dimensionality reduction using UMAP [32] ((Figure 4), colors: signal (shape); opacity: nuisance features (position)).

4.3 Image–Caption Experiments

We extended the analysis to a real image–caption dataset: MS-COCO [27], where captions are written to describe natural image content. As this is a true multimodal dataset, we used the natural caption–image pairing without artificial nuisance manipulation. We trained encoders from scratch (ResNet-18 for images, two-layer Transformer for captions) and varied modality-specific noise by applying visual style transforms of increasing strength to the image modality while keeping captions clean. As expected, we observed asymmetric performance for CP (Figure 5); $\text{CP}_{I \to T}$ strongly dominates at low noise ($\sim$10 pp over CA), and degrades monotonically as image noise increases. The asymmetry matches the linear theory: the target covariance $\mathbf{\Sigma}_{yy}$ never enters $\mathbf{A}$, so image-side perturbations affect $\text{CP}_{I \to T}$ but leave $\text{CP}_{T \to I}$ (text source, unchanged) flat. The slight rise of

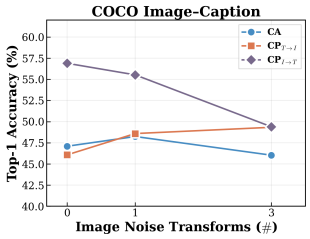


Figure 5: Top-1 accuracy vs. image style transform strength for MS-COCO experiment. CP shows an asymmetric nature: prediction of image from text results in similar performance as CA but prediction of text from an image results in much better performance. Both approaches converge to the same accuracy when image noise is high. Single run per condition.

$\mathrm{CP}_{T\to I}$ is consistent with noisy reconstruction targets acting as a regularizer, a finite-capacity effect outside the linear theory. CA is insensitive throughout, consistent with $\mathbf{\Sigma}_{yy}$ normalization absorbing modality-specific variance.

4.4 Predicting Recovery Regimes

We propose a lightweight supervised diagnostic that locates a paired dataset in the phase diagram before committing to cross-modal training. The diagnostic is supervised by design: a small labeled subsample classifies singular components of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ as signal or nuisance, allowing direct estimation of $\hat{\Delta_{\mathrm{CA}}}$ and $\hat{\Delta_{\mathrm{CP}}}$ without recovering the latent parameters $(\kappa, \gamma, \tilde{\gamma}, \eta)$. The labeled budget is small relative to the scale of cross-modal training. The diagnostic is meant to inform, and the labels need not match the downstream target: in Section 4.5, regime prediction from $\log g$ correctly orders methods on binarity and age. Under subsampling down to a few hundred labeled pairs, most draws reproduce the full-sample regime and the recovery count is stable throughout (Section C.2). The predicted regime is also unchanged under label noise, partial reassignment, and class coarsening (Section C.3).

The estimation procedure is well-posed when the representation fed to the estimator approximately satisfies the linear spiked decomposition of Section 3.2: signal and nuisance directions must be meaningfully separable in the joint covariance structure of the paired data. This condition holds most directly in *two-stage* multimodal pipelines, where each modality is first encoded independently by a unimodal model, and the cross-modal objective operates on these frozen representations. Two-stage pipelines are the dominant paradigm in scientific multimodal learning and in large foundation-model stacks where modality-specific encoders are pretrained independently. In *single-stage* pipelines, the encoder and cross-modal objective are optimized jointly from raw data, and no unimodal representation satisfying the spiked decomposition exists prior to training. Applying the estimator to compact proxies of the raw inputs works when those inputs are already low-dimensional and informative, as for the protein modality in Section 4.6.

4.5 Real Astrophysical Data (LAMOST $\times$ Kepler/TESS)

Finally, we validated regime estimation on real astrophysical data, pairing ground-based LAMOST ([44]) spectra (2048-dim encoder) with space-based photometry from two instruments: Kepler ([30]) and TESS ([38]) (1024-dim encoders), using frozen pretrained features [24] with lightweight projection/prediction heads (Section B). We estimated recovery regimes from a labeled subsample using surface gravity ($\log g$), following Algorithm 1. Downstream evaluation covered three physically distinct targets: binarity, $\log g$, and age. Binarity and age are encoded through modality-specific mechanisms (spectroscopic radial-velocity variations *vs.* photometric eclipses for binarity; isochrone *vs.* gyrochronology for age), so agreement between a $\log g$-based regime prediction and behavior on binarity and age tests whether the separation ratios capture the geometry of the modality pair rather than a task-specific artifact. The results show that the regime predictions hold as *regime-level* statements across all targets, with a revealing asymmetry in how each regime manifests (Table 1, 5 seeds per cell).

Kepler (Both). On every target, at least one of {CA, CP} matches or exceeds the best unimodal baseline, but the winner rotates with task. CP stays at the LAMOST ceiling where LAMOST dominates ($\log g$, binarity); CA captures photometric signal where LAMOST is weak (age: $+0.19$ R^2 over LAMOST). The 'Both' prediction should be read as *some cross-modal objective helps on every task*, not as both objectives helping uniformly, consistent with $\hat{\Delta_{\mathrm{CA}}}$ and $\hat{\Delta_{\mathrm{CP}}}$ being very different.

TESS (Neither). No cross-modal method exceeds LAMOST alone on any target, and the gap is well beyond seed variance. Moreover, we have $\hat{r}_{\mathrm{CA}} = \hat{r}_{\mathrm{CP}} = 0$, the complete-failure form of the Neither regime. Figure 12 visualizes the underlying singular-value decompositions of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ for both pairs.

CP direction asymmetry. CP's recovery condition Δ_{CP} never involves the target covariance $\mathbf{\Sigma}_{yy}$, which does not enter $\mathbf{A}$, so swapping source and target produces a structurally different recovery condition: the preferred direction is the one in which the source modality more directly encodes the task signal. For $\log g$ and binarity, spectra encode the signal more directly than photometry (through absorption-line broadening rather than granulation/oscillation) so CP forward (spectra $\rightarrow$ photometry) succeeds while $\mathrm{CP}_{\mathrm{rev}}$ (photometry $\rightarrow$ spectra) fails. For age, photometric rotation periods provide a more direct signal via gyrochronology than spectroscopic activity indicators, so the preferred direction reverses: $\mathrm{CP}_{\mathrm{rev}}$ on age ($R^2 = 0.497$) outperforms forward CP.

Table 1: Astrophysical cross-modal results, mean $\pm$ std over 5 seeds. Best per row in **bold** (ties within seed std co-bolded). *Kepler (Both, $\hat{\Delta}_{\mathrm{CA}} = 1.08$, $\hat{\Delta}_{\mathrm{CP}} = 2.29$):* at least one cross-modal method matches or beats the best unimodal baseline on every target; CP preserves LAMOST's ceiling where LAMOST dominates, CA captures photometric signal where LAMOST is weak. *TESS (Neither, $\hat{r}_{\mathrm{CA}} = \hat{r}_{\mathrm{CP}} = 0$):* no cross-modal method beats LAMOST-only on any target. $\mathrm{CP}_{\mathrm{rev}}$ (photometry $\rightarrow$ spectra) fails on tasks where spectra carry the more direct signal ($\log g$, binarity), but outperforms forward CP on age, where photometric rotation provides a more direct gyrochronological signal — the same source-quality principle in both directions.

Task	CA	CP	$\mathrm{CP}_{\mathrm{rev}}$	LAMOST	Photometry
LAMOST $\times$ Kepler — Both regime					
Binarity (bal. acc.)	0.802 ± 0.009	$\mathbf{0.814 \pm 0.006}$	0.751 ± 0.004	$\mathbf{0.814}$	0.731
$\log g$ (R^2)	0.956 ± 0.003	$\mathbf{0.976 \pm 0.001}$	0.639 ± 0.004	$\mathbf{0.977}$	0.542
Age (R^2)	$\mathbf{0.620 \pm 0.001}$	0.434 ± 0.006	0.497 ± 0.039	0.431	0.470
LAMOST $\times$ TESS — Neither regime					
Binarity (bal. acc.)	0.756 ± 0.022	0.763 ± 0.011	0.626 ± 0.010	$\mathbf{0.779}$	0.604
$\log g$ (R^2)	0.929 ± 0.005	0.939 ± 0.001	-0.312 ± 0.001	$\mathbf{0.942}$	-0.312
Age (R^2)	0.431 ± 0.029	0.396 ± 0.064	-0.072 ± 0.004	$\mathbf{0.503}$	-0.037

4.6 Single-Cell Multi-Omics (BMMC CITE-seq)

We apply the same protocol to a single-cell multi-omics pair with a structurally different source of correlated nuisance. BMMC CITE-seq measures RNA expression and surface-protein abundance in the same 90k bone-marrow mononuclear cells [29], with 44 annotated cell types as the downstream target; donor, site, and batch effects contaminate both assays of a cell and are therefore shared across the pair. RNA is encoded by a frozen scGPT model [15] and the protein modality enters as CLR-normalized counts (Section B).

