WHICH TASKS SURVIVE SELF-SUPERVISED LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Same-instance self-supervised learning (SSL) learns representations by enforcing consistency across two views of the same underlying instance. This principle alone, however, does not determine which downstream tasks remain recoverable from the learned representation. We study this question through *semantic recoverability*, defined as the amount of a task's posterior score captured by the represented function space. We show that, for centered and whitened representations, recoverability exactly determines directional class-distance-normalized variance (CDNV), controls few-shot nearest-centroid classification, and governs the strength of task-relevant semantic directions. The population linear probe and centroid axis coincide, and multiple well-recovered tasks approach a factorial centroid geometry. We then analyze a canonical two-view SSL objective and show that its population optimum spans the leading cross-view-stable modes of the associated two-view operator. This yields a closed-form spectral characterization of semantic recoverability: a downstream task is preserved to the extent that its posterior lies in the selected spectral subspace. We validate these predictions on synthetic and real datasets across several SSL methods, testing the predicted relationships among recoverability, directional geometry, spectral structure, and few-shot transfer. Together, these results give a task-level account of what information survives same-instance SSL and how the retained information appears in downstream geometry and transfer.

1 INTRODUCTION

Self-supervised learning (SSL) (Balestriero et al., 2023) learns representations without labels and performs well across vision, language, speech, and multimodal learning (Chen et al., 2020; He et al., 2020; Zbontar et al., 2021; He et al., 2022; Oquab et al., 2024; Gao et al., 2021; Reimers & Gurevych, 2019; Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Baevski et al., 2022; Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Tschannen et al., 2025). A common principle is simple: two views of the same instance should receive compatible representations. Yet two views can share many kinds of information: semantic content, style, background, nuisance structure, or accidental regularities. Agreement alone therefore does not reveal which downstream distinctions survive.

This leads to a basic task-level question: given a representation learned only from paired views, which semantic tasks can still be recovered from it? For example, a representation learned from color-augmented face images might preserve whether a person is smiling while losing information about hair color. Existing theory provides two pieces of the answer. Operator and spectral analyses characterize which cross-view-stable directions are preferred by idealized SSL objectives (Zhai et al., 2025; Zhai, 2025; Ermolov et al., 2021; Weng et al., 2022), while geometric analyses show that SSL representations can develop strong label-aligned directional structure without global class collapse (Luthra et al., 2025; Wang & Isola, 2020; Garrido et al., 2023; Balestriero & LeCun, 2022; Qiu et al., 2023). We therefore ask:

> ***Which downstream tasks survive same-instance SSL, and what geometry and few-shot behavior follow from their recoverability?***

Our starting point is that this question is task-dependent. A representation can preserve one semantic partition while discarding another. For a fixed downstream task, the relevant object is the part of its

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

posterior score that remains visible in the represented function space, as illustrated in Fig. 2. We call this quantity *captured posterior energy*, $B(F)$, and use it to measure semantic recoverability independently of how the representation was learned.

Our theory addresses two questions: why semantic recoverability matters, and when SSL makes a task recoverable. Our contributions are threefold. **(i)** We formulate task survival through captured posterior energy, separating recoverability from any particular SSL objective. **(ii)** For centered and whitened representations, we derive an exact relation between recoverability and directional class-distance-normalized variance (CDNV) (Luthra et al., 2025), establish probe–centroid alignment, obtain a direct finite-sample guarantee for few-shot nearest-centroid classification, and characterize the joint centroid geometry of multiple well-recovered tasks. **(iii)** For a canonical two-view population SSL objective, we show that the learned subspace is spanned by the leading cross-view-stable modes of the associated two-view operator. Consequently, semantic recoverability is exactly the spectral overlap between the task posterior and the selected subspace, yielding a criterion for which downstream tasks survive SSL. We further extend this subspace perspective to covariance-regularized and prediction-based SSL surrogates in the appendix.

1.1 Related Work

Our work connects two lines of SSL theory that are usually studied separately: task-relevant representation geometry and the population subspaces selected by two-view objectives.

Directional geometry and downstream transfer. SSL representations can support strong downstream classification and retrieval, while remaining globally anisotropic and retaining substantial variance in nuisance directions (Chen et al., 2020; Caron et al., 2021; Ben-Shaul et al., 2023; Weng et al., 2025; Oquab et al., 2024; Wang & Isola, 2020; Wang & Liu, 2021; Chen et al., 2021; HaoChen et al., 2021; Saunshi et al., 2019). This motivates task-aware geometric measures rather than global collapse. In particular, directional CDNV isolates within-class variation along the semantic decision direction and has been shown to track linear-probe and few-shot transfer across SSL methods (Luthra et al., 2025). Related work on neural collapse and few-shot transfer connects class geometry to generalization (Papyan et al., 2020; Han et al., 2022; Galanti et al., 2022b;a;c; Wang et al., 2023; Goldblum et al., 2020). Our goal is complementary: we ask what property of a label-free learned subspace predicts whether this favorable task geometry will appear.

Spectral and operator views of SSL. Many SSL methods combine cross-view agreement with mechanisms that prevent representational collapse (Wang & Isola, 2020; Zbontar et al., 2021; Bardes et al., 2022; Garrido et al., 2023; Balestriero & LeCun, 2022; Ermolov et al., 2021). At the population level, whitening-style formulations connect naturally to CCA, maximal correlation, and conditional expectation operators (Hotelling, 1936; Hannan, 1961; Breiman & Friedman, 1985; Lai & Fyfe, 2000; Hardoon et al., 2004; Fukumizu et al., 2007; Andrew et al., 2013; Michaeli et al., 2016; Asoodeh et al., 2015). Recent analyses characterize the subspace selected by such objectives through the leading spectrum of an induced two-view operator (Zhai et al., 2025; Zhai, 2025). We use this characterization for computing which task posteriors are recoverable from the selected subspace.

Relation to broader SSL theory. A broader literature studies why SSL captures useful structure through mutual-information views (Bachman et al., 2019; McAllester & Stratos, 2020; Tschannen et al., 2020), alignment and uniformity (Wang & Isola, 2020; Wang & Liu, 2021; Chen et al., 2021), latent-variable recovery and augmentation structure (Saunshi et al., 2019; Tosh et al., 2021; Zimmermann et al., 2021; Ash et al., 2022; Nozawa & Sato, 2021; HaoChen et al., 2021; Shen et al., 2022; Awasthi et al., 2022; Saunshi et al., 2022), and optimization, projection heads, and sample complexity (HaoChen & Ma, 2023; Parulekar et al., 2023; Wen & Li, 2021; Tian, 2023; Tian et al., 2020; Gupta et al., 2022; Gui et al., 2023; Alon et al., 2024). Our question is narrower and task-conditional: given the information retained by a learned representation, which downstream posterior scores survive, and what geometric and few-shot consequences follow?

2 Problem Setup

We consider a latent instance that generates an observed sample and two conditionally independent views, together with a semantic task defined at the instance level. A representation is learned from

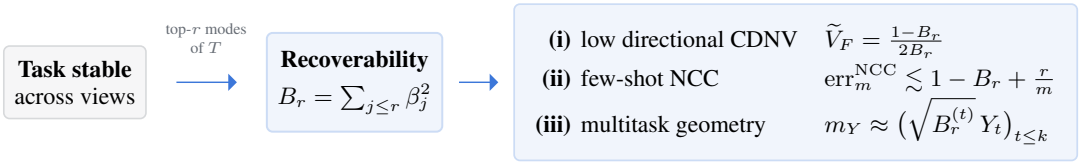


Figure 1: Conceptual flow of the theory. Same-instance SSL retains the most cross-view-stable modes of the two-view operator T (Prop. 4.1), so a task survives to the extent that its posterior overlaps that subspace (Cor. 4.2). Under centered whitening, B_r fixes the directional collapse law (Prop. 3.1) and few-shot nearest-centroid guarantee (Thm. 3.4); the task-specific $B_r^{(t)}$ determine the multitask centroid geometry (Thm. 3.5).

paired views without access to the task labels and is then evaluated by how well the task can be recovered from a small labeled support set.

2.1 Latent Instance Model and Two-View Generation

Let C denote the *latent instance* (for example, the underlying object, scene, or document) taking values in a measurable space $\mathcal{C}$, with distribution P_C. The semantic label is determined at the instance level: $Y = y(C) \in \{\pm 1\}$, $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = \frac{1}{2}$. Thus C may contain much more information than Y: many latent instances can share the same label.

We let $X^{(0)}$ denote the sample associated with C, drawn as $X^{(0)} \sim P_0(\cdot \mid C)$. A positive pair is formed by drawing two independent views, $X^{(1)}, X^{(2)} \overset{\text{i.i.d.}}{\sim} P(\cdot \mid X^{(0)})$. In vision, for example, C may encode image content, while each view is produced by a random augmentation such as cropping, color jitter, or blur. Let P_X denote the marginal law of a single view.

2.2 Downstream Few-Shot Learning

Let $F : \mathcal{X} \to \mathbb{R}^r$ be a representation. We evaluate its downstream quality through the expected m-shot error of a classifier built on top of F. Given a support set with m labeled examples per class, with all support examples sampled independently, $S_m = \{(X_{s,i}, s) : s \in \{\pm 1\}, X_{s,i} \sim P(\cdot \mid Y = s), (i \in [m])\}$, let $\hat{g}^{\mathrm{A}}_{S_m}$ denote the classifier returned by a learning rule A from the embedded support set. We define $\mathrm{err}_m^{\mathrm{A}}(F) := \mathbb{E}_{S_m}\left[\mathbb{P}_{(X,Y)}\big(\hat{g}^{\mathrm{A}}_{S_m}(F(X)) \neq Y\big)\right]$, where (X, Y) is drawn from the single-view distribution induced by the latent-instance model.

We focus on nearest-class-centroid (NCC) classification: $g^{\mathrm{NCC}}_{S_m}(z) := \arg\min_{s \in \{\pm 1\}} \|z - \widehat{\mu}_s\|_2$, where $\widehat{\mu}_s := \frac{1}{m}\sum_{i=1}^m F(X_{s,i})$. This gives an operational notion of task survival: a useful representation should make the task learnable from only a small labeled support set. The next section identifies a population quantity that controls when this happens.

3 Semantic Recoverability and Its Consequences

3.1 Captured Posterior Energy

For the balanced task, define the posterior score $\eta(x) := \mathbb{E}[Y \mid X = x]$. The posterior contains the information about Y available from a single view. For a representation $F = (F_1, \ldots, F_r)$, let $\mathcal{S}_F := \mathrm{span}\{F_1, \ldots, F_r\} \subset L^2(P_X)$ denote its represented function space. We define the *captured posterior energy* $B(F) := \|\Pi_{\mathcal{S}_F}\eta\|^2_{L^2(P_X)}$, where $\Pi_{\mathcal{S}_F}$ is the orthogonal projection onto $\mathcal{S}_F$.

This definition does not require SSL or whitening. It asks directly how much of the Bayes-relevant posterior score remains in the function space represented by F. The dependence on the downstream task is intentional: the same representation may preserve one semantic partition and discard another. If one wants a literal fraction of the single-view task signal that is retained, it is $B(F)/\|\eta\|^2_{L^2(P_X)}$. In particular, large $B(F)$ means that most of the posterior is expressible through the representation, while small $B(F)$ means that the task is largely absent from the represented subspace.

The quantity is invariant to invertible changes of coordinates in feature space because such changes leave $\mathcal{S}_F$ unchanged. Whitening enters below for another reason: it converts this intrinsic subspace quantity into exact Euclidean statements about feature geometry and few-shot classification.

3.2 Why Recoverability Matters Under Whitening

Assume for the remainder of this section that F is centered and whitened, so that $\mathbb{E}[F(X)] = 0$ and $\mathbb{E}[F(X)F(X)^\top] = I_r$. The coordinate functions of F are then orthonormal in $L^2(P_X)$, and the intrinsic recoverability quantity has the simple feature-space form $B(F) = \|\mathbb{E}[YF(X)]\|_2^2$.

Let $\mu_s := \mathbb{E}[F(X) \mid Y = s]$, $\Sigma_s := \mathrm{Cov}(F(X) \mid Y = s)$, and $\Delta := \mu_+ - \mu_-$. When $\Delta \neq 0$, let $u := \Delta/\|\Delta\|_2$. The CDNV (V_F) and directional CDNV ($\tilde{V}_F$) are

$$V_F := (\mathrm{Tr}(\Sigma_+) + \mathrm{Tr}(\Sigma_-))/\|\Delta\|_2^2, \qquad \tilde{V}_F := (u^\top \Sigma_+ u + u^\top \Sigma_- u)/\|\Delta\|_2^2. \tag{1}$$

When $\Delta = 0$, we set $V_F = \tilde{V}_F = +\infty$. Under whitening, total within-class variation can remain large even when variation along the semantic direction is small.

3.2.1 Exact directional-collapse law

Proposition 3.1. *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Then $V_F = (r - B(F))/(2B(F))$ and $\tilde{V}_F = (1 - B(F))/(2B(F))$, with both quantities equal to $+\infty$ when $B(F) = 0$.*

Prop. 3.1 gives an exact population law: directional CDNV vanishes as recoverability approaches one and diverges as recoverability vanishes. Thus the representation need not collapse globally for the task to become geometrically simple; it only needs to collapse along the semantic direction that survives in the represented subspace.

