ON THE LIMITATIONS OF IDENTIFIABILITY THEORY FOR CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent identifiability results for contrastive learning connect successful representation learning to recovery of latent factors underlying an invertible data-generating process. We study a limitation of using such recovery as an explanation of downstream performance: latent recoverability, task sufficiency, and stability are distinct properties. Focusing on classification, we first introduce class-conditional structure into standard latent-variable models and show how concentration simultaneously reduces class ambiguity and shrinks the scale of within-class variation. Thus, even exact access to the latent state does not by itself guarantee discriminability, while resolving individual instances at fixed effective precision becomes harder as within-class variation shrinks. We then derive pairwise lower bounds linking recovery error, classification error, representation geometry, and probe Lipschitz constants. These bounds expose a recovery-prediction-stability trilemma: when a learned representation compresses latent directions that a downstream probe must preserve, accuracy can be maintained only at the cost of increased sensitivity. We further characterize a directional mechanism by which variations in distinct latent factors can interfere through a shared representation. Controlled simulations and experiments on four disentanglement benchmarks exhibit the predicted accuracy-recovery tensions and coupled degradation under perturbations. Our results suggest that identifiability guarantees should be complemented by explicit assumptions on task structure, representation geometry, and stability when they are used to explain downstream performance.

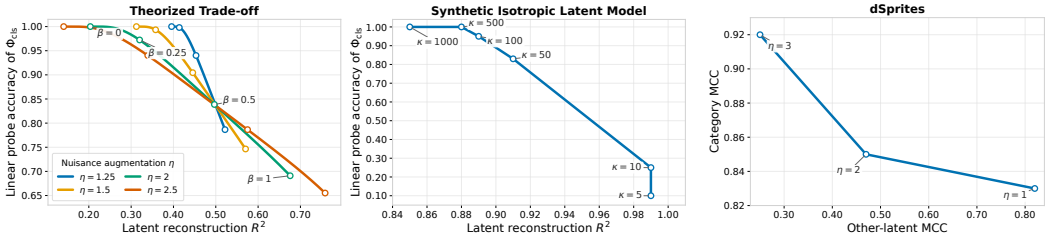


Figure 1: Expected trade-off between downstream performance and invertibility of the generating process as function of augmentation strength η/class concentration κ: (left) as idealized in this paper, (middle) as observed in numerical simulations using the non-linear generator by Reizinger et al. (2025b), and (right) measured in the dSprites disentanglement benchmark Matthey et al. (2017) corrupted with ℓ_∞-noise.

1 INTRODUCTION

Contemporary self-supervised learning (SSL) makes it possible to exploit large amounts of unlabeled data and often yields representations that transfer well to downstream tasks. Contrastive learning (CL) is a central SSL paradigm, yet the mechanism by which its representations become useful downstream remains only partially understood. Motivated by the empirical success of linear probes, recent theories connect CL to nonlinear independent component analysis and establish conditions

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.

under which contrastive objectives recover latent variables, or functions of them, up to specified indeterminacies (Hyvarinen et al., 2019; Zimmermann et al., 2021; Von Kügelgen et al., 2021; Rusak et al., 2025; Reizinger et al., 2025b).

These results characterize what information can be identified under assumptions on the data-generating process (DGP), the augmentations, and the learning objective. Downstream usefulness, however, requires additional structure. A classification task depends on how labels relate to the latent variables, which latent distinctions must be preserved, and which variations should be ignored. Invertibility of the DGP alone does not specify these relationships. Indeed, invariance to position, orientation, color, or other nuisance factors can be beneficial for classification even though those factors remain necessary to reconstruct the complete generative state (Tian et al., 2020; Chen et al., 2020; Von Kügelgen et al., 2021; Wang et al., 2023). Consistently, Rusak et al. (2025) report settings in which recovering previously collapsed information can reduce downstream accuracy. This motivates two questions: *when does latent recovery support downstream discrimination*, and *when do the geometric requirements of recovery and prediction conflict?*

A second issue is stability: even when both recovery and prediction are accurate on clean data, nearby observations or representations may have to be mapped to well-separated targets, forcing a sensitive probe. Viewed as an inverse problem, latent recovery depends on how the generator (and/or the representations) transforms distances of latent factors in observable/representation space. When separated latent states produce nearby observations, accurately recovering both states requires a sensitive inverse map. For a learned representation shared by recovery and classification probes, the consequences depend on which directions each probe uses. In particular, if label-preserving variation acts along a representation direction needed for class discrimination, the classification output cannot remain invariant to that variation. Evaluating recovery and classification only on clean samples can therefore miss dependencies that become apparent under perturbations.

In this work, we study these relationships by distinguishing latent identifiability (i.e, invertibility of the underlying generator), classification accuracy, and robustness. Building on very general anisotropic latent-variable models for contrastive learning Zimmermann et al. (2021); Von Kügelgen et al. (2021); Reizinger et al. (2025b); Rusak et al. (2025), we incorporate class-conditional structure and analyze the precision and sensitivity requirements of recovery and classification. We complement this analysis with controlled generative simulations and unsupervised experiments on four disentanglement benchmarks, where known generative factors allow us to evaluate category performance alongside recovery of other latents. Together, these analyses identify conditions under which recovery and downstream performance diverge, and motivate evaluating identifiability alongside task structure and stability. Our main contributions are:

1. **Separating identifiability from task sufficiency.** We make class-conditional concentration and overlap explicit and relate them to Bayes classification error. This clarifies why recovering the latent state alone does not guarantee downstream discrimination and why concentration can simultaneously improve class separation while shrinking the scale of within-class distinctions.
2. **A recovery-prediction-stability trilemma.** We derive pairwise bounds that jointly relate latent-recovery error, downstream prediction error, representation distances, and probe Lipschitz constants. The bounds distinguish class-preserving compression, which can be harmless for classification but costly for stable recovery, from class-changing compression, which forces either prediction error or sensitivity.
3. **Directional cross-factor interference.** For affine probes, we characterize when variations of distinct latent factors share a representation direction. In this regime, a classifier that must respond strongly to a class-relevant factor cannot remain invariant to a label-preserving factor acting along the same direction.

Experiments with numerical simulations and in four disentanglement learning benchmark datasets yield empirical results consistent with our theoretical claims.

2 Related work

Identifiability in contrastive learning. Recent theory connects contrastive learning to nonlinear independent component analysis (ICA) and establishes recovery of latent variables, or selected latent components, under assumptions on the generator, auxiliary variables, positive-pair distribution, and

admissible indeterminacies (Hyvarinen et al., 2019; Zimmermann et al., 2021; Von Kügelgen et al., 2021; Reizinger et al., 2025b; Rusak et al., 2025). These guarantees answer an identifiability question: which latent information is determined by the learning problem? They do not by themselves imply that the recovered coordinates are useful for a particular downstream task, nor that the inverse is stable. In particular, recovery of the latent state up to an arbitrary invertible transformation is weaker than recovery compatible with a linear probe with respect to a particular downstream task.

Augmentations and downstream structure. Contrastive and related joint-embedding methods encourage agreement across views while avoiding collapse, although they implement these principles differently (Chen et al., 2020; Grill et al., 2020; Chen and He, 2021; Bardes et al., 2022; Zbontar et al., 2021; Garrido et al., 2023; Jha et al., 2024). The information retained is therefore strongly shaped by the augmentation process (Tian et al., 2020; Wang and Isola, 2020; Wang et al., 2023). This creates a basic tension with full latent recovery: a useful representation may intentionally suppress variations that are irrelevant to the task. Consistently, Rusak et al. (2025) report regimes in which recovering previously collapsed information improves identifiability while reducing downstream accuracy.

Recovery versus stability. Our second connection is to inverse problems. Even when a generator is invertible, its inverse can be poorly conditioned: nearby observations or representations may correspond to well-separated latent states. Reconstruction instabilities have been documented in learned inverse problems (Antun et al., 2020), and accuracy-stability trade-offs have been formalized under related inverse-problem assumptions (Gottschling et al., 2025). We use this perspective to separate three properties of a contrastive representation-*recoverability*, *task sufficiency*, and *stability*-and to study how the corresponding probe requirements interact in a shared representation space.