The diagnostic predicts *Neither* in all five seeds ($\hat{\Delta}_{\mathrm{CA}} = 0.52 \pm 0.13$, $\hat{\Delta}_{\mathrm{CP}} = 0.33 \pm 0.10$), but unlike TESS the recovery counts are non-zero ($0 < \hat{r} < \hat{k}$ in both spectra in every seed): partial recovery in the sense of Proposition 3.2. The downstream results track this reading (Table 2): the predicted ordering $\mathrm{CA} > \mathrm{CP} > \mathrm{CP}_{\mathrm{rev}}$ holds in every seed, CP falls below the RNA-only baseline in $4/5$ seeds, consistent with the variance trap, and CA exceeds it by less than its own seed-to-seed spread. The two Neither cases thus separate the regime's two manifestations: $\hat{r} = 0$ is complete failure, where the stronger modality alone is the best available representation, while $0 < \hat{r} < \hat{k}$ leaves a partially recovered subspace whose downstream value the separation ratios do not by themselves determine. The nuisance floor is set by donor, site, and batch structure, and removing the dominant batch direction promotes the next rather than lifting $\hat{\Delta}_{\mathrm{CA}}$ above one (Section C.4).

Key takeaway. Estimating effective recovery regimes is a practical and feasible analysis for real-world multimodal problems. It can indicate whether cross-modal learning is likely to succeed, identify which modality is the informative bottleneck, and guide the choice of objective.

Table 2: BMMC CITE-seq, 44-class cell-type balanced accuracy, mean $\pm$ std over 5 seeds. Last column: within-seed difference from RNA-only, with the number of seeds in which it is positive. Predicted regime: Neither with partial recovery.

Representation	Balanced accuracy	Δ vs. RNA-only	
RNA only (strong)	0.584 ± 0.043	—	
ADT only (weak)	0.558 ± 0.040	—	
CA	0.606 ± 0.030	$+0.023 \pm 0.026$	$(4/5)$
CP	0.572 ± 0.034	-0.011 ± 0.011	$(1/5)$
$\mathrm{CP}_{\mathrm{rev}}$	0.524 ± 0.048	-0.060 ± 0.024	$(0/5)$

5 Conclusion

We studied cross-modal alignment and cross-modal prediction in a unified linear framework, deriving recovery conditions governed by separation ratios Δ_{CA} and Δ_{CP} that partition multimodal problems into four regimes, refined by the recovery count r. A data-driven estimation of these quantities identifies the preferred objective and prediction direction from a small labeled subsample, before any cross-modal training, and identifies the case $r = 0$ in which no objective in the CA/CP family recovers shared signal above the nuisance floor. Experiments span synthetic data, stereo-vision benchmarks, and image–caption pairs, with the sharpest validation on real scientific data: the same spectroscopic encoder paired with two photometric instruments of differing quality yields two distinct predicted regimes, both confirmed across multiple downstream targets, and the predicted CP direction asymmetry is confirmed on all tasks. A single-cell multi-omics pair, where the shared nuisance is donor and batch structure rather than instrument systematics, reproduces the Neither prediction in a structurally different domain.

The Neither regime is the most important open problem raised by this work, the natural habitat of complementary scientific modalities, where each view is structurally distinct yet the shared signal does not clear the nuisance floor. Our two instances show its two faces: complete failure ($r = 0$), where the stronger modality alone is the best available representation, and partial recovery ($0 < r < k$), where some signal is recovered but full recovery fails and the downstream value of the recovered subspace is not settled by the separation ratios. Escaping it likely requires objectives that go beyond pairwise cross-covariance: higher-order structure, auxiliary supervision, or modality-specific priors, rather than more expressive fusion heads, which can at best learn to rely on the stronger modality. We hope the phase diagram introduced here provides a principled starting point to solve it.

Limitations. The recovery conditions are derived for linear encoders and decoders at the population level; the nonlinear experiments test whether the mechanisms persist but do not extend the guarantees, and finite-sample thresholds remain open. The spiked model assumes orthogonal signal and nuisance blocks, which real data satisfy only approximately, and the stereo benchmarks control nuisance through camera jitter, a simplified proxy. The diagnostic requires a small labeled subsample and is most reliable in two-stage pipelines.

Acknowledgments and Disclosure of Funding

IK is supported by the Azrieli Graduate Studies Fellowship. AR is an employee of Genentech and has equity in Roche. HV is an employee of Genentech

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza

Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv e-prints*, page arXiv:2204.14198, April 2022.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1247–III–1255. JMLR.org, 2013.

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

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023.

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *arXiv e-prints*, page arXiv:2202.03555, February 2022.

[6] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *ArXiv*, abs/2205.11508, 2022.

[7] Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics. *arXiv e-prints*, page arXiv:2511.08544, November 2025.

[8] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv e-prints*, page arXiv:2105.04906, May 2021.

[9] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, May 2025.

[10] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3d-shapes, 2018.

[11] Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning*, pages 3252–3298, 2023.

[12] James Chapman, Lennie Wells, and Ana Lawry Aguila. Unconstrained Stochastic CCA: Unifying Multiview and Self-Supervised Learning. *arXiv e-prints*, page arXiv:2310.01012, October 2023.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.

[14] Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, Julio Saez-Rodriguez, Simona Cristea, Hani Goodarzi, Mohammad Lotfollahi, Fabian J. Theis, and Bo Wang. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, April 2025.

[15] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024.

[16] Claire Donnat and Elena Tuzhilina. Canonical Correlation Analysis as Reduced Rank Regression in High Dimensions. *arXiv e-prints*, page arXiv:2405.19539, May 2024.

[17] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

[18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023.

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

[20] KaiMing He, Xinlei Chen, Saining Xie, YangHao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

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

[22] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *arXiv e-prints*, page arXiv:2106.04538, 2021.

[23] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

[24] Ilay Kamai, Alex M. Bronstein, and Hagai B. Perets. Machine learning inference of stellar properties using integrated photometric and spectroscopic data. *The Astrophysical Journal*, 994, 2025.

[25] Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27, 2022.

[26] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. *arXiv e-prints*, arXiv:2203.02053, March 2022.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Zhou Lu. A theory of multimodal learning. *Neural Information Processing Systems*, arXiv:2309.12458, 2023.

[29] Malte Luecken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, BONY DE KUMAR, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Pisco, Smita Krishnaswamy, Fabian Theis, and Jonathan M Bloom. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[30] Savita Mathur, Daniel Huber, Natalie M. Batalha, David R. Ciardi, Fabienne A. Bastien, Allyson Bieryla, Lars A. Buchhave, William D. Cochran, Michael Endl, Gilbert A. Esquerdo, Elise Furlan, Andrew Howard, Steve B. Howell, Howard Isaacson, David W. Latham, Phillip J. MacQueen, and David R. Silva. Revised stellar properties of kepler targets for the q1-17 (dr25) transit detection run. *The Astrophysical Journal Supplement Series*, 229(2):30, mar 2017.

[31] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[32] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, arXiv:1802.03426, February 2018.

[33] Pierre Mergny and Lenka Zdeborová. Spectral thresholds in correlated spiked models and fundamental limits of partial least squares. *arXiv preprint arXiv:2510.17561*, 2025.

[34] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, January 1960.

[35] L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306, December 1975.

[36] Liam Holden Parker, François Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Géraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Aion-1: Omnimodal foundation model for astronomical sciences. *Advances in Neural Information Processing Systems 38*, 2025.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv e-prints*, page arXiv:2103.00020, February 2021.

[38] George R. Ricker, Joshua N. Winn, Roland Vanderspek, David W. Latham, Gáspár Á. Bakos, Jacob L. Bean, Zachory K. Berta-Thompson, Timothy M. Brown, Lars Buchhave, Nathaniel R. Butler, R. Paul Butler, William J. Chaplin, David Charbonneau, Jørgen Christensen-Dalsgaard, Mark Clampin, Drake Deming, John Doty, Nathan De Lee, Courtney Dressing, Edward W. Dunham, Michael Endl, Francois Fressin, Jian Ge, Thomas Henning, Matthew J. Holman, Andrew W. Howard, Shigeru Ida, Jon M. Jenkins, Garrett Jernigan, John Asher Johnson, Lisa Kaltenegger, Nobuyuki Kawai, Hans Kjeldsen, Gregory Laughlin, Alan M. Levine, Douglas Lin, Jack J. Lissauer, Phillip MacQueen, Geoffrey Marcy, Peter R. McCullough, Timothy D. Morton, Norio Narita, Martin Paegert, Enric Palle, Francesco Pepe, Joshua Pepper, Andreas Quirrenbach, Stephen A. Rinehart, Dimitar Sasselov, Bun'ei Sato, Sara Seager, Alessandro Sozzetti, Keivan G. Stassun, Peter Sullivan, Andrew Szentgyorgyi, Guillermo Torres, Stephane Udry, and Joel Villasenor. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003, January 2015.