Remark 3.2 (Directional vs. global collapse)**.** *Even at $B(F) = 1$, the global CDNV is $V_F = (r-1)/2$, large whenever $r \geq 3$. Whitening fixes the total representation variance, so within-class spread remains in directions orthogonal to the task axis. Global collapse is therefore neither expected nor necessary. When the label is fully determined by a single view, $Y = g(X)$, we have $B(F) = 1$ if and only if $\eta \in \mathcal{S}_F$, giving perfect directional collapse despite potentially large global CDNV.*

3.2.2 Probe–centroid alignment

A second consequence of whitening is that the population squared-loss probe and the centroid decision axis coincide.

Corollary 3.3 (MSE probe and decision-axis alignment)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Define $\mu_s := \mathbb{E}[F(X) \mid Y = s]$ for $s \in \{\pm 1\}$, and let $\Delta := \mu_+ - \mu_-$. Then the unique population MSE probe, $(w^\star, b^\star) \in \arg\min_{w \in \mathbb{R}^r, b \in \mathbb{R}} \mathbb{E}[(Y - (w^\top F(X) + b))^2]$, satisfies $b^\star = 0$ and $w^\star = \mathbb{E}[YF(X)] = \frac{1}{2}\Delta$. In particular, $\mu_- = -\mu_+$, and $\Delta = 2\mu_+ = 2w^\star$.*

Cor. 3.3 shows that linear probing and NCC use the same separating direction. Moreover, $\|w^\star\|_2^2 = B(F)$, so the strength of the population probe is itself a measurement of semantic recoverability.

3.2.3 Few-shot transfer guarantees

The geometric identities above translate recoverability into an operational downstream guarantee. In the whitened setting, the same scalar that controls directional geometry also controls the expected error of an m-shot nearest-centroid classifier.

Theorem 3.4 (Few-shot NCC bound via captured posterior energy)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Then, for every $m \geq 1$,*

$$\mathrm{err}_m^{\mathrm{NCC}}(F) \leq 1 - B(F) + \frac{r - B(F)}{m} + \frac{1 - B(F)}{1 - B(F) + 2mB(F)} = \frac{2\tilde{V}_F}{1 + 2\tilde{V}_F} + \frac{(r-1) + 2r\tilde{V}_F}{m(1 + 2\tilde{V}_F)} + \frac{\tilde{V}_F}{\tilde{V}_F + m},$$

where the second expression is understood by continuity when $B(F) = 0$.

The bound separates two effects. The leading term reflects task information not captured by the representation, while the remaining terms arise from estimating class centroids from finitely many examples. Whitening is what makes this a one-scalar Euclidean statement: without it, coordinate rescaling could change $\|\mathbb{E}[YF]\|_2^2$ and Euclidean NCC would no longer coincide with the natural population geometry. The coordinate-invariant analogue is obtained by rewhitening, or equivalently by using a Mahalanobis metric in the original feature space.

3.2.4 MULTITASK SEMANTIC GEOMETRY

Recoverability also determines how several semantic tasks are organized jointly. We use a superscript to index tasks, reserving the subscript r for representation rank. Consider k balanced binary tasks $Y_1, \dots, Y_k \in \{\pm 1\}$, and write $Y := (Y_1, \dots, Y_k)$. For each task t, let $\eta_t(x) = \mathbb{E}[Y_t \mid X = x]$, $B^{(t)}(F) = \|\Pi_{\mathcal{S}_F}\eta_t\|^2_{L^2(P_X)}$, and $\varepsilon_t = \mathbb{E}[\mathrm{Var}(Y_t \mid X)] = 1 - \|\eta_t\|^2_{L^2(P_X)}$. Let $w_t = \mathbb{E}[Y_t F(X)]$. Assuming $B^{(t)}(F) > 0$, define $u_t = w_t/\|w_t\|_2$ and $Z_t(X) = \langle u_t, F(X)\rangle$.

Theorem 3.5 (Near-orthogonality and centroid geometry)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened. Assume each $Y_t \in \{\pm 1\}$ is balanced and $B^{(t)}(F) > 0$ for all t; arbitrary dependence among $Y_1, \dots, Y_k$ is allowed. For $i \neq j$, define $\rho_{ij} := \mathbb{E}[Y_i Y_j]$. Then, for any $i \neq j$,*

$$|u_i^\top u_j| \;\le\; \left(|\rho_{ij}| + \sqrt{\varepsilon_i \varepsilon_j} + \sqrt{\big(\|\eta_i\|^2_{L^2} - B^{(i)}(F)\big)\big(\|\eta_j\|^2_{L^2} - B^{(j)}(F)\big)}\right) \Big/ \sqrt{B^{(i)}(F) B^{(j)}(F)}.$$

Moreover, letting $m_Y := \mathbb{E}[Z(X) \mid Y] \in \mathbb{R}^k$,

$$\mathbb{E}\Big[\big\|m_Y - \big(\sqrt{B^{(1)}(F)}Y_1, \dots, \sqrt{B^{(k)}(F)}Y_k\big)\big\|_2^2\Big] \;\le\; \sum_{t=1}^{k} \big(1 - B^{(t)}(F)\big).$$

Thm. 3.5 shows that well-recovered tasks induce strong semantic axes that become nearly orthogonal when task correlations and residual posterior uncertainty are small. Their joint centroids approach $\big(\sqrt{B^{(t)}(F)}Y_t\big)_{t\le k}$, yielding an approximately axis-aligned hyperrectangle whose side lengths are set by the task-specific capture values. Thus recoverability controls the strength of each semantic direction, while cross-task dependence determines how closely the overall geometry approaches a factorial product.

The results so far deliberately make no claim about how F was learned. They say that *if* a representation captures a task, then $B(F)$ determines its directional geometry, its few-shot behavior, and its role in multitask centroid structure. We now turn to the complementary question: *why should a same-instance SSL objective produce a large value of $B(F)$ for some tasks and not others?*

4 RECOVERABILITY UNDER SAME-INSTANCE SSL

The preceding results show that the recoverability quantity $B(F)$ characterizes the task-relevant geometry of a representation and controls its few-shot behavior. The next question is whether, for a representation learned by same-instance SSL, this quantity can itself be identified explicitly from the SSL objective. We show that this is indeed the case: under the whitening-based population objective, the learned representation spans the leading eigenspace of the two-view operator, and $B(F)$ becomes the spectral overlap between the downstream posterior and these view-stable modes.

4.1 A CANONICAL TWO-VIEW POPULATION OBJECTIVE

Many SSL methods (Grill et al., 2020; Zbontar et al., 2021; Bardes et al., 2022; Chen et al., 2020; Caron et al., 2021; Ermolov et al., 2021; Tao et al., 2022) combine instance-level agreement with an anti-collapse mechanism that promotes global spread, decorrelation, or isotropy (Wang & Isola, 2020; Garrido et al., 2023; Balestriero & LeCun, 2022). To obtain a clean population characterization of the represented subspace, we study the canonical whitening-based surrogate (W-MSE) (Ermolov et al., 2021; Weng et al., 2022)

$$\max_{F:\mathcal{X}\to\mathbb{R}^r} \mathbb{E}\Big[\big\langle F(X^{(1)}), F(X^{(2)})\big\rangle\Big] \quad \text{subject to} \quad \mathbb{E}[F(X)] = 0, \qquad \mathbb{E}[F(X)F(X)^\top] = I_r. \quad (2)$$

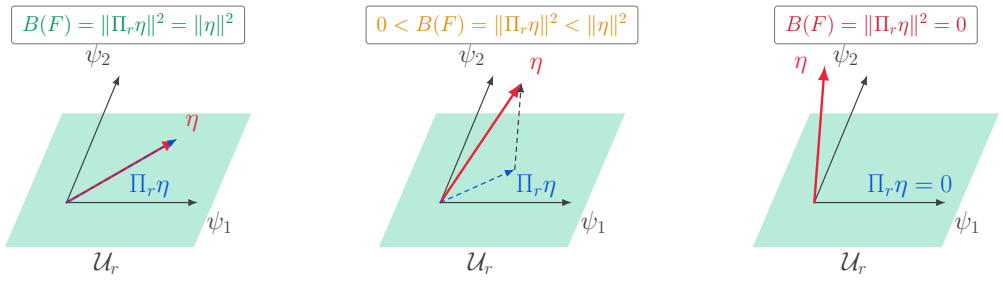


Figure 2: **Geometric interpretation of semantic recoverability.** The shaded plane is the SSL-selected subspace $\mathcal{U}_r = \mathrm{span}\{\psi_1, \psi_2\}$ of leading two-view eigenfunctions. The task posterior η decomposes into its projection $\Pi_r\eta$ onto this subspace and an orthogonal residual. From left to right, the posterior is fully captured, partially captured, or discarded. The captured posterior energy $B_r = \|\Pi_r\eta\|^2_{L^2(P_X)}$ measures the retained task-relevant information.

The objective rewards agreement across two views of the same instance, while whitening rules out collapse and fixes the feature metric. Its role in our theory is specific: it gives an analyzable rule for which r-dimensional function space is selected from the two-view distribution.

Strict whitening is the cleanest setting for this identification and for the exact geometry above. App. C studies a covariance-regularized surrogate related to Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022), while App. D gives a corresponding reduced-rank regression view of prediction-based objectives such as BYOL (Grill et al., 2020), SimSiam (Tao et al., 2022), and JEPA (Assran et al., 2023).

4.2 The Two-View Operator and the Selected Subspace

Define the two-view conditional expectation operator $T : L^2(P_X) \to L^2(P_X)$ by $(Tf)(x) := \mathbb{E}[f(X^{(2)}) \mid X^{(1)} = x]$. Thus T maps a function of one view to its conditional expectation from another view of the same instance. In the conditionally i.i.d. two-view model, exchangeability gives $\langle f, Tg\rangle = \langle Tf, g\rangle$, while conditional independence given $X^{(0)}$ gives $\langle f, Tf\rangle = \mathbb{E}[(\mathbb{E}[f(X) \mid X^{(0)}])^2] \geq 0$. Hence T is self-adjoint and positive semidefinite, and an eigenvalue measures the cross-view stability of the corresponding mode.

Proposition 4.1. *Let $L^2_0(P_X) := \{f \in L^2(P_X) : \mathbb{E}[f(X)] = 0\}$, and assume that T restricted to $L^2_0(P_X)$ admits an orthonormal eigenbasis $(\psi_j)_{j\geq 1}$, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Assume $\lambda_r > 0$ and $\lambda_r > \lambda_{r+1}$. Then $\psi_1, \ldots, \psi_r$ are feasible for* (2) *and achieve the optimal value $\sum_{j=1}^r \lambda_j$. Moreover, if $F : \mathcal{X} \to \mathbb{R}^r$ is any optimal solution of* (2)*, then its coordinates span the same r-dimensional subspace as $\psi_1, \ldots, \psi_r$.*

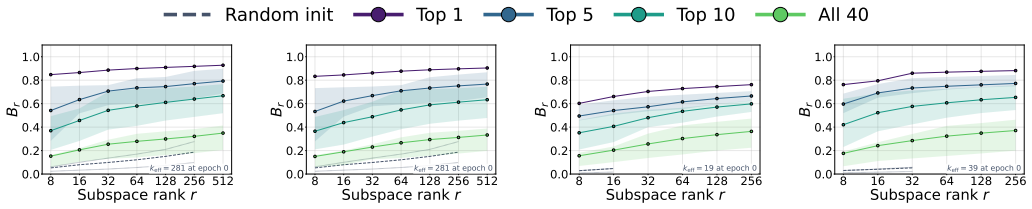
Prop. 4.1 says that the population objective keeps the r most cross-view-stable modes. The representation is determined only up to an orthogonal change of basis within this selected subspace. This result identifies what SSL preserves without referring to any downstream labels; the task enters only when we ask how much of its posterior lies in that subspace.

4.3 Closed-Form Recoverability at the SSL Optimum

Let $F^\star$ be any population-optimal solution of (2), and define $\mathcal{U}_r := \mathrm{span}\{\psi_1, \ldots, \psi_r\} \subset L^2(P_X)$, with Π_r denoting the orthogonal projection onto $\mathcal{U}_r$. For a balanced binary label with posterior score $\eta(x) = \mathbb{E}[Y \mid X = x]$, write $\beta_j := \langle \eta, \psi_j\rangle_{L^2(P_X)}$.

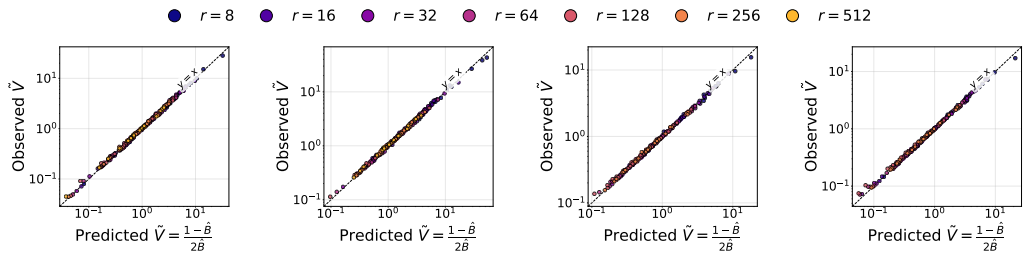
Corollary 4.2. *Under the assumptions of Prop. 4.1, $\Pi_r\eta = \sum_{j=1}^r \beta_j\psi_j$, and $B(F^\star) = \|\Pi_r\eta\|^2_{L^2(P_X)} = \sum_{j=1}^r \beta_j^2 =: B_r$.*

For multiple downstream tasks, we write $B_r^{(t)}$ for the rank-r capture of task t; equivalently, $B_r^{(t)} = B^{(t)}(F^\star)$ for an optimal rank-r representation.



