Provable Subspace Identification of Nonlinear Multi-view CCA

Zhiwei Han[1] **Stefan Matthes**[1] **Hao Shen**[1,2]

[1]Chair of Data Processing, Technical University of Munich, Germany
[2]fortiss GmbH, Munich, Germany

Abstract

We investigate the identifiability of nonlinear canonical correlation analysis (CCA) in a multi-view setup, in which each view is generated by applying an unknown nonlinear map to a linear mixture of shared latent variables plus view-private noise. Rather than pursuing exact unmixing, which is known to be ill-posed under general nonlinear mixing, we instead reframe multi-view CCA as a basis-invariant subspace identification problem. Under suitable latent priors and spectral separation conditions, we prove that the pairwise population CCA objective recovers correlated signal subspaces up to view-wise orthogonal ambiguity. For $N \geq 3$ views, their multi-view aggregation provably isolates the jointly correlated subspaces shared across all views while eliminating view-private variation. We further establish finite-sample statistical consistency guarantees by translating the concentration of empirical cross-covariances into explicit subspace error bounds via spectral perturbation theory. Experiments on synthetic and rendered image datasets support our theoretical findings and illustrate the necessity of the assumed conditions.

1 INTRODUCTION

Aligned multi-view data are ubiquitous, arising from multimodal sensing, multi-camera systems, and synchronized instruments. A primary objective of representation learning is to recover latent structures shared across views from *nonlinear* observations while discarding view-private variation, yielding representations that are interpretable, robust to nuisance perturbations, and transferable across modalities. Correlation-based methods, most notably canonical correlation analysis (CCA) and its multi-view generalizations [Kettenring, 1971], address this objective by learning whitened representations that are maximally aligned in second-order statistics. Such methods and their whitening-based variants are also widely used in self-supervised learning to prevent feature collapse and stabilize training [Zbontar et al., 2021, Weng et al., 2022, Ermolov et al., 2021, He and Ozay, 2022, Kalapos and Gyires-Tóth, 2024].

Despite their empirical success, what nonlinear CCA *identifies* under realistic nonlinear distortions remains insufficiently characterized. Given the impossibility of unsupervised source recovery from general nonlinear mixtures [Hyvärinen and Pajunen, 1999, Locatello et al., 2019], multi-view constraints have emerged as a prominent structural prior to restore identifiability. Existing CCA-based guarantees either impose post-nonlinear assumptions [Lyu and Fu, 2020] or establish equivalence up to arbitrary invertible transforms [Lyu et al., 2022]. While recent work establishes affine identifiability under general nonlinear mixing in the two-view regime [Han et al., 2026], what additional views enable in a basis-invariant sense remains unclear.

In this work, we address this question through the lens of *subspace identification*. We study an additive multi-view generative process where each view is produced by an unknown smooth invertible map applied to a view-specific source. Each source decomposes into a linearly mixed shared latent vector plus view-private noise, relaxing the coordinate-wise independence assumption standard in ICA and aligning with the content–style modeling paradigm in causal representation learning [von Kügelgen et al., 2021, Yao et al., 2024]. Because the mixing matrices are not identifiable, we target their basis-invariant *signal subspaces* instead. We show that multi-view CCA acts as an *intersection filter* over latent factors: it isolates the jointly correlated structure shared across views while eliminating view-private variation. We further provide finite-sample guarantees for the empirical estimator.

We summarize our contributions as follows:

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1947–1982.

- We propose an N-view ($N \geq 3$) additive latent model that relaxes the coordinate-wise independence assumption and casts nonlinear CCA as a basis-invariant subspace identification problem.
- We prove that generalized nonlinear CCA isolates the jointly correlated signal subspaces, formalizing the objective as an intersection filter over latent factors.
- We establish finite-sample consistency for empirical multi-view CCA, translating concentration of second-order statistics into explicit subspace recovery rates via spectral perturbation bounds.
- We validate the theory on synthetic and rendered image datasets, corroborating correlated-subspace recovery and the necessity of the stated conditions.

2 RELATED WORK

Identifiable Disentangled Representations. Identifiable representation learning seeks to recover latent structure up to unavoidable symmetries, an objective rooted in blind source separation and independent component analysis (ICA) [Barlow et al., 1989, Comon, 1994, Cardoso, 1998]. Without additional assumptions, exact recovery from single-view i.i.d. observations under general nonlinear mixing is fundamentally impossible [Hyvärinen and Pajunen, 1999, Locatello et al., 2019]. To restore identifiability, prior work has introduced symmetry-breaking inductive biases, such as temporal dynamics, nonstationarity, auxiliary variables, interventions, or multi-view constraints [Hyvärinen and Morioka, 2016, 2017, Hyvärinen et al., 2019, Khemakhem et al., 2020, Locatello et al., 2020, Shu et al., 2020].

Identifiability from Multiple Views. Multi-view observations provide cross-view constraints that mitigate the non-identifiability of single-view mixing. In linear settings, higher-order moment methods and multi-view ICA can isolate shared and view-specific sources up to standard permutation and scaling ambiguities [Podosinnikova et al., 2016, Pandeva and Forré, 2023]. Depending on the model class, contrastive objectives can identify shared factors in nonlinear mixtures up to specific transformations. Assuming independent sources and structured noise, these identifiability guarantees range from component-wise transformations to linear or block-level equivalence classes [Zimmermann et al., 2021, Gresele et al., 2020, Matthes et al., 2023, Rodríguez Gálvez et al., 2023]. Recent advances relax the independent-source assumption by exploiting causal latent structures, shifting the target to *block-identifiability*. These methods are increasingly unified through partial-observability frameworks, where identifiability follows from underlying graphical structures and suitable regularity conditions [von Kügelgen et al., 2021, Schölkopf et al., 2016, Sturma et al., 2023, Yao et al., 2024, 2025, Ahuja et al., 2022, Daunhawer et al., 2023, Xu et al., 2024, Locatello et al., 2020].

Identifiability of Nonlinear Multi-view CCA. CCA aligns views through second-order statistics after whitening. Recent research on CCA-based identifiability has made progress on several related problems, including factor-level identifiability [Lyu and Fu, 2020, Lyu et al., 2022, Karakasis and Sidiropoulos, 2023], nonlinear dynamical system identification [Sidiropoulos and Sørensen, 2022], subspace identification [Sørensen et al., 2021], and subspace clustering [Karakasis and Sidiropoulos, 2025]. However, arbitrary nonlinear mixing makes nonlinear CCA substantially more challenging: existing guarantees are typically limited to recovery up to arbitrary invertible transformations, a substantially weaker conclusion than what is available in the linear case. Recent work establishes affine identifiability for nonlinear CCA in the two-view regime [Han et al., 2026] by leveraging Lancaster distribution priors [Lancaster, 1958, Eagleson, 1964]. It does not characterize what additional views identify from a basis-invariant perspective. To bridge this gap, we shift the identifiability target from exact mixing recovery to signal subspace identification. We prove that, under suitable latent priors, generalized CCA identifies the jointly correlated subspaces shared across views, thereby yielding a basis-invariant guarantee stronger than equivalence up to arbitrary invertible transformations.

3 PROBLEM FORMULATION

Multi-view Generative Process. Let $N \geq 2$ denote the number of views. We observe M i.i.d. multi-view samples

$$\{(\mathbf{x}_1^{(m)}, \ldots, \mathbf{x}_N^{(m)})\}_{m=1}^{M}.$$

We suppress the sample index and write $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ for a generic multi-view observation. For each view $i \in [N] := \{1, \ldots, N\}$, the observation $\mathbf{x}_i \in \mathcal{X}_i \subseteq \mathbb{R}^{d_{\mathcal{X}_i}}$ is generated from an *unobserved* view-specific latent source $\mathbf{s}_i \in \mathcal{S}_i \subseteq \mathbb{R}^{d_{\mathcal{S}_i}}$ via an *unknown, generally nonlinear* generative mapping $\mathbf{g}_i : \mathcal{S}_i \to \mathcal{X}_i$, i.e., $\mathbf{x}_i = \mathbf{g}_i(\mathbf{s}_i)$. Throughout our identifiability analysis, we assume that each $\mathbf{g}_i$ is smooth and invertible on the support of $\mathbf{s}_i$. Here, $\mathcal{S}_i$ and $\mathcal{X}_i$ denote the source and observation spaces of view i, respectively, and $d_{(\cdot)}$ is the dimension of the underlying ambient space.

Formally, we define the multi-view generative process via an additive-noise latent construction at the source level as follows: for each $i \in [N]$, we have

$$\mathbf{x}_i = \mathbf{g}_i(\mathbf{s}_i), \quad \mathbf{s}_i = \mathbf{A}_i \mathbf{c} + \boldsymbol{\epsilon}_i, \quad \mathbf{c} \sim p_{\mathbf{c}}, \quad \boldsymbol{\epsilon}_i \sim p_{\boldsymbol{\epsilon}_i}, \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^{d_C}$ is a latent vector shared across views, $\boldsymbol{\epsilon}_i \in \mathbb{R}^{d_{\mathcal{S}_i}}$ captures view-private variation, and $\mathbf{A}_i \in \mathbb{R}^{d_{\mathcal{S}_i} \times d_C}$ is a view-specific mixing matrix.

Our generative process generalizes the single-view model of Lyu and Fu [2022] to N views, with view-dependent mixing matrices $\{\mathbf{A}_i\}_{i=1}^N$ and nonlinear maps $\{\mathbf{g}_i\}_{i=1}^N$. Unlike standard ICA assumptions that enforce coordinate-wise independence [Hyvärinen and Morioka, 2016, 2017, Hyvärinen et al., 2019, Khemakhem et al., 2020], the shared signal term $\mathbf{A}_i\mathbf{c}$ induces structured dependencies within each source $\mathbf{s}_i$, e.g., as in causal latent models, consistent with the checkerboard-type dependencies of Matthes et al. [2023]. Interpreting $\mathbf{c}$ and $\boldsymbol{\epsilon}_i$ as shared *content* and view-private *style*, our model differs from concatenation-based multi-view formulations in causal representation learning [Daunhawer et al., 2023, Yao et al., 2024]. Moreover, the model connects to partial observability: if $\ker(\mathbf{A}_i) \neq \{0\}$, then components of $\mathbf{c}$ lying in $\ker(\mathbf{A}_i)$ do not affect $\mathbf{s}_i$ and are therefore unobservable from view i.

Latent Structure. The additive construction induces nontrivial cross-view dependence among the source variables. Since generalized CCA aggregates pairwise correlations, our analysis first characterizes the joint distribution of each source pair $(\mathbf{s}_i, \mathbf{s}_j)$ over $\mathcal{S}_i \times \mathcal{S}_j$ with $d_{\mathcal{S}_i}, d_{\mathcal{S}_j} \geq 2$. We then combines the resulting pairwise structures across views.

Assumption 1 (Latent Distributional Prior)**.** Let $\mathbf{c}$ and $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ be as in Equation 1.

1. **Latent Factorization.** The shared latent vector and private latent vectors are mutually independent. Within each vector, coordinates are i.i.d.:

$$p_{\mathbf{c}}(\mathbf{c}) = \prod_{j=1}^{d_C} \phi(c_j), \qquad p_{\boldsymbol{\epsilon}_i}(\boldsymbol{\epsilon}_i) = \prod_{j=1}^{d_{\mathcal{S}_i}} \phi(\epsilon_{ij}),$$

$$\text{and} \qquad p(\mathbf{c}, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N) = p_{\mathbf{c}}(\mathbf{c}) \prod_{i=1}^{N} p_{\boldsymbol{\epsilon}_i}(\boldsymbol{\epsilon}_i),$$

   where ϕ denotes standardized Gaussian density.
2. **Standardization.** The latents are centered and isotropic:

$$\mathbb{E}[\mathbf{c}] = \mathbf{0},\ \mathrm{Cov}(\mathbf{c}) = \mathbf{I}_{d_C}, \quad \mathbb{E}[\boldsymbol{\epsilon}_i] = \mathbf{0},\ \mathrm{Cov}(\boldsymbol{\epsilon}_i) = \mathbf{I}_{d_{\mathcal{S}_i}}.$$

3. **Admissible Marginals.** In the main text, we use standardized Gaussian marginals, which are sufficient for our main theory. Extensions to other standardized Lancaster-type marginals are discussed in the supplementary material.

Learning Objective. The exact linear mixing matrices $\{\mathbf{A}_i\}_{i=1}^N$ are fundamentally not identifiable as bases. Instead, we target the basis-invariant column-space information encoded by these matrices, formalized later as correlated signal subspaces in Definition 1. Specifically, we learn view-specific encoders

$$\mathbf{f}_i : \mathcal{X}_i \to \mathcal{Z} \subseteq \mathbb{R}^{d_{\mathcal{Z}}}, \qquad \forall i \in [N]$$

and analyze the induced source-domain maps $\mathbf{h}_i := \mathbf{f}_i \circ \mathbf{g}_i$. Our identifiability analysis asks whether CCA forces these maps to recover the identifiable correlated signal subspaces, up to the unavoidable orthogonal ambiguity introduced by whitening. For simplicity, we use a common representation dimension $d_{\mathcal{Z}}$ across views.

4 PRELIMINARIES

Notations. Throughout, scalars are denoted by plain letters. Random vectors and vector-valued functions are denoted by bold lowercase letters. Matrices are denoted by bold uppercase letters. Sets and spaces are denoted by calligraphic letters.

4.1 NONLINEAR MULTI-VIEW CCA

CCA learns encoders whose whitened representations maximize cross-view correlation measured by singular values of normalized cross-covariances.

Generalized CCA for Multi-view Learning. Under the multi-view generative process in Equation 1, generalized CCA [Kettenring, 1971] aggregates pairwise objectives across all view pairs as

$$J := \sum_{1 \leq i < j \leq N} \left\| \boldsymbol{\Sigma}_{ii}^{-1/2} \boldsymbol{\Sigma}_{ij} \boldsymbol{\Sigma}_{jj}^{-1/2} \right\|_*, \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm, $\boldsymbol{\Sigma}_{ii} := \mathrm{Cov}(\mathbf{f}_i(\mathbf{x}_i)) \succ 0$ and $\boldsymbol{\Sigma}_{jj} := \mathrm{Cov}(\mathbf{f}_j(\mathbf{x}_j)) \succ 0$, while $\boldsymbol{\Sigma}_{ij} := \mathrm{Cov}(\mathbf{f}_i(\mathbf{x}_i), \mathbf{f}_j(\mathbf{x}_j))$ denotes the cross-covariance matrix between $\mathbf{f}_i$ and $\mathbf{f}_j$.

CCA in the source domain. Let the induced representation mapping for view $i \in [N]$ be $\mathbf{h}_i := \mathbf{f}_i \circ \mathbf{g}_i : \mathcal{S}_i \to \mathcal{Z}$. By reparameterization invariance (Lemma 6;[Han et al., 2026]), for all $1 \leq i < j \leq N$, $\mathrm{Cov}(\mathbf{f}_i(\mathbf{x}_i), \mathbf{f}_j(\mathbf{x}_j)) = \mathrm{Cov}(\mathbf{h}_i(\mathbf{s}_i), \mathbf{h}_j(\mathbf{s}_j))$ and $\mathrm{Cov}(\mathbf{f}_i(\mathbf{x}_i)) = \mathrm{Cov}(\mathbf{h}_i(\mathbf{s}_i))$. Consequently, optimizing Equation 2 over $\{\mathbf{f}_i\}$ is mathematically equivalent to optimizing it over the induced representation mapping $\mathbf{h}_i$. Under the regularity and representability conditions of Lemma 6, this reparameterization allows us to analyze population maximizers in the source domain without changing the CCA objective.

4.2 WHITENING AND CANONICALIZATION

Whitening, or its regularized variants such as soft-whitening [Zbontar et al., 2021], imposes a non-degenerate normalization [Weng et al., 2022] that makes the CCA objective well posed and prevents representational collapse [Ermolov et al., 2021]. Both properties are critical to our identifiability analysis.

Whitened Representations. For each view $i \in [N]$, let $\boldsymbol{\mu}_i' := \mathbb{E}[\mathbf{h}_i(\mathbf{s}_i)]$ and $\boldsymbol{\Sigma}_{ii}' := \mathrm{Cov}(\mathbf{h}_i(\mathbf{s}_i)) \succ 0$. A representation-whitening matrix $\mathbf{B}_i \in \mathbb{R}^{d_{\mathcal{Z}} \times d_{\mathcal{Z}}}$ satisfies $\mathbf{B}_i \boldsymbol{\Sigma}_{ii}' \mathbf{B}_i^\top = \mathbf{I}$ and defines

$$\tilde{\mathbf{h}}_i(\mathbf{s}_i) := \mathbf{B}_i\big(\mathbf{h}_i(\mathbf{s}_i) - \boldsymbol{\mu}_i'\big), \qquad \mathrm{Cov}(\tilde{\mathbf{h}}_i(\mathbf{s}_i)) = \mathbf{I}.$$

A canonical choice is the symmetric inverse square root $\mathbf{B}_i = {\boldsymbol{\Sigma}_{ii}'}^{-1/2}$. Under this whitening, the normalized cross-covariance in Equation 2 becomes the cross-covariance of the whitened source-domain representations:

$$\tilde{\boldsymbol{\Sigma}}_{ij} := \mathrm{Cov}\big(\tilde{\mathbf{h}}_i(\mathbf{s}_i), \tilde{\mathbf{h}}_j(\mathbf{s}_j)\big) = \mathbf{B}_i\, \boldsymbol{\Sigma}_{ij}'\, \mathbf{B}_j^\top,$$

where $\boldsymbol{\Sigma}_{ij}' := \mathrm{Cov}(\mathbf{h}_i(\mathbf{s}_i), \mathbf{h}_j(\mathbf{s}_j))$. Thus the pairwise CCA term equals $\|\tilde{\boldsymbol{\Sigma}}_{ij}\|_*$. While orthogonal left-multiplication of $\mathbf{B}_i$ defines the equivalence class for our identifiability theorem, the CCA objective remains strictly invariant to this ambiguity since it depends solely on the orthogonally invariant singular values of $\tilde{\boldsymbol{\Sigma}}_{ij}$.

Canonicalization via Whitening. Whitening removes scale and non-orthogonal within-view linear ambiguities, reducing the feasible set to orthonormal directions in L^2; the remaining ambiguity is orthogonal. Fix a view $i \in [N]$ and consider the whitened representation mapping $\tilde{\mathbf{h}}_i(\mathbf{s}_i)$. Let $\{\xi_{i,n}\}_{n\geq 1}$ be a fixed orthonormal basis of $L^2(P_{\mathbf{s}_i})$ restricted to centered square-integrable functions. Since $\tilde{\mathbf{h}}_i$ is centered, the constant basis component is zero and is omitted from the expansion. Each component of $\tilde{\mathbf{h}}_i$ admits the following orthonormal basis expansion: for the component $j \in [d_{\mathcal{Z}}]$,

$$\tilde{h}_{i,j}(\mathbf{s}_i) = \sum_{n\geq 1} c_{i,j,n}\, \xi_{i,n}(\mathbf{s}_i), \;\; c_{i,j,n} := \langle \tilde{h}_{i,j}, \xi_{i,n} \rangle_{L^2(\mu_i)},$$

where $\mathbf{c}_{i,j} = [c_{i,j,1}, \cdots, c_{i,j,\infty}]$ denotes the corresponding coefficient vector under the basis $\{\xi_{i,n}\}_{n\geq 1}$. The whitening constraint is equivalent to the orthonormality of these coefficient vectors across components within the same view i:

$$\mathbf{c}_{i,p}^\top \mathbf{c}_{i,q} = \delta_{pq}, \qquad p, q \in [d_{\mathcal{Z}}], \tag{3}$$

where δ_{pq} is the Kronecker-delta function, i.e., $\delta_{pq} = 1$ iff $p = q$ and 0 otherwise. Whitening canonically enforces (i) decorrelation between representation components and (ii) unit normalization of each component in function space. Equivalently, the coefficient matrix $\mathbf{C}_i$ with rows $(\mathbf{c}_{i,1}^\top, \ldots, \mathbf{c}_{i,d_{\mathcal{Z}}}^\top)$ satisfies $\mathbf{C}_i \mathbf{C}_i^\top = \mathbf{I}$.

5 THEORY

5.1 ORTHOGONAL POLYNOMIAL EXPANSION OF THE CANONICAL DENSITY

We first analyze the pairwise source distribution induced by the additive multi-view model in Equation 1. After whitening each source, the normalized cross-covariance admits an SVD whose canonical coordinates diagonalize the Gaussian dependence between two views. This yields a product factorization of the joint density and, via the Mehler–Hermite expansion, a spectral decomposition of nonlinear cross-view correlations.

Canonical Coordinate System. Under the mutual independence and i.i.d. conditions in Assumption 1.1, let

$$\mathbf{W}_i := \mathrm{Cov}(\mathbf{s}_\mathrm{i})^{-\frac{1}{2}} = \big(\mathbf{A}_\mathrm{i}\mathbf{A}_\mathrm{i}^\top + \mathbf{I}\big)^{-\frac{1}{2}}, \quad \mathrm{i} \in [\mathrm{N}]$$

be a whitening matrix for the view-specific source $\mathbf{s}_i$. For any pair of views $1 \leq i < j \leq N$, the dependency structure of the view-specific sources $(\mathbf{s}_i, \mathbf{s}_j)$ is captured by their normalized cross-covariance matrix:

$$\mathbf{R}_{ij} := \mathrm{Cov}(\mathbf{W}_i\mathbf{s}_i, \mathbf{W}_j\mathbf{s}_j) = \mathbf{W}_i\mathbf{A}_i\mathbf{A}_j^\top\mathbf{W}_j^\top. \tag{4}$$

We consider the following singular value decomposition,

$$\mathbf{R}_{ij} = \mathbf{U}_{ij}\mathbf{T}_{ij}\mathbf{V}_{ij}^\top, \quad \mathbf{U}_{ij} \in O(d_{\mathcal{S}_i}), \quad \mathbf{V}_{ij} \in O(d_{\mathcal{S}_j})$$

where $\mathbf{T}_{ij} \in \mathbb{R}^{d_{\mathcal{S}_i} \times d_{\mathcal{S}_j}}$ is a rectangular diagonal matrix. Let $r_{ij} := \mathrm{rank}(\mathbf{A}_i\mathbf{A}_j^\top)$, the nonzero singular values of $\mathbf{T}_{ij}$ satisfy

$$1 > t_{ij,1} \geq \cdots \geq t_{ij,r_{ij}} > 0,$$

and all remaining singular values are zero.

Lemma 1 (Canonical Factorization)**.** *Under the isotropic Gaussian prior in Assumption 1.2, define the canonical coordinates*

$$\mathbf{u} = \mathbf{U}_{ij}^\top\mathbf{W}_i\mathbf{s}_i, \qquad \mathbf{v} = \mathbf{V}_{ij}^\top\mathbf{W}_j\mathbf{s}_j, \tag{5}$$

where $\mathbf{U}_{ij}$ and $\mathbf{V}_{ij}$ are orthogonal canonicalizers of the whitened cross-covariance between $\mathbf{s}_i$ and $\mathbf{s}_j$. By the change-of-variables formula for this invertible linear transformation,

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) = p_{\mathbf{s}_i,\mathbf{s}_j}\big(\mathbf{W}_i^{-1}\mathbf{U}_{ij}\mathbf{u}, \mathbf{W}_j^{-1}\mathbf{V}_{ij}\mathbf{v}\big) \cdot |\det \mathbf{W}_i|^{-1} |\det \mathbf{W}_j|^{-1}.$$

Moreover, the joint density factorizes as

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) = \underbrace{\prod_{k=1}^{r_{ij}} \phi_{t_{ij,k}}(u_k, v_k)}_{\mathcal{K}_{ij}(\mathbf{u}[1:r_{ij}], \mathbf{v}[1:r_{ij}])} \prod_{m=r_{ij}+1}^{d_{\mathcal{S}_i}} \phi(u_m) \prod_{n=r_{ij}+1}^{d_{\mathcal{S}_j}} \phi(v_n)$$

where ϕ is the standard normal density and ϕ_t is the standardized bivariate normal density with correlation t.