[39] Arabind Swain, Sean Alexander Ridout, and Ilya Nemenman. Better Together: Cross and Joint Covariances Enhance Signal Detectability in Undersampled Data. *arXiv e-prints*, page arXiv:2507.22207, July 2025.

[40] Hugo Tabanelli, Pierre Mergny, Lenka Zdeborova, and Florent Krzakala. Computational Thresholds in Multi-Modal Learning via the Spiked Matrix-Tensor Model. *arXiv e-prints*, page arXiv:2506.02664, June 2025.

[41] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, 2021.

[42] Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint-embedding vs reconstruction: Provable benefits of latent space prediction for self-supervised learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[43] Meir Yossef Levi and Guy Gilboa. The Double-Ellipsoid Geometry of CLIP. *arXiv e-prints*, page arXiv:2411.14517, November 2024.

[44] Gang Zhao, Yong-Heng Zhao, Yao-Quan Chu, Yi-Peng Jing, and Li-Cai Deng. LAMOST spectral survey — An overview. *Research in Astronomy and Astrophysics*, 12(7):723–734, July 2012.

A Closed-form solutions and spiked model derivations

Both results are classical: Theorem A.1 is the CCA solution [21], and Theorem A.2 is truncated reduced-rank regression [23]. We restate them in our notation, with proofs, because the spiked-model analysis builds on their explicit form.

A.1 Full statement of closed-form solutions

Theorem A.1 (Closed-form solutions for CA)**.** *Assume $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ are positive definite, and let $k \leq \min(d_x, d_y)$. Let $\mathbf{C} = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top$ be the SVD of $\mathbf{C} \coloneqq \mathbf{\Sigma}_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1/2}$ with singular values $\phi_1 \geq \cdots \geq \phi_{\min(d_x,d_y)} \geq 0$. The minimizers of equation 1 with linear encoders are*

$$\mathbf{W}^\star = \mathbf{U}\mathbf{P}_k^\top \mathbf{\Sigma}_{xx}^{-1/2}, \quad \mathbf{V}^\star = \mathbf{U}\mathbf{Q}_k^\top \mathbf{\Sigma}_{yy}^{-1/2}, \tag{9}$$

where $\mathbf{P}_k, \mathbf{Q}_k$ contain the leading k columns and $\mathbf{U} \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix.

Theorem A.2 (Closed-form solutions for CP)**.** *Let $\mathbf{A} = \mathbf{U}_A \operatorname{diag}(\boldsymbol{\sigma})\mathbf{V}_A^\top$ be the SVD of $\mathbf{A} \coloneqq \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1/2}$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. The composed map $\mathbf{B} \coloneqq \mathbf{D}\mathbf{E}$ at a minimizer of equation 2 with linear encoder and decoder is*

$$\mathbf{B}^\star = \mathbf{U}_{A,k}\, \operatorname{diag}(\sigma_1, \ldots, \sigma_k)\, \mathbf{V}_{A,k}^\top\, \mathbf{\Sigma}_{xx}^{-1/2}. \tag{10}$$

The factorization $\mathbf{B}^\star = \mathbf{D}^\star\mathbf{E}^\star$ is non-unique: for any invertible $\mathbf{M} \in \mathbb{R}^{k\times k}$, $(\mathbf{D}^\star\mathbf{M}, \mathbf{M}^{-1}\mathbf{E}^\star)$ yields the same composed map.

A.2 Proof of Theorem A.1

For linear encoders the constraint reads $\mathbf{W}\mathbf{\Sigma}_{xx}\mathbf{W}^\top = \mathbf{V}\mathbf{\Sigma}_{yy}\mathbf{V}^\top = \mathbf{I}_k$, so $\frac{1}{n}\sum_i \|\mathbf{W}\mathbf{x}_i\|^2 = \frac{1}{n}\sum_i \|\mathbf{V}\mathbf{y}_i\|^2 = k$ and the objective reduces to

$$\min_{\mathbf{W},\mathbf{V}} \quad 2k - 2\mathrm{Tr}(\mathbf{W}\mathbf{\Sigma}_{xy}\mathbf{V}^\top) \quad \text{s.t.} \quad \mathbf{W}\mathbf{\Sigma}_{xx}\mathbf{W}^\top = \mathbf{V}\mathbf{\Sigma}_{yy}\mathbf{V}^\top = \mathbf{I}_k\,, \tag{11}$$

that is, to maximizing $\mathrm{Tr}(\mathbf{W}\mathbf{\Sigma}_{xy}\mathbf{V}^\top)$. Let $\mathbf{W}' = \mathbf{W}\mathbf{\Sigma}_{xx}^{1/2}$ and $\mathbf{V}' = \mathbf{V}\mathbf{\Sigma}_{yy}^{1/2}$, and define $\mathbf{C} \coloneqq \mathbf{\Sigma}_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1/2}$. The constraints become $\mathbf{W}'\mathbf{W}'^\top = \mathbf{V}'\mathbf{V}'^\top = \mathbf{I}_k$, so $\mathbf{W}'$ and $\mathbf{V}'$ have orthonormal rows, and the problem becomes

$$\max_{\mathbf{W}',\mathbf{V}'} \quad \mathrm{Tr}(\mathbf{W}'\mathbf{C}\mathbf{V}'^\top) = \langle \mathbf{C}, \mathbf{W}'^\top\mathbf{V}'\rangle \quad \text{s.t.} \quad \mathbf{W}'\mathbf{W}'^\top = \mathbf{V}'\mathbf{V}'^\top = \mathbf{I}_k\,. \tag{12}$$

The matrix $\mathbf{N} \coloneqq \mathbf{W}'^\top\mathbf{V}'$ is a product of two partial isometries, hence $\mathrm{rank}(\mathbf{N}) \leq k$ and all singular values of $\mathbf{N}$ are at most 1. By von Neumann's trace inequality [35],

$$\langle \mathbf{C}, \mathbf{N}\rangle \;\leq\; \sum_{i\leq k} \phi_i\, \sigma_i(\mathbf{N}) \;\leq\; \sum_{i\leq k} \phi_i\,. \tag{13}$$

Equality is attained by $\mathbf{W}' = \mathbf{U}\mathbf{P}_k^\top$ and $\mathbf{V}' = \mathbf{U}\mathbf{Q}_k^\top$ with $\mathbf{U}$ orthogonal, which satisfy both constraints and give $\mathbf{N} = \mathbf{P}_k\mathbf{Q}_k^\top$. Transforming back gives

$$\mathbf{W}^\star = \mathbf{U}\mathbf{P}_k^\top\mathbf{\Sigma}_{xx}^{-1/2}\,, \qquad \mathbf{V}^\star = \mathbf{U}\mathbf{Q}_k^\top\mathbf{\Sigma}_{yy}^{-1/2}\,. \tag{14}$$

A.3 Proof of Theorem A.2

Since $\mathbf{E} \in \mathbb{R}^{k\times d_x}$ and $\mathbf{D} \in \mathbb{R}^{d_y\times k}$, the composed map $\mathbf{B} = \mathbf{D}\mathbf{E}$ satisfies $\mathrm{rank}(\mathbf{B}) \leq k$. Conversely, any rank-$k$ matrix $\mathbf{B}$ admits such a factorization, so minimizing over $(\mathbf{D}, \mathbf{E})$ is equivalent to minimizing over rank-k matrices $\mathbf{B}$. Writing the CP objective in terms of $\mathbf{B}$ and expanding:

$$\frac{1}{n}\sum_i \|\mathbf{y}_i - \mathbf{B}\mathbf{x}_i\|^2 = \mathrm{Tr}(\mathbf{\Sigma}_{yy}) - 2\,\mathrm{Tr}(\mathbf{\Sigma}_{yx}\mathbf{B}^\top) + \mathrm{Tr}(\mathbf{B}\mathbf{\Sigma}_{xx}\mathbf{B}^\top)\,. \tag{15}$$

Substituting $\mathbf{B}' \coloneqq \mathbf{B}\boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}$ and $\mathbf{A} \coloneqq \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}$ (so that $\boldsymbol{\Sigma}_{yx}\mathbf{B}^\top = \mathbf{A}(\mathbf{B}')^\top$ and $\mathbf{B}\boldsymbol{\Sigma}_{xx}\mathbf{B}^\top = \mathbf{B}'(\mathbf{B}')^\top$):

$$= \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - 2\operatorname{Tr}(\mathbf{A}(\mathbf{B}')^\top) + \operatorname{Tr}(\mathbf{B}'(\mathbf{B}')^\top) \tag{16}$$

$$= \operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - \operatorname{Tr}(\mathbf{A}\mathbf{A}^\top) + \|\mathbf{B}' - \mathbf{A}\|_F^2 \,. \tag{17}$$

Since $\operatorname{Tr}(\boldsymbol{\Sigma}_{yy}) - \operatorname{Tr}(\mathbf{A}\mathbf{A}^\top)$ does not depend on $\mathbf{B}$, and since $\boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}$ is invertible, minimizing over rank-$\leq k$ matrices $\mathbf{B}$ is equivalent to minimizing $\|\mathbf{B}' - \mathbf{A}\|_F^2$ over rank-$\leq k$ matrices $\mathbf{B}'$. By the Eckart–Young–Mirsky theorem [17, 34], the best rank-k approximation of $\mathbf{A}$ in Frobenius norm is $(\mathbf{B}')^\star = \mathbf{U}_{A,k}\,\operatorname{diag}(\sigma_1, \ldots, \sigma_k)\,\mathbf{V}_{A,k}^\top$. Transforming back via $\mathbf{B}^\star = (\mathbf{B}')^\star\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}$ gives the stated solution.