(a) W-MSE-CelebA **(b) VICReg-CelebA** **(c) Barlow-IM1K** **(d) I-JEPA-IM1K**

Figure 3: **Captured posterior energy on CelebA.** Mean $B_r^{(t)}$ over the 1, 5, and 10 most recoverable CelebA attributes and for all 40 attributes, shown across spectral ranks r. Shaded regions denote the corresponding interquartile ranges. The dashed curve is the mean for a matched randomly initialized encoder, with dotted curves showing its interquartile range.



(a) W-MSE-CelebA **(b) VICReg-CelebA** **(c) Barlow-IM1K** **(d) I-JEPA-IM1K**

Figure 4: **Predicted versus observed directional CDNV on CelebA.** Each point is one CelebA attribute at a given rank r. The prediction $\widehat{\tilde{V}}_{\text{pred}} = (1 - \widehat{B}_A^{(t)})/(2\widehat{B}_A^{(t)})$ is computed on split A. Class centroids, the semantic direction, and observed directional CDNV are estimated independently on held-out split B. The dashed line is $y = x$, and each panel contains all 40 attributes.

This is the closed-form answer to when recoverability is large under the SSL model. The task is preserved when its posterior score is concentrated on the modes that are most stable across the two views, and it is discarded when its posterior lies primarily in view-variant directions outside the selected subspace. The two-view objective itself remains label-free; labels enter only through the overlap used to evaluate a particular downstream task.

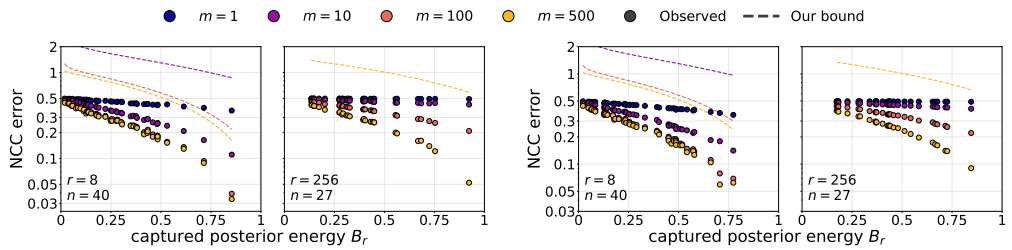
Combining this spectral identity with the previous section immediately turns B_r into task-level predictions. A large spectral overlap implies low directional CDNV, strong few-shot NCC performance, and pronounced semantic axes; across several tasks, the values $B_r^{(t)}$ determine the side lengths of the approximate hyperrectangle. In the recoverable regime where the augmentations preserve the task and the increasing view-stable subspaces capture its posterior, B_r approaches one as the representation dimension grows. Thus the overall mechanism is: SSL selects view-stable structure, a task survives according to its posterior overlap with that structure, and the amount that survives determines the downstream geometry and transfer behavior.

5 Experiments

5.1 Settings

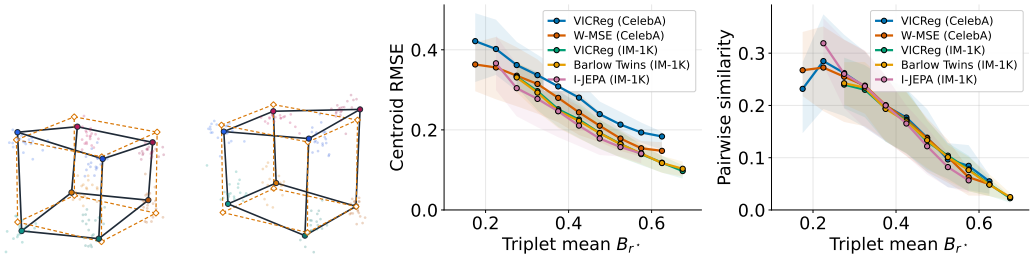
We test the main empirical consequences of the theory on natural and synthetic data. Full dataset, preprocessing, and training details are given in App. A.

Datasets. The main experiments use CelebA (Liu et al., 2015), whose 40 binary attributes provide a collection of downstream semantic tasks. For each attribute, we balance the two classes by subsampling so that the empirical problem matches the binary setting of Sec. 2. Controlled experiments on dSprites (Matthey et al., 2017) and 3DShapes (Burgess & Kim, 2018) are reported in App. B.2.



(a) W-MSE-CelebA **(b) I-JEPA-IM1K**

Figure 5: **Captured posterior energy controls few-shot NCC error.** Empirical m-shot NCC error versus task-specific $B_r^{(t)}$ across CelebA attributes, for several ranks r (panels) and support sizes m (colors). Each point corresponds to a downstream attribute, and dashed curves are the corresponding upper bounds from Thm. 3.4.



(a) W-MSE-CelebA **(b) I-JEPA-IM1K** **(c) Centroid RMSE** **(d) Task-axis overlap**

Figure 6: **Multitask centroid geometry on CelebA.** (a–b) Illustrative held-out joint centroids for three attributes, compared with the hyperrectangle predicted by Thm. 3.5; triplets are selected on the training split. (c–d) Results over all 1646 eligible triplets per encoder, without recoverability or orthogonality screening. Larger mean task recoverability is associated with lower normalized centroid RMSE and smaller maximum pairwise task-axis overlap. Curves show bin means and shaded regions show interquartile ranges; selection and normalization details are in App. B.1.

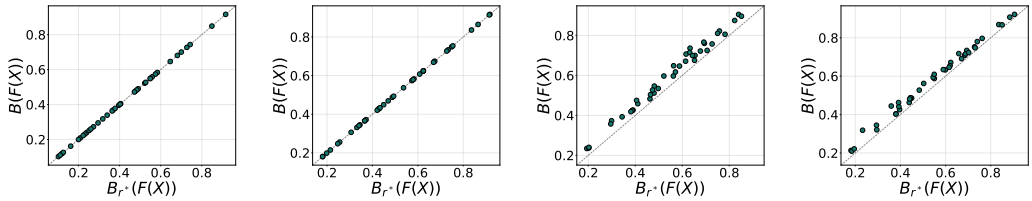
SSL methods and rank-r representations. We train W-MSE (Ermolov et al., 2021) and VICReg (Bardes et al., 2022) from scratch on CelebA with ResNet-50 (He et al., 2016) backbones. We also evaluate publicly available ImageNet-1K (Deng et al., 2009)-pretrained VICReg and Barlow Twins (Zbontar et al., 2021) ResNet-50 encoders and an I-JEPA (Assran et al., 2023) ViT-H/14 encoder. To vary representation rank, we estimate a paired-view spectral basis from unlabeled data and retain its leading r directions. The retained features are then centered and whitened before the label-dependent measurements of B_r below.

Estimating semantic recoverability. For a balanced downstream task, let $\hat{\beta} = \widehat{\mathbb{E}}[YF(X)]$ and $\widehat{G} = \widehat{\mathbb{E}}[F(X)F(X)^\top]$. We estimate captured posterior energy by $\widehat{B}(F) = \hat{\beta}^\top \widehat{G}^{-1}\hat{\beta}$. This form accounts for finite-sample deviations from exact whitening. When $\widehat{G} = I_r$, it reduces to $\widehat{B}(F) = \|\widehat{\mathbb{E}}[YF(X)]\|_2^2$. Unless noted otherwise, all reported recoverability values use this estimator.

5.2 Results

Recoverability is task-dependent. Because the rank-r subspaces are nested, $B_r^{(t)}$ is nondecreasing in r. Fig. 3 nevertheless shows substantial variation across CelebA attributes: the most recoverable tasks approach $B_r^{(t)} \approx 1$ at moderate ranks, while the average over all 40 attributes remains lower. SSL representations also exceed random-initialization baselines across a broad range of ranks. Tasks differ markedly in both saturation rate and attainable recoverability. The same pattern appears for additional pretrained encoders in App. B.1 and on dSprites and 3DShapes in App. B.2.

Recoverability predicts directional geometry. Prop. 3.1 gives the exact relation $\tilde{V}_F = (1 - B(F))/(2B(F))$ for centered, whitened representations. We test it with a split-sample protocol:



(a) W-MSE-CelebA **(b) VICReg-CelebA** **(c) Barlow-IM1K** **(d) I-JEPA-IM1K**

Figure 7: **Recoverability from the paired-view spectral subspace.** Directly measured task recoverability $\widehat{B}^{(t)}(F)$ versus its reconstruction $\widehat{B}^{(t)}_{r^\star}$ from the label-free paired-view spectral modes. Each point is one CelebA attribute; the dashed line is perfect agreement. For ImageNet-pretrained encoders, the retained subspace is reduced-rank and the median absolute discrepancy is below 0.05 across models; additional encoders are shown in App. B.1.

recoverability is estimated on split A and directional CDNV independently on split B. Fig. 4 shows close agreement over more than two orders of magnitude, across ranks and SSL representations. Additional encoder results are given in App. B.1.

Recoverability controls few-shot transfer. Thm. 3.4 predicts that, at fixed rank and shot count, larger $B^{(t)}_r$ yields lower NCC error, while increasing the support size m reduces centroid-estimation error. Fig. 5 shows both trends across CelebA attributes: empirical error decreases with recoverability, and the bound tightens as m grows. The bound remains above the observed error throughout the evaluated configurations and becomes more informative for well-recovered tasks. App. B.1 reports the remaining encoders and Tab. 1 shows comparisons with earlier neural-collapse-based transfer bounds (Galanti et al., 2023; Luthra et al., 2025; 2026).

Multiple semantic factors. Thm. 3.5 predicts two related effects: well-recovered tasks should have nearly orthogonal semantic axes when cross-task correlations and residual uncertainty are small, and their joint centroids should approach the corresponding factorial hyperrectangle. Fig. 6(a–b) shows two illustrative triplets selected on the training split using fixed recoverability and near-orthogonality thresholds (see App. B.1); the displayed centroids are then measured on held-out test images. Panels (c–d) use every eligible triplet, with no additional recoverability or orthogonality screening. Across encoders, higher mean task recoverability is associated with both lower centroid RMSE and smaller task-axis overlap, consistent with the geometry predicted by the theorem.

The spectral subspace recovers semantic energy. Cor. 4.2 identifies recoverability at the population optimum with posterior energy in the leading two-view spectral subspace. We test this empirically by comparing $\widehat{B}^{(t)}(F)$, measured directly from the learned representation, with $\widehat{B}^{(t)}_{r^\star}$, reconstructed from the estimated paired-view spectral modes at the label-free rank $r^\star$ defined in App. B.1. Fig. 7 shows close agreement. For CelebA-trained W-MSE and VICReg, the retained basis spans nearly the full projector space, so near-exact agreement is expected. For ImageNet-pretrained encoders, the lower-dimensional spectral subspace still recovers most of the full-representation semantic energy, with median absolute discrepancies below 0.05 across models. The subspace is estimated without downstream labels. For I-JEPA, the paired-view operator is a surrogate analysis kernel rather than the masking distribution used in pretraining; more details can be found in App. A.

6 Discussion and Limitations

Our results identify *semantic recoverability* as the key quantity linking same-instance SSL to downstream geometry and transfer. The task-specific captured posterior energy $B^{(t)}_r$ provides a unified account of directional neural collapse, probe–centroid alignment, multitask centroid structure, and few-shot performance. These results are derived primarily for population-optimal, centered and whitened representations. We also analyze a covariance-regularized surrogate, but extending the theory to finite-sample training and imperfect optimization remains an important direction for future work. Our main analysis also focuses on balanced binary tasks and conditionally i.i.d. two-view generation. Extending the framework to multiclass tasks and more general view dependencies is another natural direction.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT and Claude Code) to assist with developing and refining theoretical frameworks, formulating mathematical claims, discussing proof strategies, and providing feedback on experimental methodology. These tools also assisted with experimental design and implementation, including code generation and refinement, as well as research brainstorming, literature exploration, manuscript organization, drafting and editing, and LaTeX presentation.

All AI-assisted theoretical arguments, mathematical derivations, experimental designs, and generated or modified code were reviewed by the authors for correctness and consistency. Relevant literature was independently examined, and all experimental results and their interpretation were evaluated by the authors. The authors take full responsibility for the final content of the paper, including its text, claims, code, results, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies theoretical and empirical properties of representations learned through self-supervised learning. It does not involve new human-subject studies or the collection of new personal data. The experiments use existing datasets and pretrained models for research purposes. The work does not develop or evaluate a high-risk application.

REPRODUCIBILITY STATEMENT

We state the assumptions and formal results in the main text and provide complete proofs in the appendix. The experimental sections and supplementary material document the datasets, pretrained representations, evaluation protocols, preprocessing, training procedures, and implementation details needed to reproduce our empirical results. We also specify the procedures used to estimate semantic recoverability, directional CDNV, spectral structure, and few-shot transfer performance.

REFERENCES

Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NU9AYHJvYe.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/andrew13.html.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7187–7209. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/ash22a.html.

Shahab Asoodeh, Fady Alajaji, and Tamás Linder. On maximal correlation, mutual information and data privacy. In *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pp. 27–31, 2015. doi: 10.1109/CWIT.2015.7255145.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine*

Learning, volume 162 of *Proceedings of Machine Learning Research*, pp. 1101–1116. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/awasthi22b.html.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15509–15519. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=jQgsZDspz5h.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NsVEjx6YPd.

Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985. doi: 10.1080/01621459.1985.10478157. URL https://doi.org/10.1080/01621459.1985.10478157.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(14):361–383, 2007. URL http://jmlr.org/papers/v8/fukumizu07a.html.

Tomer Galanti, András György, and Marcus Hutter. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022a. URL https://openreview.net/forum?id=VrK7pKwOhT_.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=SwIp410B6aQ.

Tomer Galanti, András György, and Marcus Hutter. Generalization bounds for few-shot transfer learning with pretrained classifiers, 2022c. URL https://arxiv.org/abs/2212.12532.

Tomer Galanti, Liane Galanti, and Ido Ben-Shaul. Comparative generalization bounds for deep neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=162TqkUNPO.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kDEL91Dufpa.

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3607–3616. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/goldblum20a.html.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour, 2017. URL https://arxiv.org/abs/1706.02677.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Yu Gui, Cong Ma, and Yiqiao Zhong. Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage, 2023. URL https://arxiv.org/abs/2306.03335.

Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning. In *NeurIPS 2022 Workshop: Self-Supervised Learning – Theory and Practice*, 2022. URL https://arxiv.org/abs/2212.11491.

X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

E. J. Hannan. The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society*, 2(2):229–242, 1961. doi: 10.1017/S1446788700026707.

Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=AuEgNlEAmed.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/27debb435021eb68b3965290b5e24c49-Paper.pdf.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. doi: 10.1162/0899766042321814.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936. doi: 10.2307/2333955. URL https://doi.org/10.2307/2333955.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, October 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, October 2000. doi: 10.1142/S012906570000034X. URL https://doi.org/10.1142/S012906570000034X.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425. URL https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-supervised contrastive learning is approximately supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46671–46708, 2025. doi: 10.52202/085713-1392. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b85cbbaab0008dfc52250a857725f70-Abstract-Conference.html.

Achleshwar Luthra, Yash Salunkhe, and Tomer Galanti. Directional neural collapse explains few-shot transfer in self-supervised learning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.03530.

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

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mcallester20a.html.

Tomer Michaeli, Weiran Wang, and Karen Livescu. Nonparametric canonical correlation analysis. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1967–1976, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/michaeli16.html.

Kento Nozawa and Issei Sato. Understanding negative samples in instance discriminative self-supervised representation learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5784–5797. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2dace78f80bc92e6d7493423d729448e-Paper.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. Infonce loss provably learns cluster-preserving representations. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1914–1961. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/parulekar23a.html.

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan

Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28389–28421. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/qiu23a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19250–19286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saunshi22a.html.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pp. 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873. URL https://www.isca-archive.org/interspeech_2019/schneider19_interspeech.html.

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19847–19878. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shen22d.html.

Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14431–14440, June 2022.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9ab-Paper.pdf.

Yuandong Tian. Understanding the role of nonlinearity in training dynamics of contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=s130rTE3U_X.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2495–2504, 2021.

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

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5526–5535, 2023. doi: 10.1109/ICCV51070.2023.00511.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11112–11122. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wen21c.html.

Xi Weng, Lei Huang, Lei Zhao, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. An investigation into whitening loss for self-supervised learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=BbUxkmrstyk.

Xi Weng, Jianing An, Xudong Ma, Binhang Qi, Jie Luo, Xi Yang, Jin Song Dong, and Lei Huang. Clustering properties of self-supervised learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66597–66616. PMLR, 2025. URL https://proceedings.mlr.press/v267/weng25a.html.

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*, pp. 12310–12320. PMLR, 2021.

Runtian Zhai. *Contextures: The Mechanism of Representation Learning*. PhD thesis, Carnegie Mellon University, 2025. URL https://arxiv.org/abs/2504.19792. PhD dissertation, CMU-CS-25-104.

Runtian Zhai, Kai Yang, Burak Varıcı, Che-Ping Tsai, J Zico Kolter, and Pradeep Kumar Ravikumar. Contextures: Representations from contexts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74318–74347. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhai25c.html.

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

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*, pp. 12979–12990. PMLR, 2021.

A Implementation Details

Datasets. CelebA (Liu et al., 2015) contains 202,599 RGB face images at resolution 178×218, covering 10,177 identities and 40 binary attributes. We use the standard identity-disjoint split of 162,770 training, 19,867 validation, and 19,962 test images.

The dSprites dataset contains 737,280 binary 64×64 images generated from six latent factors. We retain two of the three shapes, the four most extreme scale levels, and the eight most extreme values of each position coordinate, leaving orientation unrestricted. This gives 20,480 images, split randomly into 18,432 training and 2,048 held-out test images. Binary downstream tasks are formed by thresholding selected generative factors.

3DShapes contains 480,000 RGB 64×64 images generated from six latent factors. We keep six of the ten object-hue levels and the four most extreme scale levels, giving 144,000 images. We use 129,600 images for training and 14,400 for held-out testing. As with dSprites, binary downstream tasks are obtained by thresholding selected latent factors; the factors used in each analysis are indicated in the corresponding figure.

Evaluation protocol. The held-out test split is used for the saturation curves and the centroid-geometry evaluations. For the recoverability and directional-CDNV analyses, we construct disjoint subsets A and B from the training data. The subsets contain 81,385 images each for CelebA, 2,048 images each for dSprites, and 14,400 images each for 3DShapes. The paired-view spectral basis, whitening transform, and recoverability estimates are computed on A; directional CDNV is then measured independently on B.

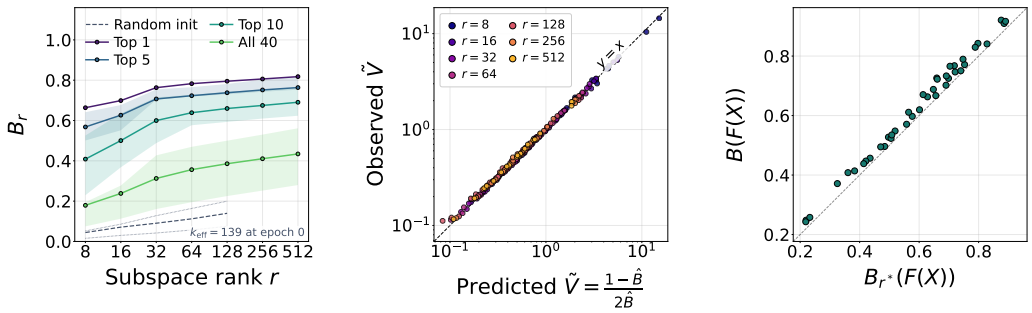
Evaluation preprocessing. All label-dependent quantities (recoverability, directional CDNV, NCC error, and linear-probe quantities) are measured on clean, single-view images. CelebA images are center-cropped to 160×160 and resized to 128×128 for the models trained from scratch. dSprites and 3DShapes are kept at their native 64×64 resolution and are bilinearly upsampled to the input resolution of externally pretrained backbones when needed. Inputs are normalized with ImageNet-1K channel statistics. On CelebA, augmented pairs are used only to estimate the paired-view spectral basis. On dSprites and 3DShapes, the same deterministic resize-and-normalize transform is used both for paired-view spectral estimation and for clean feature extraction.

VICReg and W-MSE augmentations. We use the same view distribution for the CelebA-trained VICReg and W-MSE models. Each view is generated by a random resized crop with scale $(0.2, 1.0)$ and aspect ratio $(3/4, 4/3)$, resized to 128×128, followed by horizontal flipping with probability 0.5. We apply color jitter with brightness 0.4, contrast 0.4, saturation 0.2, and hue 0.05 with probability 0.8, grayscale conversion with probability 0.1, and Gaussian blur with probability 0.5 using a 7×7 kernel and $\sigma \sim \mathcal{U}[0.1, 2.0]$. We use a minimum crop scale of 0.2, rather than the ImageNet value 0.08, to preserve more facial structure. Unlike the original VICReg pipeline (Bardes et al., 2022), we do not use solarization and use the same blur probability for both views.

Pretrained encoders. For ImageNet-1K-pretrained VICReg, Barlow Twins, and I-JEPA, CelebA images are resized to a shorter side of 256 pixels and center-cropped to 224×224. To estimate the paired-view spectral basis, we use an ImageNet-scale version of the augmentation pipeline above: random resized crops with scale $(0.08, 1.0)$, color-jitter hue 0.1, grayscale probability 0.2, and a 23×23 Gaussian blur kernel. For I-JEPA, these augmented pairs define a surrogate two-view kernel for analysis; they are not the masking-based views used during I-JEPA pretraining. For dSprites and 3DShapes, externally pretrained encoders receive bilinearly upsampled 224×224 inputs.

Training details. We train VICReg (Bardes et al., 2022) and W-MSE (Ermolov et al., 2021) from scratch on the CelebA training split for 1000 epochs. Both use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 10^{-6}. The learning rate is scaled linearly with global batch size, $\eta = \eta_{\text{base}} B/256$, with a 10-epoch linear warm-up (Goyal et al., 2017) followed by cosine decay (Loshchilov & Hutter, 2017).

For VICReg, the projection head is `Linear(2048 → 2048, bias=False) → BN → ReLU → Linear(2048 → 2048)`. The invariance, variance, and covariance coefficients are $\lambda = 25$, $\mu = 25$, and $\nu = 1$. We use a global batch size of 1024 across two NVIDIA H100 GPUs and $\eta_{\text{base}} = 1.5 \times 10^{-4}$.



(a) Captured posterior energy **(b) Predicted vs. observed $\widetilde{V}$** **(c) Spectral identification**

Figure 8: **Additional recoverability results for VICReg-IM1K on CelebA.** (a) Captured posterior energy across representation ranks, using the same attribute groups and random-initialization reference as Fig. 3. (b) Directional CDNV predicted from recoverability estimated on split A versus directional CDNV measured independently on held-out split B; the dashed line is $y = x$. (c) Recoverability measured directly from the learned representation versus its reconstruction from the paired-view spectral subspace; the dashed line denotes perfect agreement.

For W-MSE, the projection head is `Linear(2048 → 2048, bias=False) → BN → ReLU → Linear(2048 → 128)`. Following (Ermolov et al., 2021), embeddings from each view are split into sub-batches of 256 and whitened independently using the inverse Cholesky factor of the empirical covariance, with no covariance regularization ($\epsilon = 0$). The loss is normalized mean-squared error between corresponding whitened embeddings. The two views are processed in one concatenated forward pass. We use a global batch size of 512 on one NVIDIA H100 and $\eta_{\text{base}} = 10^{-3}$.

Architectures. Both CelebA-trained models use a standard ResNet-50 (He et al., 2016). At 128×128 resolution, the final residual stage produces a 4×4 feature map, which is globally averaged to a 2048-dimensional representation. Unless stated otherwise, representation analyses use these pooled backbone features rather than projection-head outputs.

We also evaluate public ImageNet-1K-pretrained VICReg and Barlow Twins (Zbontar et al., 2021) ResNet-50 encoders and an I-JEPA (Assran et al., 2023) ViT-H/14 encoder, without fine-tuning. The ResNet-50 encoders produce 2048-dimensional pooled features. I-JEPA has 32 transformer layers, 16 attention heads, hidden dimension 1280, and MLP dimension 5120. At 224×224, it produces 256 patch tokens and no classification token; we average the final-layer-normalized patch tokens to obtain a 1280-dimensional representation.

B Additional Results

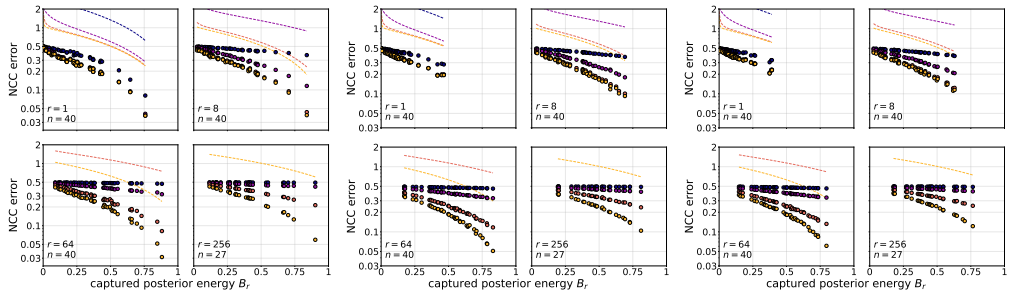
B.1 Additional CelebA Results

This section reports the remaining CelebA results using the same protocol as Sec. 5.

Semantic recoverability. Fig. 8 (a) shows the rank-dependent recoverability curves for ImageNet-1K-pretrained VICReg. It reproduces the main-text pattern: recoverability varies substantially across attributes, the strongest tasks saturate at lower rank, and the learned representations remain above their random-initialization baselines.

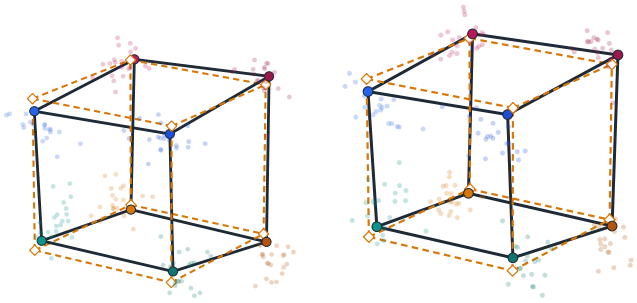
Directional geometry. Fig. 8 (b) gives the held-out test of Prop. 3.1. Predictions are computed from split A and observed directional CDNV from split B. The agreement with $y = x$ remains tight across attributes and ranks.