Remark 1 (Information-Theoretic Interpretation). The factorization in Lemma 1 implies that the cross-view interaction decouples into r_{ij} independent Gaussian coordinates:

$$v_k = t_{ij,k}\, u_k + \sqrt{1 - t_{ij,k}^2}\, \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0,1), \quad k \in [r_{ij}],$$

where the canonical correlation $t_{ij,k}$ directly governs the effective *signal-to-noise ratio* (SNR) $\gamma_{ij,k} := t_{ij,k}^2/(1-t_{ij,k}^2)$ for the pair (u_k, v_k). By definition, $r_{ij} = \operatorname{rank}(\mathbf{A}_i\mathbf{A}_j^\top) \leq \min(d_{\mathcal{S}_i}, d_{\mathcal{S}_j})$, we have $t_{ij,k} = 0$ for all $k > r_{ij}$. This indicates that only r_{ij} correlated canonical coordinates share cross-view information, i.e., $\mathrm{SNR} > 0$, while the remaining modes are independent noise.

To cast this probabilistic correlation structure into a functional-analytic framework, we expand the product bivariate Gaussian density $\mathcal{K}_{ij}$ in Lemma 1 using normalized multivariate Hermite polynomials as shown in Mehler [1866] and Han et al. [2026].

Lemma 2 (Normalized Multivariate Mehler–Hermite Expansion)**.** *Let $r := r_{ij}$ and let $\psi_n(z) := \frac{1}{\sqrt{n!}}He_n(z), n \in \mathbb{N}_0$, be the normalized probabilists' Hermite polynomials. For $\mathbf{n} = (n_1, \ldots, n_r) \in \mathbb{N}_0^r$, define normalized multivariate Hermite polynomials as*

$$\Psi_{\mathbf{n}}(\mathbf{u}) := \prod_{k=1}^{r} \psi_{n_k}(u_k), \qquad \phi_r(\mathbf{u}) := \prod_{k=1}^{r} \phi(u_k).$$

Then $\{\Psi_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}_0^r}$ is an orthonormal basis of $L^2(\nu_r)$, where $\nu_r(d\mathbf{u}) = \phi_r(\mathbf{u})d\mathbf{u}$ is the multi-dimensional Gaussian measure.

Let $(\mathbf{U}, \mathbf{V})$ be a centered Gaussian pair in canonical coordinates such that $\mathrm{Cov}(\mathbf{U}) = \mathrm{Cov}(\mathbf{V}) = I_r$ and $\mathrm{Cov}(\mathbf{U}, \mathbf{V}) = \mathrm{diag}(\rho_1, \ldots, \rho_r)$, where $\rho_k := t_{ij,k} \in (-1, 1)$. If $\mathcal{K}_{ij}$ denotes the joint density of $(\mathbf{U}, \mathbf{V})$, then

$$\mathcal{K}_{ij}(\mathbf{u}, \mathbf{v}) = \phi_r(\mathbf{u})\phi_r(\mathbf{v}) \sum_{\mathbf{n}\in\mathbb{N}_0^r} t_{\mathbf{n}}\, \Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{n}}(\mathbf{v}),$$

where $t_{\mathbf{n}} := \prod_{k=1}^{r} \rho_k^{n_k}$.

Together with reparameterization invariance (Lemma 6) and whitening-induced canonicalization (Equation 3), Lemma 2 yields a functional-analytic framework for studying nonlinear representation maps in canonical coordinates. Leveraging the orthonormality of Hermite polynomials under the Gaussian measure [Dunkl and Xu, 2014], this framework diagonalizes cross-view correlations in the Hermite basis, thereby enabling a principled separation of first-order linear modes from higher-order nonlinear Hermite components.

5.2 SUBSPACE IDENTIFIABILITY

After whitening, CCA-like objectives remain invariant to within-view orthogonal transformations, such as rotations. Since precise recovery of mixing matrices $\{\mathbf{A}_i\}_{i=1}^N$ is fundamentally impossible, we instead target the basis-invariant signal subspaces they span.

Definition 1 (Correlated Signal Subspaces)**.** For each view $i \in [N]$, the *signal subspace* is defined as $\mathcal{U}_i := \mathrm{col}(\mathbf{A}_i) \subseteq \mathbb{R}^{d_{\mathcal{S}_i}}$. For any pair of distinct views $1 \leq i < j \leq N$, let $r_{ij} := \operatorname{rank}(\mathbf{A}_i\mathbf{A}_j^\top)$. We define the *pairwise correlated signal subspace* of view i with respect to view j as

$$\mathcal{U}_{i|j} := \mathrm{col}\big(\mathbf{U}_{ij}(:, 1:r_{ij})\big), \quad \mathcal{U}_{j|i} := \mathrm{col}\big(\mathbf{V}_{ij}(:, 1:r_{ij})\big).$$

Note that $\mathcal{U}_{i|j} \subseteq \mathcal{U}_i$ and $\mathcal{U}_{j|i} \subseteq \mathcal{U}_j$, with equality holding if and only if $r_{ij} = \operatorname{rank}(\mathbf{A}_i)$ and $r_{ij} = \operatorname{rank}(\mathbf{A}_j)$, respectively. Finally, the *multi-view jointly correlated subspace* for view i is defined as the intersection of its pairwise correlated subspaces: $\mathcal{U}_i^{\mathrm{mv}} := \bigcap_{j\neq i} \mathcal{U}_{i|j}$.

We then give the identifiability of the correlated signal subspace in two-view setup.

Theorem 5.1 (Infinite-Dimensional Two-view Subspace Identifiability)**.** *Fix a pair of distinct views $i \neq j$ with $r_{ij} := \operatorname{rank}(\mathbf{A}_i\mathbf{A}_j^\top) > 0$. Consider the population two-view CCA objective for this pair in the infinite-dimensional whitened function class. Suppose the observed data are generated by the multi-view generative process defined in Equation 1, where the generating functions $\mathbf{g}_i$ and $\mathbf{g}_j$ are smooth and invertible and Assumption 1 holds. Let $(\tilde{\mathbf{f}}_i^\star, \tilde{\mathbf{f}}_j^\star)$ be any whitened population maximizer of the two-view CCA objective in Equation 2, and define the composition mappings $\tilde{\mathbf{h}}_\ell^\star := \tilde{\mathbf{f}}_\ell^\star \circ \mathbf{g}_\ell$ for $\ell \in \{i, j\}$.*

Assuming an infinite feature dimension, i.e., $d_{\mathcal{Z}} \to \infty$, there exist orthogonal matrices $\mathbf{O}_i, \mathbf{O}_j \in O(r_{ij})$ such that,

$$\begin{aligned} \mathbf{P}_{r_{ij}}\tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) &= \mathbf{O}_i\mathbf{U}_{ij}(:, 1:r_{ij})^\top\mathbf{W}_i\mathbf{s}_i, \\ \mathbf{P}_{r_{ij}}\tilde{\mathbf{h}}_j^\star(\mathbf{s}_j) &= \mathbf{O}_j\mathbf{V}_{ij}(:, 1:r_{ij})^\top\mathbf{W}_j\mathbf{s}_j, \end{aligned} \tag{6}$$

where $\mathbf{P}_{r_{ij}}$ is a rank-r_{ij} coordinate selector, defined up to the unavoidable within-view orthogonal ambiguity of CCA, that extracts the coordinates corresponding to the pairwise correlated linear Hermite modes. Consequently, the maximizers identify the r_{ij}-dimensional whitened correlated subspaces $\mathcal{U}_{i|j}$ and $\mathcal{U}_{j|i}$ up to orthogonal transformations.

Assumption 2 (First-Order Canonical Dominance)**.** For any pair of views $1 \leq i < j \leq N$ with $\operatorname{rank}(\mathbf{A}_i\mathbf{A}_j^\top) \geq 2$, the strictly positive canonical correlations satisfy $t_{ij,r_{ij}} > t_{ij,1}^2$.

Corollary 1 (Finite-Dimensional Two-view Subspace Identifiability)**.** *Suppose the conditions of Theorem 5.1 hold, along with Assumption 2. For any finite representation dimension $d_{\mathcal{Z}} \geq r_{ij}$, the identifiability guarantees in Equation 6 hold exactly, where the rank-r_{ij} coordinate selector takes the explicit form $\mathbf{P}_{r_{ij}} := [\mathbf{I}_{r_{ij}}\ \mathbf{0}] \in \mathbb{R}^{r_{ij}\times d_{\mathcal{Z}}}$.*

Remark 2 (Spectral separation of higher-order Hermite modes). The Mehler–Hermite expansion (Lemma 2) decomposes the cross-view coupling into linear and higher-order (nonlinear) polynomial modes. Assumption 2 mathematically ensures that the weakest linear correlation ($t_{ij,r_{ij}}$)

strictly exceeds the strongest possible higher-order correlation, which is bounded by $t_{ij,1}^2$. This spectral gap guarantees that in the finite-dimensional regime (Corollary 1), the generalized CCA objective strictly prioritizes the first-order linear Hermite modes, securely isolating them in the leading r_{ij} coordinates. In a canonical representative of the whitened solution, the remaining coordinates, i.e., those with indices larger than r_{ij}, span higher-order Hermite modes orthogonal to the recovered first-order linear subspace [Han et al., 2026].

Theorem 5.2 (Infinite-Dimensional Multi-View Subspace Identifiability)**.** *Suppose the conditions of Theorem 5.1 hold for $N \geq 3$ view ensembles and let $\{\tilde{\mathbf{f}}_i^\star\}_{i=1}^N$ be any whitened population maximizer of the generalized CCA objective in Equation 2. For each view i, let $r_i := \dim(\mathcal{U}_i^{\mathrm{mv}})$ and let $\mathbf{U}_i^{\mathrm{mv}} \in \mathbb{R}^{d_{\mathcal{S}_i} \times r_i}$ be an orthonormal basis spanning the multi-view jointly correlated subspace $\mathcal{U}_i^{\mathrm{mv}}$.*

Assuming an infinite feature dimension, i.e., $d_{\mathcal{Z}} \to \infty$, there exist matrices $\mathbf{L}_i \in \mathbb{R}^{r_i \times d_{\mathcal{Z}}}$ with orthonormal rows and orthogonal matrices $\mathbf{O}_i \in O(r_i)$ such that

$$\mathbf{L}_i \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_i (\mathbf{U}_i^{\mathrm{mv}})^\top \mathbf{W}_i \mathbf{s}_i, \quad \forall i \in [N]. \tag{7}$$

Here, $\mathbf{L}_i$ extracts the jointly correlated component from the whitened representation up to an allowable within-view orthogonal post-transformation. If a canonical representative of the CCA solution is chosen, $\mathbf{L}_i$ can be taken as a coordinate selector.

Remark 3 (Subspace dimensionality and partial observability). Because the multi-view jointly correlated subspace $\mathcal{U}_i^{\mathrm{mv}}$ acts as a strict intersection filter, it isolates only those latent signals that are mutually visible across the entire N-view ensemble. Consequently, its dimension r_i is strictly bounded by, and typically lower than, the pairwise ranks r_{ij}. Furthermore, for latent factors shared only among a strict subset of views, the generalized CCA objective lacks sufficient global constraints to enforce alignment, causing the corresponding unshared representation dimensions to inherently drift [Yao et al., 2024]. The finite-dimensional counterpart of Equation 7 follows analogously to Corollary 1.

5.3 FINITE-SAMPLE STATISTICAL CONSISTENCY

We now quantify the statistical error incurred when the population second-order moments of fixed population-level CCA representations are replaced by empirical estimates. Throughout this subsection, the encoders $\{\mathbf{f}_i\}_{i=1}^N$ are treated as fixed and satisfy the population identifiability conditions established above. Thus, the rank-r_{ij} singular subspace of the population normalized cross-covariance coincides with the pairwise correlated source subspace $\mathcal{U}_{i|j}$, up to the orthogonal ambiguity characterized in Corollary 1. Since our identifiability results are driven by the singular/eigenspaces of normalized cross-covariances, the analysis reduces to (i) concentration of empirical covariances in operator norm and (ii) stability of singular-/eigenspaces under perturbations. Our analysis below controls statistical estimation error only; optimization and uniform generalization error for data-trained neural encoders are outside the scope of our analysis.

Empirical Normalized Cross-Covariances. Given n i.i.d. samples $\{(\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_N^{(t)})\}_{t=1}^n$ and fixed encoders $\{\mathbf{f}_i\}_{i=1}^N$, let $\mathbf{z}_i^{(t)} := \mathbf{f}_i(\mathbf{x}_i^{(t)})$. With $\bar{\mathbf{z}}_i$ and $\bar{\mathbf{z}}_j$ denoting the empirical means of the representations for views i and j, define the sample cross-covariance matrices as $\widehat{\mathbf{\Sigma}}_{ij} := \frac{1}{n}\sum_{t=1}^n (\mathbf{z}_i^{(t)} - \bar{\mathbf{z}}_i)(\mathbf{z}_j^{(t)} - \bar{\mathbf{z}}_j)^\top$. The empirical normalized cross-covariance is

$$\widehat{\mathbf{R}}_{ij} := \widehat{\mathbf{\Sigma}}_{ii}^{-1/2} \widehat{\mathbf{\Sigma}}_{ij} \widehat{\mathbf{\Sigma}}_{jj}^{-1/2}, \quad 1 \leq i < j \leq N, \tag{8}$$

with population counterpart $\mathbf{R}_{ij} := \mathbf{\Sigma}_{ii}^{-1/2} \mathbf{\Sigma}_{ij} \mathbf{\Sigma}_{jj}^{-1/2}$. Let $\widehat{\mathcal{U}}_{i|j}$ denote the rank-r_{ij} left singular subspace of $\widehat{\mathbf{R}}_{ij}$. By Corollary 1, this subspace estimates the pairwise correlated subspace $\mathcal{U}_{i|j}$ up to orthogonal ambiguity.

Additionally, we impose a mild tail condition on the *population-whitened* representations $\tilde{\mathbf{z}}_i := \mathbf{\Sigma}_{ii}^{-1/2}(\mathbf{z}_i - \mathbb{E}[\mathbf{z}_i])$.

Assumption 3 (Sub-Gaussian whitened representations)**.** There exists $\kappa > 0$ such that for all $i \in [N]$, $\sup_{\|\mathbf{u}\|_2=1} \|\mathbf{u}^\top \tilde{\mathbf{z}}_i\|_{\psi_2} \leq \kappa$.

Assumption 3 yields operator-norm concentration for $\widehat{\mathbf{\Sigma}}_{ij}$ and controls the error of the empirical whitening factors $\widehat{\mathbf{\Sigma}}_{ii}^{-1/2}$, which together imply concentration of the normalized cross-covariances $\widehat{\mathbf{R}}_{ij}$ in Equation 8. To translate this moment error into a *subspace* error, we require a spectral separation of the target singular subspaces: for each pair (i, j), define $\Delta_{ij} := \sigma_{r_{ij}}(\mathbf{R}_{ij}) - \sigma_{r_{ij}+1}(\mathbf{R}_{ij}) > 0$. Under Assumption 2, whenever $r_{ij} \geq 2$ one has the explicit lower bound $\Delta_{ij} \geq t_{ij,r_{ij}} - t_{ij,1}^2$, reflecting the strict dominance of linear modes over higher-order Mehler–Hermite components. We then provide finite-sample guarantees for subspace recovery.

Theorem 5.3 (Finite-sample subspace recovery)**.** *Suppose Assumption 3 holds, $d_{\mathcal{Z}} < \infty$, and $n \geq C_0 \kappa^4 (d_{\mathcal{Z}} + \log(N/\delta))$ so that the empirical within-view covariance matrices are nonsingular with high probability. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, simultaneously for all $1 \leq i < j \leq N$,*

$$\left\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\right\|_2 \;\leq\; C\,\kappa^2 \sqrt{\frac{d_{\mathcal{Z}} + \log(N/\delta)}{n}}, \tag{9}$$

for a universal constant $C > 0$. Moreover, on the event

$\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 \le \Delta_{ij}/2$,

$$\left\| \sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j}) \right\|_2 \;\le\; \frac{2\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2}{\Delta_{ij}}, \qquad (10)$$

and the same bound holds for the right singular subspace $\widehat{\mathcal{U}}_{j|i}$.

Remark 4 (Scaling with n). Combining Equation 9–Equation 10 gives the parametric $O_{\mathbb{P}}(n^{-1/2})$ rate. In particular, $\|\sin\Theta\|_2 \le \varepsilon$ is achieved once $n \gtrsim \kappa^4(d_{\mathcal{Z}} + \log(N/\delta))/(\Delta_{ij}^2\varepsilon^2)$.

To recover the multi-view jointly correlated subspace $\mathcal{U}_i^{\mathrm{mv}}$ from noisy pairwise estimates, we use an averaged-projector intersection filter. Let $\mathbf{P}_{i|j}$ and $\widehat{\mathbf{P}}_{i|j}$ denote the orthogonal projectors onto $\mathcal{U}_{i|j}$ and $\widehat{\mathcal{U}}_{i|j}$, and define

$$\mathbf{S}_i := \frac{1}{N-1}\sum_{j\neq i}\mathbf{P}_{i|j}, \qquad \widehat{\mathbf{S}}_i := \frac{1}{N-1}\sum_{j\neq i}\widehat{\mathbf{P}}_{i|j}.$$

By the spectral characterization of averaged projectors, $\mathcal{U}_i^{\mathrm{mv}}$ is the eigenspace of $\mathbf{S}_i$ associated with eigenvalue 1. We assume an intersection eigengap

$$\Gamma_i := 1 - \lambda_{r_i+1}(\mathbf{S}_i) > 0,$$

and estimate $\mathcal{U}_i^{\mathrm{mv}}$ by the top-r_i eigenspace of $\widehat{\mathbf{S}}_i$.

Corollary 2 (Consistency of the intersection filter)**.** *Under the conditions of Theorem 5.3, with probability at least* $1-\delta$*, for every* $i \in [N]$*,*

$$\begin{aligned}\|\sin\Theta(\widehat{\mathcal{U}}_i^{\mathrm{mv}}, \mathcal{U}_i^{\mathrm{mv}})\|_2 &\le \frac{\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2}{\Gamma_i} \\ &\le \frac{2}{\Gamma_i}\max_{j\neq i}\|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2.\end{aligned}$$

Consequently, $\widehat{\mathcal{U}}_i^{\mathrm{mv}}$ *is consistent at rate* $O_{\mathbb{P}}(n^{-1/2})$*, governed by* $\{\Delta_{ij}\}$ *and* Γ_i.

All proofs are deferred to the supplementary material.

6 EXPERIMENTS

In this section, we empirically test the main predictions of our identifiability analysis using controlled generative experiments. The experiments are designed to assess whether correlation-based multi-view objectives recover the ground-truth canonical source subspaces under known nonlinear view transformations. We first use fully synthetic data, where the latent factors, canonical spectra, and representation dimensions can be controlled exactly. We then use 3DIdent as a rendered stress test with annotated generative factors. Our setup extends established disentanglement learning benchmarks [Zimmermann et al., 2021, Matthes et al., 2023, Han et al., 2026] to the multi-view ($N = 3$) regime considered in this work.

6.1 EXPERIMENTAL SETUP

Datasets. We evaluate on two settings that provide ground-truth access to generative factors:

- **Synthetic Data**: We extend the single-generator setup of Zimmermann et al. [2021] to an $N = 3$ multi-view regime. Observations are produced by applying independent nonlinear generating functions to view-specific sources sampled from our additive model (Equation 1), enabling exact control over the latent distributions and mixing matrices.
- **3DIdent** [Daunhawer et al., 2023]: A physically rendered benchmark with 11 annotated generative factors. Since the factor grid is discrete, 3DIdent does not exactly instantiate our continuous source-level model. We therefore use it as a controlled rendered stress test. Specifically, we first sample continuous latent vectors according to Equation 1, normalize the factor coordinates to a common scale, and then select the rendered instance whose factor tuple is nearest in Euclidean distance. This procedure preserves controlled access to ground-truth factor labels while testing the learned representations on visually nontrivial observations.

Construction of Linear Mixing Matrices. To strictly control the pairwise canonical correlations of $\mathbf{R}_{ij}$ in Equation 4, we construct the mixing matrices $\{\mathbf{A}_i\}_{i=1}^3$ to match target singular values $\{t_{ij,k}\}_{k=1}^{r_{ij}} \subset [0,1)$ for the population normalized cross-covariance matrices,

$$\mathbf{R}_{ij} = (\mathbf{A}_i\mathbf{A}_i^\top + \mathbf{I})^{-1/2}\mathbf{A}_i\mathbf{A}_j^\top(\mathbf{A}_j\mathbf{A}_j^\top + \mathbf{I})^{-1/2}.$$

By assigning a shared right singular matrix $\mathbf{P} \in O(d)$ across all $\mathbf{A}_i$, the singular values of $\mathbf{R}_{ij}$ factorize element-wise as $t_{ij,k} = g_{i,k}g_{j,k}$, where $g_{i,k} := \sigma_{i,k}/\sqrt{1+\sigma_{i,k}^2}$ and $\sigma_{i,k}$ are the singular values of $\mathbf{A}_i$. For $N = 3$ views, this algebraic system uniquely yields $g_{i,k} = (t_{ij,k}t_{ki,k}/t_{jk,k})^{1/2}$, subsequently fixing $\sigma_{i,k} = g_{i,k}/\sqrt{1-g_{i,k}^2}$. We then construct $\mathbf{A}_i := \mathbf{U}_i\mathrm{diag}(\sigma_{i,1},\ldots,\sigma_{i,r_{ij}})\mathbf{P}^\top$, sampling independent random orthogonal matrices $\mathbf{U}_i \in O(d)$. This guarantees the exact prescribed canonical spectra while randomizing the signal subspace orientations.

Methods. We compare neural GCCA, the empirical objective aligned with our population analysis, against representative self-supervised objectives using the same encoder architecture and representation dimension: Barlow Twins, W-MSE, and InfoNCE.

- **Barlow Twins** [Zbontar et al., 2021]: Enforces invariance and minimizes redundancy through cross-correlation diagonalization.
- **W-MSE** [Ermolov et al., 2021]: Aligns whitened multi-view neural representations by minimizing mean-squared error.

Table 1: Comparison of the mean and maximum principal angles (PA_{mean}, PA_{max}) on synthetic data ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 5, \forall i \in [3]$) under Gaussian prior ($p_\phi$).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	PA_{max}	PA_{mean}	PA_{max}	PA_{mean}	PA_{max}
BarlowTwins	34.12 ±0.21	86.41 ±0.95	31.85 ±0.18	81.22 ±0.82	33.40 ±0.24	84.95 ±0.91
InfoNCE	4.95 ±0.06	8.15 ±0.52	5.12 ±0.09	8.42 ±0.38	**4.88 ±0.08**	**7.98 ±0.47**
W-MSE	5.08 ±0.08	8.31 ±0.41	**4.91 ±0.05**	**7.95 ±0.58**	5.03 ±0.07	8.24 ±0.44
GCCA	**4.85 ±0.05**	**7.92 ±0.35**	4.98 ±0.08	8.11 ±0.42	4.92 ±0.06	8.05 ±0.51

Table 2: Comparison of the mean and maximum principal angles (PA_{mean}, PA_{max}) on 3DIdent ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 11, \forall i \in [3]$) under Gaussian prior ($p_\phi$).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	PA_{max}	PA_{mean}	PA_{max}	PA_{mean}	PA_{max}
BarlowTwins	34.21 ±0.45	86.42 ±1.85	36.15 ±0.52	82.75 ±2.15	35.82 ±0.61	88.14 ±1.92
InfoNCE	8.05 ±0.18	11.21 ±0.65	8.12 ±0.22	10.85 ±0.82	7.98 ±0.15	11.42 ±0.91
W-MSE	8.15 ±0.25	10.95 ±0.78	7.92 ±0.19	11.15 ±0.55	8.11 ±0.21	10.88 ±0.88
GCCA	**7.82 ±0.16**	**10.51 ±0.62**	**7.85 ±0.18**	**10.45 ±0.71**	**7.80 ±0.15**	**10.48 ±0.68**

- **InfoNCE** [van den Oord et al., 2019]: The standard contrastive learning objective, maximizing mutual information between views, shown to be identifiable under mild conditions [Zimmermann et al., 2021, Matthes et al., 2023].
- **Generalized CCA (GCCA)** [Kettenring, 1971]: The correlation-based multi-view objective studied in this work, implemented as a neural version of SUMCOR generalized CCA and corresponding to Equation 2.

To approximate the view-specific encoding functions $\{\mathbf{f}_i\}_{i=1}^N$, we parameterize them as neural networks, leveraging their universal approximation capability. We use neural-network-parameterized GCCA as a practical empirical approximation to the population GCCA objective when data are generated via online sampling from the underlying distributions. This empirical training procedure is not covered by the finite-sample perturbation theorem, which analyzes fixed finite-dimensional representations. All baselines are implemented using their official repositories when available, otherwise we re-implement following the authors' specifications.

Metrics. We assess subspace recovery via principal angles between the span of the learned whitened representation and the ground-truth canonical subspace. Specifically, we report the mean principal angle (PA_{mean}) and the maximum principal angle (PA_{max}), following standard subspace identification practice [Ma and Li, 2020, Gao et al., 2017, Cai and Zhang, 2018].