Non-uniqueness of the factorization. The product $\mathbf{B}^\star$ is uniquely determined (assuming distinct singular values), but the factorization $\mathbf{B}^\star = \mathbf{D}^\star\mathbf{E}^\star$ is not: for any invertible $\mathbf{M} \in \mathbb{R}^{k\times k}$, the pair $(\mathbf{D}^\star\mathbf{M},\, \mathbf{M}^{-1}\mathbf{E}^\star)$ yields the same composed map. Hence $\mathbf{E}^\star$ is determined only up to left-multiplication by an invertible matrix.

A.4 Full parameterization

For simplicity, assume $d_x = d_y = d$ and that there exist orthogonal matrices $\mathbf{Q}_x, \mathbf{Q}_y$ such that

$$\boldsymbol{\Sigma}_{xx} = \mathbf{Q}_x\Lambda_x\mathbf{Q}_x^\top, \qquad \Lambda_x = \operatorname{diag}\big(\mathbf{K}^2 + \boldsymbol{\Gamma}_x^{(s)},\ \boldsymbol{\Gamma}_x^{(n)}\big), \tag{18}$$

$$\boldsymbol{\Sigma}_{yy} = \mathbf{Q}_y\Lambda_y\mathbf{Q}_y^\top, \qquad \Lambda_y = \operatorname{diag}\big(\mathbf{K}^2 + \boldsymbol{\Gamma}_y^{(s)},\ \boldsymbol{\Gamma}_y^{(n)}\big), \tag{19}$$

$$\boldsymbol{\Sigma}_{xy} = \mathbf{Q}_x\Lambda_{xy}\mathbf{Q}_y^\top, \qquad \Lambda_{xy} = \operatorname{diag}\big(\mathbf{K}^2,\ \boldsymbol{\Gamma}_{xy}\big), \tag{20}$$

where $\mathbf{K} = \operatorname{diag}(\kappa_1, \ldots, \kappa_k)$ with $\kappa_1 \geq \cdots \geq \kappa_k > 0$, $\boldsymbol{\Gamma}_x^{(s)} = \operatorname{diag}(\gamma_1^x, \ldots, \gamma_k^x)$, $\boldsymbol{\Gamma}_y^{(s)} = \operatorname{diag}(\gamma_1^y, \ldots, \gamma_k^y)$, $\boldsymbol{\Gamma}_x^{(n)} = \operatorname{diag}(\tilde\gamma_1^x, \ldots, \tilde\gamma_{d-k}^x)$, $\boldsymbol{\Gamma}_y^{(n)} = \operatorname{diag}(\tilde\gamma_1^y, \ldots, \tilde\gamma_{d-k}^y)$, and $\boldsymbol{\Gamma}_{xy} = \operatorname{diag}(\eta_1, \ldots, \eta_{d-k})$ with $0 \leq \eta_j \leq \sqrt{\tilde\gamma_j^x\tilde\gamma_j^y}$.

A.5 Singular-value decompositions

Lemma A.3 (Singular values of $\mathbf{C}$ and $\mathbf{A}$)**.** *Under the spiked model, $\mathbf{C}$ and $\mathbf{A}$ are block diagonal in the bases defined by $\mathbf{Q}_x, \mathbf{Q}_y$. Their singular values are the union of the signal values ρ_i, τ_i and nuisance values ν_j, ξ_j given in equation 5.*

Proof. In the bases $\mathbf{Q}_x, \mathbf{Q}_y$, both $\mathbf{C} = \boldsymbol{\Sigma}_{xx}^{-1/2}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1/2}$ and $\mathbf{A} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1/2}$ are block diagonal with signal and nuisance blocks. Direct computation on each block yields the stated expressions.

Corollary A.4 (Recovery conditions)**.** *If $\min_i \rho_i > \max_j \nu_j$, the top-$k$ singular vectors of $\mathbf{C}$ align with the shared signal block, so the CA solution from Theorem A.1 recovers the shared signal subspace (up to rotation) and discards modality-specific noise. The analogous statement holds for $\mathbf{A}$ and CP with τ_i, ξ_j.*

A.6 Proof of Proposition 3.1

The singular-value expressions follow from Lemma A.3. For the ratio identity, we consider the homogeneous case ($\kappa_i \equiv \kappa$, $\gamma_i^x \equiv \gamma^x$, $\gamma_i^y \equiv \gamma^y$, $\tilde\gamma_j^x \equiv \tilde\gamma^x$, $\tilde\gamma_j^y \equiv \tilde\gamma^y$, $\eta_j \equiv \eta$), in which all ρ_i collapse to a single value ρ, and likewise $\tau_i \equiv \tau$, $\nu_j \equiv \nu$, $\xi_j \equiv \xi$. Substituting equation 5 into $\Delta_{\mathrm{CA}}/\Delta_{\mathrm{CP}} = (\rho/\nu)/(\tau/\xi)$ yields

$$\frac{\Delta_{\mathrm{CA}}}{\Delta_{\mathrm{CP}}} = \sqrt{\frac{\tilde\gamma^y}{\kappa^2 + \gamma^y}}. \tag{21}$$

In the heterogeneous case, the same substitution yields the upper bound

$$\frac{\Delta_{\mathrm{CA}}}{\Delta_{\mathrm{CP}}} \leq \sqrt{\frac{\max_j \tilde\gamma_j^y}{\min_i(\kappa_i^2 + \gamma_i^y)}}, \tag{22}$$

since the indices achieving $\min_i \rho_i$ and $\min_i \tau_i$ (and similarly the nuisance maxima) need not coincide. Monotonicity of $\Delta_{\mathrm{CA}}/\Delta_{\mathrm{CP}}$ in each $\tilde{\gamma}_j^y$ follows from Δ_{CP} being invariant in $\tilde{\gamma}_j^y$ (since $\xi_j = \eta_j/\sqrt{\tilde{\gamma}_j^x}$ does not depend on target nuisance variance) and Δ_{CA} being non-increasing in ν_j, with ν_j non-increasing in $\tilde{\gamma}_j^y$.

A.7 Proof of Proposition 3.2

We prove the CA case; the CP case is identical with ρ_i, ν_j replaced by τ_i, ξ_j and $\mathbf{C}$ replaced by $\mathbf{A}$.

By Lemma A.3, the singular values of $\mathbf{C}$ are the union $\{\rho_i\}_{i\in[\![k]\!]} \cup \{\nu_j\}_{j\in[\![d-k]\!]}$, with each ρ_i corresponding to a singular vector in the signal block and each ν_j to a singular vector in the nuisance block.

Let i be any signal index with $\rho_i > \max_j \nu_j$. Then ρ_i exceeds every nuisance singular value, so at most $k-1$ values of $\mathbf{C}$ (the other signal values) can exceed ρ_i, placing ρ_i among the top k singular values. Since this holds for each of the r_{CA} signal indices with $\rho_i > \max_j \nu_j$, the top-k singular vectors contain at least r_{CA} vectors from the signal block.

B Implementation Details

Linear experiments. All linear experiments use closed-form solvers on population covariances drawn from the spiked model of Section 3. We set $d = 20$, $k = 3$ shared dimensions with signal strengths $\kappa = (3.0, 2.0, 1.5)$, and consider a regime with clean signal in the target modality ($\gamma_y^{(s)} = 0.05$) but large target nuisance variance ($\gamma_y^{(n)} = 50.0$), with source-side parameters $\gamma_x^{(s)} = 0.5$, $\gamma_x^{(n)} = 1.0$. We sweep the normalized noise correlation $\nu = \eta/\sqrt{\gamma_x^{(n)}\gamma_y^{(n)}} \in [0, 0.95]$ and report subspace distance (averaged over 20 random rotations) and theoretical separation ratios. Subspace distances are computed as $\|\mathbf{P}_{\hat{U}} - \mathbf{P}_U\|_F/\sqrt{2k}$ where $\mathbf{P}$ denotes the orthogonal projector, and averaged over 20 random rotations of the signal/noise bases. Here $\mathbf{P}_U$ is the orthogonal projector onto the column span of U, so the distance is basis-invariant, lies in $[0, 1]$, and applies equally to CA and CP, whose recovered subspaces both live in $\mathbb{R}^{d_x}$. No optimization is involved; the solvers compute exact CA (CCA) and CP (truncated reduced-rank regression) solutions from the covariance matrices on a single CPU.