3 Theoretical Setup

Latent model. Let $\mathcal{Z} = \mathcal{Z}^{(1)} \times \cdots \times \mathcal{Z}^{(M)}$ be a d-dimensional latent space and let $g : \mathcal{Z} \to \mathcal{X}$ be a continuous invertible generator, with $\mathbf{x} = g(\mathbf{z})$. We follow a very generic latent model proposed by Rusak et al. (2025) that can instantiate most common models used throughout identifiability theory (Von Kügelgen et al., 2021; Zimmermann et al., 2021; Reizinger et al., 2025b;a), positive views are represented in latent space by

$$\mathbf{z}^+ \mid \mathbf{z} \sim \mathcal{N}(\mathbf{z}, \mathbf{\Lambda}), \qquad \mathbf{\Lambda} = \operatorname{diag}(\sigma_1^2 I_{d_1}, \ldots, \sigma_M^2 I_{d_M}). \tag{1}$$

The block variances encode which factors are preserved or varied by the augmentation mechanism.

Class structure. To study downstream classification, we follow the latent model of Reizinger et al. (2025b), while allowing the anisotropic structure in (1). Let $\mathcal{I} = \{1, \ldots, N\}$ index the training instances, and let $c : \mathcal{I} \to \mathcal{C}$ assign each instance to a class. We assume

$$c(i) \sim \mathcal{U}(\mathcal{C}), \qquad \mathbf{z}_i \mid c(i) \sim p(\cdot \mid \mathbf{z}_{c(i)}), \tag{2}$$

where $\mathbf{z}_{c(i)}$ is a class representative. This class-conditional structure is not part of the contrastive supervision; it is introduced only to characterize when latent recovery can support classification.

For the purposes of this paper it suffices to let $\mathbf{z}_i \sim \mathcal{N}(\mathbf{z}_{c(i)}, \mathbf{\Lambda})$, but in general we need to impose concentration and separation conditions sufficient for low classification error in latent space. For each class c, assume that there exist $\alpha_c \geq 0$ and $\beta_c \in (0, 1)$ such that

$$\Pr\big(\|\mathbf{Z} - \mathbf{z}_c\|_{\mathcal{Z}} \leq \alpha_c \mid C = c\big) \geq 1 - \beta_c.$$

For $c \neq c'$, define the separation between the corresponding concentration balls as

$$s_{cc'} = \|\mathbf{z}_c - \mathbf{z}_{c'}\|_{\mathcal{Z}} - \alpha_c - \alpha_{c'}.$$

If $s_{cc'} > 0$, these balls are disjoint, and a classifier that assigns each ball to its corresponding class can err only outside the correct class's ball. For two equally likely classes this immediately yields a Bayes-error bound

$$R^\star_{cc'} \leq \frac{1}{2} \min\{1, \beta_c + \beta_{c'}\}.$$

Thus, identifiability of $\mathbf{z}$ and discriminability of c are distinct properties: even exact latent recovery cannot eliminate ambiguity arising from overlap between the class-conditional distributions.

Data-generating process. We assume an underlying invertible generator $g : \mathcal{Z} \to \mathcal{X}$ and $\boldsymbol{T} \sim p_{\mathcal{T}}$, we write $\mathbf{x}_i^+ \sim \boldsymbol{T}(\mathbf{x}_i) = g(\mathbf{z}_i)$, expressing augmentation overlap Wang et al. (2023) in $\mathcal{Z}$.

Instance discrimination. Instance discrimination treats each training instance $i \in \mathcal{I}$ as a distinct label (Wu et al., 2018; Balestriero, 2023; Ibrahim et al., 2024). Its simple formulation facilitates both implementation and theoretical analysis (Balestriero, 2023; Gan et al., 2025; Reizinger et al., 2025b; Rusak et al., 2025). For an encoder $f : \mathcal{X} \to \mathcal{H}$, let $\mathbf{h}_i = f(\boldsymbol{T}(\mathbf{x}_i))$, and let $q_i(\mathbf{h})$ denote the probability assigned to instance i by an instance-prediction head q. With uniformly sampled instances, the population objective is

$$\mathcal{L}_{\mathrm{ID}}(f, q) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{T} \sim p_{\mathcal{T}}} \left[-\log q_i(f(\boldsymbol{T}(\mathbf{x}_i))) \right], \qquad (f^\star, q^\star) \in \arg\min_{f,q} \mathcal{L}_{\mathrm{ID}}(f, q).$$

The supervision uses instance indices i, without requiring the downstream labels $c(i)$. If the model can represent the true instance posterior, a global population optimum recovers that posterior almost surely (see Appendix H).

4 The Invertibility argument and its limitations

The *invertibility argument* attributes downstream usefulness to recovery of the underlying latent factors up to an admissible transformation (Zimmermann et al., 2021; Reizinger et al., 2025b; Rusak et al., 2025). In the affine setting, this means we can identify the underlying latent state with

$$\mathbf{z} \approx \boldsymbol{A} f(g(\mathbf{z})) + \boldsymbol{b}. \tag{3}$$

We ask what additional conditions make this recovery useful for accurate and stable prediction. Let $\mathbf{h}_i = f(\mathbf{x}_i)$ with $\mathbf{x}_i = g(\mathbf{z}_i)$. We study the trade-offs in terms of overall performance and stability of a recovery probe $\Phi_{\mathrm{id}} : \mathcal{H} \to \mathcal{Z}$ and a prediction probe $\Phi_{\mathrm{cls}} : \mathcal{H} \to \mathcal{U}$, where each class c has target $\mathbf{u}_c \in \mathcal{U}$.

4.1 Inversion does not imply downstream performance

A simple accuracy-recovery trade-off. Consider two equally likely classes with latent representatives $\mathbf{z}_{c_1} = (-\mu, 0)^\top$ and $\mathbf{z}_{c_2} = (\mu, 0)^\top$, and positive views $\mathbf{z}_i^+ \mid c(i) \sim \mathcal{N}(\mathbf{z}_{c(i)}, \mathrm{diag}(\sigma_1^2, \eta^2))$. Here η controls augmentation along the class-independent nuisance factor, while σ_1 is fixed. We contend that the unconstrained nature of CL likely results in representations where certain directions required to reconstruct $\mathbf{z}$ may compromise the possible linear probe performance downstream. For illustrative purposes, suppose that representations learned in a CL task are of the form $\mathbf{h}_i^+ = f_\beta(g(\mathbf{z}_i^+)) = z_i^{+,(1)} + \beta z_i^{+,(2)}$, with $\beta \in [0, 1]$. Both Φ_{id} and Φ_{cls} act on this representation, so increasing β mixes more nuisance information into the class-relevant coordinate. When $\eta^2 > \mu^2 + \sigma_1^2$, this leads to better affine recovery while reducing classification accuracy. Appendix C provides the complete derivation.

Concentration creates a shrinking within-class scale. Consider the isotropic case in which the class-conditional distribution is von Mises-Fisher (vMF), as commonly used in SSL identifiability models (Zimmermann et al., 2021; Reizinger et al., 2025b). Assume first the ideal case $f^\star = g^{-1}$ and $\mathcal{Z} = \mathbb{S}^{d-1}$ with prototypes $\{\mathbf{z}_c\}_{c \in \mathcal{C}}$. If all class-conditionals coincide, then even perfect recovery of $\mathbf{z}$ cannot reveal the class. Conversely, concentration around sufficiently separated prototypes reduces class overlap; the concentration-radius formulation in the setup above and Appendix J makes this statement precise.