Few-shot transfer. Fig. 9 reports the remaining few-shot results. Across models, ranks, and shot counts, empirical NCC error decreases with recoverability and the bound tracks the same rank–shot tradeoff seen in the main text.



(a) VICReg-CelebA **(b) VICReg-IM1K** **(c) Barlow-IM1K**

Figure 9: **Captured posterior energy and few-shot NCC error.** Results for VICReg-CelebA, VICReg-IM1K, and Barlow Twins-IM1K across ranks r and support sizes m. Dashed curves are the upper bound from Thm. 3.4.



(a) VICReg-IM1K **(b) Barlow-IM1K**

Figure 10: **Multitask centroid geometry on CelebA.** Held-out joint centroids and the hyperrectangles predicted by Thm. 3.5 for ImageNet-1K-pretrained VICReg and Barlow Twins. Experiment setup is same as Fig. 6

Multitask geometry and spectral identification. Each encoder is evaluated at a label-free rank $r^\star$, defined as the number of paired-view spectral directions whose eigenvalue is at least 10^{-3} times the largest eigenvalue. This gives $r^\star = 946$ for W-MSE, 2046 for VICReg-CelebA, 763 for VICReg-IM1K, 806 for Barlow-IM1K, and 263 for I-JEPA-IM1K.

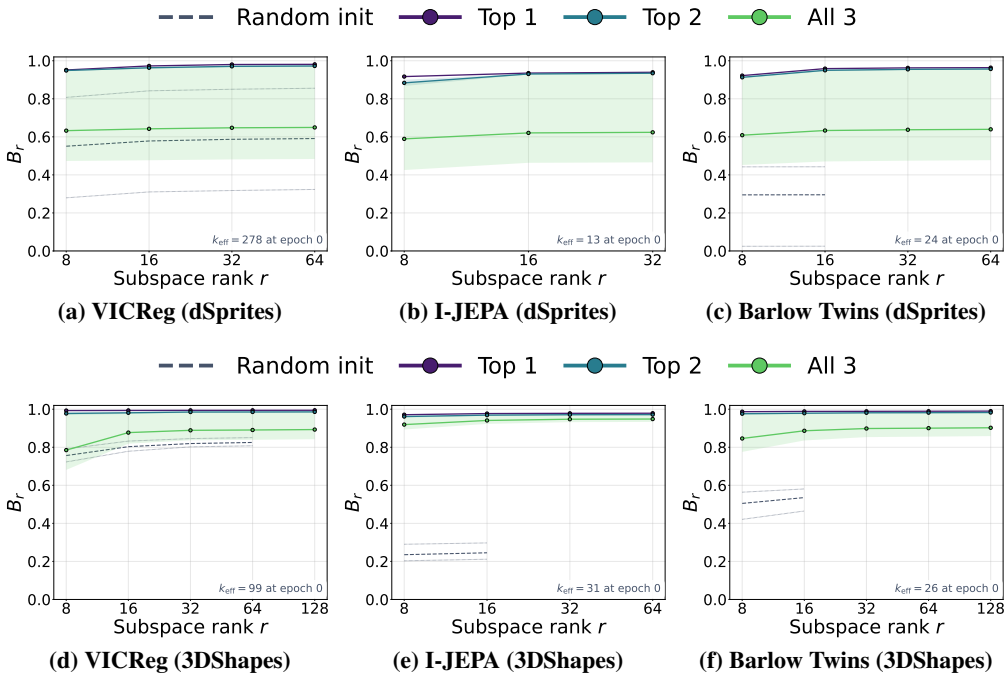
For each attribute triplet, the task-specific capture values $B^{(t)}_{r^\star}$ and the task axes are estimated on the training split, weighting the eight joint-label cells equally; the centroids are then measured on the held-out test split. We write $\overline{B}_{r^\star} := \frac{1}{3}\sum_{t=1}^{3} B^{(t)}_{r^\star}$ for the triplet mean recoverability used in Fig. 6(c–d). Normalized centroid RMSE is the root-mean-square distance between observed and predicted centroids divided by the root-mean-square radius of the predicted box. Fig. 6(c–d) use all triplets drawn from the 26 attributes for which each binary class contains at least 10% of the data and every joint-label cell contains at least 1000 training and 100 test examples. This yields 1646 triplets per encoder. No recoverability or orthogonality screening is applied to these quantitative panels. The curves are means in bins of width 0.05 containing at least 12 triplets; shaded regions show interquartile ranges. The illustrative cubes in Figs. 6(a–b) and 10(a–b) are selected on the training split only from triplets satisfying $\min_t B^{(t)}_{r^\star} \geq 0.10$ and pairwise $|\cos(u_i, u_j)| \leq 0.12$.

Fig. 10 adds held-out centroid visualizations for VICReg-IM1K and Barlow Twins-IM1K, and Fig. 8 (c) and Fig. 7 (c) reports the corresponding spectral-identification plots for VICReg-IM1K and Barlow Twins-IM1K respectively. The latter compare recoverability measured in the full feature space with recoverability reconstructed from the label-free paired-view spectral subspace.

Comparison with prior transfer bounds. Table 1 compares Thm. 3.4 with earlier neural-collapse-based transfer bounds (Galanti et al., 2023; Luthra et al., 2025; 2026). On W-MSE trained on CelebA, the earlier bounds remain above 0.5 for all 40 attributes and all five shot counts shown. Our bound is non-vacuous in several rank–shot configurations and falls below 0.5 already at $m = 1$. The

Table 1: **Comparison of few-shot transfer bounds on CelebA.** Bounds on NCC classification error for W-MSE trained on CelebA ("Male/Female" attribute) . Our bound is minimized over the evaluated ranks separately for each shot count m. Prior bounds use ℓ_2-normalized full-dimensional features, whereas ours uses whitened top-r features. Values below 0.5 are non-vacuous relative to chance classification.

Bound	$m = 1$	$m = 5$	$m = 10$	$m = 100$	$m = 500$
(Galanti et al., 2023)	2923.1	709.9	433.2	184.3	162.1
(Luthra et al., 2025)	138.1	32.1	20.6	6.70	3.53
(Luthra et al., 2026)	1544.4	282.5	126.7	10.27	2.05
Our bound	**0.490**	**0.253**	**0.223**	**0.190**	**0.161**



(a) VICReg (dSprites) **(b) I-JEPA (dSprites)** **(c) Barlow Twins (dSprites)**

(d) VICReg (3DShapes) **(e) I-JEPA (3DShapes)** **(f) Barlow Twins (3DShapes)**

Figure 11: **Semantic recoverability on dSprites and 3DShapes.** Task-specific captured posterior energy $B_r^{(t)}$ across representation ranks for ImageNet-1K-pretrained VICReg, I-JEPA, and Barlow Twins.

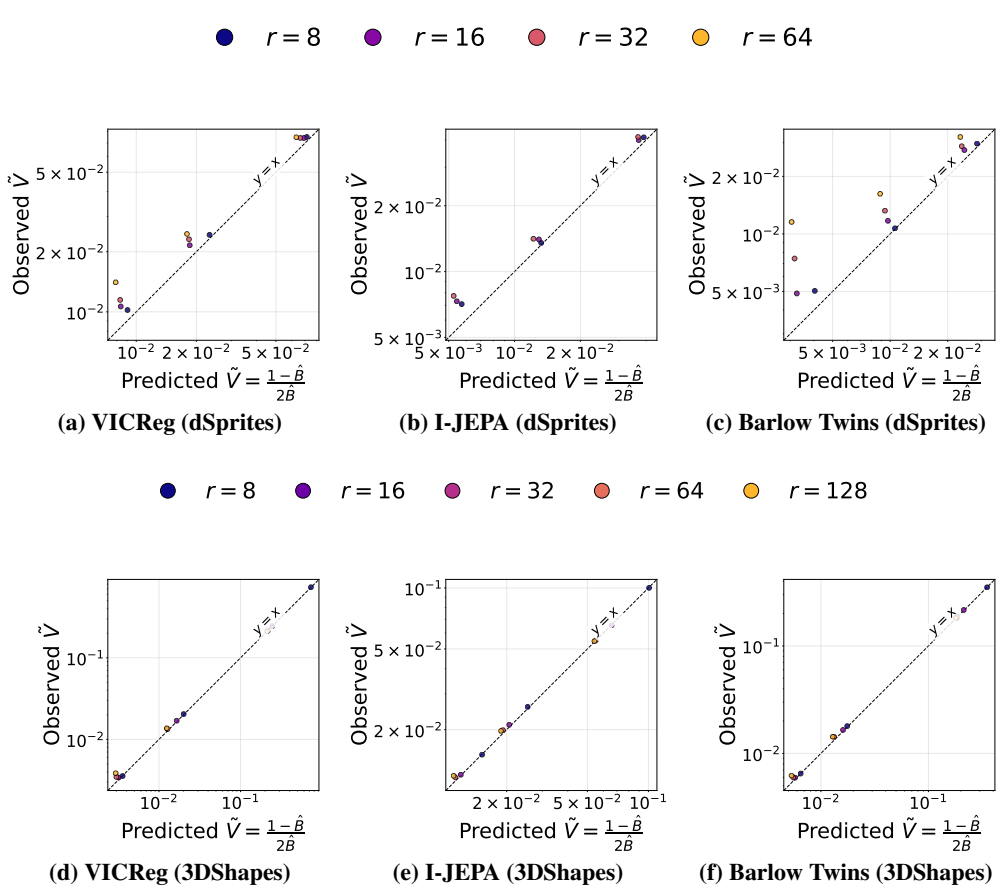
minimizing rank increases as more support examples become available, matching the rank–shot tradeoff in Thm. 3.4.

B.2 Results on Synthetic Datasets

We use dSprites and 3DShapes to check that the same recoverability–geometry relationships are visible in controlled factorized data. Figs. 11–13 report rank-dependent recoverability, the directional-CDNV identity, and held-out multitask centroid geometry. The displayed encoders are ImageNet-1K-pretrained VICReg, I-JEPA, and Barlow Twins. Across both datasets, the qualitative behavior matches the CelebA results.

C Covariance-Regularized SSL Objectives

In the main text we studied the population surrogate (2), which enforces strict whitening $\mathbb{E}[F(X)F(X)^\top] = I_r$. This yields an exact strict-whitening theory in which the main directional geometry is governed by the single semantic capture parameter $B_r = \|\Pi_r \eta\|^2_{L^2(P_X)}$. Modern



(a) VICReg (dSprites) **(b) I-JEPA (dSprites)** **(c) Barlow Twins (dSprites)**

(d) VICReg (3DShapes) **(e) I-JEPA (3DShapes)** **(f) Barlow Twins (3DShapes)**

Figure 12: **Predicted versus observed directional CDNV on synthetic datasets.** Results for VICReg, I-JEPA, and Barlow Twins on dSprites and 3DShapes. Each point is one attribute at one rank r; the prediction is computed on split A and the observed directional CDNV on held-out split B. Dashed lines show the theoretical prediction $y = x$.

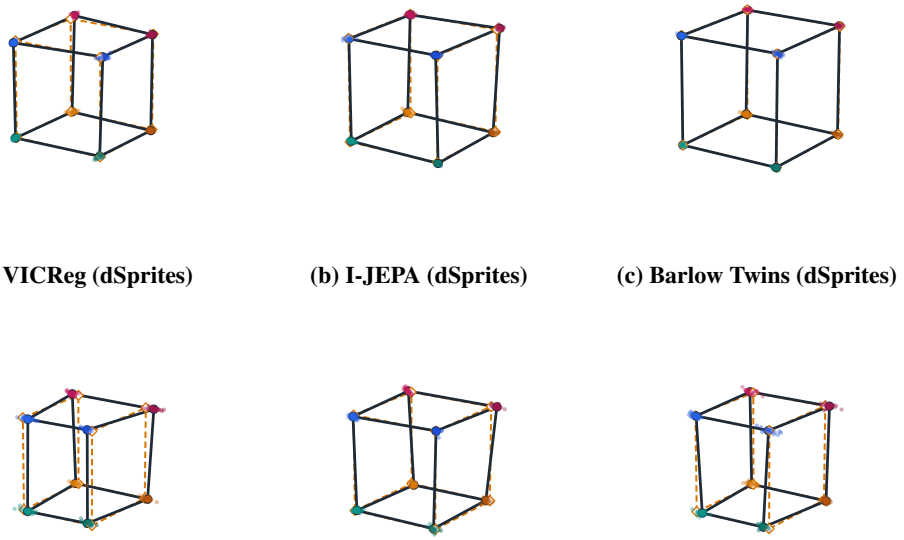
SSL objectives typically do not enforce exact whitening, but instead penalize covariance deviations through variance/covariance regularizers, as in Barlow Twins and VICReg. In this appendix we show that a natural covariance-regularized surrogate preserves the same spectral subspace identified by the two-view operator T in Prop. 4.1, but the resulting geometry is no longer governed by a single scalar. Instead, three explicit quantities $(B_\gamma, C_\gamma, S_\gamma)$ appear, with the strict-whitening formulas recovered as $\gamma \to \infty$.