Stereo-dSprites. Two virtual cameras observe a shared 2D object (Square, Ellipse, or Heart) on a 64×64 grayscale canvas. World position $P_{\text{world}} \in [-0.5, 0.5]^2$ serves as the aligned nuisance; camera jitter $\sigma_{\text{jitter}} \in \{0.0, 0.05, 0.2, 0.5\}$ controls de-alignment via per-view translation and rotation. View X receives Gaussian pixel noise $\sigma_{\text{strong}} = 0.1$; View Y receives $\sigma_{\text{weak}} \in \{0.2, 0.5, 0.9\}$. Each modality is encoded by a separate 4-layer CNN ($1 \times 64 \times 64 \to 128$-dim) with ReLU activations. CA uses VICReg ($25\times$ invariance $+25\times$ variance $+1\times$ covariance) on 32-dimensional projections. CP uses MSE reconstruction via a transposed-convolutional decoder. All models are trained with Adam (lr=10^{-3}), batch size 64, for up to 100 epochs with early stopping (patience 5). Downstream evaluation uses a linear probe on frozen encoder features for 3-class shape classification, swept over 9 probe sizes (100 to $\sim$6,000 samples) and averaged over 5 seeds. We sweep $n_{\text{samples}} \in \{10\text{k}, 50\text{k}, 100\text{k}\}$ (see Figure 11 for 10k, 50k results). The entire sweep took approximately 24 hours on one L40S GPU.

Stereo-3DShapes. Built from Google's 3DShapes dataset (480K RGB images). Canonical images, one per shape at fixed hue, scale, and orientation, are rendered into stereo pairs via affine warping with the same jitter and noise protocol as dSprites. The signal is 4-class shape (Cube, Cylinder, Sphere, Capsule); additional nuisance factors include floor hue (10 values), wall hue (10), object hue (10), scale (8), and orientation (15), all fixed per canonical image and shared across views. Encoders are 4-layer CNNs ($3 \times 64 \times 64 \to 128$-dim; channels $3 \to 32 \to 32 \to 64 \to 64$, stride 2) with FC layers $1024 \to 256 \to 128$. Training and evaluation follow the dSprites protocol with $n_{\text{samples}} = 100$k, 10 probe sizes (100 to 10,000), and 3–4 seeds. The entire sweep took approximately 24 hours on one L40S GPU.

MS-COCO image-caption. We pair each COCO 2017 image with its associated caption, using the dominant-object category (largest bounding box, 80 classes) as the downstream label. Images are $3 \times 224 \times 224$ RGB, encoded by a ResNet-18 trained from scratch; captions are word-tokenized to length 64 and encoded by a 2-layer Transformer (4 heads, $d_{\text{embed}} = 256$) followed by mean pooling and a linear projection to 128 dimensions. Neither encoder uses pretrained weights. Nuisance is injected into the image modality: each image is passed through k independent distortion groups (color cast, exposure, contrast, texture, saturation, spatial) drawn uniformly from six groups, with k controlled by a noise level $\ell \in \{0.0, 0.2, 0.5\}$ (expected $k \approx 6\ell$ groups active). Within each active group, a random transform is applied at continuous intensity $t \sim \text{Uniform}(0.3, 1.0)$, so every sample receives a unique pixel-level distortion. Training uses AdamW (lr = 10^{-3}, weight-decay = 10^{-4}) with 5-epoch linear warmup into cosine annealing, batch size 1024 per GPU across 4–6 GPUs via DDP, for 50 epochs. CA uses VICReg on projected embeddings from both encoders (invariance $25\times$, variance $25\times$, covariance $1\times$); $\text{CP}_{I\to T}$ is the image encoder feeding a caption decoder under cross-entropy loss; $\text{CP}_{T\to I}$ is the text encoder feeding a pixel decoder under MSE. Evaluation: each method is evaluated on its source /bottleneck encoder. For CA, the image encoder is probed (a symmetric choice — both encoders are equally optimized under VICReg). For $\text{CP}_{I\to T}$, the image encoder (source). For $\text{CP}_{T\to I}$, the text encoder (source). Each frozen representation is fed to a linear probe trained for 30 epochs on the 80-class label; we report top-1 accuracy. The experiment took approximately 24 hours on 4 RTX6000 GPUs.

Astrophysical cross-modal. We pair LAMOST optical spectra (DR8, resolution $R \sim 1800$, range 3690–9100 Å) with light curves from two photometric surveys — Kepler (DR25, 30-min cadence, ~ 4 year baseline; 94,876 cross-matched observations) and TESS (QLP lightcurves; 821,878 cross-matched stars). Each modality uses its own pretrained unimodal encoder, frozen during cross-modal training. LAMOST spectra are encoded by a '1d ViT' that produces a 2048-dim CLS token; light curves are encoded by a multichannel network with parallel flux and frequency (ACF and FFT/Lomb–Scargle) branches combined through a mixer, producing a 1024-dim mean-pooled embedding. Both encoders were pretrained independently on their respective modalities before any cross-modal training. On these frozen features we train lightweight heads for each method: CA uses two projection MLPs ($2048 \to 512$ and $1024 \to 512$) with a VICReg objective (invariance $25\times$, variance $25\times$, covariance $1\times$); CP uses a cross-predictor MLP with a 512-dim bottleneck and MSE loss. Optimization: AdamW (lr = 10^{-3}, weight decay = 10^{-4}) with cosine annealing, batch size 256, early stopping on validation loss. Evaluation uses a linear probe on the cross-modal representation (concatenated projections for CA, bottleneck activations for CP) against held-out stellar labels — $\log g$, age (regression), and binarity (classification). The cross-modal experiment took approximately 4 hours on one RTX6000 GPU.

BMMC CITE-seq. We pair RNA expression with surface-protein abundance measured in the same cells, using the BMMC CITE-seq dataset (90,243 bone-marrow mononuclear cells [29]), with 44 annotated cell types as the downstream label. RNA is encoded by scGPT [15] in a one-time pass over the 2,000 highly variable genes, giving cached 512-dim embeddings that are never updated. The protein modality enters as the 134 CLR-normalized values shipped with the dataset, passed through a small MLP ($134 \to 256 \to 512 \to 512$) trained jointly with the cross-modal objective; a separate encoder is trained per mode, so the ADT-only baseline is mode-specific and we report the CA-mode value. CA uses projection MLPs onto 256 dimensions per view with a VICReg objective (invariance $25\times$, variance $25\times$, covariance $1\times$); CP uses a cross-predictor MLP with a 512-dim bottleneck and MSE loss, in both directions. Optimization: AdamW (lr = 10^{-3}, weight decay = 10^{-4}) with cosine annealing, batch size 512, 200 epochs, early stopping on validation loss (patience 30). Splits are donor-based (6 train / 1 validation / 2 test donors), averaged over 5 seeds with a different donor assignment per seed. Evaluation uses a linear probe (standardization, class-balanced logistic regression) on the cross-modal representation, concatenated projections for CA, bottleneck activations for CP, reported as balanced accuracy. Regime estimation (Algorithm 1) runs on a seeded subsample of $n = 5{,}000$ cells, on the same representations the objective consumes: scGPT embeddings and raw CLR counts, each centered and reduced to 100 principal components, $\varepsilon = 10^{-6}$; PCA is used only for estimation. Across the five seeds, $\hat{r}_{\mathrm{CA}} = 13.0 \pm 3.3$ of $\hat{k}_{\mathrm{CA}} = 20.4 \pm 0.5$ and $\hat{r}_{\mathrm{CP}} = 12.6 \pm 3.2$ of $\hat{k}_{\mathrm{CP}} = 22.4 \pm 1.0$ signal components clear the nuisance floor. Training all three objectives and their probes took approximately 1.5 hours per seed on one L40S GPU.

We keep the protein branch single-stage because, with 134 measured proteins, CLR counts are already a compact representation. As a check, we repeated the experiment with the ADT encoder initialized from a unimodal VICReg pretraining stage (same architecture, trained on ADT alone with two augmented views of each CLR vector via Gaussian noise $\sigma \sim \mathcal{U}(0, 0.1)$ and random feature dropout $p \sim \mathcal{U}(0, 0.1)$; projection heads discarded), and estimated the regime from that encoder's frozen 512-dim embeddings rather than raw CLR counts. Both ratios decrease ($\hat{\Delta}_{\mathrm{CA}} = 0.34$, $\hat{\Delta}_{\mathrm{CP}} = 0.11$ versus 0.67 and 0.26 at the canonical seed), as expected from a decorrelating unimodal objective, but the regime (Neither in $5/5$ seeds), the ordering $\mathrm{CA} > \mathrm{CP} > \mathrm{CP}_{\mathrm{rev}}$, and the downstream accuracies are unchanged.