For large concentration κ, a vMF distribution admits a local tangent-space approximation (Sra, 2012; Mardia and Jupp, 2009), and thus for fixed d one can state

$$\mathbf{z}_i \approx \mathbf{z}_{c(i)} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \kappa^{-1} \mathbf{I}_{d-1}) \implies \mathbb{E}\|\mathbf{z}_i - \mathbf{z}_{c(i)}\|^2 \approx \frac{d-1}{\kappa}.$$

Hence the typical within-class scale is $O(\kappa^{-1/2})$. This fact alone does *not* imply a fundamental loss of invertibility: an injective representation may retain arbitrarily small distinctions. Rather, it implies a conditional resolution requirement. If the complete recovery pipeline has an effective absolute error ε_{eff}, then resolving individual deviations requires $\varepsilon_{\text{eff}} = o(\kappa^{-1/2})$. Equivalently, for fixed effective resolution, sufficiently concentrated classes become indistinguishable at the instance level even while class discrimination becomes easier. Table 1 (and Figure 8 in Appendix F) illustrates this opposing behavior for the DGP of Reizinger et al. (2025b).We interpret these experiments as a finite-resolution phenomenon rather than as a claim that concentration intrinsically destroys invertibility. A detailed derivation, including the role of effective numerical, statistical, and optimization resolution, is provided in Appendix K.

Table 1: Empirical linear probe accuracy and latent recovery using the vMF LVM with $|\mathcal{C}| = 100$, $|\mathcal{I}| = 1000$, $d = 5$. Linear probe trained directly on $\mathbf{z}$ to gauge the best possible empirical error. **Green** indicates the best performance; **red** indicates the worst.

κ	**Acc.** $(\Phi_{\text{cls}}(\mathbf{z}_i), c(i))$	$R^2(f(\mathbf{x}_i), \mathbf{z}_i)$
5	**0.10 ± 0.01**	**0.99 ± 0.00**
10	0.25 ± 0.01	**0.99 ± 0.00**
50	0.83 ± 0.02	0.91 ± 0.00
100	0.95 ± 0.01	0.89 ± 0.00
250	0.99 ± 0.00	0.89 ± 0.00
500	**1.00 ± 0.00**	0.88 ± 0.00
1000	**1.00 ± 0.00**	**0.85 ± 0.00**

5 Recovery, Prediction, and Stability

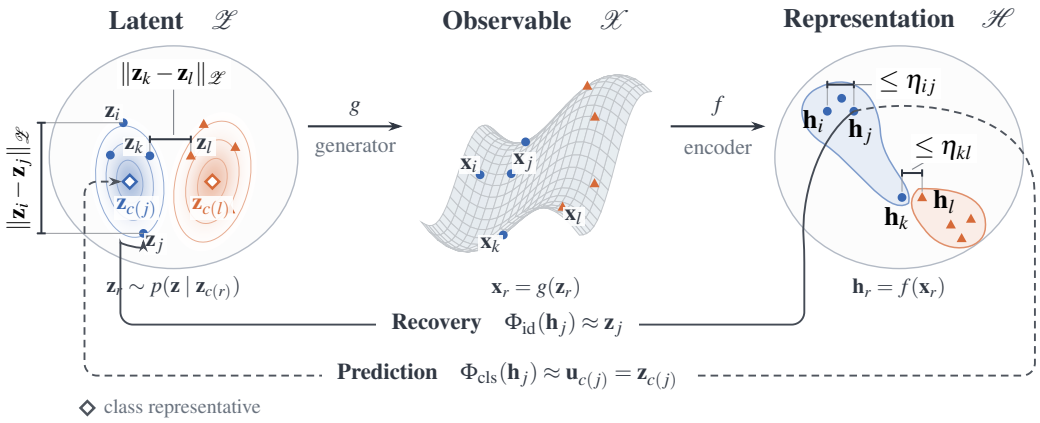


Figure 2: **Illustration of consequences of Theorem 1.** Instances i, j are far apart in latent space but have close representations $\mathbf{h}_i, \mathbf{h}_j$: accurate recovery of the latents can require higher sensitivity of Φ_{id}. Instances k and l belong to different classes yet also have nearby representations $\mathbf{h}_k, \mathbf{h}_l$, so both recovery and prediction can require sensitivity. Here $\eta_{ab} := \gamma\|\mathbf{z}_a - \mathbf{z}_b\|_{\mathcal{Z}}$, with $\gamma > 0$, and prediction uses prototype targets $\mathbf{u}_c = \mathbf{z}_c$.

We further characterize the limitations that stem from the unconstrained nature of a CL task, as illustrated in Figure 2. We start by defining the following pairwise bound isolates the geometric requirement shared by recovery and prediction.

Theorem 1 (Recovery-prediction-stability trilemma). *Let $\mathcal{H}, \mathcal{Z}, \mathcal{U}$ be normed spaces. Suppose that for two instances i, j,*

$$\|\Phi_{\text{id}}(\mathbf{h}_r) - \mathbf{z}_r\|_{\mathcal{Z}} \le \varepsilon_{\text{id}}, \qquad \|\Phi_{\text{cls}}(\mathbf{h}_r) - \mathbf{u}_{c(r)}\|_{\mathcal{U}} \le \varepsilon_{\text{cls}}, \quad r \in \{i, j\}.$$

If Φ_{id} and Φ_{cls} are respectively L_{id}- and L_{cls}-Lipschitz, then

$$L_{\text{id}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge [\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} - 2\varepsilon_{\text{id}}]_+, \tag{4}$$

$$L_{\text{cls}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge [\|\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)}\|_{\mathcal{U}} - 2\varepsilon_{\text{cls}}]_+, \tag{5}$$

where $[a]_+ = \max\{a, 0\}$. Consequently,

$$\boxed{\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge \max\left\{\frac{[\|\mathbf{z}_i - \mathbf{z}_j\| - 2\varepsilon_{\text{id}}]_+}{L_{\text{id}}}, \frac{[\|\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)}\| - 2\varepsilon_{\text{cls}}]_+}{L_{\text{cls}}}\right\}.} \tag{6}$$

Proof. By the triangle inequality,

$$\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} \le 2\varepsilon_{\mathrm{id}} + \|\Phi_{\mathrm{id}}(\mathbf{h}_i) - \Phi_{\mathrm{id}}(\mathbf{h}_j)\|_{\mathcal{Z}} \le 2\varepsilon_{\mathrm{id}} + L_{\mathrm{id}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}}.$$

Rearranging and taking the positive part gives the first inequality. Replacing $(\mathbf{z}_r, \Phi_{\mathrm{id}}, \varepsilon_{\mathrm{id}})$ by $(\mathbf{u}_{c(r)}, \Phi_{\mathrm{cls}}, \varepsilon_{\mathrm{cls}})$ gives the second. Equation (6) follows by taking their maximum. □

Interpretation. For class-preserving variation, $c(i) = c(j)$, the prediction term vanishes: a classifier may collapse this variation without penalty, while accurate stable recovery still requires representation separation whenever $\|\mathbf{z}_i - \mathbf{z}_j\| > 2\varepsilon_{\mathrm{id}}$. For class-changing variation, $c(i) \neq c(j)$, both terms can be active. Thus, if a learning objective brings representations of examples with separated latent or task targets arbitrarily close, then at least one of recovery error, prediction error, or probe sensitivity must increase. This is the sense in which recovery, prediction, and stability form a trilemma[1].

Instability as an empirical side effect. The recovery bound also clarifies the role of latent concentration. If a pair satisfies $\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} > 2\varepsilon_{\mathrm{id}}$, with $\gamma > 0$

$$\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \le \gamma\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} \implies L_{\mathrm{id}} \ge \frac{\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} - 2\varepsilon_{\mathrm{id}}}{\gamma\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}}}.$$

Thus, maintaining a fixed relative recovery error below $1/2$ requires increasingly sensitive probes as representations become more compressed relative to the latents ($\gamma \to 0$). The vMF calculation in Section 4.1 shows that typical same-class latent distances shrink as $O(\kappa^{-1/2})$. This concentration alone does not force γ to decrease. However, if the learned representation further compresses these latent distinctions, accurate recovery requires greater probe sensitivity. Finite effective resolution can additionally limit which distinctions remain recoverable. We hypothesize that these difficulties can arise in practical CL, where optimization does not explicitly enforce stable latent recovery, paralleling accuracy–stability trade-offs studied in inverse problems (Antun et al., 2021; Gottschling et al., 2025).

Directional cross-factor interference. The latent blocks $\mathcal{Z}^{(1)}, \ldots, \mathcal{Z}^{(M)}$ are generatively distinct, but the unconstrained nature of CL means that their effects need not occupy distinct directions in the (learned) representation space. The next result states a sufficient, explicitly conditional mechanism for interference.