C.1 Population covariance-regularized objective

We consider the following population surrogate:

$$\max_{F:\mathcal{X}\to\mathbb{R}^r} \mathbb{E}\Big[\langle F(X^{(1)}), F(X^{(2)})\rangle\Big] - \frac{\gamma}{2}\left\|\mathbb{E}[F(X)F(X)^\top] - I_r\right\|_F^2 \quad \text{s.t.} \quad \mathbb{E}[F(X)] = 0, \tag{3}$$

where $\gamma > 0$ controls the strength of the covariance penalty. As $\gamma \to \infty$, the penalty enforces $\mathbb{E}[F(X)F(X)^\top] \to I_r$, recovering the strict whitening setting. The centering constraint $\mathbb{E}[F(X)] = 0$ models the fact that practical implementations explicitly center features and, in a population analysis, can be imposed without loss of generality by working with centered features $\bar{F}(X) := F(X) - \mathbb{E}[F(X)]$ and rewriting the objective in terms of $\bar{F}$.



(a) VICReg (dSprites) **(b) I-JEPA (dSprites)** **(c) Barlow Twins (dSprites)**

(d) VICReg (3DShapes) **(e) I-JEPA (3DShapes)** **(f) Barlow Twins (3DShapes)**

Figure 13: **Multitask centroid geometry on synthetic datasets.** Held-out joint centroids compared with the hyperrectangles predicted by Thm. 3.5.

C.2 Operator form and canonical modes

Recall the two-view conditional expectation operator $(Tf)(x) := \mathbb{E}[f(X^{(2)}) \mid X^{(1)} = x]$, $T : L^2(P_X) \to L^2(P_X)$. For any $f, g \in L^2(P_X)$, we have

$$\mathbb{E}\big[f(X^{(1)})g(X^{(2)})\big] = \langle f, Tg\rangle_{L^2(P_X)}. \tag{4}$$

In particular,

$$\mathbb{E}\big[f(X^{(1)})f(X^{(2)})\big] = \langle f, Tf\rangle_{L^2(P_X)}.$$

Let $(\lambda_j, \psi_j)_{j\geq 1}$ be the orthonormal eigenbasis of T on $L^2_0(P_X)$ from Prop. 4.1, ordered as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0.$$

Thus

$$T\psi_j = \lambda_j\psi_j, \qquad \langle \psi_i, \psi_j\rangle_{L^2(P_X)} = \delta_{ij}, \qquad \mathbb{E}[\psi_j(X)] = 0.$$

These are exactly the canonical observable modes identified in Prop. 4.1. This appendix keeps the same operator and asks how covariance regularization rescales its top-r T-eigenspace.

C.3 Structure of the population optimum

We show that covariance regularization preserves the T-selected spectral subspace and only introduces mode-dependent scalings within that subspace.

Proposition C.1 (Structure of the regularized optimum)**.** *Assume $\lambda_r > 0$ and $\lambda_r > \lambda_{r+1}$. Then any population maximizer $F^\star_\gamma$ of* (3) *can be written as*

$$F^\star_\gamma(x) = R\left(s_1\psi_1(x), \ldots, s_r\psi_r(x)\right), \tag{5}$$

where $R \in \mathbb{R}^{r\times r}$ is orthogonal and $s_1, \ldots, s_r \geq 0$ are scalars. Moreover, for the quadratic penalty in (3)*, the optimal scalings satisfy*

$$s_j^2 = 1 + \frac{\lambda_j}{\gamma}, \qquad j = 1, \ldots, r.$$

Prop. C.1 implies that the regularized optimum can be reduced to the whitened setting by an explicit population rewhitening step, and therefore inherits the same multitask geometric structure established in the main text. Let $F_\gamma^\star$ be a population maximizer of (3), and define

$$\Sigma_\gamma := \mathrm{Cov}(F_\gamma^\star(X)), \qquad \widetilde{F}_\gamma(x) := \Sigma_\gamma^{-1/2} F_\gamma^\star(x).$$

For the maximizer in Prop. C.1, we have $s_j^2 = 1 + \lambda_j/\gamma > 0$, hence $\Sigma_\gamma \succ 0$ and $\Sigma_\gamma^{-1/2}$ is well-defined. Then $\widetilde{F}_\gamma$ is feasible for the strictly whitened objective (2) and is population-optimal. Indeed, by Prop. C.1 we may write

$$F_\gamma^\star(x) = R\,\mathrm{diag}(s_1, \dots, s_r)\,\big(\psi_1(x), \dots, \psi_r(x)\big),$$

so

$$\Sigma_\gamma = R\,\mathrm{diag}(s_1^2, \dots, s_r^2)\,R^\top, \qquad \Sigma_\gamma^{-1/2} = R\,\mathrm{diag}(s_1^{-1}, \dots, s_r^{-1})\,R^\top,$$

and therefore

$$\widetilde{F}_\gamma(x) = R\,\big(\psi_1(x), \dots, \psi_r(x)\big).$$

By Prop. 4.1, any representation of the form $R(\psi_1, \dots, \psi_r)$ is feasible for (2) and achieves the optimal value $\sum_{j=1}^r \lambda_j$, hence $\widetilde{F}_\gamma$ is a whitened population optimum.

Consequently, all main-text results that rely only on the whitened-optimum structure, including the multitask near-orthogonality and factorial-centroid geometry in Thm. 3.5, apply verbatim to $\widetilde{F}_\gamma$. In the original feature coordinates of $F_\gamma^\star$, this same structure is obtained by applying the linear map $\Sigma_\gamma^{1/2}$, so the hyperrectangle in rewhitened probe coordinates becomes a parallelepiped in the raw $F_\gamma^\star$ coordinates. This highlights that covariance regularization preserves the operator-selected semantic organization, but can distort Euclidean geometry unless one explicitly rewhitens features, or equivalently uses a Mahalanobis nearest-centroid rule with metric Σ_γ^{-1}.

C.4 How the CDNV formulas change

Let $Y = y(C) \in \{\pm 1\}$ be balanced and define the posterior score

$$\eta(x) = \mathbb{E}[Y \mid X = x].$$

Write the posterior directly in the T-eigenbasis:

$$\eta = \sum_{j \geq 1} \beta_j \psi_j, \qquad \beta_j := \langle \eta, \psi_j \rangle_{L^2(P_X)}.$$

Then, by Cor. 4.2,

$$\Pi_r \eta = \sum_{j=1}^r \beta_j \psi_j, \qquad B_r = \|\Pi_r \eta\|_{L^2(P_X)}^2 = \sum_{j=1}^r \beta_j^2.$$

Define the scalars

$$B_\gamma := \sum_{j=1}^r \beta_j^2 s_j^2, \qquad C_\gamma := \sum_{j=1}^r \beta_j^2 s_j^4, \qquad S_\gamma := \sum_{j=1}^r s_j^2. \tag{6}$$

Proposition C.2. *Let $F_\gamma^\star$ be any population maximizer of* (3)*, written as in* (5)*. Let $\mu_\pm$ and $\Sigma_\pm$ be the class means and class covariances defined as in the main text. Then $\mu_- = -\mu_+$ and $\Delta = \mu_+ - \mu_- = 2\mu_+$. Moreover, when $B_\gamma > 0$, the CDNV and directional CDNV defined in* (1) *satisfy*

$$\tilde{V}_{F_\gamma^\star} = \frac{1}{2}\left(\frac{C_\gamma}{B_\gamma^2} - 1\right), \qquad V_{F_\gamma^\star} = \frac{1}{2}\left(\frac{S_\gamma}{B_\gamma} - 1\right). \tag{7}$$

When $B_\gamma = 0$, we set both quantities to $+\infty$ by the same convention as in the main text.

For $B_\gamma > 0$, Prop. C.2 can be plugged directly into a general CDNV-based few-shot guarantee. Namely, with $F = F^\star_\gamma$, substituting (7) yields the explicit γ-dependent bound

$$\mathrm{err}_m^{\mathrm{NCC}}(F^\star_\gamma) \le 4\Big(\frac{C_\gamma}{B_\gamma^2} - 1\Big) + \frac{8}{\sqrt{m}}\sqrt{\frac{1}{2}\Big(\frac{S_\gamma}{B_\gamma} - 1\Big)} + \Big(\frac{8}{\sqrt{m}} + \frac{4}{m}\Big)\frac{1}{2}\Big(\frac{S_\gamma}{B_\gamma} - 1\Big). \tag{8}$$

In particular, this bound depends on γ only through the three scalars $B_\gamma, C_\gamma, S_\gamma$ defined in (6).

Thus γ controls a concrete tradeoff in (8). The leading directional term depends on C_γ/B_γ^2, while the finite-m terms depend on S_γ/B_γ. Under the quadratic penalty $s_j^2 = 1 + \lambda_j/\gamma$, decreasing γ inflates all selected modes, with larger inflation on more stable modes. This increases total variance through S_γ, but it also increases signal strength along the task-relevant direction through B_γ and C_γ. The resulting few-shot behavior is therefore governed by how the posterior coefficients β_j overlap with the modes that are most amplified by the regularizer.

In the strict-whitening limit $s_j \to 1$, we recover

$$B_\gamma \to B_r, \qquad C_\gamma \to B_r, \qquad S_\gamma \to r,$$

and therefore

$$\tilde{V}_{F^\star_\gamma} \to \frac{1 - B_r}{2B_r}, \qquad V_{F^\star_\gamma} \to \frac{r - B_r}{2B_r},$$

exactly as in Prop. 3.1.

D Prediction-Based SSL Objectives as Reduced-Rank Regression

This appendix gives a simple population-level connection between prediction-based SSL objectives, such as BYOL, SimSiam, and JEPA, and classical reduced-rank regression / partial least squares. We do not aim to model algorithmic details such as stop-gradient or EMA teachers. Instead, we study a natural whitened population surrogate and show that its optimizer selects the same T-spectral subspace as (2). As a result, the main-text conclusions that depend only on the selected subspace, and therefore on the induced capture value B_r, carry over directly to this prediction-based setting.

D.1 A stylized whitened prediction surrogate

Let $F : \mathcal{X} \to \mathbb{R}^r$ be a shared representation used on both views, and let $W \in \mathbb{R}^{r\times r}$ be a linear predictor. Consider the population objective

$$\min_{\substack{F:\mathcal{X}\to\mathbb{R}^r \\ W\in\mathbb{R}^{r\times r}}} \mathbb{E}\Big[\big\|F(X^{(2)}) - WF(X^{(1)})\big\|_2^2\Big] \quad \text{s.t.} \quad \mathbb{E}[F(X)] = 0,\ \mathbb{E}[F(X)F(X)^\top] = I_r, \tag{9}$$

where $(X^{(1)}, X^{(2)})$ is a same-instance pair with conditionally i.i.d. views as in Sec. 2.1.

Define

$$C_{12}(F) := \mathbb{E}\big[F(X^{(1)})F(X^{(2)})^\top\big], \qquad C_{21}(F) := C_{12}(F)^\top. \tag{10}$$

Proposition D.1 (RRR reduction to a spectral objective)**.** *Fix any feasible F in* (9)*. The unique minimizer over W is*

$$W^\star(F) = C_{21}(F) = C_{12}(F)^\top,$$

and the resulting minimum equals

$$\min_W \mathbb{E}\Big[\big\|F(X^{(2)}) - WF(X^{(1)})\big\|_2^2\Big] = r - \|C_{21}(F)\|_F^2. \tag{11}$$

Hence minimizing (9) *is equivalent to maximizing $\|C_{21}(F)\|_F^2$ over feasible F.*

Proof. Using $\mathbb{E}[F(X)F(X)^\top] = I_r$,

$$\mathbb{E}\|F(X^{(2)}) - WF(X^{(1)})\|_2^2 = r - 2\mathrm{Tr}(WC_{12}(F)) + \|W\|_F^2 = r + \|W - C_{21}(F)\|_F^2 - \|C_{21}(F)\|_F^2.$$

Thus the unique minimizer is $W^\star(F) = C_{21}(F) = C_{12}(F)^\top$, and the minimum is $r - \|C_{21}(F)\|_F^2$. □

D.2 Selected subspace and relationship to the alignment objective

For scalar $f, g \in L^2(P_X)$ with $\mathbb{E}[f(X)] = \mathbb{E}[g(X)] = 0$, the same-instance model gives

$$\mathbb{E}[f(X^{(2)})g(X^{(1)})] = \langle f, Tg\rangle_{L^2(P_X)}.$$

Thus, if $F = (f_1, \ldots, f_r)$ is feasible and the coordinates $\{f_k\}$ are orthonormal in $L^2(P_X)$, then

$$\big(C_{21}(F)\big)_{ij} = \mathbb{E}[f_i(X^{(2)})f_j(X^{(1)})] = \langle f_i, Tf_j\rangle_{L^2(P_X)}. \tag{12}$$

Consequently, $\|C_{21}(F)\|_F^2$ is the squared Hilbert–Schmidt norm of the compression $P_U T P_U$, where $U = \operatorname{span}\{f_1, \ldots, f_r\} \subset L_0^2(P_X)$.