Algorithm 1 Recovery Regime Prediction

Require: Paired embeddings $\mathbf{Z}_x \in \mathbb{R}^{n \times d_x}$, $\mathbf{Z}_y \in \mathbb{R}^{n \times d_y}$; labels $\mathbf{Y} \in \mathbb{R}^{n \times L}$
Ensure: $\hat{\Delta}_{\mathrm{CA}}$, $\hat{\Delta}_{\mathrm{CP}}$, predicted regime
{CA: read separation ratio from CCA spectrum}
Compute SVD of $\hat{\mathbf{C}} = (\hat{\mathbf{\Sigma}}_{xx} + \varepsilon \mathbf{I})^{-1/2} \hat{\mathbf{\Sigma}}_{xy} (\hat{\mathbf{\Sigma}}_{yy} + \varepsilon \mathbf{I})^{-1/2}$; obtain ϕ^{CCA}
Classify CCA components as signal/nuisance by elbow detection on per-component R^2
$\hat{\Delta}_{\mathrm{CA}} \leftarrow \min(\phi^{\mathrm{CCA}}[\mathrm{signal}]) \,/\, \max(\phi^{\mathrm{CCA}}[\mathrm{nuisance}])$
{CP: read separation ratio from $\mathbf{A}$-SVD spectrum}
Compute SVD of $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}_{yx} (\hat{\mathbf{\Sigma}}_{xx} + \varepsilon \mathbf{I})^{-1/2}$; obtain $\sigma^{\mathbf{A}}$
Classify $\mathbf{A}$-SVD components as signal/nuisance by elbow detection on per-component R^2
$\hat{\Delta}_{\mathrm{CP}} \leftarrow \min(\sigma^{\mathbf{A}}[\mathrm{signal}]) \,/\, \max(\sigma^{\mathbf{A}}[\mathrm{nuisance}])$
return $(\hat{\Delta}_{\mathrm{CA}}, \hat{\Delta}_{\mathrm{CP}})$ and regime per $\gtrless 1$

Recovery regime prediction. The pipeline takes paired embeddings $\mathbf{Z}_x \in \mathbb{R}^{n \times d_x}$, $\mathbf{Z}_y \in \mathbb{R}^{n \times d_y}$ and labels $\mathbf{Y} \in \mathbb{R}^{n \times L}$ for a labeled subsample. Under the spiked model, $\hat{\Delta}_{\mathrm{CA}}$ and $\hat{\Delta}_{\mathrm{CP}}$ are singular-value ratios of $\mathbf{C}$ and $\mathbf{A}$ respectively: signal and nuisance singular values appear together in each spectrum, distinguished only by which block of the spiked decomposition they belong to. We estimate each Δ directly from the corresponding spectrum, classifying each component as signal or nuisance by its predictive power for the labels. For each decomposition, 5-fold Ridge regression of the component scores against the labels yields a per-component R^2 value (summed across label columns for the classification statistic), and piecewise-linear elbow detection on the sorted R^2 curve identifies the breakpoint between the signal block and the nuisance floor. The labeled subsample need not cover the full training set: in our experiments, $n < 1{,}000$ samples suffice.

C Additional analyses

C.1 InfoNCE in the linear setting

CLIP-style alignment relies on InfoNCE with many negatives rather than an explicit whitening constraint, which raises the question of whether negatives resist cross-modal nuisance correlation. We add an InfoNCE-trained linear encoder (symmetric NT-Xent, linear $\mathbf{W}, \mathbf{V}$, Adam with learning rate 10^{-2}, batch size 1024) to the closed-form CA and CP solvers of Figure 6, in the same spiked model ($d = 20$, $k = 3$, $\kappa = (3.0, 2.0, 1.5)$, $\gamma_x^{(s)} = 0.5$, $\gamma_y^{(s)} = 0.05$, $\gamma_x^{(n)} = 1$, $\gamma_y^{(n)} = 50$, $n = 5{,}000$), and sweep ν over 15 points in $[0, 0.95]$ with 20 seeds per point and temperatures $\tau \in \{0.05, 0.1, 0.5\}$.

InfoNCE is somewhat worse than CA throughout but follows the same failure curve and remains far from CP (Table 3): taking failure as the first grid point where the mean distance exceeds 0.5, CP fails at $\nu \approx 0.14$, InfoNCE at $\nu \approx 0.88$, and CA at $\nu \approx 0.95$. Temperature has negligible effect: across all 15 points the other two temperatures differ from $\tau = 0.1$ by at most 0.014, below the seed standard deviation. Adding negatives therefore does not rescue alignment when nuisance is strongly cross-correlated. The reason is structural: negatives estimate the product of marginals in the InfoNCE denominator and down-weight structure shared across the batch, but correlated nuisance is a property of each matched pair and is independent across pairs, so the negatives cannot distinguish it from signal.

Table 3: Subspace distance to the true signal subspace (lower is better), mean over 20 seeds, at six of the 15 swept values of ν. InfoNCE at $\tau = 0.1$.

ν	0.14	0.41	0.68	0.81	0.88	0.95
CP (RRR)	0.57	1.00	1.00	1.00	1.00	1.00
CA (CCA)	0.05	0.05	0.08	0.14	0.50	0.84
InfoNCE	0.11	0.13	0.21	0.40	0.73	0.97

C.2 Stability under subsampling

For each pair and labeled-subsample size m, we draw 50 subsamples without replacement, re-run the full estimator (PCA to 100 components after subsampling, $\varepsilon = 10^{-6}$), and record the fraction of draws in each regime together with the median and interquartile range of the recovery counts (Table 4). At every size, most draws reproduce the full-sample regime, and the agreement grows with m: TESS and BMMC are Neither in every draw at the largest m, and LAMOST×Kepler, whose $\hat{\Delta}_{\mathrm{CA}} = 1.08$ sits close to the CA boundary, is Both in 58% of draws at $m = 250$ rising to 80% at $m = 750$. The recovery count is markedly more stable than the binary label: its median and interquartile range are 0 in every TESS cell, positive in every BMMC cell, and $\hat{r}_{\mathrm{CA}} = 1$ in every Kepler cell. The distinction between complete failure and partial recovery is therefore preserved at every subsample size, including where the regime label fluctuates.

C.3 Sensitivity to label quality

We degrade the label matrix while holding the estimator fixed (Table 5).

Label noise. Adding Gaussian noise to the $\log g$ label, calibrated so that its feature-R^2 falls to 0.75, 0.50, and 0.25 of the clean value, leaves both astro calls unchanged: Kepler stays Both with $\hat{r}_{\mathrm{CA}} = 1$, TESS stays Neither.

Label corruption and coarsening. Randomly reassigning up to 50% of BMMC cell-type labels, or coarsening the 44 cell types to 8 lineages, leaves BMMC in Neither with $\hat{r} > 0$; reassigning 25% or more reduces $\hat{r}_{\mathrm{CA}}$ from 17 to 10 and $\hat{\Delta}_{\mathrm{CA}}$ from 0.67 to 0.51. Quantizing the continuous $\log g$ label into 8 or 4 bins moves Kepler from Both to CP-only, but not by losing signal: it adds a second, weak component to the CA signal set ($\hat{k}_{\mathrm{CA}}$ from 1 to 2) that sits below the nuisance floor, pulling $\hat{\Delta}_{\mathrm{CA}}$ below one while $\hat{r}_{\mathrm{CA}}$ stays 1 — another instance of the recovery count being the more stable readout.

Unsupervised proxy labels. Replacing the labels by k-means clusters of the weaker modality's raw features ($k \in \{8, 20, 44\}$) does not reproduce the ground truth. On TESS, which has no recoverable

Table 4: Subsampling stability, 50 draws without replacement per cell, $\varepsilon = 10^{-6}$. Regime fractions list all regimes with nonzero share; recovery counts are median [IQR]. Full sample sizes: Kepler 934, TESS 153, BMMC 90,243.

Pair	m	Regime fractions	$\hat{r}_{\mathrm{CA}}$	$\hat{r}_{\mathrm{CP}}$
Kepler	250	Both .58, CP .24, Neither .10, CA .08	1 [1, 1]	2 [1, 2]
	500	Both .78, CP .18, Neither .04	1 [1, 1]	2 [2, 2]
	750	Both .80, CP .20	1 [1, 1]	2 [2, 2]
TESS	80	Neither .98, CA .02	0 [0, 0]	0 [0, 0]
	110	Neither .94, CA .06	0 [0, 0]	0 [0, 0]
	140	Neither 1.00	0 [0, 0]	0 [0, 0]
BMMC	250	Neither .78, CA .14, CP .08	6 [5, 7]	9 [6, 11]
	500	Neither .92, CA .06, CP .02	11 [10, 12]	12 [10, 14]
	1,000	Neither .82, CA .12, CP .06	14 [11, 17]	14 [12, 17]
	5,000	Neither .98, CP .02	17 [11, 17]	11 [11, 11]
	72,000	Neither 1.00	17 [17, 17]	11 [11, 11]

shared signal, every proxy produces a false positive ($\hat{r}_{\mathrm{CA}} \in \{1, 2, 3\}$), two of them with a CP-only regime call; every proxy also miscalls Kepler, and $k = 8$ yields a spurious CP-only call on BMMC. Clustering one modality tracks that modality's dominant variance, which in the complementary regime is nuisance rather than shared signal, so the labeled subsample is a genuine requirement of the diagnostic.