Corollary 2 (Cross-factor interference under shared representation directions). *Assume $\Phi_{\mathrm{cls}}(\mathbf{h}) = A_{\mathrm{cls}}\mathbf{h} + \mathbf{a}$ is affine. Suppose that $c(j) \neq c(k)$ with*

$$\|\Phi_{\mathrm{cls}}(\mathbf{h}_j) - \mathbf{u}_{c(j)}\|_{\mathcal{U}} \le \varepsilon_{\mathrm{cls}}, \qquad \|\Phi_{\mathrm{cls}}(\mathbf{h}_k) - \mathbf{u}_{c(k)}\|_{\mathcal{U}} \le \varepsilon_{\mathrm{cls}}.$$

If a label-preserving change in another latent factor produces representations $\mathbf{h}_r, \mathbf{h}_s$ satisfying

$$\mathbf{h}_r - \mathbf{h}_s = t(\mathbf{h}_j - \mathbf{h}_k), \qquad t \neq 0,$$

then

$$\|\Phi_{\mathrm{cls}}(\mathbf{h}_r) - \Phi_{\mathrm{cls}}(\mathbf{h}_s)\|_{\mathcal{U}} \ge |t|\Delta > 0,$$

provided $\Delta := \|\mathbf{u}_{c(j)} - \mathbf{u}_{c(k)}\|_{\mathcal{U}} - 2\varepsilon_{\mathrm{cls}} > 0$. Hence an affine classifier that separates the class-changing pair cannot be invariant in its output to a label-preserving factor acting along the same representation direction.

Proof. The prediction assumptions and the triangle inequality imply

$$\|A_{\mathrm{cls}}(\mathbf{h}_j - \mathbf{h}_k)\|_{\mathcal{U}} \ge \|\mathbf{u}_{c(j)} - \mathbf{u}_{c(k)}\|_{\mathcal{U}} - 2\varepsilon_{\mathrm{cls}} = \Delta.$$

Affinity and the shared-direction assumption therefore give

$$\|\Phi_{\mathrm{cls}}(\mathbf{h}_r) - \Phi_{\mathrm{cls}}(\mathbf{h}_s)\|_{\mathcal{U}} = |t|\,\|A_{\mathrm{cls}}(\mathbf{h}_j - \mathbf{h}_k)\|_{\mathcal{U}} \ge |t|\Delta > 0,$$

where the final inequality follows from $t \neq 0$ and $\Delta > 0$. □

This result is deliberately one-way: it does not claim that an attack targeting one factor must create a shared representation direction. Rather, once such alignment occurs, cross-factor response is unavoidable for an affine probe satisfying the stated margin condition.

In the experimental analysis below, we will test whether analogous trade-offs and coupled sensitivities appear in existing numerical simulations and controlled image data.

[1] Note that the theorem does not assert that every contrastive objective necessarily creates such compression.

6 NUMERICAL EXPERIMENTS

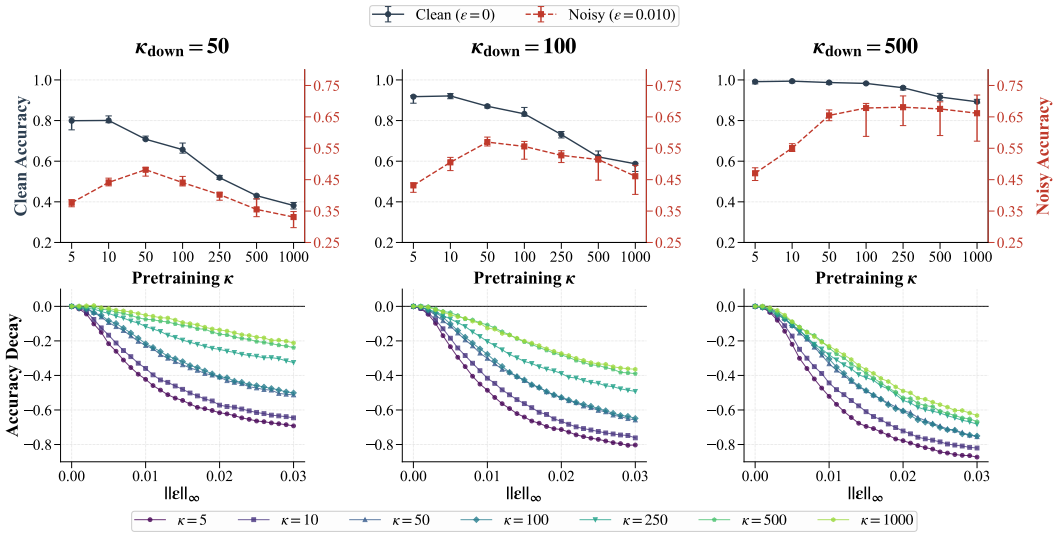


Figure 3: (Top) Downstream accuracy pairs of $(\kappa_{\text{pre}}, \kappa_{\text{down}})$. Although models that have higher R^2 scores in the task of recovering the latent $\mathbf{z}$ (Table 1) tend to be very accurate in the clean validation set, they are also sensitive to small perturbations. (Bottom) Despite being less accurate, as the perturbation increases, the decay in accuracy is less pronounced for models that are pre-trained in higher concentration regimes. On the extreme, this leads to a predictable ordering according to κ_{pre}.

Unstable but accurate linear probe classification We slightly adapt the isotropic DGP proposed by Reizinger et al. (2025b) to account for observing latent factors with varying degrees of concentration during training and validation: κ_{pre} and κ_{down}, respectively (see Fig. 3). Consistent with this picture, as seen on Fig. 3, one can train a very accurate probe even when the latent model has high-overlap (hence, low concentration) in the training stage. In fact, our experiments on clean validation sets systematically place the accuracy of these probes higher than when the model was exposed to a DGP where the instances are more concentrated around their classes. Notwithstanding, this comes at a cost of instability when ℓ_∞-noise is added to the validation set, which is consistent with the sensitivity mechanism in Theorem 1.

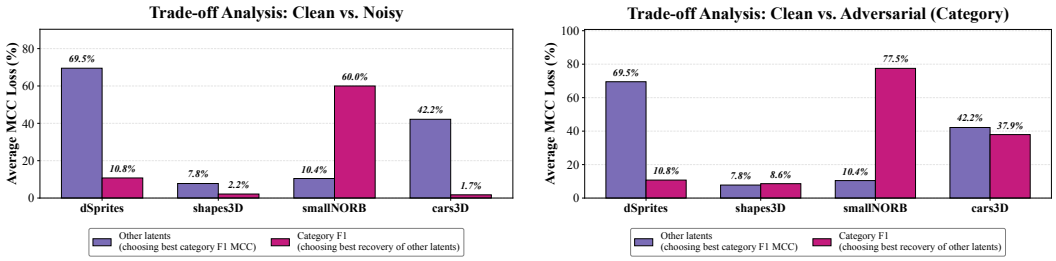
Since our main result is expressed in terms of distances $\|\mathbf{h}_i - \mathbf{h}_j\|$, we conjectured that the generating function had to be artificially pulling samples together in representation space to allow the high clean validation accuracy of low concentration pretraining regimes. We examine this hypothesis in Appendix F, where we relate it to the cone effect of neural networks Liang et al. (2022).

7 DISENTANGLEMENT BENCHMARKS

Evaluating the interplay between identifiability and downstream performance is particularly challenging, since one scarcely has access to the true generating factors of nature Rusak et al. (2025).

To have a more realistic assessment of the implications of the trade-offs we follow Von Kügelgen et al. (2021); Reizinger et al. (2025b) and use the DisLib benchmark Locatello et al. (2019) since one has access to a set of orthogonal latent factors that completely determine the image rendered in the output. We benchmark four different datasets, each with its own diverse set of unique latents and respective images. The main regularity across these tasks is that there is one latent that always dictates the shape of the observed object. The relationship between the prediction of the class factor with respect to recovery of the remaining latent factors is the main focus of this section.

