SYMMETRY-PROJECTED NEURAL COLLAPSE: ANISOTROPIC CONTRACTION OF STRUCTURED WITHIN-CLASS NUISANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Collapse (NC) describes terminal classifier geometry in which within-class features concentrate around their class means. Explicit collapse penalties encourage this NC1 component by contracting every within-class residual direction at the same rate, without distinguishing structured nuisance from task-relevant variation. We introduce *Symmetry-Projected Neural Collapse* (SP-NC), an online alternating adversarial procedure that learns a positive-semidefinite Laplacian $\bar{\Delta}_G = \frac{1}{K}\sum_k L_k^\top L_k$ from prototype-preserving skew-symmetric generators and contracts residuals along its non-kernel eigendirections. Our analysis characterizes key parts of the method: for a fixed operator, SP-NC contracts selected modes at a rate of at least $\lambda\lambda^+_{\min}$, and the NC1 reduction achievable by a fixed rank-r_G support is bounded by $r_\perp \cdot m_{r_G} \cdot \operatorname{Tr}\Sigma_W$; at a fixed representation, the one-generator adversary has an exact top-eigenspace solution with controlled prototype leakage, and finite warm-started ascent has a conditional local tracking bound under spectral separation, a dominant inner iteration, bounded drift, and suitable initialization. Controlled synthetic settings test generator recovery and selective contraction; four image and sensor benchmarks test downstream utility under injected or natural nuisance: Colored MNIST, Fashion-MNIST, Waterbirds, and UCI-HAR. SP-NC improves shifted or shortcut-sensitive accuracy over CE in every seed on all four; matched frozen PCA-SP and JTT controls do not reproduce the gains, while isotropic collapse is competitive or stronger on some datasets but does not uniformly reproduce SP's shifted-accuracy/selectivity profile. A Rotated MNIST failure analysis shows that SP-NC cannot recover a group action the encoder does not expose and demonstrates that it is complementary to equivariant architectures.

1 INTRODUCTION

Neural Collapse (NC) (Papyan et al., 2020) describes the terminal geometry of a classifier: features approach their class means (NC1), the means approach a simplex Equiangular Tight Frame (NC2), and classifier weights align with them (NC3). An explicit collapse penalty such as $\frac{1}{2}\|h - \mu_y\|^2$ encourages NC1 by shrinking every within-class residual direction equally. But a low-dimensional nuisance can dominate those residuals, while other directions may retain useful task variation. Uniform contraction cannot distinguish the two. Group-robustness methods often use group or environment labels (Sagawa et al., 2020; Kirichenko et al., 2023; Arjovsky et al., 2019), or infer biased examples from training dynamics (Nam et al., 2020; Liu et al., 2021). We ask instead whether current within-class geometry can tell a model where to concentrate its contraction.

SP-NC learns that contraction online. On each minibatch, persistent skew-symmetric generators seek directions with strong action on within-class residuals and are penalized for acting on class means. Their positive-semidefinite operator then weights a residual penalty minimized by the network alongside cross-entropy. Discovery repeats as the representation moves; no feature projection is needed at inference. Figure 1 illustrates this loop. Thus SP-NC generalizes the selective-contraction principle underlying NC1 rather than promising to accelerate every conventional NC statistic. In the controlled 3-D example of Figure 2 (Appendix B), the dominant residual eigenvalue ends $7.9\times$ lower than under matched isotropic collapse.

To understand what this procedure can and cannot do, we analyze the two players separately before considering their interaction. A fixed operator gives a mode-wise contraction rate and a bound on removable task-transverse scatter. At a fixed representation, the one-generator adversary has a closed-form optimum and a prototype-leakage bound. Under additional local conditions, finite warm-started updates track that instantaneous optimum; this does not establish that every practical trajectory does so. These results (Section 3) motivate a structural applicability diagnostic (Section 5.1) while keeping the method's guarantees distinct from its empirical behavior.

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.

Across one shared algorithmic template with dataset-specific schedules (Section 4), SP-NC improves shifted or shortcut-sensitive accuracy on Colored MNIST, Fashion-MNIST, Waterbirds, and UCI-HAR. Waterbirds uses validation groups for checkpoint selection but no training-group labels. Matched frozen PCA-SP and JTT controls do not reproduce these gains; isotropic collapse remains a strong, dataset-dependent comparator. The Rotated MNIST analysis (Section 5.2) shows the boundary: a group action must be exposed in the representation for SP-NC to act on it. The method is therefore complementary to equivariant architectures; SP-NC alone does not guarantee NC2 (Section 5.3).

Our contributions are **(i) a training mechanism:** online, prototype-aware anisotropic contraction without training-group labels or inference-time projection; **(ii) a scoped theory and diagnostic:** contraction and capacity bounds, an exact inner response with prototype-leakage control, conditional finite-step tracking, and structural triage (proofs in Appendix A); and **(iii) empirical validation and boundaries:** known-generator recovery, multi-seed utility tests under injected and natural nuisance, and failure analysis that clarifies the role of equivariant encoders.

2 Related work

Papyan et al. (2020) identified NC empirically. The unconstrained-features model (Mixon et al., 2022; Zhu et al., 2021) formalized NC as the global optimum of a regularized layer-peeled objective, extended to various losses (Lu & Steinerberger, 2022; Han et al., 2022) and analyzed under weight decay (Jacot et al., 2025; Súkeník et al., 2023). A prescriptive line inducing NC structure includes fixed-ETF classifiers (Yang et al., 2022), imbalanced-learning variants (Li et al., 2023; Yang et al., 2023; Zhong et al., 2023), and Riemannian projection onto the nearest simplex ETF at each step (Markou et al., 2024). These works shape between-class geometry (NC2/NC3). SP-NC instead targets *within-class anisotropy* (NC1) by learning which residual directions are structured nuisance and contracting them selectively; NC2 is outside its guarantee and is assessed separately (Section 5.3).

The canonical spurious-correlation literature uses group labels to mitigate shortcut learning. GroupDRO (Sagawa et al., 2020) minimizes worst-group loss; LISA (Yao et al., 2022) augments across groups; DFR (Kirichenko et al., 2023) retrains only the last layer on a group-balanced set. Methods without group labels include LfF/JTT (Nam et al., 2020; Liu et al., 2021), which upweight CE-hard examples. IRM (Arjovsky et al., 2019) and REx (Krueger et al., 2021) require environment labels. The SP-NC training objective uses neither group nor environment labels; it discovers structured within-class residual directions and contracts them via an adversarial Lie-generator objective. On Waterbirds, the reported SP-NC arm uses no training-group supervision but does use validation group metadata for checkpoint selection. It is therefore comparable to JTT/LfF as a training-group-free method, and remains distinct from GroupDRO, which uses training groups, and DFR, which uses validation groups to fit its classifier.

A popular line of work encodes known symmetries into layers: G-CNNs (Cohen & Welling, 2016), steerable CNNs (Cohen & Welling, 2017), $\mathrm{SE}(3)$-equivariant networks (Thomas et al., 2018; Fuchs et al., 2020), and Lie-group convolutions (Finzi et al., 2020). Closer to our discovery objective are methods that *learn* infinitesimal generators (Anselmi et al., 2019; Dehmamy et al., 2021; Moskalev et al., 2022; Romero & Lohit, 2022). These maximize likelihood, reconstruction, or contrastive matching under transformation. SP-NC differs structurally: it learns a positive-semidefinite Laplacian and produces *anisotropic invariance*, not equivariance. As Section 5.2 makes explicit, SP-NC is complementary to equivariant architectures rather than a substitute. On natural data, the observed effect is selective contraction of centered task-transverse residual structure, not identification of a semantic transformation group.

DANN (Ganin & Lempitsky, 2015) enforces distributional invariance across domains via an adversarial discriminator. VIB (Alemi et al., 2017) compresses representations by upper-bounding mutual information with a stochastic encoder. Center loss (Wen et al., 2016) is the mathematical progenitor of our $\mathcal{L}_{\mathrm{coll}}$; VICReg (Bardes et al., 2022) regularizes variance/invariance/covariance in self-supervised representations. These objectives address different forms of invariance and feature dispersion; none is the prototype-aware skew-generator selection rule studied here. SP-NC asks which *directions* of within-class spread should be compressed preferentially, and delegates that answer to a skew-symmetric adversary that respects a prototype-preservation constraint. Unlike DANN's domain discriminator, our adversary maximizes a Lie-parameterized quadratic on skew matrices without domain labels.

Two recent NC-inspired debiasing methods are particularly relevant. Wang et al. (2024) use designed shortcut primes guided by NC geometry to discourage biased feature learning; Lu et al. (2024) use a frozen classifier and additional representation losses for min-max fairness. In contrast, SP-NC learns a prototype-aware anisotropic residual operator online without specifying a shortcut prime or using group labels in its training objective. Structural diagnostics for assessing when a method has room to help are rare in the NC and spurious-correlation literatures. WILDS (Koh et al., 2021) standardized natural distribution-shift evaluation, and GOOD (Gui et al., 2022) did so for graph OOD. Our applicability diagnostic (Section 5.1) combines class-labeled features from a CE-trained checkpoint with nuisance measurements, when available, to assess whether anisotropic within-class contraction has structural room to help.

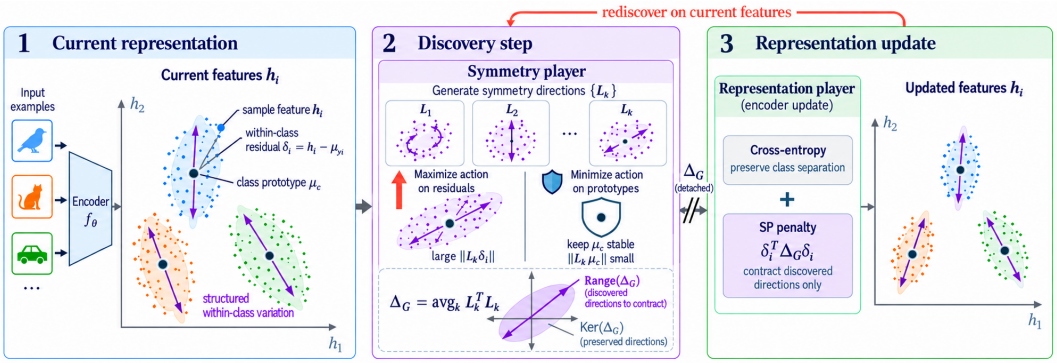


Figure 1: Core SP-NC mechanism. Current features define empirical class prototypes and within-class residuals. A persistent skew-generator player favors residual action while penalizing prototype action and forms a detached positive-semidefinite operator. The representation player then minimizes cross-entropy plus its residual quadratic penalty, and discovery repeats on the updated features. The drawing is schematic: training contracts selected residual directions through gradients; no feature-space projection is applied at inference.

3 Method and Theory

This section covers the setup, motivation and theory supporting the SP-NC mechanism. Figure 1 gives the training logic: find residual-sensitive directions in the current features, then train the network to reduce variation along them. The objectives below specify what the two players see and how the update remains selective.

3.1 Setup and Objectives

At the selected feature layer, let $h_i = \Phi_\theta(x_i) \in \mathbb{R}^d$ with label y_i. In a label-complete minibatch of N examples, μ_c is the empirical mean, or *prototype*, of class c; $\delta_i = h_i - \mu_{y_i}$ is its within-class residual. We seek directions that act strongly on these residuals without needlessly contracting the prototypes. Let $\bar{\mu} = \frac{1}{C}\sum_c \mu_c$; the centered class-mean span used in the contraction theory is $\mathcal{M} = \text{span}\{\mu_c - \bar{\mu}\}_{c=1}^C$. In a Colored-MNIST ablation, ordinary stochastic batches match balanced batches when they remain class-complete; rare-class missingness is a boundary of this estimator (Appendix K.7).

K skew-symmetric generators $L_k = A_k - A_k^\top$ define a positive-semidefinite Laplacian

$$\bar{\Delta}_G \;=\; \frac{1}{K}\sum_{k=1}^{K} L_k^\top L_k. \tag{1}$$

The inner player finds residual-sensitive directions by maximizing the *mean/mean-normalized* objective

$$\begin{aligned} J(\{L_k\}) \;=\; & \frac{1}{N}\sum_{i,k}\|L_k\delta_i\|^2 \;-\; \frac{\alpha}{C}\sum_{k,c}\|L_k\mu_c\|^2 \\ & -\; \rho_{\text{div}}\sum_{k\neq\ell}\langle L_k, L_\ell\rangle_F^2 \;-\; \rho_{\text{norm}}\sum_k(\|L_k\|_F^2 - 1)^2, \end{aligned} \tag{2}$$

with $L_k \in \mathfrak{so}(d)$ and each L_k projected to unit Frobenius norm after every ascent step. The first term rewards action on residuals; the second penalizes action on prototypes. Averaging over examples and represented classes keeps α independent of batch composition. Both terms sum over K generators; only the outer Laplacian is divided by K to keep its aggregate strength stable. The soft norm term stabilizes initialization and has zero gradient after hard projection, up to numerical error.

The prototype penalty uses raw means μ_c, not centered means $\mu_c - \bar{\mu}$. This stronger, non-translation-invariant choice means exact raw preservation $L_k\mu_c = 0$ implies $L_k\bar{\mu} = 0$ and hence $L_k(\mu_c - \bar{\mu}) = 0$, which is the centered condition used by Theorem 1 and Proposition 1. In training the penalty is soft; the exact-preservation hypothesis of those results and the strict-null control is an idealization, not a hard training constraint.

The outer player minimizes

$$\mathcal{L}_{\text{outer}} \;=\; \mathcal{L}_{\text{CE}} \;+\; \lambda\,\mathcal{R}_{\text{SP}}, \quad \mathcal{R}_{\text{SP}} \;=\; \frac{1}{2NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\|L_k\delta_i\|_2^2, \tag{3}$$

holding $\{L_k\}$ fixed. The residual form acts on within-class deviations rather than penalizing class means directly. Under exact prototype preservation $L_k(\mu_c - \bar{\mu}) = 0$, this coincides with $\frac{1}{2NK}\sum_{i,k}\|L_k(h_i - \bar{\mu})\|^2$; under soft preservation, residuals can still lie in prototype-bearing directions and be contracted there. Optional isotropic collapse $\mathcal{L}_{\text{coll}}$ and auxiliary NC2 controllers (Section 5.3) are separate from generator discovery.

3.2 Online alternating algorithm

Each minibatch updates the persistent generator and then the network:

Algorithm 1 Online alternating SP-NC

1: Initialize raw generators $\{A_k\}$ once; retain them across minibatches.
2: **for** each label-complete minibatch B_t **do**
Discover (network fixed):
3: Detach current features; compute batch means μ_c and residuals δ_i.
4: Take S ascent steps on J (Eq. 2), normalizing each $L_k = A_k - A_k^\top$ after every step.
5: Form and detach $\bar{\Delta}_G = K^{-1}\sum_k L_k^\top L_k$.
Train (operator fixed):
6: Recompute features; update (θ, W) once on CE $+\lambda_t\mathcal{R}_{\text{SP}}$ (Eq. 3) with new batch residuals.
7: **end for**

Algorithm 2 in Appendix K.1 specifies initialization, schedules, and the full update order.

SP-NC acts on a chosen feature vector, not on the raw image or a convolution kernel. In the MLP benchmarks it acts after the first affine map and before ReLU; on Waterbirds, a pretrained CNN feeds an affine-prefix head and SP-NC acts within that head. These sites give the adversary coordinates in which shared residual variation may be exposed. They are not universally suitable: an encoder can entangle nuisance with task structure, and downstream nonlinearities can remix variation. Layer choice is therefore part of SP-NC's applicability, not a universal placement rule.

The remainder of this section characterizes what this procedure does: the contraction it induces for a fixed operator (Section 3.3), how much NC1 reduction any such operator can produce (Section 3.4), the fixed-representation inner optima and their prototype leakage (Section 3.5), and a conditional bridge between the two players (Section 3.6). An ideal hard-invariance compatibility observation for simplex class means is given in Appendix A.6.

Throughout, "NC1 reduction" refers to total within-class scatter $\operatorname{Tr}\Sigma_W$ or its trace-normalized variants (e.g. $\operatorname{Tr}\Sigma_W/\operatorname{Tr}\Sigma_B$). The classical statistic $\frac{1}{C}\operatorname{Tr}(\Sigma_W\Sigma_B^\dagger)$ of Papyan et al. (2020) is *exactly invariant* under prototype-preserving fixed-operator SP-NC (Remark 8): $\Sigma_B^\dagger$ weights only scatter inside the class-mean span, which is precisely the subspace SP-NC leaves untouched, whereas SP-NC contracts variance in $\mathcal{M}^\perp$. Accelerated NC1 in this paper therefore means accelerated contraction of task-transverse within-class scatter, not of the $\Sigma_B^\dagger$-weighted statistic.

3.3 Conditional contraction (fixed-operator)

Theorem 1 (Spectral-gap contraction)**.** *Fix $\bar{\Delta}_G$, let $\lambda > 0$, and assume $\bar{\Delta}_G(\mu_c - \bar{\mu}) = 0$ for $c = 1, \dots, C$; equivalently, the centered class-mean span $\mathcal{M} = \operatorname{span}\{\mu_c - \bar{\mu}\}_{c=1}^C$ is contained in $\ker(\bar{\Delta}_G)$. In the unconstrained-features model (each h_i a free variable), under continuous-time gradient flow on the SP term $\mathcal{R}_{\text{SP}}$ with prototypes held fixed and using mean-normalized time $\dot{h}_i = -\lambda N\,\partial\mathcal{R}_{\text{SP}}/\partial h_i$ (equivalently, $t = \tau/N$ for ordinary gradient-flow time τ), residuals obey $\dot{\delta} = -\lambda\bar{\Delta}_G\delta$; components in $\ker(\bar{\Delta}_G)^\perp$ contract as $\|\delta_\perp(t)\| \le \|\delta_\perp(0)\|\exp(-\lambda\lambda_{\min}^+ t)$ where $\lambda_{\min}^+$ is the smallest positive eigenvalue of $\bar{\Delta}_G$ when its range is nonzero (the contracted-component statement is vacuous for the zero operator); components in $\ker(\bar{\Delta}_G)$ are preserved. The contracted subspace $\ker(\bar{\Delta}_G)^\perp$ lies in $\mathcal{M}^\perp$, so Σ_B is unchanged along the flow.*

Corollary 1 (Joint rate with isotropic collapse)**.** *Under the same assumptions, adding at the same feature layer $\beta_{\text{coll}}\,\mathcal{L}_{\text{coll}}$, with $\beta_{\text{coll}} \ge 0$ and $\mathcal{L}_{\text{coll}} = \frac{1}{2N}\sum_i\|\delta_i\|^2$ gives $\dot{\delta} = -(\beta_{\text{coll}}I + \lambda\bar{\Delta}_G)\delta$, hence $\|\delta_\perp(t)\| \le \|\delta_\perp(0)\|\exp\big(-(\beta_{\text{coll}} + \lambda\lambda_{\min}^+)t\big)$ and $\|\delta_0(t)\| = \|\delta_0(0)\|\exp(-\beta_{\text{coll}}t)$ for the kernel component: selected modes contract at the compounded rate, unselected modes at the isotropic rate alone.*

If an experiment divides $\mathcal{L}_{\text{coll}}$ by a fixed, detached scale $s_0 > 0$, the corollary applies with effective coefficient β_{coll}/s_0. It does not assert an additive operator or compounded rate when SP and collapse are applied at different layers.