In all experiments, view-specific encoders are trained using Adam with a fixed learning rate of 10^{-4}. The batch size is set to 1024 for synthetic data and 256 for 3DIdent. On the synthetic dataset with $d_{\mathcal{S}} = 5$, 100,001 training steps require approximately 2.5 hours, whereas 40,000 steps on 3DIdent take about 20 hours on a single RTX A5000 GPU. All tables report mean ± standard deviation over five random seeds.

6.2 VALIDATION OF THEORETICAL FINDINGS

Subspace Identifiability. We empirically validate subspace identifiability under our additive generative model in Equation 1. For simplicity, we present the results using Gaussian prior. Results for other admissible non-Gaussian distributions are deferred to the supplementary material. Setting the latent dimensions to $d_{\mathcal{C}} = d_{\mathcal{S}_i} = 5$, we independently sample the shared content $\mathbf{c}$ and view-private noise $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ from the selected distributions (details in the supplementary material). To invert this process, the view-specific encoders $\{\mathbf{f}_i\}_{i=1}^N$ are implemented as residual MLPs equipped with batch normalization to ensure stable optimization.

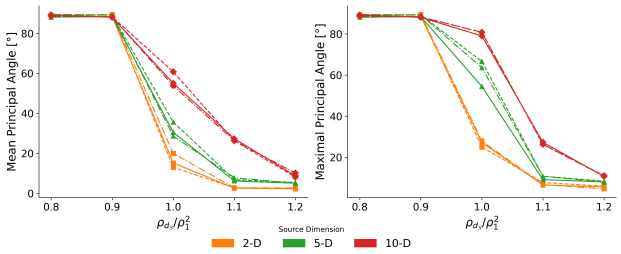

Figure 1: Ablation over the first-order canonical ratio $\rho_{d_{\mathcal{S}}}/\rho_1^2$ ($d_{\mathcal{S}} = d_{\mathcal{Z}}$). Left and right panels display the mean ($PA_{\text{mean}} \downarrow$) and maximum ($PA_{\text{max}} \downarrow$) principal angles, respectively. Colors indicate source dimensions, while solid, dashed, and dash-dotted lines denote the view-specific encoders $\tilde{\mathbf{f}}_1$, $\tilde{\mathbf{f}}_2$, and $\tilde{\mathbf{f}}_3$.

Table 1 and Table 2 show that GCCA, InfoNCE, and W-MSE all recover low-dimensional source subspaces with small principal angles, while Barlow Twins exhibits substantially larger subspace errors. On 3DIdent, GCCA achieves the lowest errors across all views and metrics. On synthetic data, GCCA is competitive with InfoNCE and W-MSE, with all three methods attaining comparable subspace recovery. These results are consistent with the view that whitening- and correlation-based objectives are well aligned with the subspace structure predicted by our theory, while the exact empirical ranking also reflects optimization and objective-specific effects outside the population identifiability analysis.

Ablation on First-order Canonical Dominance. Let $\rho_{d_{\mathcal{S}}}/\rho_1^2$ denote the first-order canonical dominance ratio. We ablate Assumption 2 by varying this ratio while keeping all other parameters fixed. Figure 1 reports the resulting mean and maximum principal angles on the synthetic dataset when the first-order canonical ratio are uniformly sampled from $[\rho_{d_{\mathcal{S}}}, \rho_1]$. When $\rho_{d_{\mathcal{S}}}/\rho_1^2 < 1$, at least

Table 3: Mean principal angle ($PA_{\text{mean}} \downarrow$) in the overcomplete regime ($d_{\mathcal{S}_i} = 5$, $d_{\mathcal{Z}} = 7, \forall i \in [3]$).

Encoder	Gaussian	Binomial	Gamma	Poisson	Hypergeometric
$\tilde{\mathbf{f}}_1$	5.42 ±0.08	5.58 ±0.12	5.47 ±0.05	5.61 ±0.14	5.53 ±0.09
$\tilde{\mathbf{f}}_2$	5.51 ±0.11	5.39 ±0.07	5.65 ±0.15	5.48 ±0.06	5.55 ±0.10
$\tilde{\mathbf{f}}_3$	5.45 ±0.09	5.62 ±0.13	5.50 ±0.08	5.41 ±0.11	5.59 ±0.07

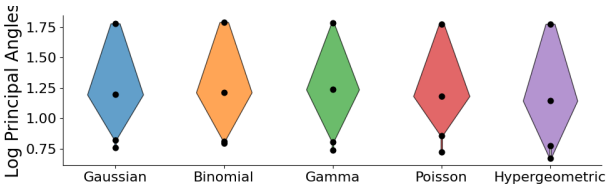

Figure 2: Log principal angles of encoder $\mathbf{f}_1$ in the undercomplete setup ($d_{\mathcal{S}_i} = 5$, $d_{\mathcal{Z}} = 4, \forall i \in [3]$). Black dots denote principal angles and the shaded region indicates the log-standard deviation.

one canonical direction violates the dominance condition, leading to persistently large $PA_{\max}$ and incomplete subspace recovery. Once the ratio exceeds 1, both PA_{mean} and $PA_{\max}$ decrease sharply, indicating full identifiability of all canonical directions. The transition is more abrupt at lower source dimensions, reflecting more concentrated recoverability in low-dimensional regimes.

Ablation on Dimension Mismatch. We study identifiability under dimension mismatch on synthetic dataset. In the under-complete regime ($d_{\mathcal{S}} > d_{\mathcal{Z}}$), Figure 2 shows the log-principal-angle distribution across source distributions. Across all cases, several directions attain small angles, while at least one direction remains bounded away from zero, indicating partial but not full subspace recovery. The variability is limited, suggesting robustness to distributional choice, yet the missing dimension prevents complete identifiability.

In the over-complete regime ($d_{\mathcal{S}} < d_{\mathcal{Z}}$), Table 3 reports consistently small mean principal angles across source distributions, indicating that the learned representations contain directions closely aligned with the source subspace. However, the additional representation coordinates are not uniquely determined by the first-order subspace criterion. Therefore, the over-complete setting supports subspace recovery but not global coordinate-level identifiability. Future work will investigate this higher-order structure more systematically.

7 DISCUSSION AND CONCLUSION

We reframed nonlinear multi-view CCA as a basis-invariant subspace identification problem. By considering an additive multi-view generative process, we proved that for $N \geq 3$ views, the population generalized CCA acts as a provable intersection filter, recovering the jointly correlated signal subspaces up to an orthogonal transformation while filtering out view-private components. The first-order canonical dominance condition theoretically guarantees the spectral separation between first-order shared Hermite components and higher-order components induced by nonlinear transformations via a multivariate Mehler–Hermite expansion. This spectral gap further enables finite-sample subspace recovery when the relevant population operators are estimated from finite-dimensional feature representations.

While our theory assumes matched representation and source dimensions and full-rank source covariance experiments demonstrate the behavior of CCA in capacity-mismatched regimes. Future work will systematically investigate the geometric isolation of higher-order Hermite components in these redundant dimensions, and extend this identifiability framework to rank-deficient source structures, e.g., partial observability, further bridging multivariate statistics with self-supervised learning.

References

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pages 15516–15528, 2022.

Horace B. Barlow, Tej P. Kaushal, and Graeme J. Mitchison. Finding minimum entropy codes. *Neural Computation*, 1(3):412–423, 1989.

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1): 60–89, 2018. doi: 10.1214/17-AOS1541.

Jean-François Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

Charles F. Dunkl and Yuan Xu. *Orthogonal polynomials of several variables*. Number 155 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2014.

Geoffrey Kennedy Eagleson. Polynomial expansions of bivariate distributions. *The Annals of Mathematical Statistics*, 35(3):1208–1215, 1964.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3015–3024. PMLR, 2021.

Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017. doi: 10.1214/16-AOS1519.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pages 217–227. PMLR, 2020.

Zhiwei Han, Stefan Matthes, and Hao Shen. Provable affine identifiability of nonlinear CCA under latent distributional priors. In *International Conference on Artificial Intelligence and Statistics*, 2026. URL `https://arxiv.org/abs/2510.04758`.

Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8613–8634. PMLR, 2022.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, pages 3765–3773, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. doi: 10.1016/S0893-6080(98)00140-3.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019.

András Kalapos and Bálint Gyires-Tóth. Whitening consistently improves self-supervised learning. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 448–453. IEEE, 2024. doi: 10.1109/ICMLA61862.2024.00066.

Paris A. Karakasis and Nicholas D. Sidiropoulos. Revisiting deep generalized canonical correlation analysis. *IEEE Transactions on Signal Processing*, 71:4392–4406, 2023. doi: 10.1109/TSP.2023.3333212.

Paris A. Karakasis and Nicholas D. Sidiropoulos. Subspace clustering of subspaces: Unifying canonical correlation analysis and subspace clustering, 2025. URL `https://arxiv.org/abs/2509.18653`.

Jon R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

Henry Oliver Lancaster. The structure of bivariate distributions. *The Annals of Mathematical Statistics*, 29(3): 719–736, 1958.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 2020.

Qi Lyu and Xiao Fu. Nonlinear multiview analysis: Identifiability and neural network-assisted implementation. *IEEE Transactions on Signal Processing*, 68:2697–2712, 2020. doi: 10.1109/TSP.2020.2986351.

Qi Lyu and Xiao Fu. Provable subspace identification under post-nonlinear mixtures. In *Advances in Neural Information Processing Systems*, volume 35, pages 1554–1567, 2022.

Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*, 2022.

Zhuang Ma and Xiaodong Li. Subspace perspective on canonical correlation analysis: Dimension reduction and minimax rates. *Bernoulli*, 26(1):432–470, 2020. doi: 10.3150/19-BEJ1131.

Stefan Matthes, Zhiwei Han, and Hao Shen. Towards a unified framework of contrastive learning for disentangled representations. In *Advances in Neural Information Processing Systems*, volume 36, pages 67459–67470, 2023.

F. G. Mehler. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *Journal für die reine und angewandte Mathematik*, 66:161–176, 1866.

Teodora Pandeva and Patrick Forré. Multi-view independent component analysis with shared and individual sources. In *Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1639–1650. PMLR, 2023.

Anastasia Podosinnikova, Francis R. Bach, and Simon Lacoste-Julien. Beyond CCA: Moment matching for multi-view models. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 458–467. PMLR, 2016.

Borja Rodríguez Gálvez, Arno Blaas, Pau Rodríguez, Adam Goliński, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29143–29160. PMLR, 2023.

Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Nicholas D. Sidiropoulos and Mikael Sørensen. Canonical identification of autoregressive nonlinear systems. In *2022 Asilomar Conference on Signals, Systems, and Computers*, pages 1021–1025. IEEE, 2022. doi: 10.1109/IEEECONF56349.2022.10052002.

Mikael Sørensen, Charilaos I. Kanatsoulis, and Nicholas D. Sidiropoulos. Generalized canonical correlation analysis: A subspace intersection approach. *IEEE Transactions on Signal Processing*, 69:2452–2467, 2021. doi: 10.1109/TSP.2021.3061218.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34465–34492, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL `https://arxiv.org/abs/1807.03748`.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467, 2021.

Xi Weng, Lei Huang, Lei Zhao, Rao Anwer, Salman H. Khan, and Fahad Shahbaz Khan. An investigation into whitening loss for self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 29748–29760, 2022.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55389–55433. PMLR, 2024.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR, 2021.

Supplementary Material

Zhiwei Han[1] **Stefan Matthes**[1] **Hao Shen**[1,2]

[1]Chair of Data Processing, Technical University of Munich, Germany
[2]fortiss GmbH, Munich, Germany

CONTENTS OF THE SUPPLEMENTARY MATERIAL

A OVERVIEW OF THE SUPPLEMENTARY MATERIAL

This supplementary material is organized as follows.

We first collect auxiliary tools used throughout the analysis, including the whitened encoder classes (Definition 2), the population CCA objective (Equation 11), pushforward identities (Lemma 3), reparameterization invariance (Lemma 6), maximizer correspondence (Lemma 6, item 2), and the canonical factorization of the pairwise source density (Lemma 1).

We then provide the proof of Lemma 2, which establishes the normalized multivariate Mehler–Hermite expansion underlying the spectral analysis.

The subsequent sections contain complete proofs of the main theoretical claims. In particular, we prove the infinite-dimensional two-view subspace identifiability result in Theorem 5.1, the finite-dimensional two-view specialization in Corollary 1, the infinite-dimensional multi-view identifiability result in Theorem 5.2, the consistency of the multi-view intersection filter in Corollary 2, and the finite-sample subspace recovery guarantee in Theorem 5.3. These proofs make explicit how reparameterization invariance, whitening-induced canonicalization, orthogonal-polynomial diagonalization, spectral separation, and standard matrix perturbation arguments combine to yield the stated identifiability and consistency guarantees.

Finally, we report additional experimental results complementing the main text. These results further evaluate the behavior of the proposed correlation-based multi-view objective across different latent distributional settings and provide additional evidence for the role of the stated spectral and dimensional assumptions.

B AUXILIARY TOOLS FROM PRIOR WORKS

B.1 REPARAMETERIZATION INVARIANCE

Definition 2 (Whitened Encoder Classes with Post-orthogonal Closure)**.** Let $\mathcal{H}_{\mathcal{X}} \subset L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ be Hilbert spaces of square-integrable vector-valued functions. For base encoders $\forall \mathbf{f} \in \mathcal{H}_{\mathcal{X}}$ with covariance matrix $\mathbf{\Sigma_f}$ symmetric positive definite, choose any $\mathbf{W_f} \in \mathrm{GL}(d_{\mathcal{Z}})$ such that $\mathbf{W_f}\,\mathbf{\Sigma_f}\,\mathbf{W_f}^\top = \mathbf{I}_{d_{\mathcal{Z}}}$. Define the whitened encoder class on domain $\mathcal{X}$ by

$$\tilde{\mathcal{F}}_{\mathcal{X}} := \Big\{ \mathbf{W_f}\big(\mathbf{f} - \mathbb{E}[\mathbf{f}(\mathbf{x})]\big) \;:\; \mathbf{f} \in \mathcal{H}_{\mathcal{X}},\ \mathbf{\Sigma_f} \succ 0 \Big\},$$

and $\mathbf{Q}\,\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \forall \tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \mathbf{Q} \in O(d_{\mathcal{Z}})$. Analogously, we define whitened encoder class $\tilde{\mathcal{F}}'_{\mathcal{X}'}$ for all base encoders $\mathbf{f}' \in \mathcal{H}_{\mathcal{X}'} \subset L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$.

Assumption 4 (Whitened Encoder Classes with Post-orthogonal Closure)**.** Let $\mathcal{H}_{\mathcal{X}} \subset L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ be Hilbert spaces of square-integrable vector-valued functions. For base encoders $\forall \mathbf{f} \in \mathcal{H}_{\mathcal{X}}$ with covariance matrix $\mathbf{\Sigma_f}$ symmetric positive definite, choose any $\mathbf{W_f} \in \mathrm{GL}(d_{\mathcal{Z}})$ such that $\mathbf{W_f}\,\mathbf{\Sigma_f}\,\mathbf{W_f}^\top = \mathbf{I}_{d_{\mathcal{Z}}}$. Define the whitened encoder class on domain $\mathcal{X}$ by

$$\tilde{\mathcal{F}}_{\mathcal{X}} := \Big\{ \mathbf{W_f}\big(\mathbf{f} - \mathbb{E}[\mathbf{f}(\mathbf{x})]\big) \;:\; \mathbf{f} \in \mathcal{H}_{\mathcal{X}},\ \mathbf{\Sigma_f} \succ 0 \Big\},$$

and $\mathbf{Q}\,\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \forall \tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \mathbf{Q} \in O(d_{\mathcal{Z}})$. Analogously, we define whitened encoder class $\tilde{\mathcal{F}}'_{\mathcal{X}'}$ for all base encoders $\mathbf{f}' \in \mathcal{H}_{\mathcal{X}'} \subset L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$.

Definition 3 (CCA Population Objective.)**.** Given the whitened encoder class in Assumption 4, nonlinear CCA solves the following population optimization problem:

$$\max_{\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}},\, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}} J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\big(\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x}),\, \tilde{\mathbf{f}}'(\mathbf{x}'))\big), \tag{11}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value and $\mathrm{Cov}(\cdot)$ the population covariance.

Lemma 3 (Pushforward identities and whitening preservation)**.** *Let ground-truth latent pair $(\mathbf{s}, \mathbf{s}')$ be square-integrable random vectors on $\mathcal{S} \times \mathcal{S}$ with joint law $P_{\mathbf{ss}'}$. Let $\mathbf{g} : \mathcal{S} \to \mathcal{X}$ and $\mathbf{g}' : \mathcal{S} \to \mathcal{X}'$ be Borel-measurable and observation pair $(\mathbf{x}, \mathbf{x}') = (g(\mathbf{s}), g'(\mathbf{s}'))$ with joint law $P_{\mathbf{xx}'} = (\mathbf{g}, \mathbf{g}')_{\#} P_{\mathbf{ss}'}$. Then the following properties hold.*

1. ***Square-integrability.*** *For any Borel-measurable $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$ with finite second moments on $\mathcal{X}$, the composition $\tilde{\mathbf{f}} \circ \mathbf{g}$ also have finite second moments on $\mathcal{S}$. Analogously, the same property also applies to all $\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$.*
2. ***Expectation and Covariance Preservation.*** *For all square-integrable $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$, $\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$,*

$$\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})] = \mathbb{E}[\tilde{\mathbf{f}} \circ \mathbf{g}(\mathbf{s})], \quad \mathbb{E}[\tilde{\mathbf{f}}'(\mathbf{x}')] = \mathbb{E}[\tilde{\mathbf{f}}' \circ \mathbf{g}'(\mathbf{s}')],$$
$$\mathrm{Cov}\big(\tilde{\mathbf{f}}(\mathbf{x})\big) = \mathrm{Cov}\big(\tilde{\mathbf{f}} \circ \mathbf{g}(\mathbf{s})\big), \quad \mathrm{Cov}\big(\tilde{\mathbf{f}}'(\mathbf{x}')\big) = \mathrm{Cov}\big(\tilde{\mathbf{f}}' \circ \mathbf{g}'(\mathbf{s}')\big),$$
$$\mathrm{Cov}\big(\tilde{\mathbf{f}}(\mathbf{x}),\, \tilde{\mathbf{f}}'(\mathbf{x}')\big) = \mathrm{Cov}\big(\tilde{\mathbf{f}} \circ \mathbf{g}(\mathbf{s}),\, \tilde{\mathbf{f}}' \circ \mathbf{g}'(\mathbf{s}')\big).$$

3. ***Whitening Preservation.*** *If $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$ and $\tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$ are whitened with respect to $P_{\mathbf{x}}$ and $P_{\mathbf{x}'}$, respectively. Then $\tilde{\mathbf{f}} \circ \mathbf{g}$ and $\tilde{\mathbf{f}}' \circ \mathbf{g}'$ are whitened under $P_{\mathbf{s}}$ and $P_{\mathbf{s}'}$.*

Proof.

1. Square-integrability.

Let $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$ be Borel-measurable and square-integrable under $P_{\mathbf{x}}$ and $P_{\mathbf{x}'}$, respectively. That is,

$$\int_{\mathcal{X}} \|\tilde{\mathbf{f}}(\mathbf{a})\|^2 \, dP_{\mathbf{x}}(\mathbf{a}) < \infty, \qquad \int_{\mathcal{X}'} \|\tilde{\mathbf{f}}'(\mathbf{a})\|^2 \, dP_{\mathbf{x}'}(\mathbf{b}) < \infty.$$

Because $P_{\mathbf{x}} = \mathbf{g}_{\#}P_{\mathbf{s}}$ and $P_{\mathbf{x}'} = \mathbf{g}'_{\#}P_{\mathbf{s}'}$ are the pushforward meausre of $P_{\mathbf{s}}$ and $P_{\mathbf{s}'}$. The pushforward identity gives, for any nonnegative measurable $\boldsymbol{\varphi} \in L^1(P_{\mathbf{x}}), \boldsymbol{\psi} \in L^1(P_{\mathbf{x}'})$,

$$\int_{\mathcal{X}} \boldsymbol{\varphi}(\mathbf{a})\, dP_{\mathbf{x}}(\mathbf{a}) = \int_{\mathcal{S}} (\boldsymbol{\varphi} \circ \mathbf{g})(\mathbf{b})\, dP_{\mathbf{s}}(\mathbf{b}) \quad \text{and} \quad \int_{\mathcal{X}'} \boldsymbol{\psi}(\mathbf{a})\, dP_{\mathbf{x}'}(\mathbf{a}) = \int_{\mathcal{S}} (\boldsymbol{\psi} \circ \mathbf{g})(\mathbf{b})\, dP_{\mathbf{s}'}(\mathbf{b}). \tag{12}$$

Applying this to $\boldsymbol{\varphi}(\mathbf{x}) = \|\mathbf{f}(\mathbf{x})\|^2$ yields

$$\mathbb{E}\big[\|(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})\|^2\big] = \int_{\mathcal{S}} \|(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{a})\|^2\, dP_{\mathbf{s}}(\mathbf{a}) = \int_{\mathcal{X}} \|\tilde{\mathbf{f}}(\mathbf{a})\|^2\, dP_{\mathbf{x}}(\mathbf{a}) < \infty,$$
$$\mathbb{E}\big[\|(\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')\|^2\big] = \int_{\mathcal{S}} \|(\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{b})\|^2\, dP_{\mathbf{s}'}(\mathbf{b}) = \int_{\mathcal{X}'} \|\tilde{\mathbf{f}}'(\mathbf{b})\|^2\, dP_{\mathbf{x}}(\mathbf{b}) < \infty.$$

Thus $\tilde{\mathbf{f}} \circ \mathbf{g} \in L^2(P_{\mathbf{s}}; \mathbb{R}^{d_{\mathcal{Z}}})$ and $\tilde{\mathbf{f}}' \circ \mathbf{g}' \in L^2(P_{\mathbf{s}'}; \mathbb{R}^{d_{\mathcal{Z}}})$, as claimed.

2. Expectation and covariance preservation.

Let $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}$ be Borel and square-integrable under $P_{\mathbf{x}}$ and $P_{\mathbf{x}'}$, respectively.

Expectations. Apply Equation 12 componentwise with $\boldsymbol{\varphi} = \tilde{\mathbf{f}}_j$ and $\boldsymbol{\psi} = \tilde{\mathbf{f}}'_j$:

$$\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})] \;=\; \mathbb{E}[(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})], \qquad \mathbb{E}[\tilde{\mathbf{f}}'(\mathbf{x}')] \;=\; \mathbb{E}[(\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')].$$

Second moments and (marginal) covariances. Using Equation 12 with $\boldsymbol{\varphi}(\mathbf{x}) = \tilde{\mathbf{f}}(\mathbf{x})\tilde{\mathbf{f}}(\mathbf{x})^\top$ (integrable by Cauchy-Schwarz), $\forall i, j \in [1, \cdots, d_{\mathcal{Z}}]$,

$$\mathbb{E}\big[\tilde{\mathbf{f}}(\mathbf{x})\tilde{\mathbf{f}}(\mathbf{x})^\top\big] \;=\; \mathbb{E}\big[(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})\,(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})^\top\big].$$

Subtracting outer products of the means yields

$$\mathrm{Cov}\big(\tilde{\mathbf{f}}(\mathbf{x})\big) = \mathbb{E}\big[\tilde{\mathbf{f}}(\mathbf{x})\tilde{\mathbf{f}}(\mathbf{x})^\top\big] - \mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})]\,\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})]^\top = \mathrm{Cov}\big((\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})\big).$$

The same argument applies to $\tilde{\mathbf{f}}'$:

$$\mathrm{Cov}\big(\tilde{\mathbf{f}}'(\mathbf{x}')\big) = \mathrm{Cov}\big((\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')\big).$$

Cross-covariances. For the joint law $P_{\mathbf{x}\mathbf{x}'} = (\mathbf{g}, \mathbf{g}')_{\#}P_{\mathbf{s}\mathbf{s}'}$ and any integrable $\Phi : \mathcal{X} \times \mathcal{X}' \to \mathbb{R}$, the well-established generalization of the pushforward measure to joint mappings is as follows:

$$\int_{\mathcal{X}\times\mathcal{X}'} \Phi(\mathbf{a}, \mathbf{b})\, dP_{\mathbf{x}\mathbf{x}'}(\mathbf{a}, \mathbf{b}) \;=\; \int_{\mathcal{S}\times\mathcal{S}'} \Phi(\mathbf{g}(\mathbf{c}), \mathbf{g}'(\mathbf{d}))\, dP_{\mathbf{s}\mathbf{s}'}(\mathbf{c}, \mathbf{d}). \tag{13}$$

By Equation 13 with the matrix-valued integrand $\Phi(\mathbf{x}) = \tilde{\mathbf{f}}(\mathbf{x})\,\tilde{\mathbf{f}}'(\mathbf{x}')^\top$ (integrable by Cauchy–Schwarz), $\forall i, j \in [1, \cdots, d_{\mathcal{Z}}]$,

$$\mathbb{E}\big[\tilde{\mathbf{f}}(\mathbf{x})\,\tilde{\mathbf{f}}'(\mathbf{x}')^\top\big] \;=\; \mathbb{E}\big[(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})\,(\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')^\top\big].$$

Subtracting the product of means (already matched above) gives

$$\mathrm{Cov}\big(\tilde{\mathbf{f}}(\mathbf{x}),\, \tilde{\mathbf{f}}'(\mathbf{x}')\big) = \mathrm{Cov}\big((\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s}),\, (\tilde{\mathbf{f}}' \circ \mathbf{g}')(\mathbf{s}')\big).$$

All three identities are thus established.