Our f will be a ResNet18 optimized using the DIET instance discrimination framework (**D**atum **I**nd**E**x as its **T**arget class) with its default hyperparameters Ibrahim et al. (2024); Gan et al. (2025). We selected DIET due to its theory-friendly formulation and prior identifiability results, as well as its



(a) ℓ_∞ noise. (b) PGD attack on category latent.

Figure 4: **Accuracy-identifiability trade-offs in disLib Benchmark.** a) Trade-off between other-latent identifiability and category F1 under ℓ_∞ noise, and b) trade-off between other-latent identifiability and category F1 under adversarial category perturbations. **The trade-off is pervasive even if the noise is unbiased.**

straightforward optimization when compared to other more intricate CL frameworks such as, among other, the objectives related to InfoNCE (see appendix H).

Since the category is defined as the shape of highly symmetrical objects across datasets, we designed our experiments to leverage a tuple of augmentations with an increasingly expansive set of compositions of transformations that impose geometrical invariances as well as color-jittering inspired by SimCLR style augmentations Chen et al. (2020) (see description of Table 2). We denote as η the number of admissible transformations $\boldsymbol{T} \in \mathcal{T}$ in each configuration, and associate it loosely with the "strength" of the augmentation set.

We then predict the correlation-based recovery score for each latent factor. For each dataset, we use the mean correlation coefficient (MCC) of the latents that do not control the category as a surrogate for an identifiability score. For each run, the validation is conducted on both a clean and two corrupted versions of the same holdout set. Specifically, we chose ℓ_∞ noise, and an adversarial projected gradient descent attack Madry et al. (2018) supervised with the categorical label for this purpose. The results are collected in Table 2 (complete version of the Pearson correlation score of each latent can be found on Table 3 in Appendix E).

We remark that Reizinger et al. (2025b) studied the identifiability of the factors of this dataset with supervision of the class latent factor, without any augmentations or noise.

Prediction-recovery trade-off. In Fig. 4 we plot the average relative drop in MCC by taking the values recorded in Table 2 and pair the configuration with the best harmonic mean (F1) between the clean and noisy validation sets in terms of the categorical and other latent factors side by side. The trade-off is quite pronounced, even for the subtlest cases like shapes3D, where the most robust model still needs a decrease of around 8% of F1 with respect to the other latents.

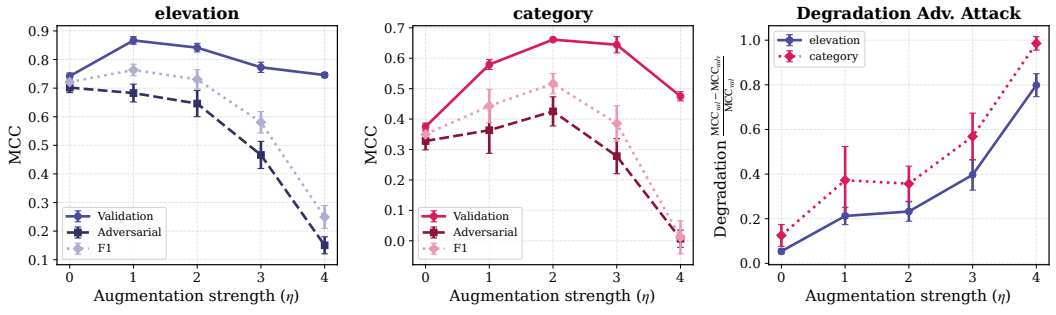


Figure 5: Two latent variables in the cars3D dataset that display poor robustness in terms of their MCC score under an adversarial attack. Their errors display a strong association over several runs and augmentation strengths.

Table 2: Identifiability and category mean correlation coefficient (MCC) on clean, ℓ_∞ and cateogry adversarial noise holdout validation sets for disLib Locatello et al. (2019). Results are reported as mean $\pm$ standard deviation by selecting distinct train/validation splits using three random seeds. The augmentation strength $\eta = i$ denotes the first i transformations from the ordered set $(\{\mathrm{id}_X\}, \{\text{crop}\}, \{\text{crop}, \mathbf{x}\text{-flip}\}, \{\text{crop}, \mathbf{x}\text{-}\mathbf{y}\text{-flip}\}, \{\text{crop}, \mathbf{x}\text{-}\mathbf{y}\text{-flip}, \{\text{color-jitter}\}\})$. **Green** indicates the best performance; **red** indicates the worst.

Dataset	Latent MCC	$\eta = 0$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
dSprites	**other latents**	0.62 ± 0.02	$\mathbf{0.82 \pm 0.01}$	0.47 ± 0.01	$\mathbf{0.25 \pm 0.00}$	N/A
	category (clean)	0.34 ± 0.00	0.93 ± 0.00	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.96 \pm 0.00}$	N/A
	category (noisy)	0.34 ± 0.00	$\mathbf{0.83 \pm 0.04}$	0.85 ± 0.05	$\mathbf{0.92 \pm 0.01}$	N/A
	category (adversarial)	0.34 ± 0.00	$\mathbf{0.83 \pm 0.04}$	0.85 ± 0.05	$\mathbf{0.92 \pm 0.01}$	N/A
shapes3D	**other latents**	0.81 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$
	category (clean)	0.79 ± 0.01	$\mathbf{0.90 \pm 0.02}$	0.92 ± 0.01	0.93 ± 0.00	$\mathbf{0.99 \pm 0.00}$
	category (noisy)	0.78 ± 0.01	$\mathbf{0.88 \pm 0.01}$	0.89 ± 0.04	0.91 ± 0.01	$\mathbf{0.95 \pm 0.01}$
	category (adversarial)	0.72 ± 0.02	$\mathbf{0.82 \pm 0.03}$	0.85 ± 0.03	0.85 ± 0.02	$\mathbf{0.93 \pm 0.03}$
smallNORB	**other latents**	0.54 ± 0.02	$\mathbf{0.67 \pm 0.01}$	0.64 ± 0.01	0.62 ± 0.01	$\mathbf{0.60 \pm 0.01}$
	category (clean)	0.62 ± 0.01	$\mathbf{0.80 \pm 0.01}$	0.82 ± 0.01	0.82 ± 0.01	$\mathbf{0.89 \pm 0.00}$
	category (noisy)	0.62 ± 0.01	0.32 ± 0.10	0.33 ± 0.09	$\mathbf{0.29 \pm 0.13}$	$\mathbf{0.66 \pm 0.03}$
	category (adversarial)	0.50 ± 0.02	0.18 ± 0.10	$\mathbf{0.10 \pm 0.06}$	0.10 ± 0.08	$\mathbf{0.57 \pm 0.09}$
cars3D	**other latents**	0.75 ± 0.01	$\mathbf{0.83 \pm 0.01}$	0.48 ± 0.02	$\mathbf{0.44 \pm 0.03}$	0.52 ± 0.02
	category (clean)	0.37 ± 0.01	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.66 \pm 0.00}$	0.64 ± 0.03	0.59 ± 0.02
	category (noisy)	0.36 ± 0.01	0.57 ± 0.02	$\mathbf{0.65 \pm 0.00}$	0.59 ± 0.04	$\mathbf{0.27 \pm 0.08}$
	category (adversarial)	0.33 ± 0.03	0.36 ± 0.08	$\mathbf{0.43 \pm 0.05}$	0.28 ± 0.06	$\mathbf{0.01 \pm 0.03}$

Instability by coupled sensitivity across factors. Interestingly, cars3D provides a clear example of coupled sensitivity: although elevation and category are both recovered accurately on clean data, a projected-gradient attack optimized only against the categorical factor is accompanied by a substantial degradation of elevation recovery (Fig. 5). We interpret this observation as *consistent with* Corollary 2. Do note that our theoretical arguments do not imply that every category-targeted attack must perturb elevation in this specific case. Rather, it offers a sufficient condition for the association of their errors.

8 Conclusion

This work separates three properties that are often conflated when identifiability theory is used to explain the success of contrastive learning: latent recoverability, downstream task sufficiency, and stability. Class-conditional concentration illustrates the first distinction. Concentration around separated prototypes can reduce class ambiguity while simultaneously shrinking the scale of within-class variation. The latter does not destroy invertibility by itself, but it increases the resolution required to recover individual instances.