Proofs are in Appendix A.2, where we also show that the "prototypes held fixed" hypothesis is automatic when μ_c is the empirical class mean of a fixed labeled cohort under the isolated SP flow (the flow preserves class means), and that the hypothesis $\mathcal{M} \subseteq \ker(\bar{\Delta}_G)$ is not used for the residual dynamics but only to locate the contracted directions in the task-transverse subspace. The theorem is conditional on a fixed operator and does not by itself prove that the online adversary discovers a semantically meaningful $\bar{\Delta}_G$.

3.4 Applicability upper bound

Proposition 1 (NC1 reduction is bounded by $r_\perp \cdot m_{r_G}$)**.** *Let Σ_W be the within-class scatter at a fixed representation checkpoint and let $P_\perp$ project onto the orthogonal complement of the centered class-mean span $\mathcal{M} = \operatorname{span}\{\mu_c - \bar{\mu}\}$. Let $\bar{\Delta}_G = \frac{1}{K}\sum_k L_k^\top L_k$ be built from prototype-preserving skew generators, $L_k(\mu_c - \bar{\mu}) = 0$ for all k, c, with $\operatorname{rank}(\bar{\Delta}_G) \le r_G$, and let Π_G be the orthogonal projector onto $\operatorname{range}(\bar{\Delta}_G)$. Consider any fixed-operator SP contraction of the residuals: the flow of Theorem 1 over any horizon, or any sequence of discrete steps $\delta \leftarrow (I - s_t\bar{\Delta}_G)\delta$ with, for a nonzero operator, $0 < s_t \le 1/\|\bar{\Delta}_G\|_{\text{op}}$ (the zero-operator case is trivial). Then the reduction in $\operatorname{Tr}\Sigma_W$ it produces satisfies*

$$\Delta[\operatorname{Tr}\Sigma_W]_{\bar{\Delta}_G} \;\le\; \operatorname{Tr}\big(\Pi_G\,\Sigma_W\big) \;\le\; \sum_{j=1}^{r_G}\lambda_j^\perp \;=\; r_\perp \cdot m_{r_G} \cdot \operatorname{Tr}\Sigma_W, \tag{4}$$

where $\lambda_j^\perp$ are the descending eigenvalues of $P_\perp\Sigma_W P_\perp$, $r_\perp = \operatorname{Tr}(P_\perp\Sigma_W)/\operatorname{Tr}\Sigma_W$, and $m_{r_G} = \sum_{j=1}^{r_G}\lambda_j^\perp/\operatorname{Tr}(P_\perp\Sigma_W)$. These normalized quantities are used when their denominators are positive; if $\operatorname{Tr}\Sigma_W = 0$ or $\operatorname{Tr}(P_\perp\Sigma_W) = 0$, the corresponding removable scatter and the right-hand side of Eq. 4 are defined as zero. The first inequality is an equality for the infinite-horizon flow, and the instantaneous rate is $-\frac{d}{dt}\operatorname{Tr}\Sigma_W = 2\lambda\operatorname{Tr}(\bar{\Delta}_G\Sigma_W)$.

The proof (Appendix A.3) reduces to two facts: any product of fixed-operator contraction maps acts on Σ_W through an operator $0 \preceq I - Q^2 \preceq \Pi_G$, and Π_G is a rank-r_G projector supported in $P_\perp\mathbb{R}^d$, so the Ky Fan inequality applies to $P_\perp\Sigma_W P_\perp$. The bound therefore holds for *every* prototype-preserving operator of rank at most r_G, including a discovered operator whenever it satisfies the exact preservation and rank hypotheses, which is what makes it a structural upper-bound diagnostic.

Proposition 1 justifies $r_\perp$ and m_{r_G} as fixed-checkpoint applicability statistics: at the measured checkpoint, a fixed prototype-preserving support on a dataset with $r_\perp = 0.014$ (ModelNet40 at the tested bridge, Appendix I) can remove at most 1.4% of the current trace-NC1; $r_\perp = 0.827$ (Colored MNIST at the SP layer) leaves up to 82.7% available to such a support. These figures do not bound a coupled trajectory whose representation and operator support change online. The canonical implementation uses dense skew generators, so $2K$ is not a hard rank bound on its finite-step operators. We report the $r = 2K$ quantity as an idealized one-plane-per-generator capture diagnostic, and use the measured hard support rank when applying the proposition to a realized dense operator; effective rank is reported only as a descriptive spectral-concentration heuristic (Remark 5).

3.5 Prototype protection and multi-generator capacity

The raw prototype-preservation term in Eq. 2 increasingly penalizes generator action on class means as α grows; at a fixed representation the inner problem is solvable in closed form, which yields the leakage rate below. The fixed-representation solution is the ideal best response used by the conditional analysis of Section 3.6; finite warm-started discovery ascends the same inner objective, while proximity to that response requires the local conditions of Theorem 2. Both terms of Eq. 2 are traces against scatter matrices (Lemma 2): $\frac{1}{N}\sum_{i,k}\|L_k\delta_i\|^2 = K\operatorname{Tr}(\bar{\Delta}_G\Sigma_W)$ and $\frac{1}{C}\sum_{k,c}\|L_k\mu_c\|^2 = K\operatorname{Tr}(\bar{\Delta}_G\Sigma_P)$, where $\Sigma_P = \frac{1}{C}\sum_c \mu_c\mu_c^\top = \Sigma_B + \bar{\mu}\bar{\mu}^\top$ is the raw prototype second moment. Thus, apart from the diversity and norm terms, dividing the complete fixed-K objective by K shows that the adversary maximizes $\operatorname{Tr}\big(\bar{\Delta}_G(\Sigma_W - \alpha\Sigma_P)\big)$ over unit-Frobenius skew matrices. Let $\mathcal{P} = \operatorname{span}\{\mu_c\}_{c=1}^C$ and let $P_{P,\perp}$ project onto $\mathcal{P}^\perp$.

Proposition 2 (Inner maximizer and prototype leakage)**.** *Assume $d \ge 2$. Let $S_\alpha = \Sigma_W - \alpha\Sigma_P$ have eigenpairs (s_a, w_a) with $s_1 \ge \cdots \ge s_d$, and let $\lambda_1 \ge \lambda_2$ and $\lambda_1^{P,\perp} \ge \lambda_2^{P,\perp}$ be the top eigenvalues of Σ_W and of $P_{P,\perp}\Sigma_W P_{P,\perp}$.*

(i) For $K = 1$, $\max_{L \in \mathfrak{so}(d),\, \|L\|_F = 1} J(L) = \frac{1}{2}(s_1 + s_2)$, attained by $L^\star = \pm(w_1 w_2^\top - w_2 w_1^\top)/\sqrt{2}$ (unique up to sign provided $d \geq 3$ and $s_2 > s_3$; for $d = 2$ the plane is unique), for which $\bar{\Delta}_G^\star = \frac{1}{2}(w_1 w_1^\top + w_2 w_2^\top)$: the displayed generator rotates in the plane of the two leading eigenvectors of S_α.

(ii) For $1 \leq K \leq \dim \mathfrak{so}(d) = d(d-1)/2$ with Frobenius-orthonormal generators (the $\rho_{\mathrm{div}} \to \infty$ limit), $\max \frac{1}{K} \sum_k \mathrm{Tr}(L_k^\top L_k S_\alpha) = \frac{1}{2K} \sum_{m=1}^{K} \sigma_m$, where $\sigma_1 \geq \sigma_2 \geq \cdots$ enumerate $\{s_a + s_b\}_{a<b}$, attained by the corresponding K coordinate rotation planes. With $\rho_{\mathrm{div}} = 0$ the maximum is $\frac{1}{2}(s_1 + s_2)$ for every K, and K copies of $L^\star$ are optimal; under spectral degeneracy, distinct leading generators can also be optimal.

(iii) (Leakage.) If $\alpha > 0$ and $\dim \mathcal{P}^\perp \geq 2$, every maximizer $L^\star$ in (i) satisfies

$$\frac{1}{C} \sum_c \|L^\star \mu_c\|^2 \;=\; \mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_P) \;\leq\; \frac{(\lambda_1 + \lambda_2) - (\lambda_1^{P,\perp} + \lambda_2^{P,\perp})}{2\alpha}. \tag{5}$$

In particular leakage is $O(1/\alpha)$, and it is exactly zero for every $\alpha > 0$ whenever $\lambda_1 + \lambda_2 = \lambda_1^{P,\perp} + \lambda_2^{P,\perp}$ (for instance when the two leading eigenvectors of Σ_W lie in $\mathcal{P}^\perp$).

Proof in Appendix A.4. Part (ii) characterizes the zero-diversity and exact-orthogonality endpoints: without diversity, duplicating a leading generator is optimal, whereas exact orthogonality can allocate generators across leading skew-space modes. Part (i) also clarifies the frozen PCA-SP baseline (Section 4.1.1): as $\alpha \to \infty$, every raw-mean inner-optimum accumulation point solves the raw-prototype-null restricted problem; with a two-dimensional complement or the spectral gap in Remark 6, its Laplacian converges to one-half the projector onto the leading two-plane of $P_{P,\perp} \Sigma_W P_{P,\perp}$. The reported frozen PCA-SP control instead projects away the centered class-mean span; it is a closely related static prototype-orthogonal control, not literally the large-α optimum of the raw-mean online objective. Online SP-NC continually updates persistent generators against the current penalized inner objective as the representation drifts (Theorem 2(i)). The residual-based $\mathcal{R}_{\mathrm{SP}}$ (Eq. 3) further insulates the outer objective from any remaining leakage into the class-mean span. Multi-generator capacity: normalizing the outer Laplacian by K separates aggregate contraction strength from generator count, but K does not identify the number of physical causal factors and cannot create sample-conditioned actions (Section 5.2).

3.6 Conditional coupling of the two players

For the $K = 1$ configuration used in the four headline dataset results, let $s_1 \geq \cdots \geq s_d$ be the eigenvalues of $S_\alpha = \Sigma_W - \alpha \Sigma_P$ and define $\Phi_\alpha := \max_{\|L\|_F = 1} \mathrm{Tr}(L^\top L S_\alpha) = \frac{1}{2}(s_1 + s_2)$ and $F_\alpha := \mathcal{L}_{\mathrm{CE}} + \frac{\lambda}{2} \Phi_\alpha$. The value Φ_α is one half of a top-two eigenvalue sum, not a Ky Fan norm in general because S_α can be indefinite.

Theorem 2 (Exact best response and conditional outer dynamics)**.** *Let $K = 1$, $d \geq 3$, $\dim \mathcal{P}^\perp \geq 2$, and $s_2 > s_3$. (i) Once the hard projection has placed $L = A - A^\top$ on the unit Frobenius sphere, one implemented ascent step in the raw parameter A induces power iteration on $I + 8\eta_G \mathcal{S}_\alpha$. If its leading eigenvalue is unique in magnitude and the initialization has a nonzero leading component, the induced Laplacian converges to $\bar{\Delta}_\alpha^\star$. (ii) With class means fixed, full-batch free features, and an exact inner best response, the SP component of the outer step is the Danskin subgradient $\frac{\lambda}{N} \bar{\Delta}_\alpha^\star \delta_i$ of $\frac{\lambda}{2} \Phi_\alpha$ on the within-class block and has zero classwise mean. (iii) Under a uniform skew-space eigengap, a dominant-magnitude inner iteration, bounded representation drift, and a local angle condition, the S warm-started inner steps track the moving exact best response at a geometric rate. (iv) If $W P_{P,\perp} = 0$, rotations fixing $\mathcal{P}$ leave $\mathcal{L}_{\mathrm{CE}}$ and Φ_α invariant. If additionally a frozen operator satisfies $\bar{\Delta}_0 = P_{P,\perp} \bar{\Delta}_0 P_{P,\perp}$ and the stated rank–dimension condition, such a rotation can drive its penalty to zero without changing the CE value or transverse covariance spectrum.*

The full statement and proofs are in Appendix A.5. These claims have distinct scopes: fixed-representation optimization identifies the exact inner response, while finite ascent tracks it only under the local conditions in (iii). The latter is not required for the practical alternating procedure to define a selective residual penalty, nor is it a global convergence claim for the minibatched neural-network optimizer. Two consequences are an $O(1/\alpha)$ operator approach to the hard-constraint optimum under spectral separation and, at matched instantaneous contraction budget, a leading-plane rate advantage of $1/(r_\perp^P m_2^P)$ over isotropic collapse, using the raw prototype-transverse statistics defined in Appendix A.5.

4 Empirical results

Controlled synthetic validation of the mechanism under identifiability (3-D anisotropic mixture, hidden 10-D SO(2), full linear SO(3)) is reported in Appendix B. The main comparison covers controlled chromatic shortcuts (Colored

		No group labels for training/refitting					Group-informed references		
Dataset	Metric	CE	CE+IC	Frozen PCA-SP	SPNC	JTT	Bal-ERM	GroupDRO	DFR
Colored MNIST	Shifted acc. (%)	47.53±2.98	42.54±3.34	46.95±1.89	**92.64±0.52**	46.55±2.86	n/e	n/e	80.55±1.08
	trace-NC1 ↓	0.622	**0.192**	0.734	0.512	0.593	n/e	n/e	0.622
	NC2 error ↓	**0.909**	1.339	0.917	0.928	0.958	n/e	n/e	0.909
	Task/nuis. ratio	0.373	0.114	0.613	**340.7**	0.370	n/e	n/e	0.373
	Group info	0	0	0	0	0	n/e	n/e	H_f
Fashion-MNIST	Shifted acc. (%)	34.37±2.98	31.97±8.88	37.22±5.70	**82.58±0.47**	22.82±1.62	n/e	n/e	66.83±5.06
	trace-NC1 ↓	0.466	**0.010**	0.474	0.324	0.365	n/e	n/e	0.466
	NC2 error ↓	1.645	2.773	1.621	**1.443**	1.810	n/e	n/e	1.645
	Task/nuis. ratio	0.520	**4259.7**	0.603	166.7	0.336	n/e	n/e	0.520
	Group info	0	0	0	0	0	n/e	n/e	H_f
Waterbirds	Shifted acc. (%)	71.45±3.69	76.60±3.30	73.67±3.67	**78.41±2.22**	73.64±2.68	78.79±1.29	86.05±1.75	84.52±2.81
	trace-NC1 ↓	1.50	**0.65**	7.82	0.79	1.37	1.08	0.93	1.41
	NC2 error	–	–	–	–	–	–	–	–
	Task/nuis. ratio	2.07	**11.83**	7.30	4.93	2.63	5.29	317.6	1.91
	Group info	V_s	V_s	V_s	V_s	V_s	$T + V_s$	$T + V_s$	V_f
UCI-HAR	Shifted acc. (%)	93.53±0.32	**95.68±0.76**	93.71±0.09	94.97±0.31	92.95±0.42	93.45±0.34	93.53±0.41	92.70±0.47
	trace-NC1 ↓	0.233	**0.025**	0.212	0.239	0.274	0.241	0.233	0.233
	NC2 error ↓	0.789	1.035	0.803	**0.518**	0.697	0.810	0.782	0.789
	Task/nuis. ratio	8.55	**102.7**	9.21	11.48	7.21	8.13	8.50	8.55
	Group info	0	0	0	0	0	T	T	V_f

Table 1: Five-seed comparison. CE+IC is CE with normalized isotropic collapse; Bal-ERM is group- or subject-balanced ERM. Accuracy entries are mean ± sample standard deviation; diagnostic entries are five-seed means. *n/e* means not evaluated; dashes mark the non-informative binary-class NC2 row. Bold and underline mark the numerical best and second-best values among the five non-group-trained/refit methods within each row; the three group-informed columns are unranked references. Energy ratios are descriptive and are not independent measures of quality. Benchmark-specific metrics, group-information codes, and SP-NC variants are defined in the text.

MNIST and Fashion-MNIST), a natural image-background shortcut (Waterbirds), and natural subject shift in sensor features (UCI-HAR). SP-NC uses its online $K{=}1$ algorithm at a pre-nonlinear affine site on all four; inner steps, outer weights, and schedules are dataset-specific. Table 1 uses five paired seeds per dataset. Accuracy is the primary shifted test metric for each dataset, whereas trace-NC1, NC2, and task/nuisance ratios are descriptive penultimate-layer diagnostics on that same evaluation split. Ratios compare methods within a dataset, not representation scales across datasets.

4.1 Benchmarks and protocol

In the group-information row of Table 1, 0 means no nuisance/group annotations, T denotes training-group labels, V_s validation-group labels used for checkpoint selection, V_f validation-group labels used to fit a classifier, and H_f a held-out nuisance-equalized classifier-refit set. Thus DFR and the group-aware baselines have stronger supervision than SP-NC. The SP-NC column uses standard online SP on Colored and Fashion-MNIST (F-MNIST), the early affine-head site on Waterbirds, and the late-decay schedule on UCI-HAR. Full paired five-seed intervals and NC/energy measurements for each method are in Appendix K.5. Its discovery adds training-only work whose relative overhead depends on the model and inner-step count; measured costs and sensitivity are reported in Appendix K.6.

4.1.1 Controlled chromatic-variance shortcuts

Colored MNIST and F-MNIST share a centered chromatic-variance shortcut with balanced within-class color angles and class-dependent radii, but differ in image classes. Both use a four-layer MLP with SP at its first pre-ReLU affine site. Their permuted tests reverse the training class–radius association; ID, equalized, and amplified views provide additional diagnostics. F-MNIST's results use the locked full official test after all five-seed checkpoints were fixed.

4.1.2 Natural background shortcut

Bird class (waterbird/landbird) is spuriously correlated with background (water/land) in training. Worst-group accuracy on the four bird×background subgroups is the primary metric. All methods share an ImageNet-pretrained ResNet-18 and affine-prefix head; SP-NC acts at an early pre-nonlinear head affine. The checkpoint maximizes val-

idation worst-group accuracy, so even methods without training-group supervision use validation group labels for selection.

4.1.3 NATURAL SUBJECT VARIATION

The six-class UCI Human Activity Recognition task uses 561 engineered sensor features and an official subject-disjoint test. We reserve six subjects for development (15 train / 6 development / 9 test), use an MLP with SP at the first pre-ReLU affine, and evaluate fixed epoch-100 checkpoints. The selected SP-NC schedule decays its coefficient after epoch 80 to avoid the observed terminal instability; CE and SP-NC receive no subject identities.

4.2 RESULTS AND INTERPRETATION

On the controlled color shortcuts, online SP-NC improves permuted accuracy over CE by $+45.11$ points on Colored MNIST and $+48.21$ points on Fashion-MNIST, with all five paired seeds positive in each study. DFR is the strongest alternative on both but uses an additional nuisance-equalized holdout; SP-NC exceeds it by 12.09 and 15.75 points, respectively. Frozen PCA-SP does not reproduce SP-NC gains, while JTT falls below CE on both, by 11.54 points on Fashion-MNIST. CE+IC lowers accuracy on the two shortcuts despite compressing NC1, and SP+IC behaves differently across them: it improves Colored-MNIST accuracy to 93.30% but falls to 39.67% on Fashion-MNIST (Appendix I). SP-NC trades 3.52 and 9.71 ID points, respectively, for this shift robustness. An independent ten-seed Colored-MNIST confirmation gives a paired $+41.64$ permuted points, 95% CI $[+40.57, +42.71]$.