3. Whitening preservation. If $\mathbb{E}[\tilde{\mathbf{f}}(\mathbf{x})] = \mathbf{0}$ and $\mathrm{Cov}(\tilde{\mathbf{f}}(\mathbf{x})) = \mathbf{I}$, then by step 2, we get $\mathbb{E}[\tilde{\mathbf{f}} \circ g(\mathbf{s})] = \mathbf{0}$ and $\mathrm{Cov}(\tilde{\mathbf{f}} \circ g(\mathbf{s})) = \mathbf{I}$. The argument for $\tilde{\mathbf{f}}'$ is identical. □

Lemma 4 (Bijection under composition)**.** *Let $\mathcal{S}, \mathcal{X}, \mathcal{X}'$ be standard Borel spaces and $\mathbf{g}, \mathbf{g}'$ be Borel-measurable and injective. Assume the pushforward measures $P_{\mathbf{x}} = \mathbf{g}_{\#}P_{\mathbf{s}}$ and $P_{\mathbf{x}'} = \mathbf{g}'_{\#}P_{\mathbf{s}'}$. Define the whitened representable latent classes as in Definition 2. Then:*

1. ***Composition isometries.*** *The composition operators*

$$C_{\mathbf{g}} : L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}}) \to L^2(P_{\mathbf{s}}; \mathbb{R}^{d_{\mathcal{Z}}}), \quad \tilde{\mathbf{f}} \mapsto \tilde{\mathbf{f}} \circ \mathbf{g},$$
$$C_{\mathbf{g}'} : L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}}) \to L^2(P_{\mathbf{s}'}; \mathbb{R}^{d_{\mathcal{Z}}}), \quad \tilde{\mathbf{f}}' \mapsto \tilde{\mathbf{f}}' \circ \mathbf{g}',$$

are linear isometries.

2. ***Bijection on whitened classes.*** *The map*

$$\Psi:\ \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'} \longrightarrow \tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}'}, \qquad (\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \longmapsto (\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}')$$

is a bijection (modulo null sets), with inverse

$$\Psi^{-1}:\ (\tilde{\mathbf{h}}, \tilde{\mathbf{h}}') \longmapsto (\tilde{\mathbf{h}} \circ \mathbf{g}^{-1},\ \tilde{\mathbf{h}}' \circ \mathbf{g}'^{-1}),$$

where $\mathbf{g}^{-1}$ and $\mathbf{g}'^{-1}$ are the Borel inverses on the Borel images $\mathbf{g}(\mathcal{S})$ and $\mathbf{g}'(\mathcal{S}')$, respectively. For any $Q, Q' \in O(d_{\mathcal{Z}})$,

$$\Psi(Q\tilde{\mathbf{f}}, Q'\tilde{\mathbf{f}}') = (Q(\tilde{\mathbf{f}} \circ \mathbf{g}),\ Q'(\tilde{\mathbf{f}}' \circ \mathbf{g}')).$$

Proof.

1. Isometry.

Since $\forall \tilde{\mathbf{f}} \in \tilde{F}_{\mathcal{X}}$,

$$\|\tilde{\mathbf{f}} \circ \mathbf{g}\|^2_{L^2(P_{\mathbf{s}})} = \int_{\mathcal{S}} \|(\tilde{\mathbf{f}} \circ \mathbf{g})(\mathbf{s})\|^2 \, dP_{\mathbf{s}}(\mathbf{s}) = \int_{\mathcal{X}} \|\tilde{\mathbf{f}}(\mathbf{x})\|^2 \, d(\mathbf{g}_{\#} P_{\mathbf{s}})(\mathbf{x}) = \|\tilde{\mathbf{f}}\|^2_{L^2(P_{\mathbf{x}})},$$

so $C_{\mathbf{g}}$ is a linear isometry. The argument for $C_{\mathbf{g}'}$ is identical.

2. Measurable inverses on images.

Since $\mathcal{S}, \mathcal{X}$ are standard Borel and $\mathbf{g}$ is Borel and injective, the Lusin-Souslin theorem implies $\mathbf{g}(\mathcal{S})$ is Borel in $\mathcal{X}$ and $\mathbf{g}^{-1} : \mathbf{g}(\mathcal{S}) \to \mathcal{S}$ is Borel. The same holds for $\mathbf{g}'$.

3. Surjectivity.

Let $\tilde{\mathbf{h}} \in \tilde{\mathcal{F}}_{\mathcal{S}}$. Define $\tilde{\mathbf{f}} := \tilde{\mathbf{h}} \circ \mathbf{g}^{-1}$ on $\mathbf{g}(\mathcal{S})$ and extend arbitrarily to a Borel function on $\mathcal{X}$. Then, by Step 2, $\tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}$ and $(\tilde{\mathbf{f}} \circ \mathbf{g})(s) = \tilde{\mathbf{h}}(s)$. The same construction gives the primed component. Thus Ψ is surjective

4. Injectivity.

If $\Phi(\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}'_1) = \Phi(\tilde{\mathbf{f}}_2, \tilde{\mathbf{f}}'_2)$, then $\tilde{\mathbf{f}}_1 \circ \mathbf{g} = \tilde{\mathbf{f}}_2 \circ \mathbf{g}$, hence $\tilde{\mathbf{f}}_1 = \tilde{\mathbf{f}}_2$. by applying $\mathbf{g}^{-1}$ on $\mathbf{g}(\mathcal{S})$; similarly for the primed side. Therefore Ψ is injective.

Combining the steps yields the stated bijection and properties. □

Lemma 5 (Continuity of the Whitening Map)**.** *Let $\{\, \mathbf{f}_n \,\}_{n \geq 1} \subset L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ be a sequence of square-integrable vector-valued functions such that $\mathbf{f}_n \to \mathbf{f}$ in $L^2(P_{\mathbf{x}})$. Denote*

$$\boldsymbol{\mu}_n := \mathbb{E}[\mathbf{f}_n(\mathbf{x})], \qquad \boldsymbol{\mu} := \mathbb{E}[\mathbf{f}(\mathbf{x})], \qquad \boldsymbol{\Sigma}_{\mathbf{f}_n} := \mathrm{Cov}(\mathbf{f}_n(\mathbf{x})), \qquad \boldsymbol{\Sigma}_{\mathbf{f}} := \mathrm{Cov}(\mathbf{f}(\mathbf{x})).$$

Assume $\boldsymbol{\Sigma}_{\mathbf{f}} \succ 0$, and let

$$\mathbf{W}_{\mathbf{f}_n} := \boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2}, \qquad \mathbf{W}_{\mathbf{f}} := \boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2}.$$

Then, for all n sufficiently large, $\boldsymbol{\Sigma}_{\mathbf{f}_n} \succ 0$ and

$$\mathbf{W}_{\mathbf{f}_n}\big(\mathbf{f}_n - \boldsymbol{\mu}_n\big) \xrightarrow[n\to\infty]{L^2} \mathbf{W}_{\mathbf{f}}\big(\mathbf{f} - \boldsymbol{\mu}\big). \tag{14}$$

Moreover, if $\mathbf{U}_{\mathbf{f}_n} \in \mathrm{GL}(d_{\mathcal{Z}})$ are any whiteners satisfying $\mathbf{U}_{\mathbf{f}_n} \boldsymbol{\Sigma}_{\mathbf{f}_n} \mathbf{U}_{\mathbf{f}_n}^{\top} = \mathbf{I}_{d_{\mathcal{Z}}}$, then there exist orthogonal matrices $\mathbf{Q}_n \in O(d_{\mathcal{Z}})$ such that

$$\mathbf{Q}_n\, \mathbf{U}_{\mathbf{f}_n}(\mathbf{f}_n - \boldsymbol{\mu}_n) \xrightarrow[n\to\infty]{L^2} \mathbf{W}_{\mathbf{f}}(\mathbf{f} - \boldsymbol{\mu}).$$

Proof.

1. Convergence of moments.

Since $\mathbf{f}_n \to \mathbf{f}$ in $L^2(P_{\mathbf{x}})$, we have $\|\mathbf{f}_n - \mathbf{f}\|^2_{L^2} = \mathbb{E}[\|\mathbf{f}_n - \mathbf{f}\|^2] \to 0$. By the Cauchy-Schwarz inequality,

$$\|\boldsymbol{\mu}_n - \boldsymbol{\mu}\| = \|\mathbb{E}[\mathbf{f}_n - \mathbf{f}]\| \leq \mathbb{E}[\|\mathbf{f}_n - \mathbf{f}\|] \leq \|\mathbf{f}_n - \mathbf{f}\|_{L^2} \longrightarrow 0.$$

For the second moment matrices, each entry of $\mathbb{E}[\mathbf{f}_n\mathbf{f}_n^\top]$ converges to the corresponding entry of $\mathbb{E}[\mathbf{f}\mathbf{f}^\top]$ because

$$\|\mathbf{f}_n\mathbf{f}_n^\top - \mathbf{f}\mathbf{f}^\top\|_{\mathrm{F}} \;\le\; (\|\mathbf{f}_n\| + \|\mathbf{f}\|)\,\|\mathbf{f}_n - \mathbf{f}\|,$$

and taking expectations yields

$$\|\mathbb{E}[\mathbf{f}_n\mathbf{f}_n^\top] - \mathbb{E}[\mathbf{f}\mathbf{f}^\top]\| \le \mathbb{E}[\|\mathbf{f}_n\mathbf{f}_n^\top - \mathbf{f}\mathbf{f}^\top\|] \le (\|\mathbf{f}_n\|_{L^2} + \|\mathbf{f}\|_{L^2})\,\|\mathbf{f}_n - \mathbf{f}\|_{L^2} \;\longrightarrow\; 0,$$

so $\mathbb{E}[\mathbf{f}_n\mathbf{f}_n^\top] \to \mathbb{E}[\mathbf{f}\mathbf{f}^\top]$ in operator norm.

Combining the above,

$$\boldsymbol{\Sigma}_{\mathbf{f}_n} = \mathbb{E}[(\mathbf{f}_n - \boldsymbol{\mu}_n)(\mathbf{f}_n - \boldsymbol{\mu}_n)^\top] = \mathbb{E}[\mathbf{f}_n\mathbf{f}_n^\top] - \boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top \xrightarrow[n\to\infty]{\|\cdot\|} \mathbb{E}[\mathbf{f}\mathbf{f}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top = \boldsymbol{\Sigma}_{\mathbf{f}}.$$

2. Positive definiteness and bounded whitening operators.

Since $\boldsymbol{\Sigma}_{\mathbf{f}} \succ 0$, all its eigenvalues are positive. Let $\sigma_{\min} > 0$ denote its smallest eigenvalue, i.e.,

$$\sigma_{\min} := \sigma_{\min}(\boldsymbol{\Sigma}_{\mathbf{f}}) \quad \text{so that} \quad \boldsymbol{\Sigma}_{\mathbf{f}} \succeq \sigma_{\min}\mathbf{I}.$$

Because $\boldsymbol{\Sigma}_{\mathbf{f}_n} \to \boldsymbol{\Sigma}_{\mathbf{f}}$ in operator norm, we have

$$\|\boldsymbol{\Sigma}_{\mathbf{f}_n} - \boldsymbol{\Sigma}_{\mathbf{f}}\| \xrightarrow[n\to\infty]{} 0.$$

By Weyl's eigenvalue perturbation inequality,

$$\forall i \in [1, \cdots, d_{\mathcal{Z}}], \quad |\sigma_i(\boldsymbol{\Sigma}_{\mathbf{f}_n}) - \sigma_i(\boldsymbol{\Sigma}_{\mathbf{f}})| \le \|\boldsymbol{\Sigma}_{\mathbf{f}_n} - \boldsymbol{\Sigma}_{\mathbf{f}}\|.$$

Hence for sufficiently large n,

$$\sigma_{\min}(\boldsymbol{\Sigma}_{\mathbf{f}_n}) \ge \sigma_{\min}(\boldsymbol{\Sigma}_{\mathbf{f}}) - \|\boldsymbol{\Sigma}_{\mathbf{f}_n} - \boldsymbol{\Sigma}_{\mathbf{f}}\| \ge \tfrac{\sigma_{\min}}{2} > 0.$$

Therefore each $\boldsymbol{\Sigma}_{\mathbf{f}_n}$ is symmetric positive definite and satisfies

$$\boldsymbol{\Sigma}_{\mathbf{f}_n} \succeq \tfrac{\sigma_{\min}}{2}\mathbf{I}.$$

Taking inverse square roots preserves the Löwner order on the SPD cone, so

$$\|\mathbf{W}_{\mathbf{f}_n}\| = \|\boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2}\| = \big(\sigma_{\min}(\boldsymbol{\Sigma}_{\mathbf{f}_n})\big)^{-1/2} \le (\sigma_{\min}/2)^{-1/2}.$$

Thus the sequence $\{\mathbf{W}_{\mathbf{f}_n}\}$ is uniformly bounded in operator norm.

3. Continuity of matrix square roots.

Since $\boldsymbol{\Sigma}_{\mathbf{f}} \succ 0$ and $\boldsymbol{\Sigma}_{\mathbf{f}_n} \to \boldsymbol{\Sigma}_{\mathbf{f}}$ in operator norm, there exist constants

$$m := \tfrac{\sigma_{\min}(\boldsymbol{\Sigma}_{\mathbf{f}})}{2} > 0, \qquad M := \|\boldsymbol{\Sigma}_{\mathbf{f}}\| + 1,$$

and an integer N such that for all $n \ge N$, the spectra satisfy $\sigma(\boldsymbol{\Sigma}_{\mathbf{f}_n}),\, \sigma(\boldsymbol{\Sigma}_{\mathbf{f}}) \subset [m, M]$.

Fix $\varepsilon > 0$. By the Weierstrass approximation theorem, there exists a polynomial p such that

$$\sup_{\sigma\in[m,M]} \big|\sigma^{-1/2} - p(\sigma)\big| < \varepsilon/3.$$

Using the continuous functional calculus for self-adjoint matrices, we have

$$\big\|\boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2} - \boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2}\big\| \;\le\; \big\|\boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2} - p(\boldsymbol{\Sigma}_{\mathbf{f}_n})\big\| + \big\|p(\boldsymbol{\Sigma}_{\mathbf{f}_n}) - p(\boldsymbol{\Sigma}_{\mathbf{f}})\big\| + \big\|p(\boldsymbol{\Sigma}_{\mathbf{f}}) - \boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2}\big\|.$$

The first and third terms are each bounded by $\sup_{\sigma\in[m,M]} |\sigma^{-1/2} - p(\sigma)| < \varepsilon/3$. For the middle term, note that $p(A) = \sum_k a_k A^k$ is a finite polynomial, hence norm-continuous; since $\boldsymbol{\Sigma}_{\mathbf{f}_n} \to \boldsymbol{\Sigma}_{\mathbf{f}}$, we have $\|p(\boldsymbol{\Sigma}_{\mathbf{f}_n}) - p(\boldsymbol{\Sigma}_{\mathbf{f}})\| < \varepsilon/3$ for all large n. Therefore $\|\boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2} - \boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2}\| < \varepsilon$ for n sufficiently large, i.e.

$$\mathbf{W}_{\mathbf{f}_n} = \boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2} \xrightarrow[n\to\infty]{\|\cdot\|} \boldsymbol{\Sigma}_{\mathbf{f}}^{-1/2} = \mathbf{W}_{\mathbf{f}}.$$

4. L^2-continuity of whitening.

Define the difference

$$\boldsymbol{\Delta}_n := \mathbf{W}_{\mathbf{f}_n}(\mathbf{f}_n - \boldsymbol{\mu}_n) - \mathbf{W}_{\mathbf{f}}(\mathbf{f} - \boldsymbol{\mu}) = (\mathbf{W}_{\mathbf{f}_n} - \mathbf{W}_{\mathbf{f}})(\mathbf{f} - \boldsymbol{\mu}) + \mathbf{W}_{\mathbf{f}_n}\big[(\mathbf{f}_n - \boldsymbol{\mu}_n) - (\mathbf{f} - \boldsymbol{\mu})\big].$$

Taking L^2 norms and applying Cauchy-Schwarz,

$$\|\boldsymbol{\Delta}_n\|_{L^2} \le \|\mathbf{W}_{\mathbf{f}_n} - \mathbf{W}_{\mathbf{f}}\|\,\|\mathbf{f} - \boldsymbol{\mu}\|_{L^2} + \|\mathbf{W}_{\mathbf{f}_n}\|\,\|(\mathbf{f}_n - \boldsymbol{\mu}_n) - (\mathbf{f} - \boldsymbol{\mu})\|_{L^2}.$$

The first term tends to 0 because $\mathbf{W}_{\mathbf{f}_n} \to \mathbf{W}_{\mathbf{f}}$ and $\|\mathbf{f} - \boldsymbol{\mu}\|_{L^2} < \infty$. For the second term, $\|\mathbf{W}_{\mathbf{f}_n}\|$ is uniformly bounded and

$$\|(\mathbf{f}_n - \boldsymbol{\mu}_n) - (\mathbf{f} - \boldsymbol{\mu})\|_{L^2} \le \|\mathbf{f}_n - \mathbf{f}\|_{L^2} + \|\boldsymbol{\mu}_n - \boldsymbol{\mu}\| \;\to\; 0.$$

Thus $\|\boldsymbol{\Delta}_n\|_{L^2} \to 0$, proving Equation 14.

5. Orthogonal alignment for general whiteners.

Let $\mathbf{U}_{\mathbf{f}_n} \in \mathrm{GL}(d_{\mathcal{Z}})$ satisfy $\mathbf{U}_{\mathbf{f}_n}\boldsymbol{\Sigma}_{\mathbf{f}_n}\mathbf{U}_{\mathbf{f}_n}^\top = \mathbf{I}_{d_{\mathcal{Z}}}$. Using the polar decomposition,

$$\mathbf{U}_{\mathbf{f}_n}\boldsymbol{\Sigma}_{\mathbf{f}_n}^{1/2} = \mathbf{Q}_n, \qquad \mathbf{Q}_n \in O(d_{\mathcal{Z}}),$$

which implies $\mathbf{Q}_n\mathbf{U}_{\mathbf{f}_n} = \boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2} = \mathbf{W}_{\mathbf{f}_n}$. Therefore,

$$\mathbf{Q}_n\,\mathbf{U}_{\mathbf{f}_n}(\mathbf{f}_n - \boldsymbol{\mu}_n) = \mathbf{W}_{\mathbf{f}_n}(\mathbf{f}_n - \boldsymbol{\mu}_n) \xrightarrow[n\to\infty]{L^2} \mathbf{W}_{\mathbf{f}}(\mathbf{f} - \boldsymbol{\mu})$$

by the first part, completing the proof. □

Lemma 6 (Reparameterization Invariance and Representational Universality of CCA)**.** *Let $\mathcal{S}, \mathcal{X}, \mathcal{X}'$ be standard Borel spaces and $\mathbf{g}, \mathbf{g}'$ be Borel-measurable and injective. In addition, let $\tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathcal{F}}'_{\mathcal{X}'}$ be the whitened encoder classes of Definition 2 and assume that the base encoders $\mathbf{f}, \mathbf{f}'$ underlying $\tilde{\mathcal{F}}_{\mathcal{X}}, \tilde{\mathcal{F}}'_{\mathcal{X}'}$ are dense in $L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}})$ and $L^2(P_{\mathbf{x}'}; \mathbb{R}^{d_{\mathcal{Z}}})$, respectively. Define whitened representable latent classes:*

$$\tilde{\mathcal{F}}_{\mathcal{S}} := \{\tilde{\mathbf{f}} \circ \mathbf{g} :\ \tilde{\mathbf{f}} \in \tilde{\mathcal{F}}_{\mathcal{X}}\}, \qquad \tilde{\mathcal{F}}'_{\mathcal{S}} := \{\tilde{\mathbf{f}}' \circ \mathbf{g}' :\ \tilde{\mathbf{f}}' \in \tilde{\mathcal{F}}'_{\mathcal{X}'}\},$$

and the whitened feasible latent classes:

$$\hat{\mathcal{F}}_{\mathcal{S}} := \{\,\hat{\mathbf{h}} \in L^2(P_{\mathbf{s}}; d_{\mathcal{Z}}) : \mathbb{E}[\hat{\mathbf{h}}(\mathbf{s})] = 0,\ \mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{s})) = \mathbf{I}_{d_{\mathcal{Z}}}\,\},$$
$$\hat{\mathcal{F}}'_{\mathcal{S}} := \{\,\hat{\mathbf{h}} \in L^2(P_{\mathbf{s}'}; d_{\mathcal{Z}}) : \mathbb{E}[\hat{\mathbf{h}}(\mathbf{s}')] = 0,\ \mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{s}')) = \mathbf{I}_{d_{\mathcal{Z}}}\,\}.$$

Further define the population CCA objective with respect to whitened feasible latent classes $\hat{\mathcal{F}}_{\mathcal{S}}, \hat{\mathcal{F}}'_{\mathcal{S}}$ on their domains $\mathcal{S} \times \mathcal{S}$:

$$J_{\mathcal{S}}(\hat{\mathbf{h}}, \hat{\mathbf{h}}') = \sum_{i=1}^{d_{\mathcal{Z}}} \sigma_i\big(\mathrm{Cov}(\hat{\mathbf{h}}(\mathbf{s}),\, \hat{\mathbf{h}}'(\mathbf{s}'))\big), \tag{15}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value. Then the following properties hold.

1. ***Objective Preservation.*** *$\forall(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \in \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$, $J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \;=\; J_{\mathcal{S}}\big(\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}'\big)$.*
2. ***Maximizer Correspondence.*** *If $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ is a maximizer of J in Equation 11 over $\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}_{\mathcal{X}'}$, then its composition $(\tilde{\mathbf{f}} \circ \mathbf{g}, \tilde{\mathbf{f}}' \circ \mathbf{g}')$ is also a maximizer of $J_{\mathcal{S}}$ in Equation 15 over $\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}_{\mathcal{S}'}$, and conversely.*
3. ***Representation Universality.*** *For any $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$, $\hat{\mathbf{h}}' \in \hat{\mathcal{F}}'_{\mathcal{S}}$, and any $\epsilon, \epsilon' > 0$, there exist $\tilde{\mathbf{h}} \in \tilde{\mathcal{F}}_{\mathcal{S}}$ and $\tilde{\mathbf{h}}' \in \tilde{\mathcal{F}}'_{\mathcal{S}}$ such that*

$$\mathbb{E}\Big[\|\hat{\mathbf{h}}(\mathbf{s}) - \tilde{\mathbf{h}}(\mathbf{s})\|^2\Big] < \epsilon^2 \quad \textit{and} \quad \mathbb{E}\Big[\|\hat{\mathbf{h}}'(\mathbf{s}') - \tilde{\mathbf{h}}'(\mathbf{s}')\|^2\Big] < \epsilon'^2.$$

Proof. We work in L^2, identifying functions that agree P-a.s. Let J be the CCA objective on $\mathcal{X} \times \mathcal{X}'$ from Equation 11 and $J_{\mathcal{S}}$ the latent CCA objective on $\mathcal{S} \times \mathcal{S}'$ from Equation 15.