Our main geometric result makes the stability issue explicit. For any pair of examples, accurate recovery and accurate prediction impose lower bounds on the representation separation relative to the corresponding probe Lipschitz constants. Class-preserving compression may therefore be harmless for classification while making stable latent recovery difficult, whereas class-changing compression can instead force prediction error or sensitivity. For affine probes, shared representation directions additionally provide a sufficient mechanism for cross-factor interference. The experiments are consistent with these mechanisms: stronger category structure can accompany poorer recovery of other factors, and attacks targeted at one factor can be associated with degradation of another.

These results do not argue against identifiability theory. Rather, they delimit what an identifiability guarantee can establish about downstream usefulness. Explaining practical representation quality additionally requires assumptions on the label structure, the geometry induced by augmentations and optimization, and the stability with which downstream probes access the retained information. Future work should quantify these terms directly, extend the analysis beyond affine probes and classification, and test the pairwise bounds using empirical representation distances and local probe gains.

Reproducibility statement. The assumptions used in the theoretical results are stated in the main text, with supporting derivations and additional examples provided in the appendices. The experimental sections specify the data-generating processes, datasets, representation architecture, training framework, augmentation regimes, probe construction, and perturbation protocols used in our evaluations; additional per-factor results are reported in the supplementary material. Preliminary code and configuration files used to generate the numerical simulations, figures, and benchmark experiments will be included in the anonymous supplementary material. A more refined version will be made public in the moment of publication.

AI use statement We used large language models to assist with editing, condensing, and organizing earlier versions of the manuscript, including consolidating the presentation of theoretical results to improve accessibility for machine learning practitioners. AI assistance also supported the refinement of mathematical statements, the drafting of portions of proofs, and the interpretation and presentation of experimental results. Additionally, these tools assisted with the formatting and visual presentation of plots, tables, and illustrations. The authors take responsibility for the final manuscript and accompanying materials, including all mathematical claims, proofs, experimental results, and code.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

Vegard Antun, Nina M Gottschling, Anders C Hansen, and Ben Adcock. Deep learning in scientific computing: Understanding the instability mystery. *SIAM NEWS MARCH*, 54, 2021.

Randall Balestriero. Unsupervised learning on a diet: Datum index as target free of self-supervision, reconstruction, projector head. *arXiv preprint arXiv:2302.10260*, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *Advances in Neural Information Processing Systems*, 35:37878–37891, 2022.

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

Ihab Bendidi, Adrien Bardes, Ethan Cohen, Alexis Lamiable, Guillaume Bollot, and Auguste Genovesio. No free lunch in self supervised representation learning. *arXiv preprint arXiv:2304.11718*, 2023.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The SSL interplay: Augmentations, inductive bias, and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3252–3298. PMLR, 2023. URL https://proceedings.mlr.press/v202/cabannes23a.html.

Ömer Veysel Çağatan, Ömer Faruk Tal, and M Emre Gursoy. Adversarial robustness of discriminative self-supervised learning in vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2313–2324, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J Maddison. Lossy compression for lossless prediction. *Advances in Neural Information Processing Systems*, 34:14014–14028, 2021.

Eric Gan, Patrik Reizinger, Alice Bizeul, Attila Juhos, Mark Ibrahim, David Klindt, Randall Balestriero, Wieland Brendel, and Baharan Mirzasoleiman. Occam's razor for SSL: Memory-efficient parametric instance discrimination. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=GFNTbsVFlP.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Nina M Gottschling, Vegard Antun, Anders C Hansen, and Ben Adcock. The troublesome kernel: On hallucinations, no free lunches, and the accuracy-stability tradeoff in inverse problems. *SIAM Review*, 67(1):73–104, 2025.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pages 859–868. PMLR, 2019.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

Mark Ibrahim, David Klindt, and Randall Balestriero. Occam's razor for self supervised learning: What is sufficient to learn good representations?, 2024. URL https://arxiv.org/abs/2406.10743.

Abhishek Jha, Matthew B. Blaschko, Yuki M. Asano, and Tinne Tuytelaars. The common stability mechanism behind most self-supervised learning approaches. *arXiv preprint arXiv:2402.14957*, 2024. URL https://arxiv.org/abs/2402.14957.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. From superposition to sparse codes: interpretable representations in neural networks. *arXiv preprint arXiv:2503.01824*, 2025.

Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4367–4378, 2023.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

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

Warren Morningstar, Alex Bijamov, Chris Duvarney, Luke Friedman, Neha Kalibhat, Luyang Liu, Philip Mansfield, Renan Rojas-Gomez, Karan Singhal, Bradley Green, et al. Augmentations vs algorithms: What works in self-supervised learning. *arXiv preprint arXiv:2403.05726*, 2024.

Théo Moutakanni, Maxime Oquab, Marc Szafraniec, Maria Vakalopoulou, and Piotr Bojanowski. You don't need data-augmentation in self-supervised learning. *CoRR*, abs/2406.09294, 2024. URL https://doi.org/10.48550/arXiv.2406.09294.

Fernando Navarro, Christopher Watanabe, Suprosanna Shit, Anjany Sekuboyina, Jan C Peeken, Stephanie E Combs, and Bjoern H Menze. Evaluating the robustness of self-supervised learning in medical imaging. *arXiv preprint arXiv:2105.06986*, 2021.

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

Patrik Reizinger, Randall Balestriero, David Klindt, and Wieland Brendel. Position: An empirically grounded identifiability theory will accelerate self supervised learning research. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82072–82090. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/reizinger25a.html.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. Cross-entropy is all you need to invert the data generating process. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=hrqNOxpItr.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4159–4167. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/rusak25a.html.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

Ravid Shwartz Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Suvrit Sra. Directional statistics in machine learning: A brief review. In *Applied Directional Statistics: Modern Methods and Case Studies*. CRC Press, 2012.

Suvrit Sra. Directional statistics in machine learning: a brief review. *arXiv preprint arXiv:1605.00316*, 2016.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations (ICLR)*, 2023.

Yifei Wang, Kaiwen Hu, Sharut Gupta, Ziyu Ye, Yisen Wang, and Stefanie Jegelka. Understanding the role of equivariance in self-supervised learning. *Advances in Neural Information Processing Systems*, 37:127483–127510, 2024.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38938–38970. PMLR, 2023. URL https://proceedings.mlr.press/v202/xue23d.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised contrastive learning more robust than supervised learning? In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=fuPVTbOEkUj.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pages 12979–12990. PMLR, 2021.

A Related Work

A.1 Contrastive learning and related joint-embedding methods

Contrastive learning (CL) learns representations by discriminating between related and unrelated observations. In the self-supervised setting, related observations typically correspond to augmented views of the same instance, while other instances supply negative examples. Methods such as SimCLR implement this principle through a contrastive objective that promotes agreement between positive views relative to negative pairs (Chen et al., 2020; Radford et al., 2021; Alayrac et al., 2022; Zhai et al., 2023).

Explicit negative pairs are nevertheless not necessary for successful joint-embedding learning Grill et al. (2020); Chen and He (2021); Gan et al. (2025); Siméoni et al. (2025).

Moreover, despite the plurality of SSL models, many popular joint-embedding methods can be understood as addressing a common underlying task: promoting agreement across views while preventing representational collapse (Garrido et al., 2023; Jha et al., 2024). Indeed, Garrido et al. (2023) establish algebraic connections between particular contrastive and covariance-based objectives, showing their equivalence under specified assumptions. From a complementary perspective, Jha et al. (2024) interpret several contrastive and non-contrastive methods through a shared mechanism that promotes consistency across augmentations of an instance while controlling the magnitude of the mean representation over the dataset.

These commonalities motivate studying the methods within a broader joint-embedding perspective, while retaining distinctions between their objectives and theoretical guarantees.