On natural shifts, early-head SP-NC raises Waterbirds worst-group accuracy by a paired $+6.96 \pm 1.73$ points and average accuracy by $+2.84$ points over CE, with all five paired worst-group gains positive. Its penultimate background energy falls from 260.35 to 24.08 ($10.8\times$), and its worst-group accuracy nearly matches group-balanced ERM (78.41% versus 78.79%) without training-group labels. GroupDRO and DFR remain stronger on worst-group accuracy with group-informed training or classifier refitting. Additional training-group-label-free AFR and LFR comparisons under the same ResNet-18 architecture are reported in Appendix Table 11. Late-decay SP-NC improves UCI-HAR subject-disjoint accuracy by $+1.44 \pm 0.58$ points in all five seeds and lowers NC2 error from 0.789 to 0.518, although CE+IC is the strongest single method there. These are qualified utility results, not evidence that a recognizable group generator was recovered on natural data.

The geometry is correspondingly non-uniform. On the primary permuted Colored-MNIST view, SP-NC lowers trace-NC1 from 0.622 to 0.512, while its ID-view trace-NC1 instead rises from 0.407 to 0.512; NC2 error does not improve on the permuted view. Fashion-MNIST and Waterbirds show lower trace-NC1 under SP-NC, but UCI-HAR's trace-NC1 is essentially unchanged. The task/nuisance ratio records strong selectivity on the color shortcuts (on Colored MNIST, 0.373 to 340.7 on the primary permuted view), yet a large ratio alone is insufficient: CE+IC nearly erases both task and nuisance energy on Fashion-MNIST and loses substantial accuracy. Waterbirds GroupDRO's ratio is also highly seed-unstable (317.6 ± 503.2) and should not be read as a separate representation-level superiority claim.

Independent five-seed mechanism controls (Appendix D) separate learned discovery from arbitrary anisotropic contraction: on permuted Colored MNIST, learned SP-NC reaches 92.51% versus 88.84% for a freshly resampled random operator and 78.65% when the online operator's eigenvalues are retained but its eigenvectors randomized. These comparisons support data-dependent selective orientation, not the stronger claim that gradient ascent is the only way to obtain it. An exact online top-two eigenspace response also strongly outperforms CE, but finite SP-NC exceeds it by 1.23 paired permuted points on Colored MNIST; the finite operator need not track that exact plane (Appendix E). The prototype-preservation ablation (Appendix F) finds a useful moderate-α regime rather than monotonic gains; the multi-generator ablation (Appendix G) shows that moderate diversity prevents duplicate generators without making larger K more accurate. Known-nuisance equalization and randomization exceed SP-NC by 1.48–2.00 permuted points on the color tasks (Appendix H), but use the injected nuisance construction unavailable to the annotation-matched methods. Other secondary and negative evidence is in Appendix I.

5 SCOPE: APPLICABILITY, FAILURE MODES, AND NC2

5.1 STRUCTURAL APPLICABILITY DIAGNOSTIC

Proposition 1 bounds achievable NC1 reduction for a fixed rank-constrained support by $r_\perp \cdot m_{r_G}$. This fixed-checkpoint capacity bound is not a guarantee of coupled online gain. Definitions and layerwise case studies are in Appendix C.

At a CE-trained checkpoint, we first inspect candidate pre-nonlinear affine layers for task-transverse within-class scatter and residual spectral concentration; these measurements need class labels, but no nuisance annotation. Where

paired transformations or development-group metadata are available, we also test whether the measured nuisance lies outside the centered class-mean span and whether its leading plane transfers across splits. High transverse capacity alone is insufficient when nuisance occupies task-bearing directions or the representation changes during training. This *structural triage* identifies plausible SP-NC intervention sites, not a prediction of gain under coupled training.

5.2 Failure analysis: Rotated MNIST

Rotated MNIST tests the boundary of SP-NC's scope. Under a 100-epoch full-circle evaluation on a four-layer MLP, SP-NC at the penultimate layer improves symmetric-OOD accuracy from 41.49% (CE) to 43.13%; a $+1.64$-point gain over CE but only $+0.23$ over normalized isotropic collapse. This is not a decisive SP-NC-specific advantage.

We conducted a five-phase repair program (inner-step and generator counts, curricula and multi-view training, controlled synthetic ladders, paired-tangent variants, learned conditional fields, classifier/representation splits, and explicit group architectures). None produced a resolved SP-specific gain. In the explicit group-architecture experiment, a C8-orbit CNN solved rotation directly ($47.96\% \rightarrow 96.63\%$ symmetric-OOD without SP-NC). On the C8 model, adding SP-NC reduced rotation energy from 12.56 to 0.007 but also reduced task energy from 573.7 to 2.06 and lowered OOD accuracy by 3.19 points.

The failure is structural: an ordinary MLP or CNN does not expose image rotation as a shared feature-space linear action. Local rotation tangents are identity-, angle-, and layer-dependent; their union overlaps task-sensitive structure; and more generators enlarge the contracted union without providing sample-conditioned assignment. At the first pre-ReLU layer of a CE-trained MLP, only $r_\perp = 0.368$ of within-class scatter lies outside the centered class-mean span, while 0.429 of measured rotation nuisance lies inside that span. Even the best rank-two projector confined to the task-transverse subspace captures only 0.369 of measured rotation nuisance. These fixed-checkpoint measurements limit safe one-plane cleanup, but do not by themselves bound the coupled online trajectory; the failure also reflects the absence of a transferable shared linear action. Further comparisons are in Appendix C, Table 2.

SP-NC is not a replacement for group-equivariant architecture. Where the architecture already handles the group (C8 for rotation), SP-NC is redundant or destructive. Where the architecture does not expose the group as a linear feature action, more generators cannot repair the deficit. Complementarity between an equivariant front end (handling a known group) and SP-NC (handling an unknown, task-transverse residual nuisance at a pre-nonlinear affine site) is the correct positioning.

5.3 NC2 scope

SP-NC directly targets NC1 through anisotropic within-class contraction. Its prototype penalty discourages generator action on class means, but does not prescribe their relative angles. Thus NC2 (the simplex-ETF configuration of centered class means) is neither guaranteed by SP-NC nor a success criterion for its selective-contraction mechanism. Explicit NC2 control is a separate objective, with exploratory SEP and raw-Gram results in Appendix J. Under ideal hard invariance, an ETF can fit inside the operator kernel only when that kernel has sufficient dimension (Appendix A.6); this is not a convergence claim.

6 Conclusion

We introduced SP-NC, a mean/mean-normalized adversarial method that learns prototype-preserving skew-symmetric generators and contracts within-class residuals along their non-kernel eigendirections. The theory characterizes the two players conditionally: the exact fixed-representation adversary selects a leading penalized skew plane, the exact-best-response, fixed-prototype free-feature outer SP step is its Danskin subgradient, and finite warm-started discovery has a conditional local tracking bound under spectral separation, a dominant inner iteration, bounded drift, and suitable initialization (Theorem 2). A spectral-gap contraction result (Theorem 1) and a hard upper bound on the NC1 reduction achievable by a fixed rank-constrained support (Proposition 1) motivate the structural applicability diagnostic. Controlled synthetic settings identify the mechanism, while four benchmark positives test its utility across image and sensor settings; three of the four are multi-class.

SP-NC fails when the encoder does not expose the nuisance as a shared feature-space linear action (Rotated MNIST, PointNet SO(3) after max-pool) and when nuisance is encoded in class means (mean-color boundary). The correct positioning is complementary to group-equivariant architectures: architectures handle known groups whose action must remain structured; SP-NC discovers unknown centered task-transverse residual structure at a pre-nonlinear affine interface. A modular geometry-exposing adapter combining an equivariant front end with a SP-NC-monitored residual channel is the natural future research direction.

AI USE STATEMENT

We used generative AI tools (Claude, ChatGPT) to assist with improving and editing the manuscript, to develop and check the proofs in Appendix A from theorem statements and proof sketches supplied by the authors, generation and editing of the main mechanism figure, and to draft the bibliography entries, which were then manually verified against the original sources. Generative AI tools were also used to code some experiments and generate a clean repository of the code for supplementary material. We did not use generative AI tools to generate research ideas. All AI-assisted text and proofs were reviewed line by line by the authors; the closed-form inner-maximizer and leakage results were additionally verified numerically. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes a training-time regularizer and evaluates it on public benchmarks (Colored MNIST, Fashion-MNIST, Waterbirds, UCI-HAR, and the datasets in Appendix I); no new data were collected and no human subjects were involved. CelebA is used only for a layerwise diagnostic in Appendix C. Methods that suppress spurious correlations can improve fairness under group shift, but SP-NC does not certify group robustness and, as Section 4.1.2 shows, group-supervised methods remain stronger when group labels are available; we caution against deploying it as a substitute for group-aware evaluation.

REPRODUCIBILITY STATEMENT

The objective, algorithm, and default hyperparameters are given in Section 3 and at the start of Section 4; per-dataset architectures, splits, seeds, and schedules are stated in the corresponding subsections of Section 4 and Appendices B–I. All theoretical claims are stated with their assumptions in Section 3 and proved in Appendix A; detailed training and measurement protocols are in Appendix K. Source code for SP-NC, the diagnostic, and the experiment configurations is provided as supplementary material.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Fabio Anselmi, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio. Symmetry-adapted representation learning. *Pattern Recognition*, 86:201–208, 2019. arXiv:1607.05164 (2016).

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. arXiv:1612.08498.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with Lie algebra convolutional network. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, 2020.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

Mahdi Ghaznavi, Hesam Asadollahzadeh, HamidReza Yaghoubi Araghi, Fahimeh Hosseini Noohdani, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Annotation-free group robustness via loss-based resampling, 2023. URL `https://arxiv.org/abs/2312.04893`.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 35, 2022.

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.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. In *International Conference on Learning Representations*, 2025. arXiv:2410.04887.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, 2021.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. arXiv:2012.08465.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Neural collapse-inspired debiased representation learning for min-max fairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2048–2059, 2024. doi: 10.1145/3637528.3671902.

Evan Markou, Thalaiyasingam Ajanthan, and Stephen Gould. Guiding neural collapse: Optimising towards the nearest simplex equiangular tight frame. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):11, 2022. arXiv:2011.11619.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. LieGG: Studying learned Lie group generators. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28448–28467. PMLR, 2023. URL https://proceedings.mlr.press/v202/qiu23c.html.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Peter Súkeník, Marco Mondelli, and Christoph H. Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang, and Min Yang. Navigate beyond shortcuts: Debiased learning through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12322–12331, 2024.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gl3D-xY7wLq.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *International Conference on Learning Representations*, 2023.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, 2022.

Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

APPENDIX CONTENTS

- **A** Proofs
- **B** Controlled validation
- **C** Structural applicability case studies
- **D** Learned discovery versus arbitrary quadratic operators
- **E** Exact online best-response control
- **F** Prototype-preservation ablation
- **G** Multi-generator diversity and capacity
- **H** Known-nuisance data controls
- **I** Secondary dataset studies
- **J** Auxiliary NC2 controls
- **K** Reproducibility details

A PROOFS

A.1 CONVENTIONS AND PRELIMINARY LEMMAS

$\langle\cdot,\cdot\rangle_F$, $\|\cdot\|_F$ denote the Frobenius inner product and norm, $\|\cdot\|_{\mathrm{op}}$ the spectral norm, and $\mathfrak{so}(d)$ the real skew-symmetric $d\times d$ matrices. Class means μ_c, global mean $\bar\mu=\frac{1}{C}\sum_c\mu_c$, residuals $\delta_i=h_i-\mu_{y_i}$, and

$$\Sigma_W=\frac{1}{N}\sum_{i=1}^{N}\delta_i\delta_i^\top,\qquad \Sigma_B=\frac{1}{C}\sum_{c=1}^{C}(\mu_c-\bar\mu)(\mu_c-\bar\mu)^\top.$$

$\Sigma_P=\frac{1}{C}\sum_c\mu_c\mu_c^\top=\Sigma_B+\bar\mu\bar\mu^\top$ denotes the raw prototype second moment used by the implemented inner objective, and $\mathcal{P}=\mathrm{span}\{\mu_c\}_c=\mathrm{range}(\Sigma_P)$, with $P_{P,\|}$ and $P_{P,\perp}$ the projectors onto $\mathcal{P}$ and $\mathcal{P}^\perp$. We write $\Sigma_{P\perp}=P_{P,\perp}\Sigma_WP_{P,\perp}$ for the raw task-transverse within-class scatter, with eigenvalues $\lambda_1^{P,\perp}\ge\lambda_2^{P,\perp}\ge\cdots$, and $\Sigma_\times=P_{P,\|}\Sigma_WP_{P,\perp}$ for the cross block, so that in the decomposition $\mathcal{P}\oplus\mathcal{P}^\perp$

$$\Sigma_W=\begin{bmatrix}P_{P,\|}\Sigma_WP_{P,\|} & \Sigma_\times\\ \Sigma_\times^\top & \Sigma_{P\perp}\end{bmatrix},\qquad S_\alpha=\Sigma_W-\alpha\Sigma_P.$$

Σ_B is class-averaged, matching the $1/C$ normalization of the centered class geometry. Let $\mathcal{M}=\mathrm{span}\{\mu_c-\bar\mu\}_c=\mathrm{range}(\Sigma_B)$ ($\dim\mathcal{M}\le C-1$), with orthogonal projectors $P_\|$ onto $\mathcal{M}$ and $P_\perp=I-P_\|$. *Raw prototype preservation* means $L_k\mu_c=0$ for all k,c, exactly the idealized zero-penalty condition of Eq. 2. It implies the centered condition $L_k(\mu_c-\bar\mu)=0$ and hence $\mathcal{M}\subseteq\ker\bar\Delta_G$ by Lemma 1(a). When a result requires only the weaker centered condition, it states that condition explicitly. For a PSD operator $\bar\Delta_G$ with eigenpairs (γ_j,v_j) we write Π_G for the orthogonal projector onto $\mathrm{range}(\bar\Delta_G)=\ker(\bar\Delta_G)^\perp$ and $P_0=I-\Pi_G$ for the projector onto the kernel.

Lemma 1 (Structure of $\bar\Delta_G$)**.** *Let $L_1,\dots,L_K\in\mathfrak{so}(d)$ and $\bar\Delta_G=\frac{1}{K}\sum_k L_k^\top L_k$.*

(a) $\bar\Delta_G$ is symmetric PSD, $v^\top\bar\Delta_Gv=\frac{1}{K}\sum_k\|L_kv\|^2$, and $\ker\bar\Delta_G=\bigcap_k\ker L_k$.

(b) The nonzero singular values of each L_k occur in equal pairs, so $\mathrm{rank}(L_k)$ is even. If $L_k=A_k-A_k^\top$ with $\mathrm{rank}(A_k)\le r$ then $\mathrm{rank}(L_k)\le 2r$ and $\mathrm{rank}(\bar\Delta_G)\le 2rK$.

(c) $\|L_k\|_{\mathrm{op}}^2\le\frac{1}{2}\|L_k\|_F^2$, with equality for a nonzero L_k iff $\mathrm{rank}(L_k)=2$ (and also for $L_k=0$). Hence unit-Frobenius generators give $0\preceq\bar\Delta_G\preceq\frac{1}{2}I$.

(d) Under centered prototype preservation, $\mathcal{M}\subseteq\ker\bar\Delta_G$; consequently $\bar\Delta_G=P_\perp\bar\Delta_GP_\perp$, $\bar\Delta_G\Sigma_B=\Sigma_B\bar\Delta_G=0$, and $\Pi_G\preceq P_\perp$. Under raw preservation the same holds with $P_{P,\perp}$ in place of $P_\perp$.

Proof. (a) $v^\top L_k^\top L_kv=\|L_kv\|^2\ge 0$; the average vanishes iff every term does. (b) A real skew matrix is normal with purely imaginary spectrum closed under conjugation, $\{\pm i\sigma_1,\dots,\pm i\sigma_m,0,\dots\}$, $\sigma_j>0$; thus $L_k^\top L_k=-L_k^2$ has

eigenvalues σ_j^2 each with multiplicity at least two, and $\operatorname{rank}(L_k) = 2m$. If $A_k = U_k V_k^\top$ with $U_k, V_k \in \mathbb{R}^{d\times r}$, then $\operatorname{range}(L_k) \subseteq \operatorname{span}(U_k, V_k)$, of dimension at most $2r$. Finally $\operatorname{rank}(\sum_k L_k^\top L_k) \le \sum_k \operatorname{rank}(L_k^\top L_k) = \sum_k \operatorname{rank}(L_k)$. (c) If $L_k \neq 0$, then $\|L_k\|_F^2 = \operatorname{Tr}(L_k^\top L_k) = 2\sum_{j=1}^m \sigma_j^2 \ge 2\sigma_{\max}^2 = 2\|L_k\|_{\mathrm{op}}^2$, with equality iff $m = 1$; the zero case is immediate. Then $\|\bar\Delta_G\|_{\mathrm{op}} \le \frac{1}{K}\sum_k \|L_k\|_{\mathrm{op}}^2 \le \frac{1}{2}$. (d) Centered prototype preservation gives $\mathcal{M} \subseteq \bigcap_k \ker L_k = \ker \bar\Delta_G$ by (a), so $\bar\Delta_G P_\parallel = 0$ and, by symmetry, $P_\parallel \bar\Delta_G = 0$; expanding $\bar\Delta_G = (P_\perp + P_\parallel)\bar\Delta_G(P_\perp + P_\parallel)$ leaves $P_\perp \bar\Delta_G P_\perp$. Since $\operatorname{range}(\Sigma_B) = \mathcal{M}$, $\bar\Delta_G \Sigma_B = 0$, and transposing gives $\Sigma_B \bar\Delta_G = 0$. Finally $\operatorname{range}(\bar\Delta_G) = (\ker \bar\Delta_G)^\perp \subseteq \mathcal{M}^\perp$, i.e. $\Pi_G \preceq P_\perp$. The raw case is identical with $\mathcal{P}$ in place of $\mathcal{M}$. □

Lemma 2 (Trace identities)**.**

$$\frac{1}{NK}\sum_{i,k}\|L_k\delta_i\|^2 = \operatorname{Tr}(\bar\Delta_G\Sigma_W) = 2\,\mathcal{R}_{\mathrm{SP}}, \qquad \frac{1}{CK}\sum_{k,c}\|L_k\mu_c\|^2 = \operatorname{Tr}(\bar\Delta_G\Sigma_P).$$

Hence $K^{-1}J(\{L_k\}) = \operatorname{Tr}\big(\bar\Delta_G(\Sigma_W - \alpha\Sigma_P)\big) - K^{-1}\rho_{\mathrm{div}}\sum_{k\neq\ell}\langle L_k, L_\ell\rangle_F^2 - K^{-1}\rho_{\mathrm{norm}}\sum_k(\|L_k\|_F^2 - 1)^2$.