1. Objective preservation.

By the whitening preservation of Lemma 3.3, the two cross-covariance matrices $\mathrm{Cov}(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}'), \mathrm{Cov}(\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}')$ coincide, hence so do their sum of the singular values:

$$J(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') \ = \ J_{\mathcal{S}}(\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}').$$

2. Maximizer correspondence (between representable classes).

Since the composition map

$$\Psi : \ \tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'} \to \tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}, \qquad \Psi(\tilde{\mathbf{f}}, \tilde{\mathbf{f}}') = (\tilde{\mathbf{f}} \circ \mathbf{g},\ \tilde{\mathbf{f}}' \circ \mathbf{g}').$$

is bijective, given by Lemma 4.2, and the objective preservation in Lemma 6.1 implies

$$\Phi\left(\underset{\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}}{\operatorname{argmax}} J \right) \ = \ \underset{\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}}{\operatorname{argmax}} J_{\mathcal{S}}.$$

If $(\tilde{\mathbf{f}}^*, \tilde{\mathbf{f}}'^*)$ is a maximizer on $\tilde{\mathcal{F}}_{\mathcal{X}} \times \tilde{\mathcal{F}}'_{\mathcal{X}'}$, then $(\tilde{\mathbf{f}}^* \circ \mathbf{g}, \tilde{\mathbf{f}}'^* \circ \mathbf{g}')$ is also a maximizer on $\tilde{\mathcal{F}}_{\mathcal{S}} \times \tilde{\mathcal{F}}'_{\mathcal{S}}$, and conversely.

3. Representation Universality.

Fix $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$ and $\epsilon > 0$. We show there exists $\tilde{\mathbf{h}} \in \tilde{\mathcal{F}}_{\mathcal{S}}$ with $\mathbb{E}\Big[\|\hat{\mathbf{h}}(\mathbf{s}) - \tilde{\mathbf{h}}(\mathbf{s})\|^2\Big] < \epsilon^2$. (The primed side is identical.)

(i) Pull back $\hat{\mathbf{h}}$ to the observation domain. By Lemma 4.2 (measurable inverses on images), the Borel inverse $\mathbf{g}^{-1} : \mathbf{g}(\mathcal{S}) \to \mathcal{S}$ exists. Define on $\mathbf{g}(\mathcal{S})$

$$\hat{\mathbf{h}}_{\mathcal{X}} := \hat{\mathbf{h}} \circ \mathbf{g}^{-1},$$

and extend it arbitrarily to a Borel function on $\mathcal{X}$ (the extension does not affect $L^2(P_{\mathbf{x}})$ since $P_{\mathbf{x}}$ is supported on $\mathbf{g}(\mathcal{S})$). By Lemma 3.2-3 (expectation/covariance preservation and whitening preservation), $\hat{\mathbf{h}} \in \hat{\mathcal{F}}_{\mathcal{S}}$ implies

$$\hat{\mathbf{h}}_{\mathcal{X}} \in \hat{\mathcal{F}}_{\mathcal{X}} := \big\{ \mathbf{u} \in L^2(P_{\mathbf{x}}; \mathbb{R}^{d_{\mathcal{Z}}}) : \ \mathbb{E}[\mathbf{u}(\mathbf{x})] = \mathbf{0},\ \mathrm{Cov}(\mathbf{u}(\mathbf{x})) = \mathbf{I} \big\}.$$

(ii) Approximate in $L^2(P_{\mathbf{x}})$ by base encoders. By the density assumption in Definition 2, there exists a sequence $\{\mathbf{f}_n\}_{n \geq 1} \subset \mathcal{H}_{\mathcal{X}}$ such that

$$\|\mathbf{f}_n - \hat{\mathbf{h}}_{\mathcal{X}}\|_{L^2(P_{\mathbf{x}})} \xrightarrow[n \to \infty]{} 0.$$

Write $\boldsymbol{\mu}_n = \mathbb{E}[\mathbf{f}_n(\mathbf{x})]$ and $\boldsymbol{\Sigma}_{\mathbf{f}_n} = \mathrm{Cov}(\mathbf{f}_n(\mathbf{x}))$. Since $\hat{\mathbf{h}}_{\mathcal{X}} \in L^2$ and $\mathbf{f}_n \to \hat{\mathbf{h}}_{\mathcal{X}}$ in L^2, Lemma 5.1 implies $\boldsymbol{\mu}_n \to \mathbf{0}$ and $\boldsymbol{\Sigma}_{\mathbf{f}_n} \to \mathbf{I}$ in operator norm; in particular, $\boldsymbol{\Sigma}_{\mathbf{f}_n} \succ 0$ for all large n.

(iii) Whiten the approximants and pass limits through whitening. Let $\mathbf{W}_{\mathbf{f}_n} = \boldsymbol{\Sigma}_{\mathbf{f}_n}^{-1/2}$ and define the whitened encoders

$$\tilde{\mathbf{f}}_n := \mathbf{W}_{\mathbf{f}_n}\big(\mathbf{f}_n - \boldsymbol{\mu}_n\big) \in \tilde{\mathcal{F}}_{\mathcal{X}}.$$

By Lemma 5.4,

$$\tilde{\mathbf{f}}_n \xrightarrow[n \to \infty]{L^2(P_{\mathbf{x}})} \boldsymbol{\Sigma}_{\hat{\mathbf{h}}_{\mathcal{X}}}^{-1/2}\big(\hat{\mathbf{h}}_{\mathcal{X}} - \mathbb{E}[\hat{\mathbf{h}}_{\mathcal{X}}]\big) = \hat{\mathbf{h}}_{\mathcal{X}},$$

since $\hat{\mathbf{h}}_{\mathcal{X}}$ is already whitened (mean 0, covariance $\mathbf{I}$). Hence there exists N such that $\|\tilde{\mathbf{f}}_N - \hat{\mathbf{h}}_{\mathcal{X}}\|_{L^2(P_{\mathbf{x}})} < \epsilon$.

(iv) Push forward to the latent domain and conclude. Set $\tilde{\mathbf{h}} := \tilde{\mathbf{f}}_N \circ \mathbf{g} \in \tilde{\mathcal{F}}_{\mathcal{S}}$ by definition of the representable class. By Lemma 4.1 (composition isometry),

$$\|\tilde{\mathbf{h}} - \hat{\mathbf{h}}\|_{L^2(P_{\mathbf{s}})} = \|\tilde{\mathbf{f}}_N \circ \mathbf{g} - \hat{\mathbf{h}}_{\mathcal{X}} \circ \mathbf{g}\|_{L^2(P_{\mathbf{s}})} = \|\tilde{\mathbf{f}}_N - \hat{\mathbf{h}}_{\mathcal{X}}\|_{L^2(P_{\mathbf{x}})} < \epsilon.$$

Therefore $\mathbb{E}[\|\hat{\mathbf{h}}(\mathbf{s}) - \tilde{\mathbf{h}}(\mathbf{s})\|^2] < \epsilon^2$. Since $\epsilon > 0$ was arbitrary, the claim follows. The proof on the primed side is identical, yielding $\tilde{\mathbf{h}}' \in \tilde{\mathcal{F}}'_{\mathcal{S}}$ with $\mathbb{E}[\|\hat{\mathbf{h}}'(\mathbf{s}') - \tilde{\mathbf{h}}'(\mathbf{s}')\|^2] < \epsilon'^2$ for any $\epsilon' > 0$.

□

B.2 NORMALIZED HERMITE-MEHLER EXPANSION OF JOINT GAUSSIAN DISTRIBUTION

Lemma 7 (Hermite–Mehler expansion of a bivariate Gaussian pair)**.** *Let (S, S') be jointly Gaussian with means $\mu_S, \mu_{S'}$, variances $\sigma_S^2, \sigma_{S'}^2$, and correlation coefficient $t \in (-1, 1)$. Define the standardized coordinates*

$$U := \frac{S - \mu_S}{\sigma_S}, \qquad V := \frac{S' - \mu_{S'}}{\sigma_{S'}}.$$

Let He_n denote the probabilists' Hermite polynomials and define

$$\psi_n(z) := \frac{1}{\sqrt{n!}}\, He_n(z), \qquad n \in \mathbb{N}_0.$$

Let

$$\nu(dz) = \phi(z)\, dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

denote the standard Gaussian measure on $\mathbb{R}$. Then $\{\psi_n\}_{n\geq 0}$ is an orthonormal basis of $L^2(\nu)$, namely,

$$\int_{\mathbb{R}} \psi_m(z)\psi_n(z)\, \nu(dz) = \delta_{mn}.$$

The joint density of (U, V) admits the Hermite–Mehler expansion

$$p_{U,V}(u, v) = \phi(u)\phi(v) \sum_{n=0}^{\infty} t^n\, \psi_n(u)\psi_n(v), \tag{16}$$

Equivalently,

$$p_{S,S'}(s, s') = \frac{1}{\sigma_S \sigma_{S'}} p_{U,V}(u, v) = \frac{1}{\sigma_S \sigma_{S'}} \phi(u)\phi(v) \sum_{n=0}^{\infty} t^n\, \psi_n(u)\psi_n(v). \tag{17}$$

Finally, the Hermite polynomials diagonalize the covariance operator of the standardized pair (U, V):

$$\mathbb{E}[\psi_m(U)\psi_n(V)] = t^n\, \delta_{mn}, \qquad m, n \in \mathbb{N}_0.$$

Proof. The classical Mehler formula for the probabilists' Hermite polynomials states that, for $|t| < 1$,

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} He_n(u)He_n(v) = \frac{1}{\sqrt{1-t^2}} \exp\left(\frac{2tuv - t^2(u^2+v^2)}{2(1-t^2)} \right).$$

Multiplying both sides by $\phi(u)\phi(v)$ gives

$$\phi(u)\phi(v) \sum_{n=0}^{\infty} \frac{t^n}{n!} He_n(u)He_n(v) = \frac{1}{2\pi\sqrt{1-t^2}} \exp\left(-\frac{u^2 - 2tuv + v^2}{2(1-t^2)} \right).$$

The right-hand side is precisely the density of a centered bivariate Gaussian vector with unit variances and correlation coefficient t. Hence,

$$p_{U,V}(u, v) = \phi(u)\phi(v) \sum_{n=0}^{\infty} \frac{t^n}{n!} He_n(u)He_n(v).$$

Since

$$\psi_n(u)\psi_n(v) = \frac{1}{n!} He_n(u)He_n(v),$$

we obtain

$$p_{U,V}(u, v) = \phi(u)\phi(v) \sum_{n=0}^{\infty} t^n\, \psi_n(u)\psi_n(v).$$

This proves the expansion for the standardized pair (U, V).

The expansion for (S, S') follows by the change of variables

$$u = \frac{s - \mu_S}{\sigma_S}, \qquad v = \frac{s' - \mu_{S'}}{\sigma_{S'}},$$

whose Jacobian determinant is

$$\left| \det \begin{pmatrix} 1/\sigma_S & 0 \\ 0 & 1/\sigma_{S'} \end{pmatrix} \right| = \frac{1}{\sigma_S \sigma_{S'}}.$$

Therefore,

$$p_{S,S'}(s, s') = \frac{1}{\sigma_S \sigma_{S'}} p_{U,V}\left(\frac{s - \mu_S}{\sigma_S}, \frac{s' - \mu_{S'}}{\sigma_{S'}} \right),$$

which gives (17).

Finally, since $\{\psi_n\}_{n \geq 0}$ is an orthonormal basis of $L^2(\nu)$,

$$\int_{\mathbb{R}} \psi_m(z) \psi_n(z)\, \nu(dz) = \delta_{mn}.$$

Using the density expansion above,

$$\begin{aligned}
\mathbb{E}[\psi_m(U)\psi_n(V)] &= \int_{\mathbb{R}^2} \psi_m(u)\psi_n(v) \frac{p_{U,V}(u,v)}{\phi(u)\phi(v)}\, \nu(du)\nu(dv) \\
&= \int_{\mathbb{R}^2} \psi_m(u)\psi_n(v) \left(\sum_{k=0}^{\infty} t^k \psi_k(u)\psi_k(v) \right) \nu(du)\nu(dv) \\
&= \sum_{k=0}^{\infty} t^k \left(\int_{\mathbb{R}} \psi_m(u)\psi_k(u)\, \nu(du) \right) \left(\int_{\mathbb{R}} \psi_n(v)\psi_k(v)\, \nu(dv) \right) \\
&= \sum_{k=0}^{\infty} t^k \delta_{mk}\delta_{nk} \\
&= t^n \delta_{mn}.
\end{aligned}$$

The interchange between summation and integration is justified, for $|t| < 1$, by the $L^2(\nu \otimes \nu)$ convergence of

$$\sum_{k=0}^{\infty} t^k \psi_k \otimes \psi_k,$$

since

$$\sum_{k=0}^{\infty} t^{2k} < \infty.$$

This completes the proof. □

Proposition 1 (Hermite–Mehler expansion of a multivariate joint Gaussian density)**.** *Let*

$$\mathbf{S} = (S_1, \ldots, S_{d_{\mathcal{S}}}), \qquad \mathbf{S}' = (S'_1, \ldots, S'_{d_{\mathcal{S}}})$$

be jointly Gaussian random vectors in $\mathbb{R}^{d_{\mathcal{S}}}$. Assume that the coordinate pairs (S_i, S'_i) are mutually independent across $i = 1, \ldots, d_{\mathcal{S}}$, and that each pair (S_i, S'_i) is bivariate Gaussian with means $\mu_{S_i}, \mu_{S'_i}$, variances $\sigma^2_{S_i}, \sigma^2_{S'_i}$, and correlation coefficient $t_i \in (-1, 1)$.

For each multi-index

$$\mathbf{n} = (n_1, \ldots, n_{d_{\mathcal{S}}}) \in \mathbb{N}_0^{d_{\mathcal{S}}},$$

define the normalized multivariate probabilists' Hermite polynomial

$$\Psi_{\mathbf{n}}(\mathbf{x}) = \prod_{i=1}^{d_{\mathcal{S}}} \psi_{n_i}(x_i), \qquad \mathbf{x} = (x_1, \ldots, x_{d_{\mathcal{S}}}) \in \mathbb{R}^{d_{\mathcal{S}}},$$

where

$$\psi_n(z) = \frac{1}{\sqrt{n!}} He_n(z).$$

For $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^{d_{\mathcal{S}}}$, define the standardized coordinates

$$u_i = \frac{s_i - \mu_{S_i}}{\sigma_{S_i}}, \qquad v_i = \frac{s_i' - \mu_{S_i'}}{\sigma_{S_i'}}, \qquad i = 1, \ldots, d_{\mathcal{S}}.$$

Then the joint density of $(\mathbf{S}, \mathbf{S}')$ admits the expansion

$$p_{\mathbf{S},\mathbf{S}'}(\mathbf{s}, \mathbf{s}') = c \, \exp\!\left(-\frac{1}{2}\big(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2\big)\right) \sum_{\mathbf{n} \in \mathbb{N}_0^{d_{\mathcal{S}}}} t_{\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v}), \tag{18}$$

where

$$c = (2\pi)^{-d_{\mathcal{S}}} \prod_{i=1}^{d_{\mathcal{S}}} \frac{1}{\sigma_{S_i} \sigma_{S_i'}}, \qquad t_{\mathbf{n}} = \prod_{i=1}^{d_{\mathcal{S}}} t_i^{n_i}.$$

Equivalently, if

$$\phi_{d_{\mathcal{S}}}(\mathbf{x}) = (2\pi)^{-d_{\mathcal{S}}/2} \exp\!\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)$$

denotes the standard Gaussian density on $\mathbb{R}^{d_{\mathcal{S}}}$, then

$$p_{\mathbf{S},\mathbf{S}'}(\mathbf{s}, \mathbf{s}') = \left(\prod_{i=1}^{d_{\mathcal{S}}} \frac{1}{\sigma_{S_i} \sigma_{S_i'}}\right) \phi_{d_{\mathcal{S}}}(\mathbf{u}) \phi_{d_{\mathcal{S}}}(\mathbf{v}) \sum_{\mathbf{n} \in \mathbb{N}_0^{d_{\mathcal{S}}}} t_{\mathbf{n}} \, \Psi_{\mathbf{n}}(\mathbf{u}) \Psi_{\mathbf{n}}(\mathbf{v}). \tag{19}$$

Proof. By the mutual independence of the coordinate pairs (S_i, S_i') across $i = 1, \ldots, d_{\mathcal{S}}$, the joint density factorizes as

$$p_{\mathbf{S},\mathbf{S}'}(\mathbf{s}, \mathbf{s}') = \prod_{i=1}^{d_{\mathcal{S}}} p_{S_i, S_i'}(s_i, s_i').$$

For each coordinate pair, applying Lemma 7 gives

$$p_{S_i, S_i'}(s_i, s_i') = \frac{1}{\sigma_{S_i} \sigma_{S_i'}} \phi(u_i) \phi(v_i) \sum_{n_i = 0}^{\infty} t_i^{n_i} \psi_{n_i}(u_i) \psi_{n_i}(v_i),$$

where

$$u_i = \frac{s_i - \mu_{S_i}}{\sigma_{S_i}}, \qquad v_i = \frac{s_i' - \mu_{S_i'}}{\sigma_{S_i'}}.$$

Therefore,

$$\begin{aligned} p_{\mathbf{S},\mathbf{S}'}(\mathbf{s}, \mathbf{s}') &= \prod_{i=1}^{d_{\mathcal{S}}} \left[\frac{1}{\sigma_{S_i} \sigma_{S_i'}} \phi(u_i) \phi(v_i) \sum_{n_i = 0}^{\infty} t_i^{n_i} \psi_{n_i}(u_i) \psi_{n_i}(v_i)\right] \\ &= \left(\prod_{i=1}^{d_{\mathcal{S}}} \frac{1}{\sigma_{S_i} \sigma_{S_i'}}\right) \left(\prod_{i=1}^{d_{\mathcal{S}}} \phi(u_i)\right) \left(\prod_{i=1}^{d_{\mathcal{S}}} \phi(v_i)\right) \\ &\qquad \times \prod_{i=1}^{d_{\mathcal{S}}} \left[\sum_{n_i = 0}^{\infty} t_i^{n_i} \psi_{n_i}(u_i) \psi_{n_i}(v_i)\right]. \end{aligned}$$

Since

$$\prod_{i=1}^{d_{\mathcal{S}}} \phi(u_i) = (2\pi)^{-d_{\mathcal{S}}/2} \exp\!\left(-\frac{1}{2}\|\mathbf{u}\|^2\right),$$

and analogously

$$\prod_{i=1}^{d_{\mathcal{S}}} \phi(v_i) = (2\pi)^{-d_{\mathcal{S}}/2} \exp\!\left(-\frac{1}{2}\|\mathbf{v}\|^2\right),$$

we obtain

$$\left(\prod_{i=1}^{d_{\mathcal{S}}} \phi(u_i)\right)\left(\prod_{i=1}^{d_{\mathcal{S}}} \phi(v_i)\right) = (2\pi)^{-d_{\mathcal{S}}} \exp\!\left(-\frac{1}{2}\big(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2\big)\right).$$

Moreover, since the product is finite,

$$\prod_{i=1}^{d_{\mathcal{S}}}\left[\sum_{n_i=0}^{\infty} t_i^{n_i}\psi_{n_i}(u_i)\psi_{n_i}(v_i)\right] = \sum_{\mathbf{n}\in\mathbb{N}_0^{d_{\mathcal{S}}}} \prod_{i=1}^{d_{\mathcal{S}}} t_i^{n_i}\psi_{n_i}(u_i)\psi_{n_i}(v_i).$$

Using

$$t_{\mathbf{n}} = \prod_{i=1}^{d_{\mathcal{S}}} t_i^{n_i}$$

and

$$\Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{n}}(\mathbf{v}) = \prod_{i=1}^{d_{\mathcal{S}}} \psi_{n_i}(u_i)\psi_{n_i}(v_i),$$

we get

$$p_{\mathbf{S},\mathbf{S}'}(\mathbf{s},\mathbf{s}') = (2\pi)^{-d_{\mathcal{S}}}\left(\prod_{i=1}^{d_{\mathcal{S}}} \frac{1}{\sigma_{S_i}\sigma_{S_i'}}\right)\exp\!\left(-\frac{1}{2}\big(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2\big)\right)\sum_{\mathbf{n}\in\mathbb{N}_0^{d_{\mathcal{S}}}} t_{\mathbf{n}}\Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{n}}(\mathbf{v}).$$

This is exactly (18).

The convergence of the tensor-product expansion is inherited from the one-dimensional Mehler expansions. Equivalently, the likelihood-ratio series converges in $L^2(\nu^{\otimes d_{\mathcal{S}}} \otimes \nu^{\otimes d_{\mathcal{S}}})$, because

$$\sum_{\mathbf{n}\in\mathbb{N}_0^{d_{\mathcal{S}}}} t_{\mathbf{n}}^2 = \prod_{i=1}^{d_{\mathcal{S}}}\left(\sum_{n_i=0}^{\infty} t_i^{2n_i}\right) = \prod_{i=1}^{d_{\mathcal{S}}} \frac{1}{1-t_i^2} < \infty.$$

This completes the proof. □

C PROOF OF LEMMA 1

Proof of Lemma 1. Fix a pair of views $1 \le i < j \le N$, and write

$$p := d_{\mathcal{S}_i}, \qquad q := d_{\mathcal{S}_j}, \qquad r := r_{ij}.$$

Under the Gaussian latent prior and the additive source model

$$\mathbf{s}_i = \mathbf{A}_i\mathbf{c} + \boldsymbol{\epsilon}_i, \qquad \mathbf{s}_j = \mathbf{A}_j\mathbf{c} + \boldsymbol{\epsilon}_j,$$

the pair $(\mathbf{s}_i, \mathbf{s}_j)$ is jointly Gaussian. Moreover, by the mutual independence and standardization of $\mathbf{c}, \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_j$,

$$\mathbb{E}[\mathbf{s}_i] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{s}_j] = \mathbf{0},$$

and

$$\begin{aligned}
\mathrm{Cov}(\mathbf{s}_i) &= \mathbf{A}_i\mathbf{A}_i^\top + \mathbf{I}_p =: \boldsymbol{\Sigma}_i,\\
\mathrm{Cov}(\mathbf{s}_j) &= \mathbf{A}_j\mathbf{A}_j^\top + \mathbf{I}_q =: \boldsymbol{\Sigma}_j,\\
\mathrm{Cov}(\mathbf{s}_i,\mathbf{s}_j) &= \mathbf{A}_i\mathbf{A}_j^\top.
\end{aligned}$$

Thus the whitening matrices

$$\mathbf{W}_i = \mathbf{\Sigma}_i^{-1/2}, \qquad \mathbf{W}_j = \mathbf{\Sigma}_j^{-1/2}$$

are invertible, and the whitened source variables

$$\mathbf{z}_i := \mathbf{W}_i \mathbf{s}_i, \qquad \mathbf{z}_j := \mathbf{W}_j \mathbf{s}_j$$

satisfy

$$\mathrm{Cov}(\mathbf{z}_i) = \mathbf{I}_p, \qquad \mathrm{Cov}(\mathbf{z}_j) = \mathbf{I}_q, \qquad \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{W}_i \mathbf{A}_i \mathbf{A}_j^\top \mathbf{W}_j^\top = \mathbf{R}_{ij}.$$

Now use the singular value decomposition

$$\mathbf{R}_{ij} = \mathbf{U}_{ij} \mathbf{T}_{ij} \mathbf{V}_{ij}^\top, \qquad \mathbf{U}_{ij} \in O(p), \quad \mathbf{V}_{ij} \in O(q),$$

and define the canonical coordinates

$$\mathbf{u} = \mathbf{U}_{ij}^\top \mathbf{z}_i = \mathbf{U}_{ij}^\top \mathbf{W}_i \mathbf{s}_i, \qquad \mathbf{v} = \mathbf{V}_{ij}^\top \mathbf{z}_j = \mathbf{V}_{ij}^\top \mathbf{W}_j \mathbf{s}_j.$$

Since $(\mathbf{z}_i, \mathbf{z}_j)$ is jointly Gaussian and the above transformation is linear, $(\mathbf{u}, \mathbf{v})$ is also jointly Gaussian. Its covariance blocks are

$$\mathrm{Cov}(\mathbf{u}) = \mathbf{I}_p, \qquad \mathrm{Cov}(\mathbf{v}) = \mathbf{I}_q,$$

and

$$\mathrm{Cov}(\mathbf{u}, \mathbf{v}) = \mathbf{U}_{ij}^\top \mathbf{R}_{ij} \mathbf{V}_{ij} = \mathbf{T}_{ij}.$$

Therefore

$$\mathrm{Cov} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{T}_{ij} \\ \mathbf{T}_{ij}^\top & \mathbf{I}_q \end{pmatrix}.$$