A.2 Identifiability in contrastive learning

Recent advancements bridging non-linear independent component analysis (ICA) (Hyvärinen et al., 2024) to contrastive learning are promising Zimmermann et al. (2021); Reizinger et al. (2025b); Rusak et al. (2025); Reizinger et al. (2025a); Klindt et al. (2025). The idea is that, under reasonable assumptions about a data-generation process (DGP), which is comprised of an invertible non-linear map that emits observations as a function of well-defined latent factors, the minimizer of a contrastive loss coincides with recovering the original latent sources of the distribution up to admissible indeterminacies (e.g., affine transformations). However, there is a lack of consensus on what the precise articulation of the DGP should be or whether these identifiably results are useful in practice (Reizinger et al., 2025a; Klindt et al., 2025). Notably, Rusak et al. (2025) did observe an identifiably-accuracy trade-off in practice, precisely because there was a mismatch of the expectations of the theory with the actual practical experimental setup.

Even going back to the classical non-linear ICA perspective, these methods do not presume any prior knowledge of the joint distribution of the observations and their latents. It is well known that in this case, contrastive models tend to favor the recovery of the sources that dominate the first moments of the data, to the detriment of the inversion of those that dictate higher-order variations (Hyvarinen et al., 2019; Khemakhem et al., 2020).

A.3 AUGMENTATIONS, INVARIANCE, AND DOWNSTREAM STRUCTURE

Contrastive learning encourages agreement between positive views while distinguishing them from negative samples. The alignment and uniformity perspective characterizes two geometric properties associated with this objective (Wang and Isola, 2020). The information retained across views depends on the augmentation process: transformations intended to preserve task-relevant information encourage invariance to other variations (Tian et al., 2020; Chen et al., 2020). Accordingly, invariance and identifiability can be compatible when the recovery target consists of the factors preserved by the augmentations, as in content–style models (Von Kügelgen et al., 2021).

Connecting this view-level objective to class-level performance requires assumptions relating augmentations to downstream labels. Wang et al. (2023) develop an augmentation-overlap perspective that avoids assuming that positive views are conditionally independent given their class. Their analysis explains how overlap between the augmented supports of different same-class instances can connect instance-level alignment to class-level clustering. This highlights the importance of how augmentations relate different instances, beyond their effect on an individual positive pair.

However, augmentation properties and contrastive loss values alone do not fully determine downstream performance. Saunshi et al. (2022) show that the function class and training algorithm must also be considered. Cabannes et al. (2023) analyze the interaction between augmentations, inductive bias, and generalization in kernel regimes, while Xue et al. (2023) show how the simplicity bias of gradient-based training can contribute to class collapse and feature suppression. Together, these results motivate studying the information retained by the learned representation alongside the geometry through which downstream probes access it.

A.4 AN INVERSE-PROBLEM PERSPECTIVE ON STABILITY

Viewing representation learning through latent recovery also connects it to inverse problems. For an observation $\mathbf{x} = g(\mathbf{z})$, an identification map attempts to recover $\mathbf{z}$ from $\mathbf{x}$. Even when g is invertible, its inverse need not satisfy a useful Lipschitz bound: nearby observations may correspond to substantially separated latent states. Accurate reconstruction of such states consequently requires sensitivity to small changes in the observations.

Studies of learned image reconstruction document instabilities under small perturbations (Antun et al., 2020), while Gottschling et al. (2025) establish accuracy–stability trade-offs under specified inverse-problem conditions. These results motivate treating recovery accuracy and stability as separate properties.

We adopt this perspective to study recovery and classification probes acting on a shared representation. Our focus is on how their sensitivity requirements interact along particular representation directions, and on the consequences when label-preserving variation affects directions used for class discrimination. This complements population identifiability analysis with an examination of the precision and stability required to use recovered information downstream.

A.5 FULL LATENT RECOVERY AND TASK-DEPENDENT INVARIANCE

The idea that identifiably could be a design principle in CL implies a degree of *equivariance* to transformations between the representations of observations and their respective latent factors, which is counterintuitive with principles drawn from information theory. A recurring principle is to design the pre-text task so that it imposes *invariance* over a set of possibly non-invertible transforms, so as any variation along the axes that are not desired for the downstream task is minimized Wang and Isola (2020); Tian et al. (2020); Chen et al. (2020); Dubois et al. (2021); Von Kügelgen et al. (2021); Shwartz Ziv and LeCun (2024).

A.6 OVERSIMPLIFIED EFFECT OF DATA AUGMENTATIONS IN IDENTIFIABLY THEORY

As outlined by Reizinger et al. (2025a), current DGPs used to study the identifiably of contrastive learning either do not account for, or are too simplistic with respect to, the effect of the data augmentations. For instance, most DGP models assume that the instances (and respective views) of the observables $\mathbf{x}_i$ are conditionally independent given their labels Zimmermann et al. (2021);

Von Kügelgen et al. (2021); Reizinger et al. (2025a); Rusak et al. (2025). However, as argued by Wang et al. (2023), this claim is easily refuted in practice: views of the same instance should be more similar to one another than to other views generated from other instances present in the dataset. Moreover, if this conditional independence holds, there is a possibility that the learned representation easily degenerates to a trivial solution, i.e., fares no better than random chance for any downstream task (Wang et al., 2023). Empirical evidence supports this: Morningstar et al. (2024) observed that the design of the data augmentations is more relevant for downstream performance than any other algorithmic choices. This sensitivity to the set of augmentations has been verified extensively on empirical studies Zhong et al. (2022); Bendidi et al. (2023); Li et al. (2023).

Furthermore, unlike supervised learning, CL does not have a supervisory signal from the ground truth labels to correct mis-specification of its pre-text task Van Assel et al. (2025). Thus, selecting the wrong set of augmentation can be catastrophic downstream, especially if these transformations are biased towards certain classes or vary as a function of spurious or adversarial regularities in the data Balestriero et al. (2022); Li et al. (2023).

In general, when considering the plethora of augmentations and inductive biases, general downstream performance guarantees of self-supervised methods remain a challenging open problem Saunshi et al. (2022), in spite of promising developments that strive to ameliorate some of these issues Ibrahim et al. (2024); Balestriero and LeCun (2025).

A.7 Toward simpler formulations of representation learning

A related effort in predictive joint-embedding learning is LeJEPA, which combines a predictive objective with Sketched Isotropic Gaussian Regularization (SIGReg) (Balestriero and LeCun, 2025). By explicitly regularizing the embedding distribution, LeJEPA removes mechanisms such as teacher–student asymmetry and stop-gradient operations. Although DIET and LeJEPA use different objectives, both illustrate how making the constraints on representations explicit can simplify training and facilitate theoretical analysis.

We follow this emphasis on fundamental learning principles, focusing on contrastive learning and its instance-discrimination formulation. The connections to related joint-embedding methods motivate the broader relevance of our questions, without implying that contrastive identifiability guarantees automatically apply to every such method. In particular, preventing collapse during training does not itself establish accurate latent recovery, downstream classification performance, or robustness to perturbations. The latter notion of stability is the focus of our analysis and is distinct from the stability of non-collapsed training dynamics.

The scope of the unification. Garrido et al. relate sample- and dimension-contrastive criteria through an algebraic identity between embedding Gram matrices, with normalization determining their equivalence. They also introduce VICReg-exp and VICReg-ctr to interpolate between VICReg and SimCLR, together with squared- and absolute-similarity variants of SimCLR and DCL (Garrido et al., 2023). Their exact results concern specified criteria; connections to asymmetric and prototype-based methods are more indirect.

Complementarily, Jha et al. (2024) analyze SimCLR, BYOL, SimSiam, DINO, SwAV, and Barlow Twins through a common mechanism that encourages consistency across an instance's augmentations while controlling the magnitude of the dataset-wide mean representation. Their experiments also examine SwAV with fixed prototypes and Barlow Twins without its off-diagonal penalty, illustrating how different mechanisms can maintain non-collapsed representations. These findings support a shared conceptual framework, rather than a universal equivalence of objectives, optimization trajectories, or recovered information.

This perspective motivates our focus on contrastive learning and its instance-discrimination formulation. The broader connections suggest that the interaction between invariance, information preservation, and downstream usefulness is relevant across joint-embedding methods. However, avoiding collapse during training is distinct from recovering particular latent factors or remaining robust to perturbations. Our analysis addresses these latter properties under explicitly stated assumptions.