Proof. $\frac{1}{NK}\sum_{i,k}\|L_k\delta_i\|^2 = \frac{1}{N}\sum_i \delta_i^\top\big(\frac{1}{K}\sum_k L_k^\top L_k\big)\delta_i = \operatorname{Tr}\big(\bar\Delta_G\,\frac{1}{N}\sum_i \delta_i\delta_i^\top\big)$. The second identity is the same computation with μ_c in place of δ_i and $1/C$ in place of $1/N$. □

Lemma 3 (Ky Fan)**.** *Let S be symmetric with eigenvalues $s_1 \ge \cdots \ge s_d$.*

(a) If $V \in \mathbb{R}^{d\times r}$ has orthonormal columns, $\operatorname{Tr}(V^\top S V) \le \sum_{j=1}^r s_j$.

(b) If $0 \preceq M \preceq I$ and $\operatorname{rank}(M) \le r$, then $\operatorname{Tr}(MS) \le \sum_{j=1}^r \max(s_j, 0)$; for $S \succeq 0$, $\operatorname{Tr}(MS) \le \sum_{j=1}^r s_j$, with equality when M is the projector onto a top-r eigenspace of S.

The same statements hold for a self-adjoint operator on any finite-dimensional real inner-product space.

Proof. (a) By Cauchy interlacing, $\lambda_i(V^\top S V) \le s_i$ for $i = 1, \dots, r$; sum over i. (b) Write $M = \sum_{i=1}^r \gamma_i v_i v_i^\top$ with $\gamma_i \in [0,1]$ and $\{v_i\}$ orthonormal. Then $\operatorname{Tr}(MS) = \sum_i \gamma_i\, v_i^\top S v_i \le \sum_{i\in I^+} v_i^\top S v_i$ where $I^+ = \{i : v_i^\top S v_i > 0\}$, $|I^+| = r' \le r$; by (a) this is at most $\sum_{j\le r'} s_j \le \sum_{j\le r}\max(s_j, 0)$. The equality case is direct. Only the spectral theorem and interlacing were used, so the abstract version follows. □

A.2 Proof of Theorem 1 and Corollary 1

Unconstrained features: each $h_i \in \mathbb{R}^d$ is a free variable (Mixon et al., 2022; Zhu et al., 2021), prototypes μ_c are held fixed, and $\bar\Delta_G$ is fixed. From Lemma 2, $\mathcal{R}_{\mathrm{SP}} = \frac{1}{2N}\sum_i \delta_i^\top \bar\Delta_G \delta_i$, so $\partial\mathcal{R}_{\mathrm{SP}}/\partial h_i = \frac{1}{N}\bar\Delta_G\delta_i$. We use the mean-normalized gradient-flow time $\dot h_i = -\lambda N\,\partial\mathcal{R}_{\mathrm{SP}}/\partial h_i$, so that the per-sample gradient is $\lambda\bar\Delta_G\delta_i$; since $\dot\delta_i = \dot h_i$,

$$\dot\delta_i = -\lambda\bar\Delta_G\delta_i. \tag{6}$$

Proof of Theorem 1. Let $\bar\Delta_G = \sum_j \gamma_j v_j v_j^\top$ with $\gamma_j \ge 0$ and $\{v_j\}$ orthonormal, and set $z_j(t) = \langle v_j, \delta(t)\rangle$ for a fixed residual $\delta = \delta_i$. Projecting Eq. 6 onto v_j gives $\dot z_j = -\lambda\gamma_j z_j$, so $z_j(t) = e^{-\lambda\gamma_j t}z_j(0)$, i.e. $\delta(t) = e^{-\lambda t\bar\Delta_G}\delta(0)$. Coordinates with $\gamma_j = 0$ are constant, which is the kernel statement. Writing $\delta_\perp = \Pi_G\delta$,

$$\|\delta_\perp(t)\|^2 = \sum_{\gamma_j>0} e^{-2\lambda\gamma_j t}z_j(0)^2 \le e^{-2\lambda\lambda_{\min}^+ t}\sum_{\gamma_j>0} z_j(0)^2 = e^{-2\lambda\lambda_{\min}^+ t}\|\delta_\perp(0)\|^2.$$

For the last sentence of the theorem: since $\mathcal{M} \subseteq \ker(\bar\Delta_G)$, $\operatorname{range}(\bar\Delta_G) = \ker(\bar\Delta_G)^\perp \subseteq \mathcal{M}^\perp$, so the contracted modes are task-transverse; prototypes are fixed, so Σ_B is constant. □

Proof of Corollary 1. $\partial\mathcal{L}_{\mathrm{coll}}/\partial h_i = \frac{1}{N}\delta_i$, so the flow becomes $\dot\delta = -(\beta_{\mathrm{coll}}I + \lambda\bar\Delta_G)\delta$. The operator $\beta_{\mathrm{coll}}I + \lambda\bar\Delta_G$ shares the eigenbasis $\{v_j\}$ with eigenvalues $\beta_{\mathrm{coll}} + \lambda\gamma_j$; repeat the previous argument with these rates. □

Remark 1 (Empirical means). If μ_c is the empirical class mean of a fixed labeled sample set rather than a fixed parameter, the theorem still holds verbatim for the isolated fixed-operator SP flow. Indeed, with $\delta_i = h_i - \frac{1}{n_{y_i}} \sum_{j:y_j=y_i} h_j$, one has $\sum_{i:y_i=c} \delta_i = 0$, and differentiating $\mathcal{R}_{\mathrm{SP}}$ through the mean gives

$$\frac{\partial \mathcal{R}_{\mathrm{SP}}}{\partial h_m} = \frac{1}{N}\Big[\bar{\Delta}_G \delta_m - \frac{1}{n_{y_m}} \sum_{i:y_i=y_m} \bar{\Delta}_G \delta_i\Big] = \frac{1}{N} \bar{\Delta}_G \delta_m,$$

so Eq. 6 is unchanged; moreover $\dot{\mu}_c = -\lambda \bar{\Delta}_G \frac{1}{n_c} \sum_{i:y_i=c} \delta_i = 0$, so the flow preserves class means. The "prototypes held fixed" hypothesis is therefore automatic in this case. This statement does not assert that means remain fixed across changing minibatches or under additional CE updates. Theorem 2(ii) below gives the corresponding zero-class-mean SP force only in the fixed-prototype free-feature block; in a parameterized network, the tangent kernel can couple that force to prototype motion as noted above.

Remark 2 (Scatter dynamics). With $E_t = e^{-\lambda t \bar{\Delta}_G}$, Eq. 6 gives $\Sigma_W(t) = E_t \Sigma_W(0) E_t$, hence

$$\operatorname{Tr} \Sigma_W(t) = \sum_j e^{-2\lambda \gamma_j t}\, v_j^\top \Sigma_W(0) v_j, \qquad \frac{d}{dt} \operatorname{Tr} \Sigma_W(t) = -2\lambda \operatorname{Tr}\big(\bar{\Delta}_G \Sigma_W(t)\big), \qquad \lim_{t\to\infty} \operatorname{Tr} \Sigma_W(t) = \operatorname{Tr}\big(P_0 \Sigma_W(0)\big).$$

The second identity uses that E_t commutes with $\bar{\Delta}_G$. These are the ingredients of Proposition 1.

Remark 3 (Parameterized features). In the mean-normalized time convention of Theorem 1, with $h_i = \Phi_\theta(x_i)$ and gradient flow on θ, $\dot{h}_i = -\lambda \sum_j \Theta_{ij}(t)\, \bar{\Delta}_G \delta_j$ where $\Theta_{ij} = \partial_\theta h_i\, (\partial_\theta h_j)^\top \in \mathbb{R}^{d\times d}$ is the (time-dependent) tangent kernel block. Theorem 1 is the identity-kernel case $\Theta_{ij} = \delta_{ij} I$; for a general kernel the residuals are coupled across samples and no per-sample contraction rate follows without further assumptions.

A.3 Proof of Proposition 1

Proof. Step 1 (reduction operator). Under either dynamics, the residuals after the horizon are $\delta_i(T) = Q\delta_i(0)$ for a single matrix Q that is a polynomial or exponential in $\bar{\Delta}_G$: $Q = e^{-\lambda T \bar{\Delta}_G}$ for the flow, or $Q = \prod_t (I - s_t \bar{\Delta}_G)$ for discrete steps. In the eigenbasis of $\bar{\Delta}_G$, Q has eigenvalues $q_j = e^{-\lambda T \gamma_j}$ or $q_j = \prod_t (1 - s_t \gamma_j)$; in both cases $q_j \in [0,1]$ (for discrete steps because $s_t \gamma_j \le s_t \|\bar{\Delta}_G\|_{\mathrm{op}} \le 1$) and $q_j = 1$ whenever $\gamma_j = 0$. Then $\Sigma_W(T) = Q \Sigma_W(0) Q$ and

$$\Delta[\operatorname{Tr} \Sigma_W]_{\bar{\Delta}_G} = \operatorname{Tr} \Sigma_W(0) - \operatorname{Tr}\big(Q \Sigma_W(0) Q\big) = \operatorname{Tr}\big((I - Q^2)\Sigma_W(0)\big).$$

The operator $I - Q^2$ has eigenvalues $1 - q_j^2 \in [0,1]$ on $\operatorname{range}(\bar{\Delta}_G)$ and 0 on $\ker(\bar{\Delta}_G)$, so $0 \preceq I - Q^2 \preceq \Pi_G$ and therefore $\Delta[\operatorname{Tr} \Sigma_W]_{\bar{\Delta}_G} \le \operatorname{Tr}(\Pi_G \Sigma_W(0))$, since $\Sigma_W(0) \succeq 0$. For the flow, $q_j \to 0$ on $\operatorname{range}(\bar{\Delta}_G)$ as $T \to \infty$, so $I - Q^2 \to \Pi_G$ and the inequality becomes an equality. The rate identity is Remark 2.

Step 2 (Ky Fan on the task-transverse scatter). By Lemma 1(d), $\Pi_G \preceq P_\perp$, and since both are orthogonal projectors this means $\Pi_G = P_\perp \Pi_G P_\perp$. Hence $\operatorname{Tr}(\Pi_G \Sigma_W) = \operatorname{Tr}(P_\perp \Pi_G P_\perp \Sigma_W) = \operatorname{Tr}(\Pi_G\, P_\perp \Sigma_W P_\perp)$. Now Π_G is an orthogonal projector of rank $\operatorname{rank}(\bar{\Delta}_G) \le r_G$ and $P_\perp \Sigma_W P_\perp \succeq 0$ has eigenvalues $\lambda_1^\perp \ge \lambda_2^\perp \ge \cdots$, so Lemma 3(b) gives $\operatorname{Tr}(\Pi_G \Sigma_W) \le \sum_{j=1}^{r_G} \lambda_j^\perp$.

Step 3 (normalization). $\operatorname{Tr}(P_\perp \Sigma_W P_\perp) = \operatorname{Tr}(P_\perp \Sigma_W)$ and, when both normalizing traces are positive, the definitions of m_{r_G} and $r_\perp$ give $\sum_{j \le r_G} \lambda_j^\perp = m_{r_G} \operatorname{Tr}(P_\perp \Sigma_W) = m_{r_G}\, r_\perp\, \operatorname{Tr} \Sigma_W$. If either trace vanishes, positive semidefiniteness makes the corresponding removable scatter zero, agreeing with the convention in the proposition. □

Remark 4 (Tightness and rank budget). One sufficient way to attain equality throughout Eq. 4 is infinite-horizon contraction with Π_G the projector onto a top-r_G eigenspace of $P_\perp \Sigma_W P_\perp$. These conditions are not necessary: finite discrete steps can annihilate supported modes, incomplete modes may carry zero scatter, and spectral degeneracies permit multiple maximizing supports. In the $\alpha \to \infty$ limit, the raw-mean $K = 1$ inner optimum characterized by Remark 6 instead selects a restricted optimum in the possibly smaller space $\mathcal{P}^\perp$; under its spectral-gap condition this is the leading plane and attains the centered bound only when it is also a top plane of $P_\perp \Sigma_W P_\perp$. The rank budget r_G is a condition on the realized operator. Lemma 1(b) gives $r_G \le 2rK$ for a genuinely rank-r factorization and $r_G \le 2K$ for the one-plane control, but neither is a hard bound for the dense canonical parameterization. No spectral cap on $\bar{\Delta}_G$ is used in Steps 1–3; $\|\bar{\Delta}_G\|_{\mathrm{op}}$ enters only through the step-size condition $s_t \le 1/\|\bar{\Delta}_G\|_{\mathrm{op}}$, which Lemma 1(c) guarantees for $s_t \le 2$. The proposition is a statement about a *frozen* support; a changing online sequence is not bounded by any one frozen support, although each individual realized operator remains subject to the proposition.

Remark 5 (Rank in practice). With the standard skew-symmetric parameterization $L_k = A_k - A_k^\top$, each L_k has even rank at most $2\lfloor d/2 \rfloor$, so an unrestricted K-generator family can in principle have full support on $P_\perp$. In our

canonical implementation each $A_k \in \mathbb{R}^{d\times d}$ is dense; there is no default rank-4 factorization. Consequently, applying Proposition 1 to a realized dense operator uses its measured support rank rather than a formal $2K$ bound. A participation-ratio effective rank is a descriptive approximation, not the hard support-rank parameter of Proposition 1; a rigorous lower-rank approximation additionally requires an explicit spectral truncation and tail term. A separate one-plane-per-generator control $L_k = a_k b_k^\top - b_k a_k^\top$ enforces $\mathrm{rank}(L_k) = 2$ by construction and gives an exact m_{2K} bound. We use m_{2K} only as an idealized one-plane capture diagnostic for the dense method, not as a theorem about its finite-step rank. At a unique exact inner optimum with $K = 1$ and $s_2 > s_3$, the operator has rank two (Proposition 2(i)); the dense parameterization matters for the inexact finite-step iterate. Theorem 2(iii) controls its distance from the exact optimum only under the stated local eigengap, drift, and angle assumptions.

A.4 Proof of Proposition 2

Let $S = S_\alpha = \Sigma_W - \alpha\Sigma_P = \sum_a s_a w_a w_a^\top$ with $s_1 \geq \cdots \geq s_d$ and $\{w_a\}$ orthonormal. For $a < b$ define $E_{ab} = (w_a w_b^\top - w_b w_a^\top)/\sqrt{2}$; the family $\{E_{ab}\}_{a<b}$ is a Frobenius-orthonormal basis of $\mathfrak{so}(d)$.

Lemma 4. *The map $\mathcal{S} : \mathfrak{so}(d) \to \mathfrak{so}(d)$, $\mathcal{S}(L) = \frac{1}{2}(LS + SL)$, is well defined and self-adjoint for $\langle\cdot,\cdot\rangle_F$, satisfies $\langle L, \mathcal{S}L\rangle_F = \mathrm{Tr}(L^\top LS)$, and has $\mathcal{S}E_{ab} = \frac{1}{2}(s_a + s_b)E_{ab}$.*

Proof. $(LS+SL)^\top = SL^\top + L^\top S = -(SL+LS)$, so $\mathcal{S}L \in \mathfrak{so}(d)$. Self-adjointness: $\langle L', \mathcal{S}L\rangle_F = \frac{1}{2}\mathrm{Tr}(L'^\top LS) + \frac{1}{2}\mathrm{Tr}(L'^\top SL)$, and by cyclicity and transposition each term is symmetric under $L \leftrightarrow L'$. For the quadratic form, $\mathrm{Tr}(L^\top SL) = \mathrm{Tr}(LL^\top S) = \mathrm{Tr}(L^\top LS)$ because $LL^\top = L^\top L = -L^2$ for skew L, hence $\langle L, \mathcal{S}L\rangle_F = \mathrm{Tr}(L^\top LS)$. Finally $E_{ab}S = \frac{1}{\sqrt{2}}(s_b\, w_a w_b^\top - s_a\, w_b w_a^\top)$ and $SE_{ab} = \frac{1}{\sqrt{2}}(s_a\, w_a w_b^\top - s_b\, w_b w_a^\top)$; adding gives $E_{ab}S + SE_{ab} = \frac{1}{\sqrt{2}}(s_a + s_b)(w_a w_b^\top - w_b w_a^\top) = (s_a + s_b)E_{ab}$, so $\mathcal{S}E_{ab} = \frac{1}{2}(s_a + s_b)E_{ab}$. □

Proof of Proposition 2. (i) For $K = 1$ the diversity term is absent and, by Lemma 2, $J(L) = \mathrm{Tr}(L^\top LS) = \langle L, \mathcal{S}L\rangle_F$. By Lemma 4, $\mathcal{S}$ is self-adjoint on $\mathfrak{so}(d)$ with eigenvectors E_{ab} and eigenvalues $\frac{1}{2}(s_a + s_b)$, the largest of which is $\frac{1}{2}(s_1 + s_2)$ at E_{12}. The Rayleigh principle on the unit Frobenius sphere gives the maximum and the displayed maximizers $\pm E_{12}$. For $d \geq 3$, these are the only maximizers iff the top eigenvalue is simple, equivalently $s_1 + s_2 > s_1 + s_3$, i.e. $s_2 > s_3$; for $d = 2$ the unit skew sphere consists of this sign pair. Finally $\bar{\Delta}_G^\star = E_{12}^\top E_{12} = -E_{12}^2 = -\frac{1}{2}(w_1 w_2^\top - w_2 w_1^\top)^2 = \frac{1}{2}(w_1 w_1^\top + w_2 w_2^\top)$.

(ii) With Frobenius-orthonormal $L_1, \ldots, L_K$, $\frac{1}{K}\sum_k \mathrm{Tr}(L_k^\top L_k S) = \frac{1}{K}\sum_k \langle L_k, \mathcal{S}L_k\rangle_F$, and Lemma 3(a) applied to the self-adjoint operator $\mathcal{S}$ on the inner-product space $\mathfrak{so}(d)$ bounds the sum by the K largest eigenvalues of $\mathcal{S}$, namely $\frac{1}{2}\sigma_1, \ldots, \frac{1}{2}\sigma_K$; equality holds at the corresponding eigenvectors $E_{a_k b_k}$. If instead $\rho_{\mathrm{div}} = 0$, each summand is at most the top eigenvalue $\frac{1}{2}(s_1 + s_2)$, attained by $L_k = L^\star$ for every k.