Table 5: Regime and recovery counts $(\hat{r}_{\mathrm{CA}}, \hat{r}_{\mathrm{CP}})$ under label degradation, $\varepsilon = 10^{-6}$. Ground truth: Kepler Both $(1, 2)$, TESS Neither $(0, 0)$, BMMC Neither $(17, 11)$. Proxy rows use k-means on the weaker modality's raw features.

Condition	Kepler	TESS	BMMC
Noise, $R^2 \to 0.25\times$	Both $(1, 2)$	Neither $(0, 0)$	—
Reassign 50%	—	—	Neither $(10, 11)$
$44 \to 8$ lineages	—	—	Neither $(4, 5)$
$\log g$ in 8 bins	CP-only $(1, 2)$	Neither $(0, 0)$	—
k-means, $k = 8$	CP-only $(1, 3)$	CP-only $(1, 4)$	CP-only $(5, 8)$
k-means, $k = 20$	Neither $(5, 2)$	CP-only $(2, 2)$	Neither $(12, 10)$
k-means, $k = 44$	Neither $(5, 2)$	Neither $(3, 2)$	Neither $(14, 17)$

C.4 Origin of the BMMC nuisance floor

$\hat{\Delta}_{\mathrm{CA}}$ is a ratio whose denominator is the largest nuisance canonical correlation, so a single component sets the nuisance floor. To identify it, we regress the floor-setting component's score on cell-level metadata (canonical seed, $n = 5{,}000$). It is explained by technical rather than biological covariates: batch ($R^2 = 0.35$), site (0.25), and donor (0.21), with sequencing depth, mitochondrial fraction, cell-cycle phase, and lineage all near zero. The floor component of the $\mathbf{A}$-SVD is likewise batch-loaded ($R^2 = 0.25$).

We then ask whether removing this structure lifts the pair out of Neither. Regressing batch and sequencing depth out of both modalities lowers $\hat{\Delta}_{\mathrm{CA}}$ from 0.67 to 0.52, and centering each modality within donor gives 0.61; both remain Neither. In the first case the floor moves to a different component that is itself batch-loaded: removing the dominant batch direction promotes the next. The floor is thus a multi-dimensional technical subspace rather than a single removable direction, which is why linear batch correction does not move the pair across the boundary.

D Additional Figures

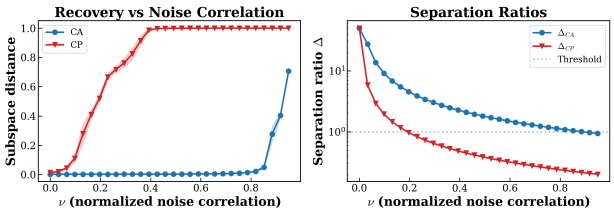


Figure 6: Recovery as a function of normalized noise correlation ν. *(Left)* Subspace distance between the estimated and true signal subspace (lower is better; shading shows ± 1 std over 20 trials). CP fails at $\nu \approx 0.15$ while CA remains robust until $\nu \approx 0.75$. *(Right)* Theoretical separation ratios Δ_{CA} and Δ_{CP} (log scale). The dashed line marks $\Delta = 1$; recovery succeeds above this threshold.

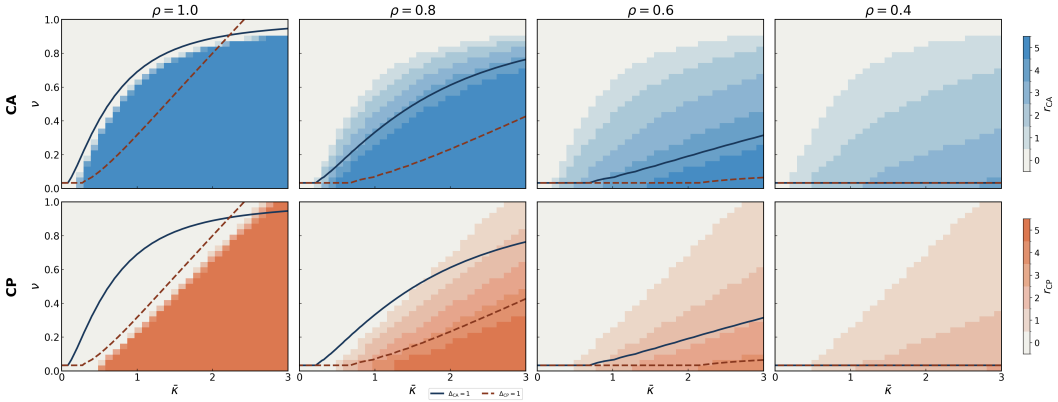


Figure 7: **Partial recovery under heterogeneous signal spectra.** Empirical recovery count r (number of signal directions with squared projection ≥ 0.8 onto the top-k recovered subspace) across the $(\bar{\kappa}, \nu)$ plane, for signal spread $\rho \in \{1.0, 0.8, 0.6, 0.4\}$ (columns, homogeneous $\rightarrow$ heterogeneous). *(Top)*: CA. *(Bottom)*: CP. Signal strengths follow a geometric decay $\kappa_i = \bar{\kappa}\,\rho^{i-1}$ with $k = 5$, $d = 20$; noise parameters $\gamma_x = 1$, $\gamma_y = 0.05$, $\tilde{\gamma}_y = 5$. Overlaid curves: $\Delta_{\mathrm{CA}} = 1$ (solid) and $\Delta_{\mathrm{CP}} = 1$ (dashed) from the homogeneous theory of figure 2. At $\rho = 1$ the transition is sharp ($r \in \{0, k\}$) and the $\Delta = 1$ contours align with the empirical boundary, reproducing figure 2. As ρ decreases, intermediate counts $r \in \{1, \ldots, k-1\}$ fill a widening band in which stronger signal directions are recovered first; the four-region phase diagram smears into a graded continuum. Averaged over 10 seeds with $n = 5{,}000$.

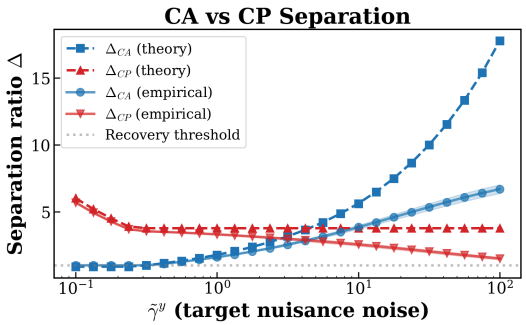


Figure 8: Separation ratios Δ_{CA} and Δ_{CP} as a function of target nuisance variance $\tilde{\gamma}^y$, validating Proposition 3.1. Theory curves (dashed) are computed from the closed-form expressions in Theorems A.1 and A.2; empirical curves (solid) are estimated from finite-sample covariances averaged over 20 random rotations. As $\tilde{\gamma}^y$ grows, Δ_{CA} increases unboundedly — CA's symmetric whitening suppresses high-variance nuisance on both sides — while Δ_{CP} remains approximately constant, since $\xi_j = \eta_j / \sqrt{\tilde{\gamma}_j^x}$ does not depend on target nuisance variance: source-side whitening operates only on the source modality. Theory and empirical curves are in close agreement throughout, confirming the accuracy of the closed-form predictions at finite sample sizes.

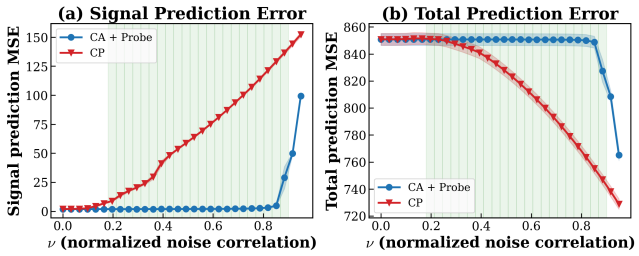
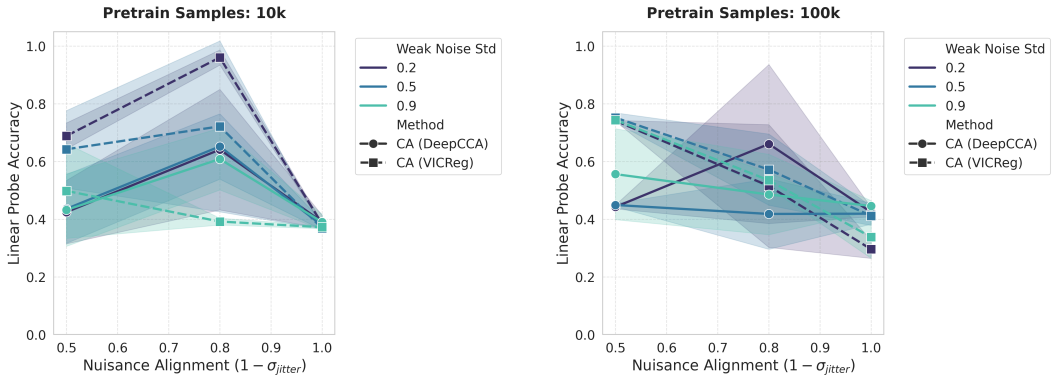