Because $\mathbf{T}_{ij}$ is rectangular diagonal with nonzero diagonal entries

$$1 > t_{ij,1} \geq \cdots \geq t_{ij,r} > 0$$

and all remaining singular values equal to zero, the only nonzero cross-covariances between canonical coordinates are

$$\mathrm{Cov}(u_k, v_k) = t_{ij,k}, \qquad k = 1, \ldots, r.$$

Equivalently, after permuting coordinates as

$$(u_1, v_1, \ldots, u_r, v_r, u_{r+1}, \ldots, u_p, v_{r+1}, \ldots, v_q),$$

the covariance matrix is block diagonal:

$$\bigoplus_{k=1}^{r} \begin{pmatrix} 1 & t_{ij,k} \\ t_{ij,k} & 1 \end{pmatrix} \oplus \mathbf{I}_{p-r} \oplus \mathbf{I}_{q-r}.$$

For jointly Gaussian random variables, block-diagonal covariance is equivalent to independence across the corresponding blocks. Hence the pairs (u_k, v_k), $k = 1, \ldots, r$, and the remaining coordinates $\{u_m : m > r\}$ and $\{v_n : n > r\}$, are mutually independent. Each (u_k, v_k) is a standardized bivariate Gaussian with correlation $t_{ij,k}$, whose density is

$$\phi_{t_{ij,k}}(u_k, v_k) = \frac{1}{2\pi\sqrt{1 - t_{ij,k}^2}} \exp\left(-\frac{u_k^2 - 2t_{ij,k} u_k v_k + v_k^2}{2(1 - t_{ij,k}^2)} \right),$$

whereas each unpaired coordinate has the standard normal density ϕ. Consequently,

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u}, \mathbf{v}) = \prod_{k=1}^{r} \phi_{t_{ij,k}}(u_k, v_k) \prod_{m=r+1}^{p} \phi(u_m) \prod_{n=r+1}^{q} \phi(v_n).$$

Substituting back $p = d_{\mathcal{S}_i}$, $q = d_{\mathcal{S}_j}$, and $r = r_{ij}$ gives

$$p_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) = \underbrace{\prod_{k=1}^{r_{ij}} \phi_{t_{ij,k}}(u_k, v_k)}_{\mathcal{K}_{ij}(\mathbf{u}[1:r_{ij}],\mathbf{v}[1:r_{ij}])} \prod_{m=r_{ij}+1}^{d_{\mathcal{S}_i}} \phi(u_m) \prod_{n=r_{ij}+1}^{d_{\mathcal{S}_j}} \phi(v_n).$$

It remains to verify the stated change-of-variables identity. The inverse linear transformation is

$$\mathbf{s}_i = \mathbf{W}_i^{-1}\mathbf{U}_{ij}\mathbf{u}, \qquad \mathbf{s}_j = \mathbf{W}_j^{-1}\mathbf{V}_{ij}\mathbf{v}.$$

Hence the absolute Jacobian determinant of the inverse map $(\mathbf{u},\mathbf{v}) \mapsto (\mathbf{s}_i,\mathbf{s}_j)$ is

$$\left|\det(\mathbf{W}_i^{-1}\mathbf{U}_{ij})\right| \left|\det(\mathbf{W}_j^{-1}\mathbf{V}_{ij})\right| = |\det \mathbf{W}_i|^{-1} |\det \mathbf{W}_j|^{-1},$$

because $\mathbf{U}_{ij}$ and $\mathbf{V}_{ij}$ are orthogonal. Therefore, by the change-of-variables formula,

$$\begin{aligned} p_{\mathbf{u},\mathbf{v}}(\mathbf{u},\mathbf{v}) = p_{\mathbf{s}_i,\mathbf{s}_j}\left(\mathbf{W}_i^{-1}\mathbf{U}_{ij}\mathbf{u}, \mathbf{W}_j^{-1}\mathbf{V}_{ij}\mathbf{v}\right) \\ \cdot |\det \mathbf{W}_i|^{-1} |\det \mathbf{W}_j|^{-1}. \end{aligned}$$

This proves both the change-of-variables identity and the claimed canonical factorization. □

D PROOF OF LEMMA 2

Proof. Recall Lemma 7. The tensor-product basis is orthonormal since

$$\int_{\mathbb{R}^r} \Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{m}}(\mathbf{u})\phi_r(\mathbf{u})\, d\mathbf{u} = \prod_{k=1}^{r} \int_{\mathbb{R}} \psi_{n_k}(u)\psi_{m_k}(u)\phi(u)\, du = \delta_{\mathbf{nm}}.$$

In canonical Gaussian coordinates, the pairs (U_k, V_k) are mutually independent standard bivariate Gaussian pairs with correlations ρ_k. By the one-dimensional Mehler–Hermite expansion,

$$p_{U_k,V_k}(u_k, v_k) = \phi(u_k)\phi(v_k) \sum_{n_k=0}^{\infty} \rho_k^{n_k} \psi_{n_k}(u_k)\psi_{n_k}(v_k).$$

Hence

$$\begin{aligned} \mathcal{K}_{ij}(\mathbf{u},\mathbf{v}) &= \prod_{k=1}^{r} p_{U_k,V_k}(u_k, v_k) \\ &= \phi_r(\mathbf{u})\phi_r(\mathbf{v}) \prod_{k=1}^{r} \sum_{n_k=0}^{\infty} \rho_k^{n_k} \psi_{n_k}(u_k)\psi_{n_k}(v_k) \\ &= \phi_r(\mathbf{u})\phi_r(\mathbf{v}) \sum_{\mathbf{n}\in\mathbb{N}_0^r} \left(\prod_{k=1}^{r} \rho_k^{n_k}\right) \Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{n}}(\mathbf{v}). \end{aligned}$$

This proves the claim. The expansion converges in $L^2(\gamma_r \otimes \gamma_r)$ because

$$\sum_{\mathbf{n}\in\mathbb{N}_0^r} t_{\mathbf{n}}^2 = \prod_{k=1}^{r} \sum_{n_k=0}^{\infty} \rho_k^{2n_k} = \prod_{k=1}^{r} \frac{1}{1-\rho_k^2} < \infty.$$

□

E PROOF OF THEOREM 5.1

Proof of Theorem 5.1. Fix a pair of distinct views $i \neq j$ and write $r := r_{ij} = \operatorname{rank}(\mathbf{A}_i\mathbf{A}_j^\top)$. Throughout the proof we work at the population level and assume $d_{\mathcal{Z}} \to \infty$.

Step 1: Reduce the problem to the source domain. Let $\tilde{\mathbf{h}}_\ell := \tilde{\mathbf{f}}_\ell \circ \mathbf{g}_\ell$ for $\ell \in \{i, j\}$. By reparameterization invariance (cf. Lemma 6), for any encoders $\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j$ we have

$$\mathrm{Cov}\big(\tilde{\mathbf{f}}_i(\mathbf{x}_i), \tilde{\mathbf{f}}_j(\mathbf{x}_j)\big) = \mathrm{Cov}\big(\tilde{\mathbf{h}}_i(\mathbf{s}_i), \tilde{\mathbf{h}}_j(\mathbf{s}_j)\big), \qquad \mathrm{Cov}\big(\tilde{\mathbf{f}}_\ell(\mathbf{x}_\ell)\big) = \mathrm{Cov}\big(\tilde{\mathbf{h}}_\ell(\mathbf{s}_\ell)\big),$$

and therefore optimizing the two-view term of Equation 2 over $(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j)$ is equivalent to optimizing it over $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j)$. Since $(\tilde{\mathbf{f}}_i^\star, \tilde{\mathbf{f}}_j^\star)$ is a *whitened* population maximizer, its source-domain counterpart $(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star)$ satisfies

$$\mathbb{E}[\tilde{\mathbf{h}}_\ell^\star(\mathbf{s}_\ell)] = \mathbf{0}, \qquad \mathrm{Cov}(\tilde{\mathbf{h}}_\ell^\star(\mathbf{s}_\ell)) = \mathbf{I}, \qquad \ell \in \{i, j\}.$$

Thus the two-view population objective reduces to

$$J_{ij}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j) := \big\|\mathrm{Cov}(\tilde{\mathbf{h}}_i(\mathbf{s}_i), \tilde{\mathbf{h}}_j(\mathbf{s}_j))\big\|_*.$$

Step 2: Canonical coordinates for the sources. Under Assumption 1 (Gaussian case), each $(\mathbf{s}_i, \mathbf{s}_j)$ is jointly Gaussian. Let

$$\mathbf{W}_i := \mathrm{Cov}(\mathbf{s}_i)^{-1/2} = (\mathbf{A}_i\mathbf{A}_i^\top + \mathbf{I})^{-1/2}, \qquad \mathbf{W}_j := \mathrm{Cov}(\mathbf{s}_j)^{-1/2}.$$

Define the normalized cross-covariance (cf. Equation 4)

$$\mathbf{R}_{ij} := \mathrm{Cov}(\mathbf{W}_i\mathbf{s}_i, \mathbf{W}_j\mathbf{s}_j) = \mathbf{W}_i\mathbf{A}_i\mathbf{A}_j^\top\mathbf{W}_j^\top,$$

and take an SVD $\mathbf{R}_{ij} = \mathbf{U}_{ij}\mathbf{T}_{ij}\mathbf{V}_{ij}^\top$, where $\mathbf{T}_{ij}$ is rectangular diagonal with singular values $1 > t_{ij,1} \geq \cdots \geq t_{ij,r} > 0$ on its diagonal and zeros afterwards. Introduce canonical coordinates (cf. Equation 5)

$$\mathbf{u} := \mathbf{U}_{ij}^\top\mathbf{W}_i\mathbf{s}_i, \qquad \mathbf{v} := \mathbf{V}_{ij}^\top\mathbf{W}_j\mathbf{s}_j.$$

Then $\mathrm{Cov}(\mathbf{u}) = \mathrm{Cov}(\mathbf{v}) = \mathbf{I}$ and $\mathrm{Cov}(\mathbf{u}, \mathbf{v}) = \mathbf{T}_{ij}$. Moreover, by Lemma 1, the joint density factorizes so that only the first r coordinate pairs (u_k, v_k) are correlated, while all remaining coordinates are independent standard normals.

Step 3: Hermite diagonalization of cross-view correlations. Let $\{\psi_n\}_{n\geq 0}$ be the normalized univariate Hermite polynomials. For $\mathbf{n} = (n_1, \ldots, n_r) \in \mathbb{N}^r$ define

$$\Psi_{\mathbf{n}}(\mathbf{u}) := \prod_{k=1}^{r} \psi_{n_k}(u_k), \qquad \Psi_{\mathbf{n}}(\mathbf{v}) := \prod_{k=1}^{r} \psi_{n_k}(v_k).$$

By Lemma 2 (Mehler–Hermite expansion) and orthonormality of Hermites, for all $\mathbf{n}, \mathbf{m} \in \mathbb{N}^r$,

$$\mathbb{E}[\Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{m}}(\mathbf{v})] = t_{\mathbf{n}}\,\delta_{\mathbf{nm}}, \qquad t_{\mathbf{n}} := \prod_{k=1}^{r} t_{ij,k}^{n_k}. \tag{20}$$

In particular, for the first-order multi-indices $\mathbf{e}_k$ we have $\Psi_{\mathbf{e}_k}(\mathbf{u}) = \psi_1(u_k) = u_k$ and $\Psi_{\mathbf{e}_k}(\mathbf{v}) = v_k$, with $\mathbb{E}[u_k v_k] = t_{ij,k}$.

Step 4: Rewrite the CCA objective in coefficient form. Define the whitened source-domain maximizers as functions of $(\mathbf{u}, \mathbf{v})$:

$$\tilde{\mathbf{f}}^\star(\mathbf{u}) := \tilde{\mathbf{h}}_i^\star(\mathbf{W}_i^{-1}\mathbf{U}_{ij}\mathbf{u}), \qquad \tilde{\mathbf{g}}^\star(\mathbf{v}) := \tilde{\mathbf{h}}_j^\star(\mathbf{W}_j^{-1}\mathbf{V}_{ij}\mathbf{v}).$$

These satisfy $\mathrm{Cov}(\tilde{\mathbf{f}}^\star(\mathbf{u})) = \mathrm{Cov}(\tilde{\mathbf{g}}^\star(\mathbf{v})) = \mathbf{I}$. For each coordinate $p \in [d_{\mathcal{Z}}]$, expand the (mean-zero) component functions in the orthonormal Hermite basis on the correlated coordinates:

$$\tilde{f}_p^\star(\mathbf{u}) = \sum_{\mathbf{n}\neq\mathbf{0}} \alpha_{p,\mathbf{n}}\,\Psi_{\mathbf{n}}(\mathbf{u}), \qquad \tilde{g}_q^\star(\mathbf{v}) = \sum_{\mathbf{n}\neq\mathbf{0}} \beta_{q,\mathbf{n}}\,\Psi_{\mathbf{n}}(\mathbf{v}).$$

Let $\mathbf{C}_i := (\alpha_{p,\mathbf{n}})_{p,\mathbf{n}}$ and $\mathbf{C}_j := (\beta_{q,\mathbf{n}})_{q,\mathbf{n}}$ denote the (possibly infinite) coefficient matrices whose rows collect the coefficient vectors. Whitening implies row-orthonormality (cf. the whitening-induced canonicalization Equation 3):

$$\mathbf{C}_i\mathbf{C}_i^\top = \mathbf{I}, \qquad \mathbf{C}_j\mathbf{C}_j^\top = \mathbf{I}.$$

Using Equation 20, the cross-covariance of the whitened representations becomes

$$\tilde{\mathbf{\Sigma}}_{ij}^\star := \mathrm{Cov}(\tilde{\mathbf{h}}_i^\star(\mathbf{s}_i), \tilde{\mathbf{h}}_j^\star(\mathbf{s}_j)) = \mathrm{Cov}(\tilde{\mathbf{f}}^\star(\mathbf{u}), \tilde{\mathbf{g}}^\star(\mathbf{v})) = \mathbf{C}_i\,\mathbf{D}\,\mathbf{C}_j^\top,$$

where $\mathbf{D} := \mathrm{diag}\big((t_{\mathbf{n}})_{\mathbf{n}\neq\mathbf{0}}\big)$ is diagonal (and trace-class since $t_{ij,k} \in (0,1)$ implies $\sum_{\mathbf{n}\neq\mathbf{0}} t_{\mathbf{n}} < \infty$). Thus,

$$J_{ij}(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star) = \|\tilde{\mathbf{\Sigma}}_{ij}^\star\|_* = \|\mathbf{C}_i\mathbf{D}\mathbf{C}_j^\top\|_*.$$

Step 5: Optimality and recovery of the linear canonical subspaces. We first upper bound $\|\mathbf{C}_i\mathbf{D}\mathbf{C}_j^\top\|_*$. Using nuclear norm duality $\|\mathbf{M}\|_* = \sup_{\|\mathbf{Q}\|_2\le 1} \operatorname{tr}(\mathbf{Q}^\top\mathbf{M})$, together with $\|\mathbf{C}_i\|_2 = \|\mathbf{C}_j\|_2 = 1$ (since they have orthonormal rows), we have

$$\|\mathbf{C}_i\mathbf{D}\mathbf{C}_j^\top\|_* = \sup_{\|\mathbf{Q}\|_2\le 1} \operatorname{tr}\big((\mathbf{C}_i^\top\mathbf{Q}\mathbf{C}_j)^\top\mathbf{D}\big) \;\le\; \sup_{\|\mathbf{H}\|_2\le 1} \operatorname{tr}(\mathbf{H}^\top\mathbf{D}) \;=\; \|\mathbf{D}\|_* \;=\; \sum_{\mathbf{n}\neq\mathbf{0}} t_{\mathbf{n}}.$$

This upper bound is achievable in the limit $d_{\mathcal{Z}} \to \infty$ by taking the representations to be (an enumeration of) the Hermite basis itself, i.e., $\tilde{\mathbf{f}}(\mathbf{u}) = (\Psi_{\mathbf{n}}(\mathbf{u}))_{\mathbf{n}\neq\mathbf{0}}$ and $\tilde{\mathbf{g}}(\mathbf{v}) = (\Psi_{\mathbf{n}}(\mathbf{v}))_{\mathbf{n}\neq\mathbf{0}}$, for which $\mathbf{C}_i = \mathbf{C}_j = \mathbf{I}$ and hence $\tilde{\mathbf{\Sigma}}_{ij} = \mathbf{D}$. Therefore the population optimum equals $\|\mathbf{D}\|_*$, and any whitened population maximizer must achieve the same value.

At this point we invoke the standard CCA canonicalization convention: because the objective and whitening constraints are invariant under within-view orthogonal post-transformations of the *feature coordinates*, we may choose an equivalent maximizer (within the unavoidable CCA orthogonal ambiguity) whose cross-covariance is diagonalized and whose coordinates form a singular basis of the diagonal operator $\mathbf{D}$ (i.e., a basis of Hermite modes; see Equation 20). Under this canonical representative, the r first-order Hermite modes $\{\Psi_{\mathbf{e}_k}\}_{k=1}^r$ appear among the feature coordinates, and $\Psi_{\mathbf{e}_k}(\mathbf{u}) = u_k$, $\Psi_{\mathbf{e}_k}(\mathbf{v}) = v_k$.

Let $\mathbf{P}_r$ be the coordinate-selection matrix that extracts exactly those r coordinates corresponding to $\{\Psi_{\mathbf{e}_k}\}_{k=1}^r$. Since the extracted coordinates form an orthonormal basis of the linear subspace $\operatorname{span}\{u_1,\dots,u_r\}$ (resp. $\operatorname{span}\{v_1,\dots,v_r\}$), there exist $\mathbf{O}_i, \mathbf{O}_j \in O(r)$ such that

$$\mathbf{P}_r\,\tilde{\mathbf{f}}^\star(\mathbf{u}) = \mathbf{O}_i\,\mathbf{u}[1:r], \qquad \mathbf{P}_r\,\tilde{\mathbf{g}}^\star(\mathbf{v}) = \mathbf{O}_j\,\mathbf{v}[1:r].$$

Finally, substituting back $\mathbf{u} = \mathbf{U}_{ij}^\top\mathbf{W}_i\mathbf{s}_i$ and $\mathbf{v} = \mathbf{V}_{ij}^\top\mathbf{W}_j\mathbf{s}_j$ yields

$$\mathbf{P}_r\,\tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_i\,\mathbf{U}_{ij}(:,1:r)^\top\mathbf{W}_i\mathbf{s}_i, \qquad \mathbf{P}_r\,\tilde{\mathbf{h}}_j^\star(\mathbf{s}_j) = \mathbf{O}_j\,\mathbf{V}_{ij}(:,1:r)^\top\mathbf{W}_j\mathbf{s}_j,$$

which is exactly Equation 6.

Consequence (subspace identification). The right-hand sides lie in $\operatorname{col}(\mathbf{U}_{ij}(:,1:r)) = \mathcal{U}_{i|j}$ and $\operatorname{col}(\mathbf{V}_{ij}(:,1:r)) = \mathcal{U}_{j|i}$, respectively, and the only remaining ambiguity is the orthogonal mixing $\mathbf{O}_i, \mathbf{O}_j$ inside these r-dimensional subspaces. Hence the maximizers identify the whitened correlated subspaces up to orthogonal transformations. □

F PROOF OF COROLLARY 1

Proof of Corollary 1. Fix a pair of views $i \neq j$ and write $r := r_{ij}$ and $d := d_{\mathcal{Z}}$, with $d \ge r$. By reparameterization invariance (Lemma 6), it suffices to analyze the *source-domain* problem: maximize the two-view CCA objective over whitened mappings $\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j$ with $\mathbb{E}[\tilde{\mathbf{h}}_\ell] = \mathbf{0}$ and $\operatorname{Cov}(\tilde{\mathbf{h}}_\ell) = \mathbf{I}_d$ for $\ell \in \{i,j\}$.

Let $(\mathbf{u},\mathbf{v})$ be the canonical coordinates from Lemma 1,

$$\mathbf{u} = \mathbf{U}_{ij}^\top\mathbf{W}_i\mathbf{s}_i, \qquad \mathbf{v} = \mathbf{V}_{ij}^\top\mathbf{W}_j\mathbf{s}_j,$$

and denote $\mathbf{u}_{1:r} := (u_1,\dots,u_r)^\top$ and $\mathbf{v}_{1:r} := (v_1,\dots,v_r)^\top$. Let $\{\Psi_{\mathbf{n}}\}_{\mathbf{n}\in\mathbb{N}^r}$ be the normalized multivariate Hermite basis from Lemma 2, and define $t_{\mathbf{n}} := \prod_{k=1}^r t_{ij,k}^{n_k}$. (We always exclude $\mathbf{n} = \mathbf{0}$ below, because mean-zero constraints remove the constant mode.)

Step 1: Diagonal cross-moments in the Hermite basis. From Lemma 2 and the orthonormality of $\{\Psi_{\mathbf{n}}\}$ in $L^2(\phi)$, we obtain for any $\mathbf{n},\mathbf{m} \in \mathbb{N}^r$,

$$\begin{aligned}
\mathbb{E}[\Psi_{\mathbf{n}}(\mathbf{u}_{1:r})\,\Psi_{\mathbf{m}}(\mathbf{v}_{1:r})] &= \int \Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\mathbf{m}}(\mathbf{v})\,\phi(\mathbf{u})\phi(\mathbf{v}) \sum_{\boldsymbol{\ell}\in\mathbb{N}^r} t_{\boldsymbol{\ell}}\Psi_{\boldsymbol{\ell}}(\mathbf{u})\Psi_{\boldsymbol{\ell}}(\mathbf{v})\,d\mathbf{u}\,d\mathbf{v} \\
&= \sum_{\boldsymbol{\ell}\in\mathbb{N}^r} t_{\boldsymbol{\ell}}\Big(\int \Psi_{\mathbf{n}}(\mathbf{u})\Psi_{\boldsymbol{\ell}}(\mathbf{u})\phi(\mathbf{u})\,d\mathbf{u}\Big)\Big(\int \Psi_{\mathbf{m}}(\mathbf{v})\Psi_{\boldsymbol{\ell}}(\mathbf{v})\phi(\mathbf{v})\,d\mathbf{v}\Big) \\
&= t_{\mathbf{n}}\,\delta_{\mathbf{nm}}. \qquad (21)
\end{aligned}$$

Step 2: Coefficient representation of feasible whitened maps. Let $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j)$ be any feasible whitened pair in L^2 with output dimension d. Expand each coordinate in the orthonormal basis: for $p, q \in [d]$,

$$\tilde{h}_{i,p}(\mathbf{u}_{1:r}) = \sum_{\mathbf{n}\neq\mathbf{0}} a_{p,\mathbf{n}}\, \Psi_{\mathbf{n}}(\mathbf{u}_{1:r}), \qquad \tilde{h}_{j,q}(\mathbf{v}_{1:r}) = \sum_{\mathbf{n}\neq\mathbf{0}} b_{q,\mathbf{n}}\, \Psi_{\mathbf{n}}(\mathbf{v}_{1:r}),$$

where $\mathbf{n} = \mathbf{0}$ is absent because $\mathbb{E}[\tilde{h}_{i,p}] = \mathbb{E}[\tilde{h}_{j,q}] = 0$. Let $\mathbf{A} = [a_{p,\mathbf{n}}]_{p,\mathbf{n}}$ and $\mathbf{B} = [b_{q,\mathbf{n}}]_{q,\mathbf{n}}$. By whitening-induced canonicalization (Equation 3), the coefficient rows are orthonormal:

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}_d, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I}_d. \tag{22}$$

Using Equation 21, the cross-covariance matrix $\mathbf{C} := \mathrm{Cov}(\tilde{\mathbf{h}}_i(\mathbf{u}_{1:r}), \tilde{\mathbf{h}}_j(\mathbf{v}_{1:r}))$ satisfies

$$C_{pq} = \mathbb{E}\Big[\tilde{h}_{i,p}(\mathbf{u}_{1:r})\, \tilde{h}_{j,q}(\mathbf{v}_{1:r})\Big] = \sum_{\mathbf{n}\neq\mathbf{0}} a_{p,\mathbf{n}}\, b_{q,\mathbf{n}}\, t_{\mathbf{n}}, \qquad \Longleftrightarrow \qquad \mathbf{C} = \mathbf{A}\,\mathbf{T}\,\mathbf{B}^\top, \tag{23}$$

where $\mathbf{T}$ is diagonal with diagonal entries $\{t_{\mathbf{n}}\}_{\mathbf{n}\neq\mathbf{0}}$.