(iii) Let $u_1, u_2 \in \mathcal{P}^\perp$ be orthonormal eigenvectors of the operator $\Sigma_{P\perp}$ restricted to $\mathcal{P}^\perp$, for $\lambda_1^{P,\perp}, \lambda_2^{P,\perp}$; they lie in $\mathcal{P}^\perp$, which has dimension at least two by assumption. Put $L_\perp = (u_1 u_2^\top - u_2 u_1^\top)/\sqrt{2}$, a unit-Frobenius skew matrix. Since $u_1, u_2 \perp \mathcal{P}$, $L_\perp \mu_c = 0$ for all c, so $\mathrm{Tr}(L_\perp^\top L_\perp \Sigma_P) = 0$ and

$$J(L_\perp) = \mathrm{Tr}(L_\perp^\top L_\perp \Sigma_W) = \tfrac{1}{2}\big(u_1^\top \Sigma_W u_1 + u_2^\top \Sigma_W u_2\big) = \tfrac{1}{2}(\lambda_1^{P,\perp} + \lambda_2^{P,\perp}),$$

using $u_i^\top \Sigma_W u_i = u_i^\top \Sigma_{P\perp} u_i = \lambda_i^{P,\perp}$. Optimality of $L^\star$ and Lemma 2 give

$$\mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_W) - \alpha\,\mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_P) = J(L^\star) \geq J(L_\perp) = \tfrac{1}{2}(\lambda_1^{P,\perp} + \lambda_2^{P,\perp}).$$

On the other hand, applying the one-generator Rayleigh bound from (i) with $S = \Sigma_W$ gives, for every unit-Frobenius skew matrix and hence for $L^\star$, $\mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_W) \leq \frac{1}{2}(\lambda_1 + \lambda_2)$; this does not require $\bar{\Delta}_G^\star$ to have rank two under spectral degeneracy. Combining, $\alpha\,\mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_P) \leq \frac{1}{2}\big[(\lambda_1 + \lambda_2) - (\lambda_1^{P,\perp} + \lambda_2^{P,\perp})\big]$, which is Eq. 5 after dividing by α and applying Lemma 2 to the left-hand side. The numerator is nonnegative by Lemma 3(a), applied to the two-dimensional subspace $\mathrm{span}\{u_1, u_2\} \subseteq \mathcal{P}^\perp$. If $\lambda_1 + \lambda_2 = \lambda_1^{P,\perp} + \lambda_2^{P,\perp}$ (as when the two leading eigenvectors of Σ_W lie in $\mathcal{P}^\perp$), the right-hand side is zero, and $\mathrm{Tr}(\bar{\Delta}_G^\star \Sigma_P) = 0$ forces $L^\star \mu_c = 0$ for every c. □

Remark 6 (The $\alpha \to \infty$ limit). Assume $\dim \mathcal{P}^\perp \geq 2$. For each $\alpha > 0$, let $L_\alpha^\star$ globally maximize the fixed-representation one-generator objective, and set $\bar{\Delta}_\alpha^\star = L_\alpha^{\star\top} L_\alpha^\star$. The unit Frobenius sphere in $\mathfrak{so}(d)$ is compact. Hence every sequence $\alpha_n \to \infty$ has a subsequence such that $L_{\alpha_n}^\star \to L^\infty$ and $\bar{\Delta}_{\alpha_n}^\star \to \bar{\Delta}^\infty = L^{\infty\top} L^\infty$. Proposition 2(iii) and the trace identity give $0 = \mathrm{Tr}(\bar{\Delta}^\infty \Sigma_P) = \frac{1}{C}\sum_c \|L^\infty \mu_c\|^2$. Thus $\mathcal{P} \subseteq \ker L^\infty$, and skew symmetry implies $\mathrm{range}(\bar{\Delta}^\infty) \subseteq \mathcal{P}^\perp$. Put $B = \frac{1}{2}(\lambda_1^{P,\perp} + \lambda_2^{P,\perp})$. The raw-prototype-null comparator from Proposition 2(iii) has

objective value B, so optimality and nonnegativity of the prototype penalty give $\mathrm{Tr}(\bar{\Delta}_\alpha^\star \Sigma_W) \geq B$; the limit therefore has value at least B. Conversely, the one-generator Rayleigh characterization in Proposition 2(i), applied to the covariance restricted to $\mathcal{P}^\perp$, bounds its value by B. Hence every accumulation point solves the raw-prototype-null restricted problem. If $\dim \mathcal{P}^\perp = 2$, the restricted Laplacian is unique. If $\dim \mathcal{P}^\perp \geq 3$ and $\lambda_2^{P,\perp} > \lambda_3^{P,\perp}$, the leading skew-space eigenvalue is simple and, for restricted leading eigenvectors u_1, u_2, $\bar{\Delta}^\infty = \frac{1}{2}(u_1 u_1^\top + u_2 u_2^\top) =: \bar{\Delta}_\infty^\star$; every accumulation point then agrees, so $\bar{\Delta}_\alpha^\star \to \bar{\Delta}_\infty^\star$. Without the gap, accumulation points remain restricted optima but need not be unique or rank two. A quantitative rate for this convergence is Corollary 2. Since $\mathcal{P}^\perp \subseteq \mathcal{M}^\perp$, this operator obeys the centered applicability bound but need not attain it: raw preservation additionally annihilates $\bar{\mu}$. It is related to, but generally distinct from, the centered frozen PCA-SP control.

A.5 Coupled dynamics of Algorithm 1

This section gives the full statement and proof of Theorem 2 (Section 3.6) for the $K = 1$ configuration used in the four headline dataset results. It separates three levels of analysis: the implemented raw-parameter inner update, an exact best-response identity in a fixed-prototype free-feature block, and a local tracking bound for a moving representation. It does not claim global convergence of the minibatched parameterized network optimizer.

Notation. With $K = 1$ the diversity term is absent, and because the unit-Frobenius projection is applied after every step the ρ_{norm} term has zero gradient on the constraint set, so the inner player ascends $q_\alpha(L) = \mathrm{Tr}(L^\top L S_\alpha) = \langle L, \mathcal{S}_\alpha L \rangle_F$ on the unit sphere of $\mathfrak{so}(d)$, where $\mathcal{S}_\alpha$ is the operator of Lemma 4 for $S = S_\alpha$. Let $\sigma_P > 0$ be the least positive eigenvalue of Σ_P on $\mathcal{P} = \mathrm{range}(\Sigma_P)$. Write $h_i = \mu_{y_i} + \delta_i$ and call $\delta = (\delta_i)_i$, subject to $\sum_{i:y_i=c} \delta_i = 0$, the *within-class block* and $\mu = (\mu_c)_c$ the *prototype block*. Define the value of the inner problem and the outer objective

$$\Phi_\alpha(\delta; \mu) := \max_{L \in \mathfrak{so}(d),\ \|L\|_F = 1} \mathrm{Tr}\left(L^\top L\, S_\alpha\right), \qquad F_\alpha(\delta; \mu) := \mathcal{L}_{\mathrm{CE}} + \tfrac{\lambda}{2}\, \Phi_\alpha(\delta; \mu). \tag{7}$$

The factor $\frac{1}{2}$ matches $\mathcal{R}_{\mathrm{SP}} = \frac{1}{2}\mathrm{Tr}(\bar{\Delta}_G \Sigma_W)$ (Lemma 2). The function Φ_α is one half of the sum of the two largest eigenvalues of S_α and is convex in S_α; it is not generally a Ky Fan norm because S_α can be indefinite.

Lemma 5 (Perturbation of the top eigenvector)**.** *Let $\mathcal{T}, \mathcal{T}'$ be self-adjoint operators on a finite-dimensional inner-product space with $g = \lambda_1(\mathcal{T}) - \lambda_2(\mathcal{T}) > 0$ and $\|\mathcal{T}' - \mathcal{T}\|_{\mathrm{op}} \leq \varepsilon \leq g/2$. Then the top eigenvectors satisfy $\sin \angle(u, u') \leq 2\varepsilon/g$. Let $\mathcal{A} = I + c\mathcal{T}$ have eigenvalues $\gamma_1, \gamma_2, \ldots$ in a common eigenbasis and assume $|\gamma_1| > \Gamma_{\mathrm{sub}} := \max_{j \geq 2} |\gamma_j|$. Power iteration $L \leftarrow \mathcal{A}L/\|\mathcal{A}L\|$ contracts $\tan \angle(L, u)$ by $\rho := \Gamma_{\mathrm{sub}}/|\gamma_1| < 1$ per step whenever L has a nonzero component along u. For unit-norm L, L' with signs aligned,*

$$\|L^\top L - L'^\top L'\|_{\mathrm{op}} \leq 2\|L - L'\|_F \leq 2\sqrt{2}\, \sin \angle(L, L').$$

Proof. The first claim is the Davis–Kahan $\sin\theta$ theorem. For the second, expand in the eigenbasis of $\mathcal{A}$: the leading coordinate is multiplied by γ_1 and the norm of all remaining coordinates by at most Γ_{sub}, giving the tangent ratio. For the last display, $L^\top L - L'^\top L' = L^\top (L - L') + (L - L')^\top L'$ and unit Frobenius norms give the first inequality. Sign alignment gives $\|L - L'\|_F = 2\sin(\angle(L, L')/2) \leq \sqrt{2} \sin \angle(L, L')$ for angles at most $\pi/2$. □

Theorem 2 (full statement). *Let $K = 1$, $d \geq 3$, $\dim \mathcal{P}^\perp \geq 2$, and $s_2 > s_3$.*

(i) ***Inner problem and implemented ascent.*** *$\Phi_\alpha = \frac{1}{2}(s_1 + s_2)$. Once hard projection has placed $L = A - A^\top$ on the unit Frobenius sphere, an ascent step in the raw parameter A induces $L^+ = L + 8\eta_G \mathcal{S}_\alpha(L)$ before renormalization. Thus the subsequent inner steps are power iteration on $I + 8\eta_G \mathcal{S}_\alpha$. If*

$$\Gamma_\star := |1 + 4\eta_G(s_1 + s_2)| \ > \ \Gamma_{\mathrm{sub}} := \max_{(a,b) \neq (1,2)} |1 + 4\eta_G(s_a + s_b)|, \tag{8}$$

and the projected initialization has a nonzero E_{12} component, the induced Laplacian converges to $\bar{\Delta}_\alpha^\star = \frac{1}{2}(w_1 w_1^\top + w_2 w_2^\top)$ of Proposition 2(i). The first update from a non-unit random initialization can also contain the norm-penalty gradient; the power iteration statement applies after its first hard projection.

(ii) ***Conditional outer step.*** *With μ fixed and full-batch free features, $\partial_{\delta_i}\left[\frac{\lambda}{2}\Phi_\alpha\right] = \frac{\lambda}{N}\bar{\Delta}_\alpha^\star \delta_i$ and $\sum_{i:y_i=c} \bar{\Delta}_\alpha^\star \delta_i = 0$. Hence the SP component is the exact Danskin subgradient of $\frac{\lambda}{2}\Phi_\alpha(\cdot; \mu)$ on the class-centered within-class block and moves no class mean in these free-feature coordinates. This does not identify the parameter-space network update with free-feature subgradient descent.*

(iii) ***Local tracking.*** *Let $g_t = \frac{1}{2}(s_2 - s_3)(t) > 0$ be the leading gap of $\mathcal{S}_{\alpha,t}$, let $\rho_t = \Gamma_{\mathrm{sub},t}/\Gamma_{\star,t} < 1$ be defined from the corrected iteration matrix in (i), and suppose $\sqrt{2}\rho_t^S < 1$ and $\|\mathcal{S}_{\alpha,t+1} - \mathcal{S}_{\alpha,t}\|_{\mathrm{op}} \le \varepsilon_t \le g_{t+1}/2$. Then $e_t := \sin\angle(L_t, L_t^\star)$ after the S warm-started steps at outer iteration t satisfies, provided $e_t + 2\varepsilon_t/g_{t+1} \le 1/\sqrt{2}$,*

$$e_{t+1} \le \sqrt{2}\,\rho_{t+1}^S\big(e_t + 2\varepsilon_t/g_{t+1}\big), \qquad \|\bar\Delta_{G,t} - \bar\Delta_{\alpha,t}^\star\|_{\mathrm{op}} \le 2\sqrt{2}\,e_t,$$

so the implemented outer update differs from the exact subgradient of F_α by at most $2\sqrt{2}\lambda e_t\|\delta_i\|/N$ for sample i.

(iv) ***Invariance.*** *If $WP_{P,\perp} = 0$, then for every orthogonal R fixing $\mathcal{P}$ pointwise the map $h_i \mapsto Rh_i$ leaves $\mathcal{L}_{\mathrm{CE}}$, μ_c, Σ_P, $\operatorname{Tr}\Sigma_W$ and Φ_α invariant. If a* frozen *operator additionally satisfies $\bar\Delta_0 = P_{P,\perp}\bar\Delta_0 P_{P,\perp}$, has rank r_0, and $\operatorname{rank}\Sigma_{P\perp} \le \dim\mathcal{P}^\perp - r_0$, then $\inf_R \mathcal{R}_{\mathrm{SP}}^{\bar\Delta_0}(Rh) = 0$: its penalty can vanish on the orbit while the CE value and transverse covariance spectrum remain unchanged; F_α itself is constant on that orbit.*

Proof. *(i)* $\Phi_\alpha = \frac{1}{2}(s_1 + s_2)$ and the form of $\bar\Delta_\alpha^\star$ are Proposition 2(i). Let $q_\alpha(L) = \operatorname{Tr}(L^\top L S_\alpha)$ and $L = A - A^\top$. Direct differentiation gives $\nabla_A q_\alpha = 2(LS_\alpha + S_\alpha L) = 4\mathcal{S}_\alpha(L)$. Because this gradient is skew, updating A by $\eta_G\nabla_A q_\alpha$ changes $L = A - A^\top$ by twice that increment: $L^+ = L + 8\eta_G\mathcal{S}_\alpha(L)$ before normalization. By Lemma 4, the eigenvalue of the iteration matrix on E_{ab} is $1 + 4\eta_G(s_a + s_b)$. Condition equation 8 and a nonzero leading component therefore give power-method convergence to $\pm E_{12}$ by Lemma 5; the sign is invisible in $L^\top L$. On the unit sphere the norm-penalty derivative vanishes. The first random, non-unit step is excluded exactly as stated.

(ii) With μ fixed, Σ_P and hence the penalty in Eq. 2 are constant, and S_α depends on δ only through Σ_W. Φ_α is the maximum over the compact unit sphere of functions $\delta \mapsto \operatorname{Tr}(L^\top L\Sigma_W(\delta))$ that are C^1 in δ, so Danskin's theorem gives $\partial_\delta\Phi_\alpha = \operatorname{conv}\{\partial_\delta \operatorname{Tr}(L^{\star\top}L^\star\Sigma_W)\}$ over maximizers, and $\partial_{\delta_i}\operatorname{Tr}(\bar\Delta\Sigma_W) = \frac{2}{N}\bar\Delta\delta_i$ at fixed $\bar\Delta$. Differentiating $\mathcal{R}_{\mathrm{SP}}$ through the empirical means produces the extra term $-\frac{1}{N}\bar\Delta\frac{1}{n_c}\sum_{i\in c}\delta_i = 0$ (Remark 1), so with an exact inner best response in this fixed-prototype free-feature block, the outer SP gradient coincides with the Danskin subgradient. Moreover, $\sum_{i\in c}\bar\Delta_\alpha^\star\delta_i = \bar\Delta_\alpha^\star\sum_{i\in c}\delta_i = 0$. The tangent gradient of CE on the constrained within-class block is its classwise centered feature gradient; the assertion here concerns the SP component and does not replace the parameterized-feature qualification above. A finite-step inner iterate gives the gradient of its selected residual penalty, without an automatic Danskin identity.

(iii) At outer iteration $t+1$ the drift moves the top eigenvector of $\mathcal{S}_\alpha$ by at most $2\varepsilon_t/g_{t+1}$ (Lemma 5). The sine of the angle between the warm start and the new optimum is at most $e_t + 2\varepsilon_t/g_{t+1}$. In the stated $\pi/4$ basin, $\tan\theta \le \sqrt{2}\sin\theta$; S corrected power steps contract the tangent by ρ_{t+1}^S, which proves the recurrence. Aligning signs and applying the final bound of Lemma 5 gives the Laplacian bound. The per-sample outer bias is then at most $2\sqrt{2}\lambda e_t\|\delta_i\|/N$.

(iv) $Wh_i = WP_{P,\|}h_i + WP_{P,\perp}h_i = WP_{P,\|}h_i$ is unchanged by R (which fixes $\mathcal{P}$), so $\mathcal{L}_{\mathrm{CE}}$ and μ_c are invariant, hence Σ_P; $\Sigma_W \mapsto R\Sigma_W R^\top$, so $\operatorname{Tr}\Sigma_W$ and the spectrum of $S_\alpha \mapsto RS_\alpha R^\top$ are invariant, hence Φ_α. For the frozen operator, its assumed support in $\mathcal{P}^\perp$ removes the prototype and cross blocks from the trace. Moreover, $\ker\bar\Delta_0 \cap \mathcal{P}^\perp$ has dimension at least $\dim\mathcal{P}^\perp - r_0 \ge \operatorname{rank}\Sigma_{P\perp}$, so some R maps the support of $\Sigma_{P\perp}$ into it and $\operatorname{Tr}(\bar\Delta_0 R\Sigma_W R^\top) = 0$. The rotation preserves the existing CE value; no claim that this value is the global CE infimum is needed. □

Remark 7 (Conditional objective and prototype block). The exact μ-gradient of Φ_α is $-\frac{2\alpha}{C}\bar\Delta_\alpha^\star\mu_c$, a force that would move prototypes *into* the adversary's plane to lower the inner value. Algorithm 1 omits it by construction: the outer player pays $\frac{1}{2}\operatorname{Tr}(\bar\Delta_G\Sigma_W)$ but not the prototype penalty. The game is therefore non-zero-sum and has no single potential; the fixed-prototype within-class block nevertheless has the exact subgradient identity in part (ii). By Corollary 2(iii), the omitted gradient of $\frac{\lambda}{2}\Phi_\alpha$ has norm $O(\lambda\|\Sigma_\times\|\|\mu_c\|/(C\sigma_P))$, independent of α; it is the limiting multiplier force associated with $L\mu_c = 0$. In a parameterized network, neither this block decomposition nor zero prototype motion is asserted exactly.

The following corollaries describe how the implemented soft penalty relates to the idealized hard constraint $L\mu_c = 0$ and compare the resulting instantaneous contraction allocation with isotropic collapse.