Figure 9: The variance trap: signal recovery vs. reconstruction quality for CA+Probe and direct CP, using the same parameter regime as Figure 6 ($\kappa = (3.0, 2.0, 1.5)$, $\tilde{\gamma}^y = 50.0$, $\tilde{\gamma}^x = 1.0$). Green shading marks the regime where $\Delta_{\text{CA}} > 1 > \Delta_{\text{CP}}$. **(a) Signal prediction MSE**: in the green region, CA+Probe achieves near-zero signal MSE while CP's signal error climbs steeply, confirming that CP encodes the wrong subspace. **(b) Total prediction MSE**: CP achieves *lower* total MSE than CA+Probe across the same region, because it successfully reconstructs the high-variance nuisance components — a task at which CA+Probe, having discarded nuisance directions, cannot compete. Together, the two panels illustrate the core danger of using reconstruction error as a proxy for signal recovery: CP can achieve lower loss than CA while failing to recover the signal, precisely because the MSE objective does not distinguish signal from nuisance.



(a) dSprites (b) Shape3D

Figure 10: Comparison between VICReg and DeepCCA for dSprites and Shape3D experiments

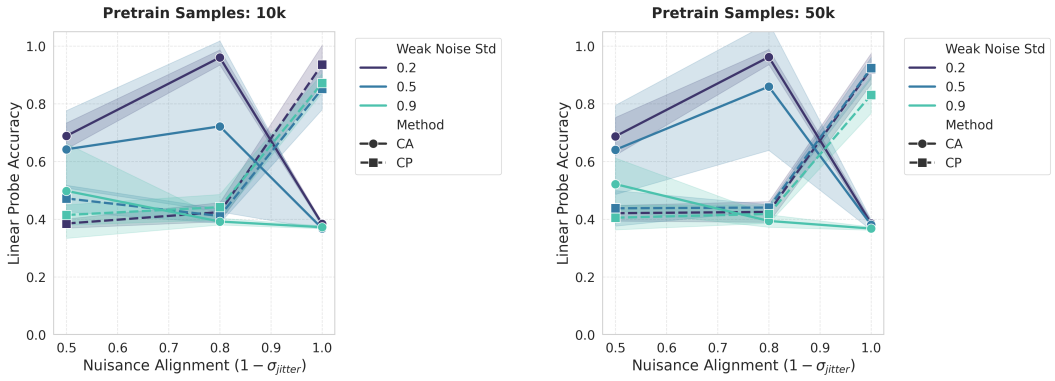


Figure 11: Stereo-dSprites accuracy vs. nuisance alignment at 10k *(Left)* and 50k *(Right)* pretraining samples. The CA–CP crossover from Figure 3 persists across dataset scales.

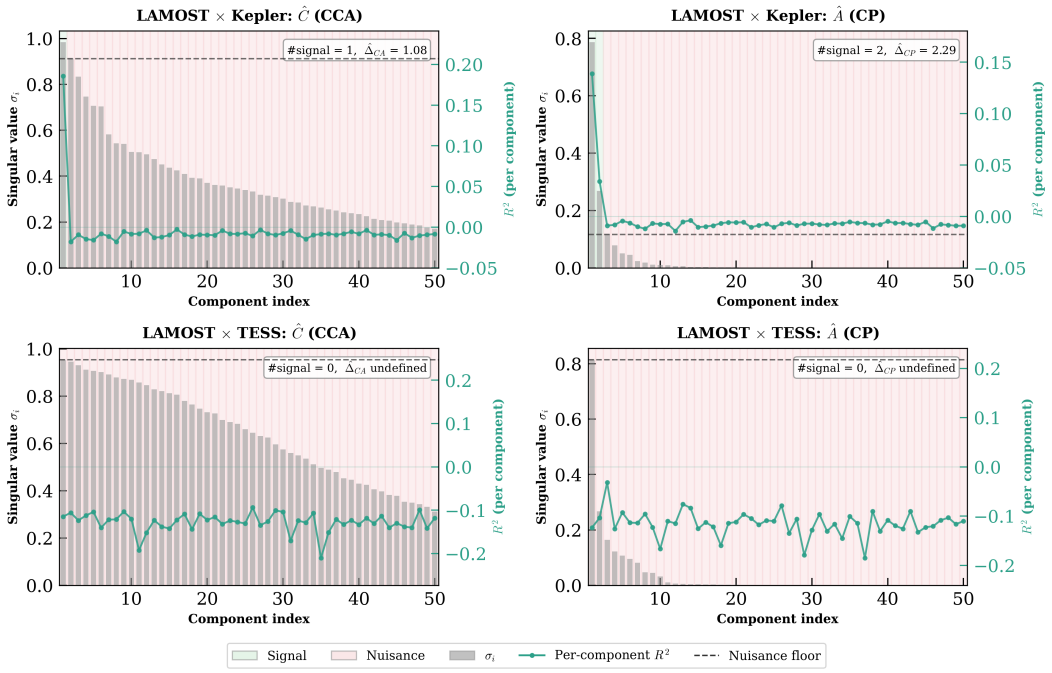


Figure 12: **Signal–nuisance decomposition underlying the regime predictions of Table 1.** Each panel shows the sorted singular values (gray bars, left axis) and per-component R^2 against $\log g$ (teal line, right axis) used by Algorithm 1 to classify components as signal (green shading) or nuisance (red shading). Dashed horizontal line: nuisance floor $\max_j \hat{\nu}_j$ (CCA panels) or $\max_j \hat{\xi}_j$ (**A**-SVD panels); $\hat{\Delta} > 1$ iff every classified signal singular value exceeds the nuisance floor. *(Top)*: LAMOST $\times$ Kepler — both decompositions have signal components above the nuisance floor ($\hat{\Delta}_{\mathrm{CA}} = 1.08$, $\hat{\Delta}_{\mathrm{CP}} = 2.29$; Both regime). *(Bottom)*: LAMOST $\times$ TESS — no CCA component predicts $\log g$ above noise ($R^2 \approx 0$ across all components, zero signal detected); the **A**-SVD has one dominant singular value, but no component predicts $\log g$, so it too is classified as nuisance. No signal direction clears the floor in either spectrum ($\hat{r}_{\mathrm{CA}} = \hat{r}_{\mathrm{CP}} = 0$), the complete-failure form of the Neither regime. The contrast between the two rows — same LAMOST encoder, same protocol, different photometric instrument — shows that instrument quality determines the signal–nuisance separation and hence the regime placement.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The four contributions stated in the introduction: closed-form solutions, separation ratios with phase diagrams, data-driven estimation, and experimental validation are delivered in Section 3, and Section 4. Limitations of the linear analysis for single-stage pipelines are discussed in Section 4.4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed where they arise: applicability of the linear analysis to nonlinear regimes (Section 4), the scope of the diagnostic to two-stage pipelines (Section 4.4), and the open Neither regime (Section 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the assumptions and proofs are provided in Section A. Closed form solutions are given in Section A.1, proofs are given in Section A.2 and Section A.3. Proofs to Proposition 3.1 and Proposition 3.2 are given in Section A.6 and Section A.7.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section B provides architectures, training hyperparameters, optimization details, evaluation protocols, and sample sizes for all five experiments. Algorithm 1 is stated in full. Code will be released at the anonymous repository linked on page 1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized repository with code, configurations, and per-experiment READMEs is linked on page 1. All synthetic data is generated by included scripts; public datasets (MS-COCO, LAMOST, Kepler, TESS) are accessed via documented procedures. The repository will be open-sourced upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section B provides architectures, optimizers, learning rates, batch sizes, training durations, and evaluation protocols for all five experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Linear experiments report std over 20 rotations; stereo-vision experiments show shaded bands across multiple seeds. Astrophysical experiments report mean $\pm$ std over 5 seeds; The MS-COCO experiment reports a single run per condition due to computational cost.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For each experiment, we specify the GPU types, number of GPUs, and training times in Section B. Total compute for the project, including preliminary experiments not reported in the paper, is approximately 1000 GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves no human subjects, no scraped data, and no models with foreseeable misuse risks. All datasets used are public and used within their licensing terms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is a theoretical analysis of multimodal representation learning objectives. It introduces no new datasets, models, or applications with direct societal consequences.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no pretrained models, datasets, or generative systems with high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used (MS-COCO, dSprites, 3DShapes, LAMOST, Kepler, TESS) are publicly available and properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper introduces no new datasets or pretrained models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for editing and writing assistance (code and text).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.