Step 3: A Ky Fan (nuclear-norm) upper bound. Let $t_{(1)} \geq t_{(2)} \geq \cdots$ be the nonincreasing rearrangement of $\{t_{\mathbf{n}}\}_{\mathbf{n}\neq\mathbf{0}}$. Then for any $\mathbf{A}, \mathbf{B}$ satisfying Equation 22,

$$\|\mathbf{C}\|_* = \|\mathbf{A}\mathbf{T}\mathbf{B}^\top\|_* \leq \sum_{k=1}^{d} t_{(k)}. \tag{24}$$

Indeed, by the Ky Fan variational characterization of the nuclear norm,

$$\|\mathbf{A}\mathbf{T}\mathbf{B}^\top\|_* = \max_{\mathbf{U},\mathbf{V}\in O(d)} \mathrm{Tr}\big(\mathbf{U}^\top\mathbf{A}\mathbf{T}\mathbf{B}^\top\mathbf{V}\big) = \max_{\mathbf{U},\mathbf{V}\in O(d)} \mathrm{Tr}\big((\mathbf{A}^\top\mathbf{U})^\top\mathbf{T}(\mathbf{B}^\top\mathbf{V})\big)\,.$$

By Equation 22, both $\mathbf{A}^\top\mathbf{U}$ and $\mathbf{B}^\top\mathbf{V}$ have orthonormal columns, so the last quantity is bounded by the Ky Fan d-norm of $\mathbf{T}$, i.e. $\sum_{k=1}^{d} \sigma_k(\mathbf{T}) = \sum_{k=1}^{d} t_{(k)}$, proving Equation 24. Moreover, the bound is achievable by choosing $\mathbf{A}, \mathbf{B}$ to *select* the d Hermite modes with largest $t_{\mathbf{n}}$.

Step 4: First-order dominance separates linear from higher-order modes. For any multi-index $\mathbf{n} \neq \mathbf{0}$ with total degree $|\mathbf{n}| := \sum_{k=1}^{r} n_k \geq 2$,

$$t_{\mathbf{n}} = \prod_{k=1}^{r} t_{ij,k}^{n_k} \leq t_{ij,1}^{|\mathbf{n}|} \leq t_{ij,1}^2,$$

since $0 \leq t_{ij,k} \leq t_{ij,1} < 1$. On the other hand, for the r first-order indices $\mathbf{e}_k$ we have $t_{\mathbf{e}_k} = t_{ij,k}$. If $r \geq 2$, Assumption 2 gives $t_{ij,r} > t_{ij,1}^2$; if $r = 1$, the strict inequality $t_{ij,1} > t_{ij,1}^2$ holds automatically because $t_{ij,1} \in (0, 1)$. Therefore,

$$t_{ij,r} > t_{ij,1}^2 \;\geq\; t_{\mathbf{n}} \qquad \forall\, \mathbf{n} \neq \mathbf{0} \text{ with } |\mathbf{n}| \geq 2.$$

Consequently, the r largest values among $\{t_{\mathbf{n}}\}_{\mathbf{n}\neq\mathbf{0}}$ are exactly $\{t_{ij,1}, \ldots, t_{ij,r}\}$, attained by the first-order Hermite modes $\{\Psi_{\mathbf{e}_1}, \ldots, \Psi_{\mathbf{e}_r}\}$ (which are linear functions of $\mathbf{u}_{1:r}$).

Step 5: Conclusion and explicit selector. Let $(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star)$ be any population maximizer with finite $d \geq r$ and let $\mathbf{C}^\star = \mathrm{Cov}(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star)$. By Equation 24 and attainability, the optimal value equals $\sum_{k=1}^{d} t_{(k)}$. By Step 4, the top r singular values of the optimal correlation structure must therefore come from the first-order modes, and these modes are strictly separated from all higher-order ones by the gap $t_{ij,r} - t_{ij,1}^2 > 0$.

Now use the post-orthogonal closure of the whitened class (Definition 2): take an SVD $\mathbf{C}^\star = \mathbf{Q}_i\, \mathrm{diag}(\sigma_1, \ldots, \sigma_d)\, \mathbf{Q}_j^\top$ and replace $\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star$ by $\mathbf{Q}_i^\top\tilde{\mathbf{h}}_i^\star$ and $\mathbf{Q}_j^\top\tilde{\mathbf{h}}_j^\star$. This produces another maximizer (same objective value) whose cross-covariance is diagonal with $\sigma_1 \geq \cdots \geq \sigma_d$. By Step 4 and $d \geq r$, the first r diagonal entries correspond to the first-order Hermite subspaces, hence there exist $\mathbf{O}_i, \mathbf{O}_j \in O(r)$ such that, P-a.s.,

$$\mathbf{P}_r\, \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_i\, \mathbf{u}_{1:r}, \qquad \mathbf{P}_r\, \tilde{\mathbf{h}}_j^\star(\mathbf{s}_j) = \mathbf{O}_j\, \mathbf{v}_{1:r},$$

with the explicit selector

$$\mathbf{P}_r := [\mathbf{I}_r\ \mathbf{0}] \in \mathbb{R}^{r\times d}.$$

Finally, substituting $\mathbf{u}_{1:r} = \mathbf{U}_{ij}(:, 1:r)^\top\mathbf{W}_i\mathbf{s}_i$ and $\mathbf{v}_{1:r} = \mathbf{V}_{ij}(:, 1:r)^\top\mathbf{W}_j\mathbf{s}_j$ yields exactly Equation 6 with $\mathbf{P}_{r_{ij}} = \mathbf{P}_r$. □

G PROOF OF THEOREM 5.2

Proof of Theorem 5.2. We work entirely in the source domain using the induced whitened mappings

$$\tilde{\mathbf{h}}_i := \tilde{\mathbf{f}}_i \circ \mathbf{g}_i, \qquad i \in [N].$$

By reparameterization invariance (Lemma 6), optimizing Equation 2 over $\{\tilde{\mathbf{f}}_i\}$ is equivalent to optimizing it over $\{\tilde{\mathbf{h}}_i\}$, and each $\tilde{\mathbf{h}}_i$ satisfies

$$\mathbb{E}[\tilde{\mathbf{h}}_i(\mathbf{s}_i)] = \mathbf{0}, \qquad \mathrm{Cov}(\tilde{\mathbf{h}}_i(\mathbf{s}_i)) = \mathbf{I}.$$

Define the pairwise objectives

$$J_{ij}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j) := \left\| \mathrm{Cov}\big(\tilde{\mathbf{h}}_i(\mathbf{s}_i), \tilde{\mathbf{h}}_j(\mathbf{s}_j)\big) \right\|_*, \qquad 1 \le i < j \le N,$$

so that

$$J(\tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_N) = \sum_{1 \le i < j \le N} J_{ij}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j).$$

Step 1: Pairwise optimality of a multi-view maximizer.

Let

$$M_{ij} := \sup_{\tilde{\mathbf{h}}_i,\, \tilde{\mathbf{h}}_j \text{ whitened}} J_{ij}(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j)$$

be the two-view population optimum (with $d_{\mathcal{Z}} \to \infty$). For any feasible collection,

$$J(\tilde{\mathbf{h}}_1, \dots, \tilde{\mathbf{h}}_N) \le \sum_{i<j} M_{ij}.$$

Under item 1.2, by the canonical factorization (Lemma 1) and the Mehler–Hermite expansion (Lemma 2), the cross-covariance operator between two views admits a diagonal spectral decomposition in the multivariate Hermite basis, with singular values $\{t_{\mathbf{n}}\}$. The corresponding singular functions achieve

$$J_{ij} = \sum_{\mathbf{n}} t_{\mathbf{n}} = M_{ij}.$$

Thus the upper bound $\sum_{i<j} M_{ij}$ is attainable.

Therefore, if $\{\tilde{\mathbf{h}}_i^\star\}_{i=1}^N$ is a global maximizer of J, we must have

$$J_{ij}(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star) = M_{ij}, \qquad \forall\, 1 \le i < j \le N. \tag{25}$$

Hence every pair $(\tilde{\mathbf{h}}_i^\star, \tilde{\mathbf{h}}_j^\star)$ is a two-view population maximizer.

Step 2: Apply two-view subspace identifiability.

Fix $i \neq j$ and let $r_{ij} = \mathrm{rank}(\mathbf{A}_i \mathbf{A}_j^\top)$. By Theorem 5.1, there exists $\mathbf{O}_{i|j} \in O(r_{ij})$ and a selection matrix $\mathbf{P}_{r_{ij}}^{(i|j)}$ such that

$$\mathbf{P}_{r_{ij}}^{(i|j)} \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_{i|j} \mathbf{U}_{ij}(:, 1 : r_{ij})^\top \mathbf{W}_i \mathbf{s}_i. \tag{26}$$

Thus, for each fixed i and every $j \neq i$, the representation contains an r_{ij}-dimensional block identifying the pairwise correlated subspace $\mathcal{U}_{i|j}$.

Step 3: Extract the multi-view intersection.

By Definition 1,

$$\mathcal{U}_i^{\mathrm{mv}} = \bigcap_{j \neq i} \mathcal{U}_{i|j}, \qquad r_i = \dim(\mathcal{U}_i^{\mathrm{mv}}).$$

Let $\mathbf{U}_i^{\mathrm{mv}} \in \mathbb{R}^{d_{\mathcal{S}_i} \times r_i}$ be an orthonormal basis of $\mathcal{U}_i^{\mathrm{mv}}$.

Since $\mathcal{U}_i^{\mathrm{mv}} \subseteq \mathcal{U}_{i|j_0}$ for any fixed $j_0 \neq i$, there exists $\mathbf{B}_i \in \mathbb{R}^{r_{ij_0} \times r_i}$ with orthonormal columns such that

$$\mathbf{U}_i^{\mathrm{mv}} = \mathbf{U}_{ij_0}(:, 1 : r_{ij_0})\mathbf{B}_i.$$

Multiplying Equation 26 (on pair (i, j_0)) by $\mathbf{B}_i^\top \mathbf{O}_{i|j_0}^\top$ gives

$$\mathbf{B}_i^\top \mathbf{O}_{i|j_0}^\top \mathbf{P}_{r_{ij_0}}^{(i|j_0)} \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = (\mathbf{U}_i^{\mathrm{mv}})^\top \mathbf{W}_i \mathbf{s}_i.$$

Define

$$\mathbf{P}_{r_i} := \mathbf{B}_i^\top \mathbf{O}_{i|j_0}^\top \mathbf{P}_{r_{ij_0}}^{(i|j_0)} \in \mathbb{R}^{r_i \times d_Z}.$$

Then

$$\mathbf{P}_{r_i} \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = (\mathbf{U}_i^{\mathrm{mv}})^\top \mathbf{W}_i \mathbf{s}_i. \tag{27}$$

Since $\mathbf{U}_i^{\mathrm{mv}}$ is defined only up to an orthogonal change of basis, absorbing that change into $\mathbf{O}_i \in O(r_i)$ yields

$$\mathbf{P}_{r_i} \tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_i (\mathbf{U}_i^{\mathrm{mv}})^\top \mathbf{W}_i \mathbf{s}_i.$$

Conclusion.

Equation 7 holds for every $i \in [N]$. Hence the multi-view population maximizers identify the jointly correlated subspaces $\mathcal{U}_i^{\mathrm{mv}}$ up to view-specific orthogonal transformations. □

H PROOF OF COROLLARY 2

Proof of Corollary 2. Fix an arbitrary $i \in [N]$. Recall that $\mathbf{P}_{i|j}$ and $\widehat{\mathbf{P}}_{i|j}$ denote the orthogonal projectors onto $\mathcal{U}_{i|j}$ and $\widehat{\mathcal{U}}_{i|j}$, respectively, and

$$\mathbf{S}_i := \frac{1}{N-1} \sum_{j \neq i} \mathbf{P}_{i|j}, \qquad \widehat{\mathbf{S}}_i := \frac{1}{N-1} \sum_{j \neq i} \widehat{\mathbf{P}}_{i|j}.$$

Step 1: Spectral characterization of $\mathcal{U}_i^{\mathrm{mv}}$. We first show that $\mathcal{U}_i^{\mathrm{mv}} = \bigcap_{j \neq i} \mathcal{U}_{i|j}$ coincides with the eigenspace of $\mathbf{S}_i$ associated with eigenvalue 1.

($\subseteq$) If $\mathbf{v} \in \mathcal{U}_i^{\mathrm{mv}}$, then $\mathbf{P}_{i|j}\mathbf{v} = \mathbf{v}$ for all $j \neq i$, hence

$$\mathbf{S}_i \mathbf{v} = \frac{1}{N-1} \sum_{j \neq i} \mathbf{P}_{i|j}\mathbf{v} = \frac{1}{N-1} \sum_{j \neq i} \mathbf{v} = \mathbf{v},$$

so $\mathbf{v}$ lies in the 1-eigenspace of $\mathbf{S}_i$.

($\supseteq$) Conversely, suppose $\mathbf{S}_i \mathbf{v} = \mathbf{v}$. Taking inner products,

$$\|\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{S}_i \mathbf{v} = \frac{1}{N-1} \sum_{j \neq i} \mathbf{v}^\top \mathbf{P}_{i|j}\mathbf{v} = \frac{1}{N-1} \sum_{j \neq i} \|\mathbf{P}_{i|j}\mathbf{v}\|_2^2. \tag{28}$$

Since $\mathbf{P}_{i|j}$ is an orthogonal projector, $\|\mathbf{P}_{i|j}\mathbf{v}\|_2 \leq \|\mathbf{v}\|_2$ for every j. Thus the average in Equation 28 can equal $\|\mathbf{v}\|_2^2$ only if $\|\mathbf{P}_{i|j}\mathbf{v}\|_2 = \|\mathbf{v}\|_2$ for all $j \neq i$, which forces $\mathbf{P}_{i|j}\mathbf{v} = \mathbf{v}$ for all $j \neq i$. Hence $\mathbf{v} \in \bigcap_{j \neq i} \mathcal{U}_{i|j} = \mathcal{U}_i^{\mathrm{mv}}$.

Therefore, the 1-eigenspace of $\mathbf{S}_i$ is exactly $\mathcal{U}_i^{\mathrm{mv}}$. In particular, if $r_i = \dim(\mathcal{U}_i^{\mathrm{mv}})$, then $\lambda_1(\mathbf{S}_i) = \cdots = \lambda_{r_i}(\mathbf{S}_i) = 1$ and

$$\Gamma_i := 1 - \lambda_{r_i+1}(\mathbf{S}_i) = \lambda_{r_i}(\mathbf{S}_i) - \lambda_{r_i+1}(\mathbf{S}_i) > 0.$$

Step 2: Davis–Kahan for the first inequality. Both $\mathbf{S}_i$ and $\widehat{\mathbf{S}}_i$ are symmetric. Let $\mathcal{U}_i^{\mathrm{mv}}$ be the invariant subspace of $\mathbf{S}_i$ corresponding to the eigenvalue cluster at 1 (multiplicity r_i), and let $\widehat{\mathcal{U}}_i^{\mathrm{mv}}$ be the top-r_i eigenspace of $\widehat{\mathbf{S}}_i$. By the Davis–Kahan $\sin\Theta$ theorem for symmetric matrices applied to the eigen-gap $\Gamma_i = \lambda_{r_i}(\mathbf{S}_i) - \lambda_{r_i+1}(\mathbf{S}_i)$,

$$\left\| \sin\Theta(\widehat{\mathcal{U}}_i^{\mathrm{mv}}, \mathcal{U}_i^{\mathrm{mv}}) \right\|_2 \leq \frac{\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2}{\Gamma_i},$$

which is the first displayed inequality.

Step 3: Bounding $\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2$ by pairwise errors. By linearity and the triangle inequality,

$$\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2 = \left\| \frac{1}{N-1} \sum_{j \neq i} \left(\widehat{\mathbf{P}}_{i|j} - \mathbf{P}_{i|j} \right) \right\|_2 \leq \frac{1}{N-1} \sum_{j \neq i} \|\widehat{\mathbf{P}}_{i|j} - \mathbf{P}_{i|j}\|_2 \leq \max_{j \neq i} \|\widehat{\mathbf{P}}_{i|j} - \mathbf{P}_{i|j}\|_2.$$

It remains to relate projector error to principal-angle error. Fix $j \neq i$ and let $\mathbf{U}, \widehat{\mathbf{U}}$ be orthonormal basis matrices for $\mathcal{U}_{i|j}$ and $\widehat{\mathcal{U}}_{i|j}$, respectively, so that $\mathbf{P} := \mathbf{U}\mathbf{U}^\top$ and $\widehat{\mathbf{P}} := \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$. Use the identity

$$\widehat{\mathbf{P}} - \mathbf{P} = (\mathbf{I} - \mathbf{P})\widehat{\mathbf{P}} - \mathbf{P}(\mathbf{I} - \widehat{\mathbf{P}}),$$

hence

$$\|\widehat{\mathbf{P}} - \mathbf{P}\|_2 \leq \|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{P}}\|_2 + \|\mathbf{P}(\mathbf{I} - \widehat{\mathbf{P}})\|_2.$$

Since $\widehat{\mathbf{P}}$ is a projector onto $\mathrm{col}(\widehat{\mathbf{U}})$,

$$\|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{P}}\|_2 = \|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{U}}\|_2 = \|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2,$$

and similarly

$$\|\mathbf{P}(\mathbf{I} - \widehat{\mathbf{P}})\|_2 = \|(\mathbf{I} - \widehat{\mathbf{P}})\mathbf{U}\|_2 = \|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2.$$

Therefore,

$$\|\widehat{\mathbf{P}}_{i|j} - \mathbf{P}_{i|j}\|_2 \leq 2\,\|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2.$$

Taking the maximum over $j \neq i$ and combining with the previous display yields

$$\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2 \leq 2 \max_{j \neq i} \|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2.$$

Plugging this into Step 2 proves the second displayed inequality:

$$\left\|\sin\Theta(\widehat{\mathcal{U}}_i^{\mathrm{mv}}, \mathcal{U}_i^{\mathrm{mv}})\right\|_2 \leq \frac{\|\widehat{\mathbf{S}}_i - \mathbf{S}_i\|_2}{\Gamma_i} \leq \frac{2}{\Gamma_i} \max_{j \neq i} \|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2.$$

Step 4: $O_{\mathbb{P}}(n^{-1/2})$ consistency. By Theorem 5.3, $\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 = O_{\mathbb{P}}(n^{-1/2})$ uniformly over pairs, and on the event $\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 \leq \Delta_{ij}/2$,

$$\|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2 \leq \frac{2\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2}{\Delta_{ij}} = O_{\mathbb{P}}(n^{-1/2}),$$

with constants governed by Δ_{ij}. Substituting into the inequality proved above shows that

$$\|\sin\Theta(\widehat{\mathcal{U}}_i^{\mathrm{mv}}, \mathcal{U}_i^{\mathrm{mv}})\|_2 = O_{\mathbb{P}}(n^{-1/2}),$$

with the pre-constant depending on Γ_i^{-1} and $\{\Delta_{ij}^{-1}\}_{j \neq i}$, as claimed. □

I PROOF OF THEOREM 5.3

Lemma 8 (Sub-Gaussian Covariance Concentration)**.** *Let $\{\mathbf{y}^{(t)}\}_{t=1}^n$ be i.i.d. centered sub-Gaussian random vectors in $\mathbb{R}^m$ with covariance $\mathbf{\Sigma}$, satisfying $\mathbb{E}\exp\left((\mathbf{v}^\top \mathbf{y}^{(1)})^2/(K^2 \mathbf{v}^\top \mathbf{\Sigma}\mathbf{v})\right) \leq 2$ for all $\mathbf{v} \in \mathbb{S}^{m-1}$. For the empirical covariance $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{t=1}^n \mathbf{y}^{(t)}\mathbf{y}^{(t)\top}$ and any $\delta \in (0,1)$, it holds with probability at least $1 - \delta$ that:*

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2 \leq cK^2 \|\mathbf{\Sigma}\|_2 \left(\sqrt{\frac{m + \log(2/\delta)}{n}} + \frac{m + \log(2/\delta)}{n} \right),$$

for a universal constant $c > 0$.

Proof of Theorem 5.3. Fix $d := d_{\mathcal{Z}}$ and $\delta \in (0,1)$. Recall the *population-whitened* representations

$$\tilde{\mathbf{z}}_i := \mathbf{\Sigma}_{ii}^{-1/2}\big(\mathbf{z}_i - \mathbb{E}[\mathbf{z}_i]\big), \qquad \mathrm{Cov}(\tilde{\mathbf{z}}_i) = \mathbf{I}_d,$$

and note that the population normalized cross-covariance is precisely

$$\mathbf{R}_{ij} = \mathbf{\Sigma}_{ii}^{-1/2}\mathbf{\Sigma}_{ij}\mathbf{\Sigma}_{jj}^{-1/2} = \mathrm{Cov}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j).$$

Throughout the proof, we work in these population-whitened coordinates (i.e., we replace $\tilde{\mathbf{z}}_i$ by $\mathbf{z}_i$ for notational simplicity). Thus we may assume

$$\mathbb{E}[\mathbf{z}_i] = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{z}_i) = \mathbf{I}_d, \qquad \mathbf{R}_{ij} = \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j). \tag{29}$$

The empirical (cross-)covariances are defined as in the main text,

$$\widehat{\mathbf{\Sigma}}_{ij} := \frac{1}{n}\sum_{t=1}^{n}(\mathbf{z}_i^{(t)} - \bar{\mathbf{z}}_i)(\mathbf{z}_j^{(t)} - \bar{\mathbf{z}}_j)^\top, \qquad \widehat{\mathbf{R}}_{ij} := \widehat{\mathbf{\Sigma}}_{ii}^{-1/2}\widehat{\mathbf{\Sigma}}_{ij}\widehat{\mathbf{\Sigma}}_{jj}^{-1/2}.$$

Step 1: sub-Gaussian concentration for pairwise (cross-)covariances. Fix a pair $1 \le i < j \le N$ and define the stacked vector

$$\mathbf{y}^{(t)} := \begin{bmatrix} \mathbf{z}_i^{(t)} \\ \mathbf{z}_j^{(t)} \end{bmatrix} \in \mathbb{R}^{2d}.$$

Let $\mathbf{v} = [\mathbf{v}_1^\top, \mathbf{v}_2^\top]^\top \in \mathbb{S}^{2d-1}$. By the triangle inequality for the Orlicz ψ_2 norm and Assumption 3,

$$\|\mathbf{v}^\top\mathbf{y}^{(t)}\|_{\psi_2} = \|\mathbf{v}_1^\top\mathbf{z}_i^{(t)} + \mathbf{v}_2^\top\mathbf{z}_j^{(t)}\|_{\psi_2} \le \|\mathbf{v}_1^\top\mathbf{z}_i^{(t)}\|_{\psi_2} + \|\mathbf{v}_2^\top\mathbf{z}_j^{(t)}\|_{\psi_2} \le \kappa(\|\mathbf{v}_1\|_2 + \|\mathbf{v}_2\|_2) \le \sqrt{2}\,\kappa.$$

Hence $\mathbf{y}^{(t)}$ is sub-Gaussian in $\mathbb{R}^{2d}$ with parameter $\lesssim \kappa$. Its population covariance is the block matrix

$$\mathbf{\Sigma}_{\mathbf{y}} := \mathrm{Cov}(\mathbf{y}^{(1)}) = \begin{pmatrix} \mathbf{I}_d & \mathbf{R}_{ij} \\ \mathbf{R}_{ij}^\top & \mathbf{I}_d \end{pmatrix}.$$

Moreover, $\|\mathbf{R}_{ij}\|_2 \le 1$ by Cauchy–Schwarz: for any unit $\mathbf{a}, \mathbf{b}$,

$$|\mathbf{a}^\top\mathbf{R}_{ij}\mathbf{b}| = \big|\mathbb{E}[(\mathbf{a}^\top\mathbf{z}_i)(\mathbf{b}^\top\mathbf{z}_j)]\big| \le \sqrt{\mathbb{E}(\mathbf{a}^\top\mathbf{z}_i)^2}\sqrt{\mathbb{E}(\mathbf{b}^\top\mathbf{z}_j)^2} = 1,$$

so $\|\mathbf{\Sigma}_{\mathbf{y}}\|_2 \le 2$.

Define the *uncentered* empirical covariance $\widetilde{\mathbf{\Sigma}}_{\mathbf{y}} := \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}^{(t)}\mathbf{y}^{(t)\top}$ and the centered one $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} := \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}^{(t)} - \bar{\mathbf{y}})(\mathbf{y}^{(t)} - \bar{\mathbf{y}})^\top = \widetilde{\mathbf{\Sigma}}_{\mathbf{y}} - \bar{\mathbf{y}}\bar{\mathbf{y}}^\top$. Applying Lemma 8 to $\{\mathbf{y}^{(t)}\}_{t=1}^n$ (with $m = 2d$ and $K \asymp \kappa$) yields that, with probability at least $1 - \delta_{ij}/2$,

$$\big\|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}}\big\|_2 \le c_1\kappa^2\left(\sqrt{\frac{2d + \log(4/\delta_{ij})}{n}} + \frac{2d + \log(4/\delta_{ij})}{n}\right), \tag{30}$$

for a universal constant $c_1 > 0$.