Corollary 2 (α-consistency)**.** *Assume $\dim\mathcal{P}^\perp \ge 3$. Let $g = \frac{1}{2}(\lambda_2^{P,\perp} - \lambda_3^{P,\perp}) > 0$ and let $\bar\Delta_\infty^\star = \frac{1}{2}(u_1u_1^\top + u_2u_2^\top)$ be built from the top-two eigenvectors of $\Sigma_{P\perp}$ (the hard-constraint optimum of Remark 6). There is $\alpha_0(\|P_{P,\|}\Sigma_W P_{P,\|}\|, \|\Sigma_\times\|, \sigma_P, g)$ such that for $\alpha \ge \alpha_0$:*

(i) $0 \le s_j - \lambda_j^{P,\perp} \le c_1\|\Sigma_\times\|^2/(\alpha\sigma_P)$ *for* $j = 1, 2$;

(ii) $\|\bar\Delta_\alpha^\star - \bar\Delta_\infty^\star\|_{\mathrm{op}} \le c_2\big[\|\Sigma_\times\|/(\alpha\sigma_P) + \|\Sigma_\times\|^2/(\alpha\sigma_P g)\big]$;

(iii) $\|\bar{\Delta}_\alpha^\star P_{P,\|}\|_{\mathrm{op}} \le c_3\|\Sigma_\times\|/(\alpha\sigma_P)$*, hence* $\frac{1}{C}\sum_c \|L^\star\mu_c\|^2 = \mathrm{Tr}(\bar{\Delta}_\alpha^\star\Sigma_P) = O(\alpha^{-2})$*, sharpening Eq. 5.*

Proof. In the decomposition $\mathcal{P}\oplus\mathcal{P}^\perp$, $S_\alpha = \left[\begin{smallmatrix} A_\alpha & \Sigma_\times \\ \Sigma_\times^\top & \Sigma_{P\perp}\end{smallmatrix}\right]$ with $A_\alpha = P_{P,\|}\Sigma_W P_{P,\|} - \alpha\Sigma_P$. For $\alpha \ge \alpha_0$, put $a_\alpha = \alpha\sigma_P - \|P_{P,\|}\Sigma_W P_{P,\|}\| > 0$ and choose α_0 so that $a_\alpha \ge \alpha\sigma_P/2$ and $\varepsilon_\alpha := \|\Sigma_\times\|^2/a_\alpha \le g/2$. Completing the square in $x^\top A_\alpha x + 2x^\top\Sigma_\times v + v^\top\Sigma_{P\perp}v$ and applying the min–max principle gives, for $j = 1, 2$,

$$\lambda_j^{P,\perp} \le s_j \le \lambda_j^{P,\perp} + \varepsilon_\alpha,$$

which proves (i). Equivalently, for any such eigenvalue $s \ge 0$, the block eigenvalue equation gives

$$x = -(A_\alpha - sI)^{-1}\Sigma_\times v, \qquad \left[\Sigma_{P\perp} - \Sigma_\times^\top(A_\alpha - sI)^{-1}\Sigma_\times\right]v = sv.$$

Because $A_\alpha \preceq -a_\alpha I$, $\|(A_\alpha - sI)^{-1}\| \le a_\alpha^{-1}$: hence the prototype component is at most $\|\Sigma_\times\|/a_\alpha$ and the effective transverse perturbation is at most ε_α. The sin–theta bound with transverse gap $2g$ rotates the top-two transverse plane by $O(\varepsilon_\alpha/g)$. Combining this rotation with the graph component in $\mathcal{P}$ proves (ii) and (iii), using $a_\alpha \ge \alpha\sigma_P/2$. Finally, $\mathrm{Tr}(\bar{\Delta}_\alpha^\star\Sigma_P) \le \mathrm{rank}(\Sigma_P)\,\|\Sigma_P\|\,\|\bar{\Delta}_\alpha^\star P_{P,\|}\|_{\mathrm{op}}^2$ gives $O(\alpha^{-2})$. □

Corollary 3 (Speed-up at matched contraction budget)**.** *Schedule isotropic collapse* $\dot\delta = -\beta\delta$ *so that its instantaneous rate of decrease of* $\mathrm{Tr}\,\Sigma_W$ *equals that of the best-response flow. Then* $\beta = \frac{\lambda}{2}\, r_\perp^P m_2^P + O(\alpha^{-1})$*, where* $r_\perp^P = \mathrm{Tr}\,\Sigma_{P\perp}/\,\mathrm{Tr}\,\Sigma_W$ *and* $m_2^P = (\lambda_1^{P,\perp} + \lambda_2^{P,\perp})/\,\mathrm{Tr}\,\Sigma_{P\perp}$ *are the raw analogues of the statistics in Proposition 1, and the leading task-transverse plane contracts* $1/(r_\perp^P m_2^P)$ *times faster under SP-NC than under isotropic collapse at equal budget.*

Proof. $\frac{d}{dt}\,\mathrm{Tr}\,\Sigma_W = -2\lambda\,\mathrm{Tr}(\bar{\Delta}_\alpha^\star\Sigma_W) = -\lambda(\lambda_1^{P,\perp} + \lambda_2^{P,\perp}) + O(\alpha^{-1})$ under the best-response flow (Remark 2 and Corollary 2) and $-2\beta\,\mathrm{Tr}\,\Sigma_W$ under isotropic collapse; equate. On the leading plane $\bar{\Delta}_\alpha^\star = \frac{1}{2}I + O(\alpha^{-1})$, so a selected residual obeys $\dot\delta = -\frac{\lambda}{2}\delta$ and its covariance eigenvalue $\dot\nu = -\lambda\nu$; under isotropic collapse $\dot\nu = -2\beta\nu$. The ratio is $\lambda/(2\beta) = 1/(r_\perp^P m_2^P)$ (the same ratio at the amplitude level, $(\lambda/2)/\beta$). □

A.6 NC2 COMPATIBILITY UNDER HARD INVARIANCE

Lemma 6 (NC2 feasibility in an invariant subspace)**.** *Let* $C \ge 2$*,* $V = \ker(\bar{\Delta}_G)$*, and* $\rho > 0$*. There exist centered vectors* $m_1, \dots, m_C \in V$ *forming a simplex ETF of radius* ρ *if and only if* $\dim V \ge C - 1$*.*

Proof. Their Gram matrix is $\frac{C\rho^2}{C-1}(I - \frac{1}{C}\mathbf{1}\mathbf{1}^\top)$, which has rank $C-1$, proving necessity. If $\dim V \ge C-1$, embed a canonical centered simplex isometrically in V. □

Taking $m_c = \mu_c - \bar\mu$ gives the centered-prototype NC2 compatibility condition. If exact invariance is also required of the raw prototypes μ_c, then $\bar\mu \in V$ is additionally necessary; when the prototypes are free, one may choose $\bar\mu = 0$. This is a geometric existence result only: it does not assert that the finite-α, finite-step algorithm produces such a kernel or converges to NC2.

A.7 REMARK ON THE NC1 STATISTIC

Remark 8. Under centered prototype preservation and the fixed-operator flow of Theorem 1, the statistic $\mathrm{NC1}_\dagger = \frac{1}{C}\,\mathrm{Tr}(\Sigma_W\Sigma_B^\dagger)$ of Papyan et al. (2020) is exactly invariant. Indeed Σ_B is constant and, with $E_t = e^{-\lambda t\bar{\Delta}_G}$, $\mathrm{range}(\Sigma_B^\dagger) = \mathcal{M} \subseteq \ker\bar{\Delta}_G$ gives $E_t\Sigma_B^\dagger = \Sigma_B^\dagger = \Sigma_B^\dagger E_t$, so

$$\mathrm{Tr}\left(\Sigma_W(t)\Sigma_B^\dagger\right) = \mathrm{Tr}\left(E_t\Sigma_W(0)E_t\Sigma_B^\dagger\right) = \mathrm{Tr}\left(\Sigma_W(0)\,E_t\Sigma_B^\dagger E_t\right) = \mathrm{Tr}\left(\Sigma_W(0)\Sigma_B^\dagger\right).$$

$\mathrm{NC1}_\dagger$ sees within-class scatter only inside $\mathcal{M}$, which is precisely the subspace a prototype-preserving operator does not touch. The quantities controlled by Theorem 1 and Proposition 1 are $\mathrm{Tr}\,\Sigma_W$ and trace-ratio statistics such as $\mathrm{Tr}\,\Sigma_W/\,\mathrm{Tr}\,\Sigma_B$. Reported trace-NC1 values in this paper should be read as trace-ratio statistics; in the coupled game the CE term moves Σ_B and centered prototype preservation holds only approximately, so $\mathrm{NC1}_\dagger$ can change, but not through the mechanism the theory isolates.

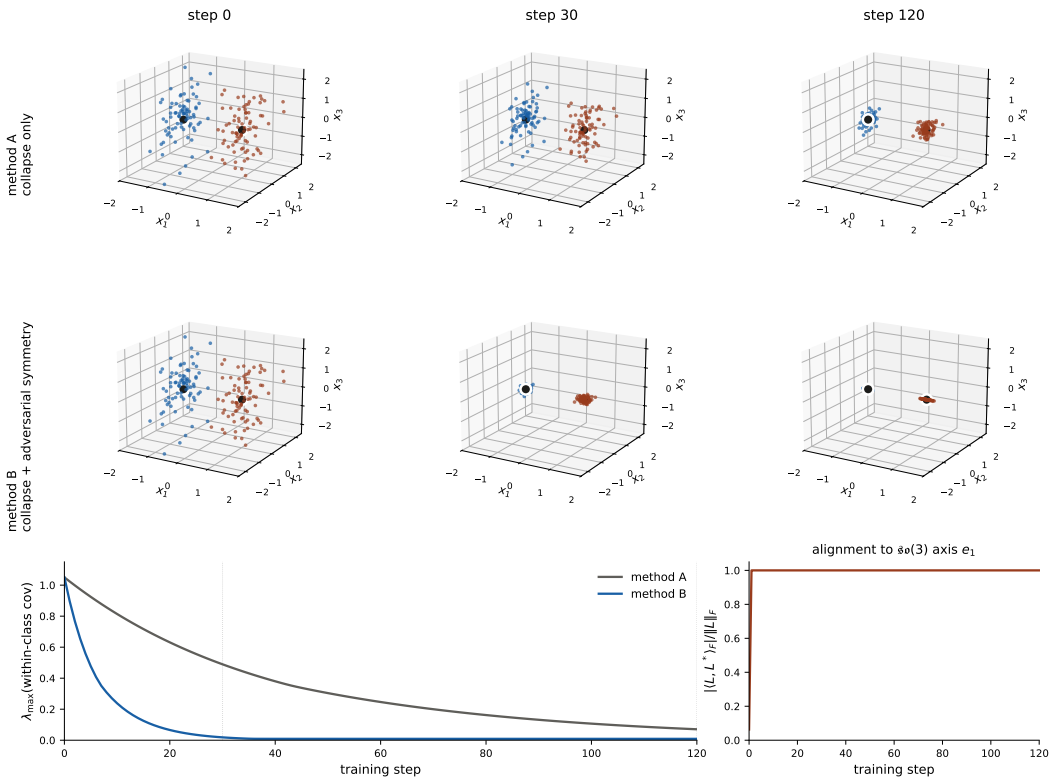


Figure 2: 3-D motivating simulation. Three class prototypes lie on a protected axis; within-class variation is dominated by a two-dimensional nuisance plane. Under matched schedules, isotropic collapse contracts the dominant residual eigenvalue from 1.05 to 0.071; SP-NC contracts it to 0.0089 (7.9× lower). The learned generator has cosine alignment 1.000 with the nuisance generator throughout; in $\mathbb{R}^3$ this is forced by prototype preservation (the rotation about the protected axis is the only element of $\mathfrak{so}(3)$ that fixes it), so the panel isolates the contraction-rate effect. Nontrivial discovery is tested in the hidden 10-D SO(2) setting below.

B CONTROLLED VALIDATION

Three synthetic settings validate the mechanism under identifiability.

Motivating 3-D anisotropic mixture. Three class prototypes on a protected axis in $\mathbb{R}^3$; nuisance is a two-dimensional rotation plane. Under matched schedules, isotropic collapse contracts the dominant residual eigenvalue to 0.071; SP-NC reaches 0.0089 (7.9× improvement; Figure 2). The generator alignment of 1.000 is forced by prototype preservation in $\mathbb{R}^3$, as the caption notes; this setting isolates the contraction rate.

Hidden 10-D SO(2) recovery. A random orthogonal transform hides a tetrahedral task subspace and a rank-2 nuisance plane in $\mathbb{R}^{10}$. With $K{=}1$, $S{=}15$, the minimum informative generator alignment is 0.99999998, prototype leakage is $8.0{\times}10^{-14}$, and final residual nuisance energy is $6.0{\times}10^{-15}$. This rules out visibility-of-canonical-axes as an explanation for recovery.

Full linear SO(3). With three fixed rotation generators acting on a task-orthogonal block, $K{=}3$ recovers the entire $\mathfrak{so}(3)$ span at alignment 1.000 and matches the oracle's nuisance suppression. Noncommutativity and multi-generator optimization are not intrinsic barriers.

The controlled ladder also includes finite-mixture, class-dependent, and continuous context-dependent nuisance fields: a global multi-generator operator succeeds whenever the union of nuisance directions is low-rank and task-orthogonal. Data dependence or continuous transformation parameters alone do not break SP-NC.

C STRUCTURAL APPLICABILITY CASE STUDIES

The CE-checkpoint quantities used in Section 5.1 are defined below. Items (i)–(ii) require class-labeled features; (iii)–(v) additionally require paired views or group metadata from training/development data:

(i) $r_\perp$: fraction of within-class covariance outside the class-mean span.

(ii) $\kappa_{1:4} = \lambda_1^\perp / \lambda_4^\perp$: spectral concentration (descriptive; not a universal gate).

(iii) Task-span overlap: the fraction of measured nuisance covariance that lies inside the class-mean span, $\mathrm{Tr}(P_\parallel \Sigma_{\mathrm{nuis}}) / \mathrm{Tr}\,\Sigma_{\mathrm{nuis}}$.

(iv) Transfer: the overlap between the leading nuisance plane estimated on the training split and on a held-out split.

(v) Oracle capture: the fraction of held-out nuisance covariance captured by the best rank-two projector supported in $\mathcal{M}^\perp$, i.e. $(\lambda_1^\perp + \lambda_2^\perp) / \mathrm{Tr}\,\Sigma_{\mathrm{nuis}}$ evaluated on Σ_{nuis}; this is the infinite-horizon projector upper bound of Proposition 1 with $r_G = 2$. The instantaneous energy of the associated unit-Frobenius skew operator is one half of this quantity.

Paired transformations estimate Σ_{nuis} when available; where nuisance-group metadata exist instead, class-conditional group offsets serve as a narrower proxy. The compact table omits transfer, which requires comparable nuisance estimates on two splits. The oracle uses nuisance information for diagnosis only, never for SP-NC training. The fixed-checkpoint case studies are summarized in Table 2.

Setting / layer	$r_\perp$	$\kappa_{1:4}$	Task-span overlap	Oracle capture (K=1)	Outcome
Centered color / pre1	**0.827**	127.0	0.044	**0.956**	positive (§4.1.1)
Rotation / pre1	0.368	3.91	0.429	0.369	boundary (§5.2)
GOOD-Motif / h1	0.147	12.72	0.845	0.760	modest positive
GOOD-CMNIST / h1	0.015	140,112	0.987	0.013	null
CelebA / pooled	0.602	8.34	0.144	0.729	separable at pool
CelebA / h1 (after head)	0.046	7.61	0.922	0.050	destroyed by head

Table 2: One-seed structural case studies on CE-trained checkpoints. In these cases $r_\perp$, interpreted together with nuisance-informed task-span overlap and oracle capture, is consistent with the clearest positive and null outcomes; κ alone does not (GOOD-CMNIST has $\kappa \sim 10^5$ but $r_\perp = 0.015$). The last two rows show a CelebA architecture-exposure failure: the pooled ResNet feature has separable gender geometry ($r_\perp = 0.602$, oracle capture 0.729) but a randomly-initialized affine head converts it into task-aligned structure before the first admissible SP site.

Brief frozen-feature online discovery can under-report a nuisance plane that becomes separable only during coupled training; a one-seed coupled probe may still be needed. The CelebA rows illustrate a distinct diagnostic use case: layerwise measurement identified that the architecture destroys otherwise-separable geometry before the intended SP site, providing an actionable architecture fix rather than a null verdict on the method.

D LEARNED DISCOVERY VERSUS ARBITRARY QUADRATIC OPERATORS

An independent five-seed Colored-MNIST control campaign kept the MLP, warm start, K=1, trace-one operator, first pre-ReLU site, coefficient 10, five-epoch ramp, balanced minibatch stream, and 100-epoch method horizon matched. The random and frozen controls change only the operator source; the current-coordinate oracle knows the injected color plane and is not an annotation-matched method.

Table 3 shows that the learned minus resampled-random accuracy gain is 3.67 paired points, 95% interval $[2.04, 5.31]$, and learned SP-NC retains 11.44 more local task-energy units while leaving statistically indistinguishable local nuisance energy. With eigenvalues preserved but orientation randomized, the learned advantage is 13.87 points, interval $[0.76, 26.97]$. This identifies a useful coupled data-dependent orientation, not a unique requirement for adversarial gradient ascent or a fixed-trajectory direct orientation effect.

E EXACT ONLINE BEST-RESPONSE CONTROL

For $K = 1$, Proposition 2 gives the exact response to the current minibatch's $S_\alpha = \Sigma_W - \alpha\Sigma_P$: $\bar{\Delta}^\star = \frac{1}{2}(w_1 w_1^\top + w_2 w_2^\top)$ for its top two eigenvectors. We recomputed this detached rank-two response at every update and compared

Operator source	Permuted acc. (%)	Local task energy	Local color energy
CE (no SP operator)	47.53	97.936	520.505
Fixed random skew	43.93	7.083	20.748
Frozen covariance/PCA	46.95	103.104	682.217
Resampled random skew	88.84	0.624	0.076
Online spectrum-randomized	78.65	1.321	0.538
Online learned SP-NC	**92.51**	12.066	0.070
Current-coordinate color oracle	92.24	2.383	0.003

Table 3: Five-seed operator controls on permuted Colored MNIST; energies are measured at the first pre-ReLU affine site. The oracle uses the known injected nuisance. Spectrum randomization preserves the contemporaneous learned eigenvalues at each update, but changes subsequent coupled training dynamics. This is an independent matched campaign, so its learned-SP mean is not substituted for the separate headline-table estimate.

Table 4: Five-seed online exact-response comparison on permuted Colored MNIST. Energies are at the first pre-ReLU site. Objective gap and principal angle compare each finite operator with the exact response at its own current representation, averaged over discovery events; the exact row is zero by construction.

Discovery	Acc. (%)	Task energy	Color energy	Obj. gap	Angle
CE	47.53	97.94	520.51	–	–
Exact top-two	91.28	65.78	5.700	≈ 0	0°
Finite $S = 5$	92.62	13.06	0.089	10.68	86.74°
Finite $S = 10$	92.65	12.31	0.083	10.85	85.89°
Finite $S = 15$	92.51	12.07	0.070	10.89	85.34°

it with persistent finite generator ascent using $S \in \{5, 10, 15\}$. The five paired Colored-MNIST arms share each seed's data, CE warm start, outer minibatches, $\alpha = 4$, coefficient and ramp, first pre-ReLU site, and fixed 100-epoch horizon; only discovery changes. Neither the exact arm nor the finite arms receive the injected nuisance basis. The exact formula was checked against the canonical inner objective. This is the exact *current inner-game* response, not an oracle for the true nuisance operator.

In Table 4, finite $S = 15$ exceeds exact by 1.23 paired permuted-accuracy points, 95% interval $[0.73, 1.74]$, with all five seeds positive; both are far above CE. The exact arm retains more local task energy, but leaves more local color energy. These are different coupled representation trajectories, so their endpoint energies do not isolate the effect of operator rank. Moreover, finite ascent is not close to the instantaneous exact plane here: in the final quarter the mean angle is 87.54° and none of 5,005 measured events enters the 45° local-tracking basin. The late relative eigengap averages 0.00093; the assay does not measure every drift and dominant-magnitude condition of Theorem 2. Thus the theorem remains conditional but does not explain this observed finite-step trajectory. This control does not show that iterative ascent is uniquely necessary, or that its audited wall times are comparable production costs: finite arms compute an extra exact eigendecomposition for diagnostics at every discovery event.

F Prototype-preservation ablation

The five-seed α assay used the same Colored-MNIST model and training stream, with K=1, S=15, mean/mean discovery, and otherwise matched outer SP training. Exact prototype-null projection is a mechanistic reference, not the proposed soft-penalty method.

Table 5 shows that, relative to α=0, the canonical α=4 retains far more task energy without turning off nuisance contraction and improves permuted accuracy by 2.57 points. α=1 gives the lowest time-averaged normalized leakage among soft penalties, whereas α=4 gives the best shifted accuracy. The α=16 reversal rules out a monotone protection claim. Exact nulling protects more task energy but leaves more nuisance and requires a different, more expensive constraint.