Let $(\lambda_j, \psi_j)_{j\geq 1}$ be the orthonormal eigenbasis of T from Prop. 4.1, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

Proposition D.2 (Prediction objective selects the top-r operator subspace)**.** *Assume $\lambda_r > 0$ and $\lambda_r > \lambda_{r+1}$. Then the maximum of $\|C_{21}(F)\|_F^2$ over feasible F equals $\sum_{j=1}^r \lambda_j^2$, and any maximizer $F^\star$ has coordinates spanning $\operatorname{span}\{\psi_1, \ldots, \psi_r\}$. In particular, there exists an orthogonal $R \in \mathbb{R}^{r\times r}$ such that*

$$F^\star(x) = R\,(\psi_1(x), \ldots, \psi_r(x)). \tag{13}$$

Proof. Let $U = \operatorname{span}\{f_1, \ldots, f_r\}$ and let P_U be the orthogonal projector onto U. In the orthonormal basis $\{f_j\}$, $C_{21}(F)$ is the matrix of $P_U T P_U$, and hence

$$\|C_{21}(F)\|_F^2 = \|P_U T P_U\|_{\mathrm{HS}}^2 \leq \|T P_U\|_{\mathrm{HS}}^2 = \operatorname{Tr}(P_U T^2) \leq \sum_{j=1}^r \lambda_j^2.$$

The last inequality is Ky Fan's variational principle applied to the positive semidefinite operator T^2. Equality is attained for $U = \operatorname{span}\{\psi_1, \ldots, \psi_r\}$, and the eigengap $\lambda_r > \lambda_{r+1}$ makes this maximizing subspace unique. □

D.3 Consequences for recoverability, CDNV, and downstream guarantees

Prop. D.2 shows that the whitened prediction surrogate (9) selects the same r-dimensional T-eigenspace as (2). In particular, any maximizer $F^\star$ of (9) has coordinates spanning $\operatorname{span}\{\psi_1, \ldots, \psi_r\}$, so $F^\star$ is feasible for (2). Since the strictly whitened alignment objective (2) depends only on the chosen r-dimensional subspace and is maximized by any orthonormal basis of the top-r eigenspace, as shown in Prop. 4.1, $F^\star$ is also population-optimal for (2) and achieves the optimal value $\sum_{j=1}^r \lambda_j$. Therefore, all main-text results that apply to whitened population optima of (2) apply directly to $F^\star$.

In particular, for any downstream binary label $Y = y(C)$ with posterior score $\eta(x) = \mathbb{E}[Y \mid X = x]$, writing

$$\eta = \sum_{j\geq 1} \beta_j \psi_j,$$

the capture value remains

$$B_r = \sum_{j=1}^r \beta_j^2,$$

and the closed-form directional CDNV and CDNV laws for whitened optima from Prop. 3.1 hold unchanged for $F^\star$. Consequently, the same few-shot NCC guarantees from Thm. 3.4 and the same multitask geometry conclusions from Thm. 3.5 follow immediately for this stylized prediction-based surrogate.

E PROOFS

E.1 PROOFS OF THE RESULTS IN THE MAIN TEXT

Proposition 3.1. *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Then $V_F = (r - B(F))/(2B(F))$ and $\tilde{V}_F = (1 - B(F))/(2B(F))$, with both quantities equal to $+\infty$ when $B(F) = 0$.*

Proof. Let $\mu_s := \mathbb{E}[F(X) \mid Y = s]$ and $\Sigma_s := \mathrm{Cov}(F(X) \mid Y = s)$ for $s \in \{\pm 1\}$, and set $\Delta := \mu_+ - \mu_-$ with $u := \Delta/\|\Delta\|_2$ when $\Delta \neq 0$. Balancedness of Y and $\mathbb{E}[F(X)] = 0$ give $\frac{1}{2}\mu_+ + \frac{1}{2}\mu_- = 0$, so $\mu_- = -\mu_+$ and $\Delta = 2\mu_+$. Moreover, $\mathbb{E}[YF(X)] = \frac{1}{2}\mu_+ - \frac{1}{2}\mu_- = \mu_+$, and by iterated expectation this also equals $\mathbb{E}[\eta(X)F(X)]$.

Write $F = (f_1, \ldots, f_r)$. Centering and whitening imply that $\{f_1, \ldots, f_r\}$ is orthonormal in $L^2(P_X)$, so $\mathcal{S}_F = \mathrm{span}\{f_1, \ldots, f_r\}$ is an r-dimensional subspace and $(\mu_+)_i = \mathbb{E}[\eta(X)f_i(X)] = \langle \eta, f_i \rangle$. Hence $\Pi_{\mathcal{S}_F}\eta = \sum_i (\mu_+)_i f_i$, and by orthonormality $\|\mu_+\|_2^2 = \|\Pi_{\mathcal{S}_F}\eta\|_{L^2}^2 = B(F)$, giving $\|\Delta\|_2^2 = 4B(F)$. If $B(F) = 0$ then $\Delta = 0$ and both ratios are $+\infty$; assume $B(F) > 0$ below.

By the law of total covariance and whitening, $I_r = \frac{1}{2}(\Sigma_+ + \Sigma_-) + \mathrm{Cov}(\mu_Y)$, where $\mu_Y \in \{\mu_+, \mu_-\}$ uniformly. Since $\mu_- = -\mu_+$, $\mathrm{Cov}(\mu_Y) = \mu_+\mu_+^\top$, yielding

$$\Sigma_+ + \Sigma_- = 2(I_r - \mu_+\mu_+^\top). \tag{14}$$

Taking traces in (14) gives $\mathrm{Tr}(\Sigma_+) + \mathrm{Tr}(\Sigma_-) = 2(r - B(F))$, so $V_F = (r - B(F))/(2B(F))$ after dividing by $\|\Delta\|_2^2 = 4B(F)$. For the directional ratio, $u = \mu_+/\|\mu_+\|_2$ gives $u^\top(\Sigma_+ + \Sigma_-)u = 2(1 - \|\mu_+\|_2^2) = 2(1 - B(F))$ by (14), so $\tilde{V}_F = (1 - B(F))/(2B(F))$. □

Corollary 3.3 (MSE probe and decision-axis alignment)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Define $\mu_s := \mathbb{E}[F(X) \mid Y = s]$ for $s \in \{\pm 1\}$, and let $\Delta := \mu_+ - \mu_-$. Then the unique population MSE probe, $(w^\star, b^\star) \in \arg\min_{w \in \mathbb{R}^r, b \in \mathbb{R}} \mathbb{E}[(Y - (w^\top F(X) + b))^2]$, satisfies $b^\star = 0$ and $w^\star = \mathbb{E}[YF(X)] = \frac{1}{2}\Delta$. In particular, $\mu_- = -\mu_+$, and $\Delta = 2\mu_+ = 2w^\star$.*

Proof. Let $Z := F(X)$, so that $\mathbb{E}[Z] = 0$, $\mathbb{E}[ZZ^\top] = I_r$, and $\mathbb{E}[Y] = 0$. Expanding the squared loss and using these identities gives

$$\mathcal{R}(w, b) := \mathbb{E}\big[(Y - (w^\top Z + b))^2\big] = 1 - 2w^\top \mathbb{E}[YZ] + \|w\|_2^2 + b^2,$$

a strictly convex quadratic in (w, b) whose unique minimizer is $w^\star = \mathbb{E}[YZ]$ and $b^\star = 0$. Let $\mu_s := \mathbb{E}[Z \mid Y = s]$. Balancedness gives $\mathbb{E}[YZ] = \frac{1}{2}(\mu_+ - \mu_-) = \frac{1}{2}\Delta$ and $\mathbb{E}[Z] = \frac{1}{2}\mu_+ + \frac{1}{2}\mu_- = 0$, so $\mu_- = -\mu_+$ and hence $\Delta = 2\mu_+ = 2w^\star$. □

Theorem 3.4 (Few-shot NCC bound via captured posterior energy)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened, and let $Y \in \{\pm 1\}$ be balanced. Then, for every $m \geq 1$,*

$$\mathrm{err}_m^{\mathrm{NCC}}(F) \leq 1 - B(F) + \frac{r - B(F)}{m} + \frac{1 - B(F)}{1 - B(F) + 2mB(F)} = \frac{2\tilde{V}_F}{1 + 2\tilde{V}_F} + \frac{(r-1) + 2r\tilde{V}_F}{m(1 + 2\tilde{V}_F)} + \frac{\tilde{V}_F}{\tilde{V}_F + m},$$

where the second expression is understood by continuity when $B(F) = 0$.

Proof. If $B(F) = 0$, the stated bound is trivial since its right-hand side is at least 1. Assume henceforth that $B(F) > 0$. Let $Z := F(X)$ and $w := \mathbb{E}[YZ]$, so $\mathbb{E}[Z] = 0$, $\mathbb{E}[ZZ^\top] = I_r$, and $\|w\|_2^2 = B(F)$. Balancedness and $\mathbb{E}[Z] = 0$ give $\mu_\pm := \mathbb{E}[Z \mid Y = \pm 1] = \pm w$, hence $\Delta = 2w$ and $\|\Delta\|_2^2 = 4B(F)$. As in the proof of Prop. 3.1, the law of total covariance yields

$$\Sigma_+ + \Sigma_- = 2(I_r - ww^\top), \tag{15}$$

where $\Sigma_s := \mathrm{Cov}(Z \mid Y = s)$; taking traces gives $\mathrm{Tr}(\Sigma_+) + \mathrm{Tr}(\Sigma_-) = 2(r - B(F))$, and contracting with $u := w/\|w\|_2$ gives $u^\top(\Sigma_+ + \Sigma_-)u = 2(1 - B(F))$.

NCC rule as a margin. Draw an m-shot support set with empirical class means $\hat{\mu}_\pm := \frac{1}{m}\sum_{i=1}^m Z_{\pm,i}$, and set $\hat{w} := (\hat{\mu}_+ - \hat{\mu}_-)/2$ and $c := (\hat{\mu}_+ + \hat{\mu}_-)/2$. The NCC rule is $g_{S_m}^{\mathrm{NCC}}(z) = \mathrm{sign}(\langle \hat{w}, z - c \rangle)$, since its decision boundary is the perpendicular bisector of $\hat{\mu}_\pm$. Define the signed margin $M_{S_m} := 2Y\langle \hat{w}, Z - c \rangle$. Regardless of how ties are broken, $\{g_{S_m}^{\mathrm{NCC}}(Z) \neq Y\} \subseteq \{M_{S_m} \leq 0\}$.

Cantelli bound conditional on the support set. Since (Z, Y) is independent of S_m with $\mathbb{E}[YZ] = w$, $\mathbb{E}[Z] = 0$, $\mathbb{E}[ZZ^\top] = I_r$,

$$\mathbb{E}[M_{S_m} \mid S_m] = 2\langle \hat{w}, w\rangle, \qquad \mathrm{Var}(M_{S_m} \mid S_m) = 4\big(\|\hat{w}\|_2^2 + \langle \hat{w}, c\rangle^2 - \langle \hat{w}, w\rangle^2\big).$$

On the event $G := \{\langle \hat{w}, w\rangle > 0\}$, Cantelli's inequality gives $\mathbb{P}(M_{S_m} \le 0 \mid S_m) \le 1 - f(S_m)$, where

$$f(S_m) := \begin{cases} \dfrac{\langle \hat{w}, w\rangle^2}{\|\hat{w}\|_2^2 + \langle \hat{w}, c\rangle^2}, & \hat{w} \neq 0, \\ 0, & \hat{w} = 0. \end{cases}$$

Since $0 \le f(S_m) \le 1$, bounding $f(S_m)\mathbf{1}_G \ge f(S_m) - \mathbf{1}_{G^c}$ and taking expectations gives

$$\mathrm{err}_m^{\mathrm{NCC}}(F) \le 1 - \mathbb{E}[f(S_m)] + \mathbb{P}(G^c). \tag{16}$$

Lower bound on $\mathbb{E}[f(S_m)]$. When $\hat{w} \neq 0$, set $p := \langle \hat{w}, w\rangle^2/\|\hat{w}\|_2^2 = \|P_{\mathrm{span}(\hat{w})} w\|_2^2$; then $f(S_m) = p/(1 + \langle \hat{w}/\|\hat{w}\|_2, c\rangle^2)$. The inequality $x/(1+y) \ge x - y$ for $0 \le x \le 1$, $y \ge 0$, combined with $p \le \|w\|_2^2 = B(F) \le 1$, yields $f(S_m) \ge p - \|c\|_2^2$. Bounding the distance from w to $\mathrm{span}(\hat{w})$ by $\|w - \hat{w}\|_2$ gives $p \ge B(F) - \|\hat{w} - w\|_2^2$; when $\hat{w} = 0$ the inequality $f(S_m) \ge B(F) - \|\hat{w} - w\|_2^2 - \|c\|_2^2$ still holds since then $\|w - \hat{w}\|_2^2 = B(F)$. Writing $\hat{w} - w = (\varepsilon_+ - \varepsilon_-)/2$ and $c = (\varepsilon_+ + \varepsilon_-)/2$ with $\varepsilon_\pm := \hat{\mu}_\pm - \mu_\pm$, class independence gives

$$\mathbb{E}\|\hat{w} - w\|_2^2 = \mathbb{E}\|c\|_2^2 = \frac{\mathrm{Tr}(\Sigma_+) + \mathrm{Tr}(\Sigma_-)}{4m} = \frac{r - B(F)}{2m},$$

so $\mathbb{E}[f(S_m)] \ge B(F) - (r - B(F))/m$.

Bound on $\mathbb{P}(G^c)$. Since $\langle \hat{w}, w\rangle = B(F) + \langle \hat{w} - w, w\rangle$, we have $\mathbb{E}[\langle \hat{w}, w\rangle] = B(F)$, and

$$\mathrm{Var}(\langle \hat{w}, w\rangle) = w^\top \mathrm{Cov}(\hat{w} - w) w = \frac{B(F)}{4m}\, u^\top(\Sigma_+ + \Sigma_-)u = \frac{B(F)(1 - B(F))}{2m}$$

using (15). Cantelli gives $\mathbb{P}(G^c) \le (1 - B(F))/(1 - B(F) + 2mB(F))$.