We also need to control the mean correction term $\bar{\mathbf{y}}\bar{\mathbf{y}}^\top$. By standard sub-Gaussian mean concentration (e.g., Vershynin, 2018), with probability at least $1 - \delta_{ij}/2$,

$$\|\bar{\mathbf{y}}\|_2 \le c_2\kappa\sqrt{\frac{2d + \log(2/\delta_{ij})}{n}}, \tag{31}$$

for a universal $c_2 > 0$, hence $\|\bar{\mathbf{y}}\bar{\mathbf{y}}^\top\|_2 = \|\bar{\mathbf{y}}\|_2^2 \le c_2^2\kappa^2\frac{2d+\log(2/\delta_{ij})}{n}$. Combining Equation 30–Equation 31 and absorbing the quadratic term into the square-root term (and using a larger universal constant when the square-root term exceeds 1), we obtain that with probability at least $1 - \delta_{ij}$,

$$\big\|\widehat{\mathbf{\Sigma}}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}}\big\|_2 \le c_3\kappa^2\sqrt{\frac{2d + \log(4/\delta_{ij})}{n}}, \tag{32}$$

for a universal $c_3 > 0$.

Since each block of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}}$ is a submatrix of the full difference, we have (on the same event)

$$\|\widehat{\mathbf{\Sigma}}_{ii} - \mathbf{I}\|_2 \le \varepsilon_{ij}, \qquad \|\widehat{\mathbf{\Sigma}}_{jj} - \mathbf{I}\|_2 \le \varepsilon_{ij}, \qquad \|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{R}_{ij}\|_2 \le \varepsilon_{ij}, \tag{33}$$

where

$$\varepsilon_{ij} := c_3\kappa^2\sqrt{\frac{2d + \log(4/\delta_{ij})}{n}}.$$

Step 2: control of empirical whitening matrices. Assume $\varepsilon_{ij} < 1$. Then Weyl's inequality gives $\lambda_{\min}(\widehat{\mathbf{\Sigma}}_{ii}) \geq 1-\varepsilon_{ij} > 0$ and similarly for j, so $\widehat{\mathbf{\Sigma}}_{ii}^{-1/2}$ and $\widehat{\mathbf{\Sigma}}_{jj}^{-1/2}$ exist. Moreover, all eigenvalues of $\widehat{\mathbf{\Sigma}}_{ii}$ lie in $[1-\varepsilon_{ij}, 1+\varepsilon_{ij}]$, hence those of $\widehat{\mathbf{\Sigma}}_{ii}^{-1/2}$ lie in $[(1+\varepsilon_{ij})^{-1/2}, (1-\varepsilon_{ij})^{-1/2}]$, and therefore

$$\|\widehat{\mathbf{\Sigma}}_{ii}^{-1/2}\|_2 \leq (1-\varepsilon_{ij})^{-1/2}, \qquad \|\widehat{\mathbf{\Sigma}}_{ii}^{-1/2} - \mathbf{I}\|_2 \leq (1-\varepsilon_{ij})^{-1/2} - 1 \leq \frac{\varepsilon_{ij}}{1-\varepsilon_{ij}}. \tag{34}$$

The same bounds hold for $\widehat{\mathbf{\Sigma}}_{jj}^{-1/2}$.

Step 3: concentration of $\widehat{\mathbf{R}}_{ij}$. Write $\mathbf{A}_i := \widehat{\mathbf{\Sigma}}_{ii}^{-1/2}$ and $\mathbf{A}_j := \widehat{\mathbf{\Sigma}}_{jj}^{-1/2}$. Then $\widehat{\mathbf{R}}_{ij} = \mathbf{A}_i \widehat{\mathbf{\Sigma}}_{ij} \mathbf{A}_j$ and

$$\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij} = \mathbf{A}_i(\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{R}_{ij})\mathbf{A}_j + \big(\mathbf{A}_i \mathbf{R}_{ij} \mathbf{A}_j - \mathbf{R}_{ij}\big).$$

Using $\|\mathbf{R}_{ij}\|_2 \leq 1$, the triangle inequality, and Equation 34, on the event $\varepsilon_{ij} \leq 1/2$ we obtain

$$\begin{aligned}
\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 &\leq \|\mathbf{A}_i\|_2 \, \|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{R}_{ij}\|_2 \, \|\mathbf{A}_j\|_2 + \|(\mathbf{A}_i - \mathbf{I})\mathbf{R}_{ij}\mathbf{A}_j\|_2 + \|\mathbf{R}_{ij}(\mathbf{A}_j - \mathbf{I})\|_2 \\
&\leq (1-\varepsilon_{ij})^{-1}\varepsilon_{ij} + \|\mathbf{A}_i - \mathbf{I}\|_2 \|\mathbf{A}_j\|_2 + \|\mathbf{A}_j - \mathbf{I}\|_2 \\
&\leq 2\varepsilon_{ij} + 2\varepsilon_{ij}\sqrt{2} + 2\varepsilon_{ij} \;\leq\; 7\,\varepsilon_{ij}.
\end{aligned}$$

If $\varepsilon_{ij} > 1/2$, then the desired bound is trivial after enlarging constants, because both $\|\widehat{\mathbf{R}}_{ij}\|_2 \leq 1$ and $\|\mathbf{R}_{ij}\|_2 \leq 1$ (immediate from Cauchy–Schwarz on the empirical and population distributions), so $\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 \leq 2$.

Step 4: union bound over all pairs. Set $\delta_{ij} := \delta/N^2$ and take a union bound over all $\binom{N}{2} \leq N^2/2$ pairs. Using $2d + \log(4/\delta_{ij}) = 2d + \log(4N^2/\delta) \lesssim d + \log(N/\delta)$ and absorbing constants, we conclude that with probability at least $1-\delta$, simultaneously for all $1 \leq i < j \leq N$,

$$\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2 \leq C\,\kappa^2 \sqrt{\frac{d + \log(N/\delta)}{n}},$$

for a universal constant $C > 0$. This proves Equation 9.

Step 5: subspace perturbation bound (Wedin). Fix $i < j$ and let $\mathbf{E} := \widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}$. Let $r := r_{ij}$ and let $\mathcal{U}_{i|j}$ (resp. $\widehat{\mathcal{U}}_{i|j}$) be the rank-$r$ left singular subspace of $\mathbf{R}_{ij}$ (resp. $\widehat{\mathbf{R}}_{ij}$). Let $\Delta_{ij} := \sigma_r(\mathbf{R}_{ij}) - \sigma_{r+1}(\mathbf{R}_{ij}) > 0$. By Weyl's inequality, for all k, $|\sigma_k(\widehat{\mathbf{R}}_{ij}) - \sigma_k(\mathbf{R}_{ij})| \leq \|\mathbf{E}\|_2$. Hence on the event $\|\mathbf{E}\|_2 \leq \Delta_{ij}/2$,

$$\sigma_r(\widehat{\mathbf{R}}_{ij}) - \sigma_{r+1}(\widehat{\mathbf{R}}_{ij}) \geq \big(\sigma_r(\mathbf{R}_{ij}) - \|\mathbf{E}\|_2\big) - \big(\sigma_{r+1}(\mathbf{R}_{ij}) + \|\mathbf{E}\|_2\big) = \Delta_{ij} - 2\|\mathbf{E}\|_2 \geq \Delta_{ij}/2.$$

Wedin's $\sin\Theta$ theorem for singular subspaces then gives

$$\|\sin\Theta(\widehat{\mathcal{U}}_{i|j}, \mathcal{U}_{i|j})\|_2 \leq \frac{\|\mathbf{E}\|_2}{\Delta_{ij}/2} = \frac{2\|\widehat{\mathbf{R}}_{ij} - \mathbf{R}_{ij}\|_2}{\Delta_{ij}},$$

which is Equation 10. Applying the same argument to $\widehat{\mathbf{R}}_{ij}^\top$ and $\mathbf{R}_{ij}^\top$ yields the identical bound for the right singular subspace $\widehat{\mathcal{U}}_{j|i}$. □

J PROOF OF ADDITIONAL LATENT PRIORS

The Gaussian proof relies on the canonical Gaussian factorization in Lemma 1 and the Mehler–Hermite expansion in Lemma 2. For the additional latent priors considered in the experiments, the same argument applies after replacing the Hermite system by the corresponding Lancaster orthogonal-polynomial system [Lancaster, 1958, Eagleson, 1964]. We state this extension at the level needed for the subspace-identifiability results.

Remark 5 (Scope of the non-Gaussian extension). For non-Gaussian priors, the statement below should be read as an admissible pairwise Lancaster condition in the canonical coordinates. Unlike the Gaussian case, arbitrary real-valued whitening and orthogonal canonicalization do not in general preserve Poisson, negative-binomial, Gamma, or hypergeometric product structure. Thus the extension requires that, for each pair of views, the canonical source coordinates themselves admit the Lancaster expansion below.

Assumption 5 (Pairwise Lancaster canonical coordinates)**.** Fix $1 \le i < j \le N$, write $r := r_{ij}$, and define the canonical coordinates as in Equation 5,

$$\mathbf{u} = \mathbf{U}_{ij}^\top \mathbf{W}_i \mathbf{s}_i, \qquad \mathbf{v} = \mathbf{V}_{ij}^\top \mathbf{W}_j \mathbf{s}_j.$$

For each correlated coordinate $k \in [r]$, assume that the pair (u_k, v_k) has a bivariate Lancaster law with common standardized marginal $\nu_{ij,k}$, orthonormal polynomial basis

$$\{\psi_{ij,k,n}\}_{n \in I_{ij,k}} \subset L^2(\nu_{ij,k}), \qquad \psi_{ij,k,0} \equiv 1, \qquad \psi_{ij,k,1}(x) = x,$$

and Lancaster coefficients $\{\lambda_{ij,k,n}\}_{n \in I_{ij,k}}$, with $\lambda_{ij,k,0} = 1$, such that

$$dP_{u_k,v_k}(a,b) = \left(\sum_{n \in I_{ij,k}} \lambda_{ij,k,n}\, \psi_{ij,k,n}(a) \psi_{ij,k,n}(b) \right) d\nu_{ij,k}(a) d\nu_{ij,k}(b). \tag{35}$$

Assume further that the correlated coordinate pairs $\{(u_k, v_k)\}_{k=1}^r$ are mutually independent, and that all coordinates with index larger than r are cross-view independent. For the infinite-dimensional statement, assume the resulting diagonal operator is trace-class:

$$\sum_{\mathbf{n} \neq \mathbf{0}} |\lambda_{ij,\mathbf{n}}| < \infty.$$

For a multi-index $\mathbf{n} = (n_1, \ldots, n_r) \in \mathcal{I}_{ij} := \prod_{k=1}^r I_{ij,k}$, define

$$\Psi_{ij,\mathbf{n}}(\mathbf{u}) := \prod_{k=1}^r \psi_{ij,k,n_k}(u_k), \qquad \Psi_{ij,\mathbf{n}}(\mathbf{v}) := \prod_{k=1}^r \psi_{ij,k,n_k}(v_k),$$

and

$$\lambda_{ij,\mathbf{n}} := \prod_{k=1}^r \lambda_{ij,k,n_k}.$$

The constant mode $\mathbf{n} = \mathbf{0}$ is omitted below because the representations are centered.

Lemma 9 (Lancaster diagonalization of pairwise correlations)**.** *Under Assumption 5, the tensor-product basis $\{\Psi_{ij,\mathbf{n}}\}_{\mathbf{n} \in \mathcal{I}_{ij}}$ is orthonormal, and for all $\mathbf{m}, \mathbf{n} \in \mathcal{I}_{ij}$,*

$$\mathbb{E}[\Psi_{ij,\mathbf{m}}(\mathbf{u}) \Psi_{ij,\mathbf{n}}(\mathbf{v})] = \lambda_{ij,\mathbf{n}} \delta_{\mathbf{mn}}. \tag{36}$$

Proof sketch. The tensor-product basis is orthonormal by independence across canonical coordinates and by the one-dimensional orthonormality of each $\{\psi_{ij,k,n}\}_{n \in I_{ij,k}}$. Multiplying the scalar Lancaster expansions in (35) over $k = 1, \ldots, r$ gives the product expansion of the joint law of $(\mathbf{u}[1:r], \mathbf{v}[1:r])$. Integrating $\Psi_{ij,\mathbf{m}}(\mathbf{u}) \Psi_{ij,\mathbf{n}}(\mathbf{v})$ against this product expansion and using orthonormality in each coordinate leaves only the diagonal term $\mathbf{m} = \mathbf{n}$, yielding (36). □

Proposition 2 (Two-view subspace recovery under Lancaster priors)**.** *Fix $1 \le i < j \le N$ and suppose Assumption 5 holds for this pair. Let $(\tilde{\mathbf{h}}_i, \tilde{\mathbf{h}}_j)$ be any feasible whitened source-domain representation pair. Expanding each coordinate in the Lancaster tensor basis gives*

$$\tilde{h}_{i,p}(\mathbf{u}) = \sum_{\mathbf{n} \neq \mathbf{0}} a_{p,\mathbf{n}} \Psi_{ij,\mathbf{n}}(\mathbf{u}), \qquad \tilde{h}_{j,q}(\mathbf{v}) = \sum_{\mathbf{n} \neq \mathbf{0}} b_{q,\mathbf{n}} \Psi_{ij,\mathbf{n}}(\mathbf{v}).$$

Let

$$\mathbf{A} = (a_{p,\mathbf{n}})_{p,\mathbf{n}}, \qquad \mathbf{B} = (b_{q,\mathbf{n}})_{q,\mathbf{n}}, \qquad \mathbf{D}_{ij} := \operatorname{diag}\big((\lambda_{ij,\mathbf{n}})_{\mathbf{n} \neq \mathbf{0}}\big).$$

Then whitening implies

$$\mathbf{A}\mathbf{A}^\top = \mathbf{I}, \qquad \mathbf{B}\mathbf{B}^\top = \mathbf{I},$$

and the whitened pairwise cross-covariance diagonalizes as

$$\operatorname{Cov}\Big(\tilde{\mathbf{h}}_i(\mathbf{s}_i), \tilde{\mathbf{h}}_j(\mathbf{s}_j)\Big) = \mathbf{A}\mathbf{D}_{ij}\mathbf{B}^\top. \tag{37}$$

Consequently, the pairwise CCA objective selects the largest absolute Lancaster coefficients.

If, in addition, the first-order coefficients satisfy the dominance condition

$$\min_{k\in[r]} |\lambda_{ij,k,1}| > \sup_{\substack{\mathbf{n}\in\mathcal{I}_{ij}\setminus\{\mathbf{0}\}\\ \mathbf{n}\notin\{\mathbf{e}_1,\dots,\mathbf{e}_r\}}} |\lambda_{ij,\mathbf{n}}|, \tag{38}$$

then the leading r selected modes are exactly the first-order modes $\Psi_{ij,\mathbf{e}_k}$, $k \in [r]$. Since $\Psi_{ij,\mathbf{e}_k}(\mathbf{u}) = u_k$ and $\Psi_{ij,\mathbf{e}_k}(\mathbf{v}) = v_k$, the conclusion of Theorem 5.1 holds with the Hermite modes replaced by the first-order Lancaster modes. In particular, there exist orthogonal matrices $\mathbf{O}_i, \mathbf{O}_j \in O(r)$ such that

$$\mathbf{P}_r\tilde{\mathbf{h}}_i^\star(\mathbf{s}_i) = \mathbf{O}_i\mathbf{U}_{ij}(:,1:r)^\top\mathbf{W}_i\mathbf{s}_i, \qquad \mathbf{P}_r\tilde{\mathbf{h}}_j^\star(\mathbf{s}_j) = \mathbf{O}_j\mathbf{V}_{ij}(:,1:r)^\top\mathbf{W}_j\mathbf{s}_j,$$

up to the same CCA post-orthogonal ambiguity as in Theorem 5.1.

Proof sketch. By reparameterization invariance (Lemma 6), it suffices to work in the source domain. After expanding the centered whitened maps in the orthonormal Lancaster tensor basis, whitening gives row-orthonormal coefficient matrices $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ and $\mathbf{B}\mathbf{B}^\top = \mathbf{I}$, exactly as in Equation 3. Using Lemma 9, the pairwise cross-covariance takes the diagonal coefficient form (37).

The nuclear-norm CCA objective is therefore bounded by the Ky Fan norm of the diagonal operator $\mathbf{D}_{ij}$, and this bound is achieved by selecting the diagonal modes with largest absolute coefficients. Under (38), the leading r coefficients are precisely those indexed by $\mathbf{e}_1, \dots, \mathbf{e}_r$. These first-order basis functions are the standardized canonical coordinates themselves. Hence the selected coordinates span

$$\operatorname{span}(u_1, \dots, u_r) = \operatorname{span}\big(\mathbf{U}_{ij}(:,1:r)^\top\mathbf{W}_i\mathbf{s}_i\big)$$

and analogously for view j. The remaining ambiguity is only an orthogonal transformation inside the selected r-dimensional block, which proves pairwise correlated subspace recovery. □

Corollary 3 (Multi-view extension under Lancaster priors)**.** *Suppose Assumption 5 and* (38) *hold for every pair $1 \le i < j \le N$. Then the multi-view generalized CCA maximizer identifies the jointly correlated subspace*

$$\mathcal{U}_i^{\mathrm{mv}} = \bigcap_{j\neq i} \mathcal{U}_{i|j}$$

in the sense of Theorem 5.2.

Proof sketch. The generalized objective in Equation 2 is the sum of pairwise nuclear-norm CCA objectives. The pairwise Lancaster argument above gives the same pairwise saturation property used in the proof of Theorem 5.2: every globally optimal multi-view solution must be pairwise optimal for each view pair. Applying Proposition 2 to every pair yields the collection of pairwise subspaces $\{\mathcal{U}_{i|j} : j \neq i\}$. Intersecting these subspaces gives the jointly correlated subspace $\mathcal{U}_i^{\mathrm{mv}}$, exactly as in Definition 1. □

Poisson prior. For a raw Poisson marginal $X \sim \mathrm{Poisson}(\lambda)$,

$$\nu(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x \in \mathbb{N}_0.$$

The corresponding orthogonal polynomials are the Charlier polynomials. The first-degree normalized polynomial is

$$\psi_1(x) = \frac{x-\lambda}{\sqrt{\lambda}}.$$

Thus, after standardization, the first-degree Lancaster mode is the canonical coordinate itself. Under (38), CCA therefore selects the Poisson first-order modes before any higher-order Charlier modes, yielding the same pairwise and multi-view subspace recovery claims as above.

Negative-binomial prior. Using the convention

$$\nu(x) = \binom{r+x-1}{x} p^r (1-p)^x, \qquad x \in \mathbb{N}_0, \quad 0 < p < 1,$$

the associated orthogonal polynomials are the Meixner polynomials. The mean and variance are

$$\mu = \frac{r(1-p)}{p}, \qquad \sigma^2 = \frac{r(1-p)}{p^2},$$

so the first-degree normalized polynomial is

$$\psi_1(x) = \frac{x-\mu}{\sigma}.$$

Hence the first-order Meixner modes are affine in the latent coordinate. Once (38) excludes higher-degree Meixner modes, the proof of Proposition 2 applies verbatim.

Hypergeometric prior. For $X \sim \mathrm{Hypergeometric}(N, K, n)$,

$$\nu(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \qquad x = \max\{0, n-(N-K)\}, \ldots, \min\{n, K\}.$$

The associated orthogonal polynomials are the Hahn polynomials. Since the support is finite, the polynomial system has finite maximal degree. The mean and variance are

$$\mu = n\frac{K}{N}, \qquad \sigma^2 = n\frac{K}{N}\left(1-\frac{K}{N}\right)\frac{N-n}{N-1},$$

and

$$\psi_1(x) = \frac{x-\mu}{\sigma}.$$

The Lancaster diagonalization is therefore finite-dimensional in this case. The sign of the degree-one coefficient may be negative, e.g. under sampling-without-replacement dependence, but the CCA objective uses singular values; hence only $|\lambda_{ij,k,1}|$ matters, and the sign is absorbed into the orthogonal ambiguity.

Gamma prior. For $X \sim \mathrm{Gamma}(k, \theta)$,

$$d\nu(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \mathbf{1}_{(0,\infty)}(x)\, dx.$$

The corresponding orthogonal polynomials are the generalized Laguerre polynomials $L_m^{(k-1)}(x/\theta)$. The mean and variance are

$$\mu = k\theta, \qquad \sigma^2 = k\theta^2,$$

so

$$\psi_1(x) = \frac{x-k\theta}{\sqrt{k}\theta}.$$

Thus the first-degree Laguerre mode is affine in the raw Gamma coordinate and equals the coordinate itself after standardization. Under (38), the CCA optimum selects these first-order Laguerre modes before all higher-order modes, giving the same subspace-identifiability conclusion.

K ADDITIONAL EXPERIMENTAL RESULTS

Table 4: Comparison of the mean and maximum principal angles (PA_{mean}, $PA_{\max}$) on synthetic data ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 5, \forall i \in [3]$) under Poisson distribution (p_ϕ).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$
BarlowTwins	34.47 ±0.20	86.87 ±1.10	31.68 ±0.16	82.34 ±0.90	33.07 ±0.29	84.62 ±0.86
InfoNCE	5.12 ±0.13	**6.93 ±0.46**	**4.40 ±0.12**	7.78 ±0.24	5.92 ±0.06	8.03 ±0.33
W-MSE	**4.70 ±0.09**	7.50 ±0.45	4.49 ±0.02	**7.52 ±0.77**	5.02 ±0.04	8.82 ±0.32
GCCA	5.00 ±0.15	6.98 ±0.37	5.50 ±0.10	8.03 ±0.39	**3.87 ±0.01**	**7.72 ±0.62**

Table 5: Comparison of the mean and maximum principal angles (PA_{mean}, $PA_{\max}$) on synthetic data ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 5, \forall i \in [3]$) under negative binomial distribution (p_ϕ).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$
BarlowTwins	34.42 ±0.19	86.24 ±1.00	32.31 ±0.30	79.76 ±0.86	33.56 ±0.42	84.23 ±0.94
InfoNCE	3.98 ±0.06	9.15 ±0.66	5.49 ±0.21	7.54 ±0.28	**4.21 ±0.07**	**8.38 ±0.57**
W-MSE	4.13 ±0.11	6.84 ±0.34	**5.76 ±0.08**	**8.22 ±0.67**	5.00 ±0.17	8.48 ±0.49
GCCA	**4.89 ±0.14**	**8.00 ±0.42**	6.18 ±0.31	8.26 ±0.26	6.55 ±0.09	8.17 ±0.40

Table 6: Comparison of the mean and maximum principal angles (PA_{mean}, $PA_{\max}$) on synthetic data ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 5, \forall i \in [3]$) under Gamma distribution (p_ϕ).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$
BarlowTwins	33.81±0.07	86.63±0.95	32.88±0.21	80.63±0.66	33.83±0.21	84.90±0.85
InfoNCE	**3.54±0.15**	8.49±0.54	5.59±0.01	9.81±0.35	5.13±0.15	**6.05±0.64**
W-MSE	5.12±0.01	8.05±0.68	**4.74±0.01**	**7.83±0.62**	**4.59±0.01**	9.08±0.52
GCCA	4.85±0.04	**8.02±0.31**	4.97±0.09	9.15±0.30	5.58±0.23	7.44±0.53

Table 7: Comparison of the mean and maximum principal angles (PA_{mean}, $PA_{\max}$) on synthetic data ($d_{\mathcal{S}_i} = d_{\mathcal{Z}} = 5, \forall i \in [3]$) under Gaussian prior ($p_\phi$).

Methods	$\tilde{\mathbf{f}}_1$		$\tilde{\mathbf{f}}_2$		$\tilde{\mathbf{f}}_3$	
	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$	PA_{mean}	$PA_{\max}$
BarlowTwins	34.05±0.28	85.39±0.88	31.78±0.39	82.21±0.80	33.94±0.26	84.21±0.83
InfoNCE	**4.20±0.05**	8.26±0.34	5.87±0.09	8.30±0.35	6.00±0.06	**7.85±0.54**
W-MSE	5.08±0.16	8.37±0.42	5.76±0.04	**7.31±0.45**	5.07±0.01	8.77±0.47
GCCA	4.63±0.01	**7.87±0.41**	**5.59±0.01**	9.45±0.36	**4.08±0.22**	8.91±0.60