G Multi-generator diversity and capacity

With dense skew generators, a single L_k can contain multiple orthogonal rotation blocks, so K counts generator basis elements, not physical nuisance planes. The controlled two-plane and Colored-MNIST five-seed assays varied $K \in \{1, 2, 4\}$ and $\rho_{\text{div}} \in \{0, 1, 10\}$ while keeping mean-over-K outer normalization. On the two-plane benchmark,

Condition	Permuted acc. (%)	Local task	Local nuisance	Mean normalized leakage
Online, α=0	89.95	0.221	0.107	0.03582
Online, α=1	91.23	24.38	0.084	0.00711
Online, α=4	**92.51**	12.07	0.070	0.02026
Online, α=16	90.51	1.59	0.104	0.11539
Exact prototype null	92.39	77.13	5.400	≈ 0

Table 5: Prototype protection is a Pareto tradeoff, not a monotone function of its coefficient. Energies are intervention-layer measurements on the permuted view; leakage is averaged over discovery steps, and accuracy is the five-seed mean.

Table 6: Five-seed known-nuisance training controls. Colored-MNIST results use the matched control campaign; Fashion-MNIST results use the locked official test. Accuracy is in percent. Dashes indicate views not included in this comparison table.

Dataset	Training arm	ID	Permuted	Unbiased	Amplified
Colored MNIST	CE	96.21	47.53	–	–
	Online SP-NC	92.45	92.51	–	–
	Oracle equalized CE	94.43	94.51	–	–
	Oracle randomized CE	94.50	94.51	–	–
Fashion-MNIST	CE	93.01	34.37	66.23	67.50
	Online SP-NC	83.30	82.58	83.08	83.23
	Oracle equalized CE	83.84	84.06	84.07	83.01
	Oracle randomized CE	84.02	84.28	84.26	83.69

K=4 without diversity reached maximum absolute off-diagonal generator Gram 0.944, versus 1.11×10^{-8} at ρ_{div}=1; both recovered the full nuisance support. On Colored MNIST, the stable K=1, 2, 4 arms with $\rho_{\text{div}} \in \{0, 1\}$ all reached roughly 92.4–92.6% permuted accuracy, with no meaningful gain from increasing K. At K=4, ρ_{div}=10 instead produced duplicate generators, poor selectivity, and 48.42% permuted accuracy. Moderate diversity prevents redundancy; large K or large diversity is not a universal performance benefit.

H KNOWN-NUISANCE DATA CONTROLS

The equalized and randomized CE controls know the injected chromatic nuisance basis and radius construction. Equalized CE removes the class-conditional radius imbalance from training; randomized CE reshuffles radii within each training minibatch, including warmup. Evaluation images are unchanged. They test an information-rich upper comparison, not an annotation-matched replacement for online SP-NC. All arms use five paired seeds and the corresponding model, optimizer, and training horizon. Fashion-MNIST checkpoints were hash-locked before one evaluation on the full official test; no test-based model selection was performed.

In Table 6, both oracles beat SP-NC on permuted accuracy in every paired seed: by 2.00 points on Colored MNIST, and by 1.48 and 1.70 points, respectively, on Fashion-MNIST's official test. The controls show the headroom available with direct nuisance knowledge. SP-NC still gives large gains over CE without that information; the oracle comparison does not make nuisance-energy minimization synonymous with optimal classification.

I SECONDARY DATASET STUDIES

Auxiliary SP variants. SP-NC plus normalized isotropic collapse reaches $93.30 \pm 0.58\%$ permuted accuracy on Colored MNIST, versus $92.64 \pm 0.52\%$ for SP-NC alone; in a separate ten-seed confirmation its paired increment is $+0.20$ points. On the earlier ID-view geometry diagnostic, the combination lowers trace-NC1 to 0.244 and NC2 error to 0.622, versus 0.512 and 0.927 for SP-NC alone; these values are not directly comparable with the primary permuted diagnostic rows in Table 1. The same combination over-contracts Fashion-MNIST and reaches only $39.67 \pm 3.94\%$, compared with $82.58 \pm 0.47\%$ for SP-NC alone. On Waterbirds, late-head SP-NC reaches $75.44 \pm 1.55\%$ worst-group accuracy versus $78.41 \pm 2.22\%$ at the selected early-head site. On UCI-HAR, late-decay SP-NC plus isotropic collapse reaches $95.48 \pm 0.62\%$ terminal accuracy versus $94.97 \pm 0.31\%$ for SP-NC alone.

Additional energy diagnostics. On the amplified Colored-MNIST view, SP-NC reduces penultimate nuisance energy from 1795.9 under CE to 0.745 (99.96% removal), while task energy changes from 236.1 to 85.5 ($\sim 36\%$

retained). The corresponding task/nuisance ratio rises from 0.119 to 122.9; on Fashion-MNIST's amplified view it rises from 0.27 to 62.01. Waterbirds CE+IC contracts background energy below early-head SP-NC, but their five-seed worst-group accuracy difference is not statistically resolved.

GOOD-Motif. A three-layer GIN with a post-pool all-affine SP head. Under basis/covariate shift, late-SP-NC improves the unseen-basis-path OOD test by $+3.81$ points across all five seeds. Under basis/concept shift, early-SP-NC improves the stopped-checkpoint OOD accuracy by $+0.64$ points, trace-NC1 by -14.5%, and task/nuisance selectivity by $+52\%$. This establishes graph-modality compatibility through the same pre-nonlinear affine placement.

Gas Sensor Array Drift (2/5 seeds positive). UCI Gas Sensor Array Drift dataset (six gases, sixteen sensors, 128 engineered features per exposure, chronological train/dev/test split across ten temporal batches). Backbone: 4-layer MLP with SP-NC at the first pre-ReLU affine. Across five seeds, SP-NC is positive in $2/5$; the first three seeds alone gave $+3.46 \pm 3.76$ points (per-seed $+5.01, +6.19, -0.84$), which the remaining two seeds did not sustain. We report Gas Drift as a *cautionary example* of seed sensitivity, precisely the risk the paper's other results ($n = 5, 10$) are designed to guard against. The one-seed diagnostic on this dataset flagged the task/nuisance overlap as high, so the outcome is consistent with the diagnostic.

ImageNet-9 backgrounds. One-seed protocol-equivalent reconstruction of the Backgrounds Challenge of Xiao et al. (2021) (the authors' training archives were unavailable; we regenerated the training corpus from ILSVRC + object annotations following the published protocol). Backbone: ResNet-18 trained from scratch with an all-affine head, $C = 9$. Four arms with matched schedule, initialization, and optimizer:

Method	Original	Mixed-Same	Mixed-Rand	Mixed-Next
CE	83.88%	76.57%	61.01%	57.68%
CE + isotropic collapse	30.84%	30.77%	23.21%	20.15%
CE + online SP-NC	**85.01%**	76.91%	60.25%	56.91%
CE + online SP-NC + isotropic	84.99%	**77.21%**	**62.94%**	**59.26%**

Table 7: ImageNet-9 backgrounds, one-seed reconstruction. The CE + isotropic control collapses to poor performance by epoch 2 under identical calibrated scale, initialization, schedule, and optimizer to the CE + SP-NC + isotropic arm; isotropic collapse alone is not viable at this scale (balanced nine-class chance is 11.11%). SP-NC + isotropic gives the four-way gain: $+1.11$ Original, $+0.64$ Mixed-Same, $+1.93$ Mixed-Rand, $+1.58$ Mixed-Next over CE.

Table 7 has two consequences. First, isotropic collapse alone does *not* explain the combined-arm gain in this run and at this calibration: it performs poorly without SP-NC, whereas the combined arm remains competitive with CE. This is evidence of an SP–isotropic interaction, not a general proof that SP is necessary for isotropic training stability. Second, the combined-arm gain over CE ($+1.93$ Mixed-Rand, $+1.58$ Mixed-Next) does not by itself demonstrate that *learned online discovery* is uniquely responsible: random-operator + isotropic and frozen-SP + isotropic controls would be needed to distinguish nuisance-aware discovery from generic anisotropic regularization, and are outside the scope of this study. We therefore present ImageNet-9 as secondary supporting evidence with a resolved isotropic-alone ablation, not as a headline.

Camelyon17. 20-seed screen on a ResNet-18 + late-affine bottleneck. Bridge SP-NC improves hospital-1 OOD validation by $+0.82$ points, CI $[+0.12, +1.52]$ (resolved above zero); trace-NC1 improves in 20/20 paired seeds; hospital energy is strongly suppressed. However, the effect does not transfer to the locked hospital-2 test (-0.22 points, CI crosses zero). Camelyon17 supplies *mechanistic evidence of hospital-selective representation cleanup* but does not support a cross-hospital utility claim. It is included as a natural-domain example where the mechanism operates as designed even when downstream utility is uncertain.

CIFAR-10 multi-plane nuisance. With three injected nuisance planes, one-seed SP-NC improves permuted and unbiased accuracy by 5–6 points over CE and beats random SP, but contracts task energy and worsens trace-NC1/NC4. $K \in \{1, 2, 3\}$ differences are small and non-monotonic. This is qualified capacity evidence rather than a headline positive.

GOOD-CMNIST. Low-capacity geometry and observed null. $r_\perp = 0.015$ at the first affine and 99% of measured color-group variation lies in the task span. SP-NC produces occasional NC2 or selectivity gains but no reliable unseen-color accuracy improvement. This outcome is consistent with the structural triage reading, without making it a universal filter.

ModelNet40 with PointNet. At the tested bridge, $r_\perp = 0.014$; at the penultimate feature it is 1.7×10^{-4}. Both indicate little fixed-checkpoint capacity. PointNet's max-pooling destroys the shared feature-space SO(3) action even though $K{=}3$ matches the input algebra dimension. The observed outcome is null.

J AUXILIARY NC2 CONTROLS

These Colored-MNIST experiments explore separate class-mean objectives; neither objective is part of the canonical SP-NC training algorithm or establishes NC2 convergence under SP-NC. The scale-free empirical-prototype ETF (SEP) objective centers the empirical class means, normalizes each centered mean to a direction, and applies a pseudo-Huber penalty ($\delta = 0.1$) to strict-upper-triangle pairwise cosine errors relative to the simplex value $-1/(C-1)$. An optional log-radius term can encourage equal relative prototype norms; SEP selects neither an absolute feature radius nor an absolute orientation. In the reported Colored-MNIST arm the radius term is off, and the SEP coefficient is adapted so that its penultimate representation-gradient norm targets a fraction q of the CE representation-gradient norm, using an exponential moving average and a ramp over the first 10% of training. With $q = 0.0125$, SEP reduces NC2 error by 26.9% without accuracy loss but worsens trace-NC1 through between-class contraction. In this separate three-seed NC2-control campaign, SP+SEP+collapse reaches trace-NC1 0.267, NC2 0.455, and 93.10% permuted-color accuracy, versus 0.511, 0.947, and 92.42% for the matched SP-NC baseline.

The raw-Gram target $\|MM^\top - \rho^2(CI - \mathbf{1}\mathbf{1}^\top)/(C-1)\|_F^2$ can enforce the intended simplex geometry (prototype ETF error 0.11; observed cosine -0.110 ± 0.012 versus ideal -0.111), but produces sharp trace-NC1 degradation ($0.252 \to 3.084$) and accuracy loss. These results illustrate an NC1–NC2–accuracy tradeoff in this controlled setting; they do not establish a generally preferred NC2 controller. The full raw-Gram zero-set characterization is reserved outside this manuscript.

K REPRODUCIBILITY DETAILS

K.1 FULL ONLINE TRAINING LOOP

Algorithm 2 expands the alternating step in Algorithm 1. The optional CE warm start and the schedule λ_t are specified by each dataset protocol below; neither changes the generator-discovery objective.

Class means in both phases are empirical means of the current batch, not separately optimized prototype parameters. Isotropic collapse, ETF control, and post-training prototype projection are not steps of this canonical loop.

For the SP residual term, differentiating through exact empirical class means or treating those same means as fixed gives the same feature gradient: the residuals sum to zero within every class. The Colored-/Fashion-MNIST and Waterbirds runners use detached batch means in the outer penalty; the UCI-HAR runner differentiates through its batch means. The operator and inner features are detached in all cases.

K.2 SHARED TRAINING AND SUPERVISION CONVENTIONS

The four headline comparisons use five paired seeds per dataset: 0–4 for Colored MNIST, Fashion-MNIST, and UCI-HAR, and 500–504 for Waterbirds. The intervention is always the output of an affine map before the next nonlinearity, never a convolutional feature map. The online arms use a persistent dense skew generator with $K{=}1$, mean reductions over examples and represented classes, prototype coefficient $\alpha{=}4$, generator learning rate 0.015, and hard unit-Frobenius projection after each ascent step. The outer operator is detached and rediscovered on each training minibatch; the model performs no SP operation at inference. Dataset-specific inner steps, outer coefficients, and schedules are below. These are a shared algorithmic template, not identical hyperparameters. The canonical prototype penalty acts on raw empirical class means; centered-span projection is a distinct diagnostic/control and is not interchangeable with that training penalty.

The group-information codes in Table 1 have different roles. Colored/Fashion CE, CE+IC, Frozen PCA-SP, SP-NC, and JTT use no nuisance annotation; their DFR controls fit a classifier on an additional nuisance-equalized holdout. On Waterbirds all deep-model checkpoints use validation group labels for worst-group selection, while group-balanced ERM and GroupDRO additionally use training groups and DFR uses validation groups for classifier refitting. On UCI-HAR the selected SP-NC and CE models use no subject identities; subject-balanced ERM and GroupDRO use training subjects, while DFR uses development-subject labels.

Algorithm 2 Full online SP-NC training loop

Require: Feature map Φ_θ, classifier W, label-complete minibatches, $K, S, \eta_G, \alpha, \rho_{\mathrm{div}}, \rho_{\mathrm{norm}}$, outer optimizer and schedule λ_t.

If prescribed, warm-start (θ, W) with CE; sample raw matrices $\{A_k\}_{k=1}^K$ once and retain them across minibatches.
for each training minibatch B_t of N examples **do**
  Set λ_t from the dataset's outer schedule.
  Phase 1: discover with the network fixed.
  Compute detached features $\tilde{h}_i = \mathrm{stopgrad}(\Phi_\theta(x_i))$; form current raw class means $\tilde{\mu}_c$ and residuals $\tilde{\delta}_i = \tilde{h}_i - \tilde{\mu}_{y_i}$ on B_t.
  for $s = 1, \ldots, S$ **do**
    Form $L_k = A_k - A_k^\top$ and evaluate $J(\{L_k\}; \tilde{\delta}, \tilde{\mu})$ from Eq. 2.
    Compute all $g_k = \nabla_{A_k} J$ before updating any A_k.
    for $k = 1, \ldots, K$ **do**
      $A_k \leftarrow A_k + \eta_G g_k$; $L_k \leftarrow A_k - A_k^\top$.
      if $\|L_k\|_F > 10^{-10}$ **then**
        $A_k \leftarrow A_k / \|L_k\|_F$.
      end if
    end for
  end for
  Form the updated $L_k = A_k - A_k^\top$ and $\bar{\Delta}_G = \mathrm{stopgrad}\big(K^{-1} \sum_k L_k^\top L_k\big)$.
  Phase 2: train with the operator fixed.
  Recompute $h_i = \Phi_\theta(x_i)$ on a fresh graph; form current-batch means μ_c and residuals $\delta_i = h_i - \mu_{y_i}$ on B_t.
  Set $\mathcal{R}_{\mathrm{SP}} = (2N)^{-1} \sum_i \delta_i^\top \bar{\Delta}_G \delta_i$ and update (θ, W) with the outer optimizer on $\mathcal{L}_{\mathrm{CE}} + \lambda_t \mathcal{R}_{\mathrm{SP}}$.
  Keep $\{A_k\}$ for the next minibatch; do not differentiate the outer loss through the discovery steps.
end for

Ensure: At inference, use (θ, W) only; apply no SP operator.

K.3 Headline datasets, architecture, and checkpoint rules

Colored MNIST. We sample 800 official MNIST training images per class and 300 official test images per class. Flattened RGB inputs have dimension 2352. A shared two-dimensional chromatic plane supplies zero-mean within-class color variation with class-dependent radii from 2 to 8. The primary permuted evaluation reassigns the radius–class association; ID, equalized, and amplified views use the same underlying test identities. The MLP is $2352 \to 64 \to 128 \to 64 \to 10$, with ReLU after each hidden affine; SP-NC acts at the first 64-dimensional pre-ReLU output, while NC metrics use the 64-dimensional penultimate layer. A common 12-epoch CE warm start is followed by 100 method-training epochs (4000 class-balanced batch-200 updates). The method phase uses Adam at 2×10^{-4}, weight decay 10^{-5}, S=5 inner steps, outer SP coefficient 10, and a 200-update/five-epoch ramp. The reported primary endpoint is the final method checkpoint; there is no test-based checkpoint selection.

Fashion-MNIST. The chromatic construction and $2352 \to 64 \to 128 \to 64 \to 10$ MLP are the same, but the task images come from Fashion-MNIST. Per class, 800 fitting identities and 300 disjoint development identities are drawn from the official training pool. A separate nuisance-equalized holdout is used only by DFR. The headline evaluation applies the fixed color conditions to all 10,000 official test images after the five-seed terminal checkpoints were locked. The method training starts cold and runs 100 epochs/4000 balanced batch-200 updates with Adam at 2×10^{-4}, weight decay 10^{-5}, K=1, S=5, α=4, SP coefficient 1, and a five-epoch ramp. SP-NC acts at the first pre-ReLU affine; NC metrics are penultimate. The terminal checkpoint, not official-test performance, determines the reported model.

Waterbirds. We use the official WILDS train, validation, and test splits (4795, 1199, and 5794 images). The ImageNet-pretrained ResNet-18 is followed by global pooling to 512 dimensions, affine maps $512 \to 256 \to 256 \to 128$, one ReLU, and a $128 \to 2$ classifier. Head affines have no intervening nonlinearity; the headline SP-NC site is the first 256-dimensional head output and the penultimate diagnostic site is the post-ReLU 128-dimensional feature. AdamW uses backbone/head learning rates $10^{-4}/10^{-3}$, weight decay 10^{-4}, batch size 128, a 150-epoch cosine schedule, and standard random-resized-crop/horizontal-flip training transforms. Online SP uses S=15, coefficient 0.3, and a five-epoch ramp. All methods run 150 epochs; the primary checkpoint maximizes validation worst-group accuracy. Worst-group test accuracy uses all four bird-label×background groups on the full official test split, with no test-based selection. The matched JTT baseline uses a one-epoch CE identification model to mark training errors, then trains a freshly initialized model with 20-fold repetition of those examples for the full 150-epoch schedule. Its identifier epoch

and repetition rule were chosen on validation data; neither stage uses training background labels. This is a ResNet-18 adaptation under the shared architecture and checkpoint protocol, not a reproduction of the published ResNet-50 setup.

UCI-HAR. We use 561 engineered sensor features for six activities, with the official nine-subject test untouched. The official training subjects are divided into 15 fitting subjects (5066 windows) and six development subjects (2286 windows); the official test has 2947 windows. Standardization uses fitting features only. The MLP is $561 \to 256 \to 256 \to 128 \to 6$, with ReLU after each hidden affine; SP-NC acts at the first 256-dimensional pre-ReLU output and NC metrics use the penultimate feature. All methods use five seeds, class-balanced batches of 600, Adam at 10^{-3}, weight decay 10^{-5}, and 100 epochs. The selected late-decay SP-NC arm uses $S{=}15$, coefficient 1 after a five-epoch ramp, then linearly releases that coefficient from epoch 80 to zero at epoch 100. The headline table uses fixed epoch-100 checkpoints for both CE and SP-NC, not each method's best development-selected checkpoint.