Combining this with (16) and the lower bound on $\mathbb{E}[f(S_m)]$ yields $\mathrm{err}_m^{\mathrm{NCC}}(F) \le 1 - B(F) + (r - B(F))/m + (1 - B(F))/(1 - B(F) + 2mB(F))$. □

Theorem 3.5 (Near-orthogonality and centroid geometry)**.** *Let $F : \mathcal{X} \to \mathbb{R}^r$ be centered and whitened. Assume each $Y_t \in \{\pm 1\}$ is balanced and $B^{(t)}(F) > 0$ for all t; arbitrary dependence among $Y_1, \ldots, Y_k$ is allowed. For $i \neq j$, define $\rho_{ij} := \mathbb{E}[Y_i Y_j]$. Then, for any $i \neq j$,*

$$|u_i^\top u_j| \;\le\; \left(|\rho_{ij}| + \sqrt{\varepsilon_i \varepsilon_j} + \sqrt{\big(\|\eta_i\|_{L^2}^2 - B^{(i)}(F)\big)\big(\|\eta_j\|_{L^2}^2 - B^{(j)}(F)\big)}\right) \Big/ \sqrt{B^{(i)}(F) B^{(j)}(F)}.$$

Moreover, letting $m_Y := \mathbb{E}[Z(X) \mid Y] \in \mathbb{R}^k$,

$$\mathbb{E}\Big[\Big\|m_Y - \big(\sqrt{B^{(1)}(F)}Y_1, \ldots, \sqrt{B^{(k)}(F)}Y_k\big)\Big\|_2^2\Big] \;\le\; \sum_{t=1}^{k} \big(1 - B^{(t)}(F)\big).$$

Proof. Write $F = (f_1, \ldots, f_r)$. Centering and whitening make $\{f_1, \ldots, f_r\}$ orthonormal in $L^2(P_X)$, so $\mathcal{S}_F := \mathrm{span}\{f_1, \ldots, f_r\}$ is r-dimensional with orthogonal projector $\Pi_{\mathcal{S}_F}$. For each task t, let $w_t := \mathbb{E}[Y_t F(X)] \in \mathbb{R}^r$ with coordinates $(w_t)_\ell = \mathbb{E}[\eta_t(X) f_\ell(X)] = \langle \eta_t, f_\ell\rangle$, so $\Pi_{\mathcal{S}_F}\eta_t = \sum_\ell (w_t)_\ell f_\ell$ and hence

$$\|w_t\|_2^2 = \|\Pi_{\mathcal{S}_F}\eta_t\|_{L^2}^2 = B^{(t)}(F). \tag{17}$$

Write $\eta_t^\perp := (I - \Pi_{\mathcal{S}_F})\eta_t$, so $\|\eta_t^\perp\|_{L^2}^2 = \|\eta_t\|_{L^2}^2 - B^{(t)}(F)$.

Near-orthogonality. By (17), $u_i^\top u_j = \langle \Pi_{\mathcal{S}_F}\eta_i, \Pi_{\mathcal{S}_F}\eta_j\rangle / \sqrt{B^{(i)}(F)B^{(j)}(F)}$. Since $\langle \Pi_{\mathcal{S}_F}\eta_i, \Pi_{\mathcal{S}_F}\eta_j\rangle = \langle \eta_i, \eta_j\rangle - \langle \eta_i^\perp, \eta_j^\perp\rangle$, Cauchy–Schwarz on the residual term gives

$$\big|\langle \Pi_{\mathcal{S}_F}\eta_i, \Pi_{\mathcal{S}_F}\eta_j\rangle\big| \le |\langle \eta_i, \eta_j\rangle| + \sqrt{(\|\eta_i\|_{L^2}^2 - B^{(i)}(F))(\|\eta_j\|_{L^2}^2 - B^{(j)}(F))}. \tag{18}$$

To bound the first term, set $\xi_t := Y_t - \eta_t(X)$; then $\mathbb{E}[\xi_t \mid X] = 0$ and $\mathbb{E}[\xi_t^2] = \varepsilon_t$. Expanding $Y_i Y_j = (\eta_i + \xi_i)(\eta_j + \xi_j)$ and using conditional mean zero on the cross terms, $\rho_{ij} = \mathbb{E}[\eta_i(X)\eta_j(X)] + \mathbb{E}[\xi_i \xi_j]$,

so $|\langle \eta_i, \eta_j \rangle| \le |\rho_{ij}| + \sqrt{\varepsilon_i \varepsilon_j}$ by Cauchy–Schwarz. Substituting into (18) yields the stated bound on $|u_i^\top u_j|$; no independence among the Y_t was used.

Centroid geometry. Let $Z(X) = (Z_1(X), \dots, Z_k(X))$ and $m_Y^{(t)} := \mathbb{E}[Z_t(X) \mid Y]$. By (17), $\mathbb{E}[Y_t Z_t(X)] = u_t^\top w_t = \|w_t\|_2 = \sqrt{B^{(t)}(F)}$, and iterated expectation gives $\mathbb{E}[Y_t m_Y^{(t)}] = \sqrt{B^{(t)}(F)}$. Expanding the square and using $Y_t^2 = 1$,

$$\mathbb{E}\big[(m_Y^{(t)} - \sqrt{B^{(t)}(F)}\, Y_t)^2\big] = \mathbb{E}[(m_Y^{(t)})^2] - B^{(t)}(F) \le \mathbb{E}[Z_t^2] - B^{(t)}(F) = 1 - B^{(t)}(F),$$

where Jensen gives the inequality and whitening gives $\mathbb{E}[Z_t^2] = u_t^\top I_r u_t = 1$; this step also used no joint structure of the Y_t. Summing over t yields the claimed centroid bound. □

Proposition 4.1. *Let $L_0^2(P_X) := \{f \in L^2(P_X) : \mathbb{E}[f(X)] = 0\}$, and assume that T restricted to $L_0^2(P_X)$ admits an orthonormal eigenbasis $(\psi_j)_{j \ge 1}$, with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$. Assume $\lambda_r > 0$ and $\lambda_r > \lambda_{r+1}$. Then $\psi_1, \dots, \psi_r$ are feasible for* (2) *and achieve the optimal value $\sum_{j=1}^r \lambda_j$. Moreover, if $F : \mathcal{X} \to \mathbb{R}^r$ is any optimal solution of* (2)*, then its coordinates span the same r-dimensional subspace as $\psi_1, \dots, \psi_r$.*

Proof. Write $F = (f_1, \dots, f_r)$. The constraints $\mathbb{E}[F(X)] = 0$ and $\mathbb{E}[F(X)F(X)^\top] = I_r$ are equivalent to requiring $\{f_1, \dots, f_r\}$ to be an orthonormal family in $L_0^2(P_X)$. Using (4), the objective is therefore

$$\mathbb{E}\Big[\langle F(X^{(1)}), F(X^{(2)}) \rangle\Big] = \sum_{\ell=1}^r \langle f_\ell, T f_\ell \rangle_{L^2(P_X)}.$$

Expand $f_\ell = \sum_{j \ge 1} a_{\ell j} \psi_j$ in the orthonormal eigenbasis of T, and define $s_j := \sum_{\ell=1}^r a_{\ell j}^2$. Since $T\psi_j = \lambda_j \psi_j$,

$$\sum_{\ell=1}^r \langle f_\ell, T f_\ell \rangle = \sum_{j \ge 1} \lambda_j s_j, \qquad 0 \le s_j \le 1, \qquad \sum_{j \ge 1} s_j = r.$$

The inequalities follow because the coefficient matrix $A = (a_{\ell j})$ has orthonormal rows. Since $\lambda_1 \ge \lambda_2 \ge \cdots$, the maximum is attained by $s_1 = \cdots = s_r = 1$ and $s_j = 0$ for $j > r$, with optimal value $\sum_{j=1}^r \lambda_j$.

Taking $F_0 = (\psi_1, \dots, \psi_r)$ attains this value, so $\psi_1, \dots, \psi_r$ are feasible and optimal. Finally, the eigengap $\lambda_r > \lambda_{r+1}$ makes the maximizing allocation unique. Hence any optimal F has coordinate span $\operatorname{span}\{\psi_1, \dots, \psi_r\}$. □

Corollary 4.2. *Under the assumptions of Prop. 4.1, $\Pi_r \eta = \sum_{j=1}^r \beta_j \psi_j$, and $B(F^\star) = \|\Pi_r \eta\|_{L^2(P_X)}^2 = \sum_{j=1}^r \beta_j^2 =: B_r$.*

Proof. By Prop. 4.1, $\mathcal{S}_{F^\star} = \mathcal{U}_r$. Since $(\psi_j)_{j \ge 1}$ is an orthonormal eigenbasis of T on $L_0^2(P_X)$ and $\eta \in L_0^2(P_X)$, write $\eta = \sum_{j \ge 1} \beta_j \psi_j$, where $\beta_j = \langle \eta, \psi_j \rangle_{L^2(P_X)}$. Hence $\Pi_r \eta = \sum_{j=1}^r \beta_j \psi_j$, and orthonormality gives $B(F^\star) = \|\Pi_r \eta\|_{L^2(P_X)}^2 = \sum_{j=1}^r \beta_j^2$. □

E.2 Proofs for Appendix C

Proposition C.1 (Structure of the regularized optimum)**.** *Assume $\lambda_r > 0$ and $\lambda_r > \lambda_{r+1}$. Then any population maximizer $F_\gamma^\star$ of* (3) *can be written as*

$$F_\gamma^\star(x) = R\,\big(s_1 \psi_1(x), \dots, s_r \psi_r(x)\big), \tag{5}$$

where $R \in \mathbb{R}^{r \times r}$ is orthogonal and $s_1, \dots, s_r \ge 0$ are scalars. Moreover, for the quadratic penalty in (3)*, the optimal scalings satisfy*

$$s_j^2 = 1 + \frac{\lambda_j}{\gamma}, \qquad j = 1, \dots, r.$$

Proof. Let $G = \mathbb{E}[F(X)F(X)^\top]$. Both terms in (3) are invariant under orthogonal changes of feature coordinates, so diagonalize G and write the resulting coordinates as $g_j = \sqrt{t_j}\, h_j$, where $t_j \geq 0$ and the nonzero h_j's are orthonormal in $L_0^2(P_X)$; complete them arbitrarily to an orthonormal family when some $t_j = 0$. Reorder so that $t_1 \geq \cdots \geq t_r$. The objective becomes

$$\sum_{j=1}^{r} t_j \langle h_j, Th_j \rangle - \frac{\gamma}{2} \sum_{j=1}^{r} (t_j - 1)^2.$$

By the weighted Ky Fan principle, $\sum_j t_j \langle h_j, Th_j \rangle \leq \sum_j t_j \lambda_j$. Hence the objective is at most

$$\sum_{j=1}^{r} \left[t_j \lambda_j - \frac{\gamma}{2} (t_j - 1)^2 \right],$$

whose j-th term is uniquely maximized at $t_j = 1 + \lambda_j/\gamma$. Equality is attained by taking $h_j = \psi_j$. The eigengap fixes the maximizing r-dimensional subspace, while rotations within eigenspaces of repeated eigenvalues are absorbed into an orthogonal output matrix. Thus every maximizer has the stated form with $s_j^2 = t_j = 1 + \lambda_j/\gamma$. □

Proposition C.2. *Let $F_\gamma^\star$ be any population maximizer of* (3)*, written as in* (5)*. Let $\mu_\pm$ and $\Sigma_\pm$ be the class means and class covariances defined as in the main text. Then $\mu_- = -\mu_+$ and $\Delta = \mu_+ - \mu_- = 2\mu_+$. Moreover, when $B_\gamma > 0$, the CDNV and directional CDNV defined in* (1) *satisfy*

$$\tilde{V}_{F_\gamma^\star} = \frac{1}{2} \left(\frac{C_\gamma}{B_\gamma^2} - 1 \right), \qquad V_{F_\gamma^\star} = \frac{1}{2} \left(\frac{S_\gamma}{B_\gamma} - 1 \right). \tag{7}$$

When $B_\gamma = 0$, we set both quantities to $+\infty$ by the same convention as in the main text.

Proof. Let $D = \mathrm{diag}(s_1^2, \ldots, s_r^2)$ and $b = (s_1\beta_1, \ldots, s_r\beta_r)^\top$. Then $\mathrm{Cov}(F_\gamma^\star(X)) = RDR^\top$, while balancedness and centering give $\mu_+ = Rb$, $\mu_- = -\mu_+$, and $\|\mu_+\|_2^2 = B_\gamma$. By total covariance,

$$\Sigma_+ + \Sigma_- = 2\big(RDR^\top - \mu_+\mu_+^\top\big).$$

Taking traces and using $\|\Delta\|_2^2 = 4B_\gamma$ yields $V_{F_\gamma^\star} = \frac{1}{2}(S_\gamma/B_\gamma - 1)$. For $u = \mu_+/\sqrt{B_\gamma}$, we have $u^\top RDR^\top u = C_\gamma/B_\gamma$, so contraction along u gives $\tilde{V}_{F_\gamma^\star} = \frac{1}{2}(C_\gamma/B_\gamma^2 - 1)$. The case $B_\gamma = 0$ follows from the stated convention. □