K.4 Geometry and energy measurement

The empirical prototype is the class mean at the stated layer; its centered span is the *task span* in the fixed-operator theory. We use *task energy* only for a measured label-associated between-class feature statistic, not as a claim that every component is causally necessary for classification. Likewise *nuisance energy* is defined by the measured color, background, or subject variation, not by an unobserved universal nuisance subspace. All trace-NC1 and NC2–NC4 metrics in the headline comparison come from the penultimate layer on the same primary evaluation environment as its accuracy entry. Trace-NC1 is within-class covariance trace divided by between-class covariance trace; it is distinct from the pseudoinverse-weighted classical NC1 statistic. NC2 error is the normalized distance between the Gram matrix of unit centered class means and the simplex target. NC3 error compares normalized centered classifier weights and class means; NC4 is classifier/nearest-class-center prediction agreement. For binary Waterbirds, centered means are necessarily antipodal, so the NC2 Gram diagnostic is not informative and is omitted.

For the color tasks, task energy is the mean squared norm of centered class means of the *clean* representation; nuisance energy is the mean squared feature difference between paired colorized and clean views of the same image. Waterbirds diagnostics use an equal-size sample from each bird$\times$background test group: task energy is the trace of the covariance of balanced bird-class means, and background energy is the trace of balanced within-class background-mean offsets. The full official test, not this balanced diagnostic sample, supplies accuracy. UCI-HAR uses a sample-weighted decomposition into class-mean energy and class-conditional subject-mean energy. Ratios are formed per seed and then averaged; the mean ratio is not generally the ratio of the two displayed mean energies. Feature scale and nuisance construction differ across datasets, so absolute energies are compared only within a dataset. The five-seed values and paired intervals for the primary endpoints and the harmonized geometry statistics are tabulated in Appendix K.5.

K.5 Harmonized five-seed statistics and geometry

Tables 8–12 use the original five paired seeds and each dataset's primary shifted test endpoint. Values are means across seeds; $\pm$ denotes sample standard deviation and brackets denote two-sided Student-t 95% intervals. Paired differences use same-seed SP/control results rather than subtracting independent intervals. These intervals describe seed variation under fixed protocols, not uncertainty across datasets. NC values and energies are means; task/nuisance is the mean of per-seed ratios, not the ratio of displayed means. Absolute energies are only comparable within a dataset; near-zero nuisance denominators can make mean ratios unstable. NC2 is omitted for binary Waterbirds, where centered class means are necessarily antipodal. Source aggregates retain individual seed values and full precision.

Colored MNIST. The endpoint is permuted-color accuracy at the final locked checkpoint. All methods use a four-layer raw-pixel MLP; SP acts after the first affine map, before ReLU. Only DFR uses extra nuisance information, through a held-out nuisance-equalized classifier-refit set; the other listed methods use no group labels.

Fashion-MNIST. Table 9 reports permuted-color accuracy on the official test at the locked checkpoint. All methods use a four-layer raw-pixel MLP, with SP at the first pre-ReLU affine output. Only DFR uses a held-out nuisance-equalized classifier-refit set.

Waterbirds. Table 10 reports worst-group accuracy on the full official test, with checkpoints selected by validation worst-group accuracy. The backbone is ImageNet-pretrained ResNet-18 with an affine-prefix head; SP acts at its early affine output. CE, SP-NC, CE+IC, Frozen PCA-SP, and JTT use validation-group labels for checkpoint selection but no training-group labels. Balanced ERM and GroupDRO use training-group labels; DFR uses validation-group labels for its post-hoc classifier refit. The energies below use a balanced bird-by-background diagnostic sample, not the full accuracy set.

Table 8: Colored-MNIST harmonized five-seed statistics. Endpoint and paired gaps are percentage points; geometry and energies use the penultimate layer except the clean-view task and paired-color nuisance energy measured as defined in Appendix K.4.

Method	Permuted acc. ± SD [95% CI]	Paired Δ vs CE [95% CI]
CE	47.527 ± 2.981 [43.826,51.228]	reference
CE+IC	42.540 ± 3.340 [38.393,46.687]	−4.987 [-6.333,-3.640]
Frozen PCA-SP	46.947 ± 1.888 [44.603,49.291]	−0.580 [-2.268,1.108]
SPNC	92.640 ± 0.523 [91.991,93.289]	45.113 [41.511,48.716]
JTT	46.547 ± 2.863 [42.992,50.102]	−0.980 [-2.318,0.358]
DFR	80.547 ± 1.083 [79.202,81.892]	33.020 [29.479,36.561]

Method	trace-NC1	NC2	NC3	NC4	Task	Nuisance	Task/nuis.
CE	0.622	0.909	1.130	0.478	565.882	1532.503	0.373
CE+IC	0.192	1.339	1.285	0.437	7.656	67.996	0.114
Frozen PCA-SP	0.734	0.917	1.136	0.476	522.199	874.185	0.613
SPNC	0.512	0.928	0.592	0.938	182.272	0.605	340.724
JTT	0.593	0.958	1.142	0.470	446.893	1229.920	0.370
DFR	0.622	0.909	1.108	0.793	565.882	1532.503	0.373

Table 9: Fashion-MNIST harmonized five-seed statistics. Endpoint and paired gaps are percentage points; NC metrics are penultimate-layer statistics and energies use the clean/paired-color diagnostic.

Method	Permuted acc. ± SD [95% CI]	Paired Δ vs CE [95% CI]
CE	34.366 ± 2.984 [30.661,38.071]	reference
CE+IC	31.970 ± 8.883 [20.941,42.999]	−2.396 [-12.720,7.928]
Frozen PCA-SP	37.218 ± 5.703 [30.138,44.298]	2.852 [-2.045,7.749]
SPNC	82.578 ± 0.467 [81.998,83.158]	48.212 [44.870,51.554]
JTT	22.824 ± 1.620 [20.813,24.835]	−11.542 [-16.286,-6.798]
DFR	66.828 ± 5.061 [60.545,73.111]	32.462 [24.767,40.157]

Method	trace-NC1	NC2	NC3	NC4	Task	Nuisance	Task/nuis.
CE	0.466	1.645	1.200	0.332	645.968	1265.268	0.520
CE+IC	0.010	2.773	0.366	0.183	1.029	0.000	4259.722
Frozen PCA-SP	0.474	1.621	1.176	0.359	627.520	1067.338	0.603
SPNC	0.324	1.443	0.667	0.851	257.466	1.699	166.670
JTT	0.365	1.810	1.292	0.231	547.851	1649.694	0.336
DFR	0.466	1.645	1.235	0.598	645.968	1265.268	0.520

Additional Waterbirds last-layer controls. We adapt AFR (Qiu et al., 2023) and LFR (Ghaznavi et al., 2023) to the shared ResNet-18 and affine-head architecture. For each seed, a class-stratified 80% of the official training images fits the CE encoder for 150 epochs; the remaining 20% fits the last-layer classifier. This allocation is part of these methods: CE and SP-NC instead use the full official training pool end to end. No method in Table 11 uses training background labels. AFR refits the classifier with 500 full-batch SGD steps and confidence weights $\exp[-\gamma p(y \mid x)]$, selecting $\gamma \in \{1, 3, 10\}$. LFR forms equal-size class-conditional pseudo-groups from high-loss errors and low-loss correct examples; validation selects a group-size fraction in $\{0.25, 0.5, 1\}$ and either a 500-step SGD head or a regularized logistic head with $C \in \{0.001, 0.01, 0.1, 1, 10, 100\}$. For both, validation worst-group accuracy selects the final or validation-best encoder source and head configuration. The test set is evaluated only after selection. These are matched-architecture adaptations, not reproductions of the original ResNet-50 protocols.

UCI-HAR. Table 12 reports subject-disjoint official-test accuracy at fixed epoch 100. The model is a 561-feature MLP with SP at its first pre-ReLU affine output and a late-decay SP coefficient. Balanced ERM and GroupDRO use training-subject labels; DFR uses validation-subject labels for classifier refitting. Other methods use no subject identities. Task and nuisance energy respectively measure class-mean and class-conditional subject-mean variation.

Table 10: Waterbirds harmonized five-seed statistics. Endpoint and paired gaps are worst-group percentage points. NC2 is omitted because the binary-class Gram diagnostic is trivial.

Method	Worst-group acc. ± SD [95% CI]	Paired Δ vs CE [95% CI]
CE	71.449 ± 3.695 [66.862,76.036]	reference
SPNC	78.409 ± 2.222 [75.651,81.167]	6.960 [4.818,9.102]
Bal-ERM	78.791 ± 1.293 [77.186,80.396]	7.342 [3.672,11.012]
GroupDRO	86.053 ± 1.749 [83.881,88.225]	14.604 [10.156,19.053]
CE+IC	76.599 ± 3.299 [72.503,80.694]	5.150 [0.462,9.838]
Frozen PCA-SP	73.671 ± 3.666 [69.120,78.222]	2.222 [-4.758,9.203]
JTT	73.642 ± 2.682 [70.312,76.973]	2.194 [-5.213,9.601]
DFR	84.522 ± 2.809 [81.035,88.009]	13.074 [5.339,20.808]

Method	trace-NC1	NC2	NC3	NC4	Task	Background	Task/bg
CE	1.499	–	0.566	0.939	526.172	260.350	2.072
SPNC	0.788	–	0.427	0.980	114.868	24.082	4.927
Bal-ERM	1.081	–	0.497	0.953	526.371	117.827	5.288
GroupDRO	0.932	–	0.639	0.946	373.577	43.554	317.577
CE+IC	0.647	–	0.372	0.987	2.047	0.187	11.832
Frozen PCA-SP	7.822	–	0.606	0.840	1982.768	287.511	7.295
JTT	1.369	–	0.491	0.957	466.317	187.547	2.627
DFR	1.410	–	1.414	0.858	1688.866	885.302	1.914

Table 11: Additional Waterbirds controls with no training-group labels. Mean ± sample standard deviation over seeds 500–504; the endpoint is official-test worst-group accuracy. All methods have access to the same official training pool, allocated according to their protocol.

Method	Average accuracy (%)	Worst-group accuracy (%)
CE (full-train)	86.32 ± 1.38	71.45 ± 3.69
ERM (80%-train encoder)	85.07 ± 0.78	70.16 ± 2.94
AFR	85.11 ± 0.71	71.00 ± 2.43
LFR	84.12 ± 3.66	69.40 ± 2.84
SPNC (full-train)	89.16 ± 1.13	78.41 ± 2.22

Table 12: UCI-HAR harmonized five-seed statistics. Endpoint and paired gaps are subject-disjoint accuracy percentage points.

Method	Test acc. ± SD [95% CI]	Paired Δ vs CE [95% CI]
CE	93.532 ± 0.317 [93.139,93.926]	reference
CE+IC	95.684 ± 0.764 [94.735,96.632]	2.151 [1.050,3.253]
Frozen PCA-SP	93.709 ± 0.092 [93.595,93.823]	0.176 [-0.204,0.557]
SPNC	94.971 ± 0.307 [94.590,95.352]	1.439 [0.721,2.157]
JTT	92.949 ± 0.425 [92.421,93.476]	−0.584 [-1.387,0.220]
DFR	92.704 ± 0.466 [92.125,93.284]	−0.828 [-1.533,-0.123]
Bal-ERM	93.451 ± 0.342 [93.027,93.875]	−0.081 [-0.349,0.186]
GroupDRO	93.526 ± 0.406 [93.022,94.030]	−0.007 [-0.163,0.149]

Method	trace-NC1	NC2	NC3	NC4	Task	Nuisance	Task/nuis.
CE	0.233	0.789	0.750	0.960	1459.157	171.457	8.552
CE+IC	0.025	1.035	0.778	0.991	40.234	0.396	102.738
Frozen PCA-SP	0.212	0.803	0.750	0.956	1532.197	166.669	9.210
SPNC	0.239	0.518	0.452	0.981	199.845	17.580	11.480
JTT	0.274	0.697	0.778	0.956	1376.908	192.026	7.205
DFR	0.233	0.789	1.265	0.939	1459.157	171.457	8.552
Bal-ERM	0.241	0.810	0.738	0.959	1454.124	179.931	8.134
GroupDRO	0.233	0.782	0.743	0.958	1435.290	169.260	8.498

Table 13: Training-step compute at the headline SP settings. Absolute overhead depends on the architecture and inner-step count; its relative multiplier is smaller for the heavier CE model.

Model	S	CE (ms)	SP (ms)	Add (ms)	SP/CE
Colored-MNIST MLP	5	1.781	8.735	6.954	4.90×
Waterbirds ResNet-18 + head	15	21.601	47.432	25.831	2.20×

Table 14: Four-seed S and λ sensitivity on Colored-MNIST development data. The grid is descriptive, not a selected-test result.

S	λ	Permuted acc.	Nuisance	Task/nuis.	trace-NC1
5	10	89.18%	0.282	67.5	0.588
10	10	89.32%	0.202	102.9	0.574
15	10	89.22%	0.316	83.6	0.579
15	2.5	89.89%	0.428	42.9	0.555
15	5	89.69%	0.265	68.5	0.564
15	20	89.11%	0.152	114.7	0.580

K.6 Inner-step sensitivity and training cost

The online operator adds a detached discovery forward pass and S generator updates per training step. This is an additive, configuration-dependent cost, not a universal multiplier or a fixed number of milliseconds across architectures. Table 13 times the headline model and batch shapes on an H200, one arm per process. Values are medians of three repeats, each with 50 warm-up and 150 CUDA-synchronized measured steps on device-resident inputs. The assay excludes data loading, evaluation, checkpointing, W&B, and discovery diagnostics; it measures training compute rather than full-epoch wall time. SP adds no inference-time operation.

At $S = 15$ on the Colored-MNIST MLP, SP takes 18.257 ms ($10.25\times$ CE); at $S = 5$ on the Waterbirds architecture it takes 36.540 ms ($1.69\times$ CE). A separate diagnostic-inclusive assay found near-identical allocated peak memory across CE and SP settings: 75.5–75.7 MiB for Colored-MNIST and 1805.9–1806.1 MiB for Waterbirds. Those allocations are not from the lean timing assay and do not establish a general peak-memory bound; dense generator storage still scales as $O(Kd^2)$.

Table 14 gives four-seed Colored-MNIST development results at the first pre-ReLU site, with $K = 1$, $\alpha = 4$, a common 12-epoch CE warm start, and 500 outer updates. Accuracy, nuisance energy, and task/nuisance ratio use the permuted development view; the energies are measured at the SP site, while Trace-NC1 uses penultimate features.

At $\lambda = 10$, $S = 5$ and 10 match the $S = 15$ anchor closely in permuted accuracy (paired gaps -0.04 and $+0.10$ points; 95% intervals $[-0.24, 0.16]$ and $[-0.30, 0.50]$). Larger S is not monotonically better, and stronger λ suppresses more measured nuisance without necessarily improving accuracy. This inspected grid does not identify a uniquely optimal setting.

K.7 Ordinary minibatches and center estimators

The canonical discovery step uses current empirical class means in label-complete batches. Here "current centers" means batch empirical prototypes, not the exact spectral best-response operator. Balanced and ordinary batches contain 200 examples; the rare-class stress uses batches of 40 and samples class 9 with probability 0.002. The tested exponential-moving-average (EMA) center has decay 0.9. If a rare batch lacks a class, the current-center arm reuses the last discovered operator, applying no SP penalty before the first complete batch. CE controls use the corresponding sampling streams.

Table 15 shows that every ordinary batch in this balanced-population assay contained all classes. Balanced-current minus ordinary-current SP has a paired shifted-accuracy gap of $+0.11$ points, 95% interval $[-0.40, +0.61]$; explicit balancing is unnecessary in this tested regime. Current centers outperform the tested EMA setting, but this does not rule out other memory estimators. In the rare-class stress, 92.9% of batches omit a class and both SP arms leave class 9 essentially unlearned. This is a limitation of the present discovery scheme under severe imbalance.

K.8 Configuration and artifact provenance

The exact configurations, checkpoints, and evaluation aggregates are retained with the experiment artifacts:

Table 15: Four-seed Colored-MNIST minibatch/center-estimator ablation. All metrics use the permuted development view except ID accuracy. Nuisance and task/nuisance ratio are measured at the SP site.

Sampler	Method	Perm. acc.	ID acc.	Nuisance	Ratio	Class-9 acc.
Balanced	CE	48.14%	95.38%	472.3	0.18	0%
Balanced	SP, current	89.22%	89.31%	0.316	83.6	82.58%
Ordinary	CE	48.12%	95.43%	470.4	0.18	0%
Ordinary	SP, current	89.11%	89.19%	0.176	104.4	83.33%
Ordinary	SP, EMA	85.30%	86.61%	1.256	9.0	71.33%
Rare	CE	48.36%	92.78%	457.7	0.18	0%
Rare	SP, current	76.62%	79.72%	4.510	3.46	0.58%
Rare	SP, EMA	65.96%	79.80%	18.678	1.26	0.58%

- Colored MNIST: `outputs/colored_mnist_baselines/runs/seed_*/manifest.json` and `outputs/colored_mnist_baselines/aggregate.json`.
- Fashion-MNIST: `outputs/nc2_transfer_3seed/fashion_mnist/online_sp/seed_*/manifest.json`, `outputs/fashion_mnist_baselines/runs/seed_*/manifest.json`, and `outputs/fashion_mnist_sealed_test/summary.json`.
- Waterbirds: `outputs/waterbirds_v2/early_head_sp/seed_*/manifest.json`, `outputs/waterbirds_v2/test_summary.json`, and `outputs/waterbirds_baselines/aggregate.json`; the revised JTT and AFR/LFR comparison is in `outputs/waterbirds_review03/aggregate.json`, with head-selection manifests at `outputs/waterbirds_last_layer_controls/head_selection/seed_*/selection.json`.
- UCI-HAR: `outputs/sp_native_mlp/uci_har/online_sp_late_decay/seed_*/manifest.json` and `outputs/uci_har_baselines/aggregate.json`.
- Cost and batch-estimator ablations: `outputs/manuscript_b1_compute_cost/`, `outputs/manuscript_b1_sensitivity/`, and `outputs/manuscript_b2_ordinary_centers/`.
- Harmonized five-seed statistics: `outputs/manuscript_b3_harmonized_stats/harmonized_stats.json`.
- Exact online response and known-nuisance controls: `outputs/b4_exact_best_response/summary.json` and `outputs/b5_oracle_nuisance_controls/summary.json`.

The shared metric implementations are in `src/spnc/experiments/emergent_collapse.py`, `src/spnc/experiments/colored_mnist_sp.py`, `src/spnc/experiments/waterbirds.py`, and `src/spnc/experiments/sp_native_mlp.py`. The dense skew parameterization has $O(Kd^2)$ storage and inner-update cost at an intervention dimension d. Low-rank generators, periodic discovery, and partial-class memory prototypes are possible scaling variants, but are not validated by the headline runs.