B PROOFS

B.1 PROOF OF THEOREM 1

Proof. We first establish the recovery bound, then apply the same argument to prediction and combine the two inequalities.

Step 1: Relate latent separation to recovery error. Insert and subtract the recovered latent vectors to write

$$\begin{aligned}\mathbf{z}_i - \mathbf{z}_j = \big(\mathbf{z}_i - \Phi_{\mathrm{id}}(\mathbf{h}_i)\big) + \big(\Phi_{\mathrm{id}}(\mathbf{h}_i) - \Phi_{\mathrm{id}}(\mathbf{h}_j)\big) \\ + \big(\Phi_{\mathrm{id}}(\mathbf{h}_j) - \mathbf{z}_j\big).\end{aligned}$$

The triangle inequality therefore gives

$$\begin{aligned}\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} \le \|\mathbf{z}_i - \Phi_{\mathrm{id}}(\mathbf{h}_i)\|_{\mathcal{Z}} + \|\Phi_{\mathrm{id}}(\mathbf{h}_i) - \Phi_{\mathrm{id}}(\mathbf{h}_j)\|_{\mathcal{Z}} \\ + \|\Phi_{\mathrm{id}}(\mathbf{h}_j) - \mathbf{z}_j\|_{\mathcal{Z}}.\end{aligned}$$

Each endpoint recovery error is at most $\varepsilon_{\mathrm{id}}$. Consequently,

$$\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} \le 2\varepsilon_{\mathrm{id}} + \|\Phi_{\mathrm{id}}(\mathbf{h}_i) - \Phi_{\mathrm{id}}(\mathbf{h}_j)\|_{\mathcal{Z}}. \tag{7}$$

Step 2: Use the stability assumption on the recovery probe. Since Φ_{id} is L_{id}-Lipschitz,

$$\|\Phi_{\mathrm{id}}(\mathbf{h}_i) - \Phi_{\mathrm{id}}(\mathbf{h}_j)\|_{\mathcal{Z}} \le L_{\mathrm{id}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}}.$$

Substituting into (7) and subtracting $2\varepsilon_{\mathrm{id}}$ yields

$$L_{\mathrm{id}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge \|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} - 2\varepsilon_{\mathrm{id}}. \tag{8}$$

Step 3: Establish the corresponding prediction bound. Likewise, insert and subtract the two prediction outputs:

$$\begin{aligned}\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)} = \big(\mathbf{u}_{c(i)} - \Phi_{\mathrm{cls}}(\mathbf{h}_i)\big) + \big(\Phi_{\mathrm{cls}}(\mathbf{h}_i) - \Phi_{\mathrm{cls}}(\mathbf{h}_j)\big) \\ + \big(\Phi_{\mathrm{cls}}(\mathbf{h}_j) - \mathbf{u}_{c(j)}\big).\end{aligned}$$

Applying the triangle inequality, bounding each endpoint prediction error by $\varepsilon_{\mathrm{cls}}$, and using the Lipschitz assumption on Φ_{cls} gives

$$\begin{aligned}\|\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)}\|_{\mathcal{U}} &\le 2\varepsilon_{\mathrm{cls}} + \|\Phi_{\mathrm{cls}}(\mathbf{h}_i) - \Phi_{\mathrm{cls}}(\mathbf{h}_j)\|_{\mathcal{U}} \\ &\le 2\varepsilon_{\mathrm{cls}} + L_{\mathrm{cls}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}}.\end{aligned}$$

Rearranging yields

$$L_{\mathrm{cls}}\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge \|\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)}\|_{\mathcal{U}} - 2\varepsilon_{\mathrm{cls}}. \tag{9}$$

Step 4: Combine the two requirements. For $L_{\mathrm{id}}, L_{\mathrm{cls}} > 0$, divide (8) and (9) by their respective Lipschitz constants. Both lower bounds hold for the same representation distance, so

$$\|\mathbf{h}_i - \mathbf{h}_j\|_{\mathcal{H}} \ge \max\left\{\frac{\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} - 2\varepsilon_{\mathrm{id}}}{L_{\mathrm{id}}}, \frac{\|\mathbf{u}_{c(i)} - \mathbf{u}_{c(j)}\|_{\mathcal{U}} - 2\varepsilon_{\mathrm{cls}}}{L_{\mathrm{cls}}}\right\}. \tag{10}$$

The left-hand sides of (8) and (9) are nonnegative. Thus, whenever a right-hand side is negative, it may also be replaced by zero. This proves the positive-part formulation in the theorem as well. □

B.2 PROOF OF COROLLARY 2

Proof. Recall that

$$\Phi_{\mathrm{cls}}(\mathbf{h}) = A_{\mathrm{cls}}\mathbf{h} + \mathbf{a}, \qquad \Delta = \|\mathbf{u}_{c(j)} - \mathbf{u}_{c(k)}\|_{\mathcal{U}} - 2\varepsilon_{\mathrm{cls}} > 0.$$

Step 1: Lower-bound the output separation of the class-changing pair. Insert and subtract the predictions at $\mathbf{h}_j$ and $\mathbf{h}_k$, and apply the triangle inequality:

$$\|\mathbf{u}_{c(j)} - \mathbf{u}_{c(k)}\|_{\mathcal{U}} \leq \|\mathbf{u}_{c(j)} - \Phi_{\text{cls}}(\mathbf{h}_j)\|_{\mathcal{U}} + \|\Phi_{\text{cls}}(\mathbf{h}_j) - \Phi_{\text{cls}}(\mathbf{h}_k)\|_{\mathcal{U}} + \|\Phi_{\text{cls}}(\mathbf{h}_k) - \mathbf{u}_{c(k)}\|_{\mathcal{U}}.$$

The first and third terms are each at most ε_{cls}. Subtracting their upper bounds therefore gives

$$\|\Phi_{\text{cls}}(\mathbf{h}_j) - \Phi_{\text{cls}}(\mathbf{h}_k)\|_{\mathcal{U}} \geq \|\mathbf{u}_{c(j)} - \mathbf{u}_{c(k)}\|_{\mathcal{U}} - 2\varepsilon_{\text{cls}} = \Delta.$$

Because the probe is affine, its offset cancels in the difference:

$$\Phi_{\text{cls}}(\mathbf{h}_j) - \Phi_{\text{cls}}(\mathbf{h}_k) = A_{\text{cls}}(\mathbf{h}_j - \mathbf{h}_k).$$

Hence

$$\|A_{\text{cls}}(\mathbf{h}_j - \mathbf{h}_k)\|_{\mathcal{U}} \geq \Delta. \tag{11}$$

Step 2: Transfer this response to the shared direction. For the label-preserving pair, the assumed alignment is

$$\mathbf{h}_r - \mathbf{h}_s = t(\mathbf{h}_j - \mathbf{h}_k), \qquad t \neq 0.$$

Using affinity, followed by linearity of A_{cls}, yields

$$\begin{aligned}\Phi_{\text{cls}}(\mathbf{h}_r) - \Phi_{\text{cls}}(\mathbf{h}_s) &= A_{\text{cls}}(\mathbf{h}_r - \mathbf{h}_s) \\ &= A_{\text{cls}}\big(t(\mathbf{h}_j - \mathbf{h}_k)\big) \\ &= tA_{\text{cls}}(\mathbf{h}_j - \mathbf{h}_k).\end{aligned}$$

Step 3: Use homogeneity of the norm. Taking norms and applying (11),

$$\begin{aligned}\|\Phi_{\text{cls}}(\mathbf{h}_r) - \Phi_{\text{cls}}(\mathbf{h}_s)\|_{\mathcal{U}} &= |t|\,\|A_{\text{cls}}(\mathbf{h}_j - \mathbf{h}_k)\|_{\mathcal{U}} \\ &\geq |t|\Delta > 0.\end{aligned}$$

The final strict inequality follows from $t \neq 0$ and $\Delta > 0$. Thus the outputs at $\mathbf{h}_r$ and $\mathbf{h}_s$ are different, even though $c(r) = c(s)$. The affine probe cannot be invariant in its output to this label-preserving change. □

C AN INTUITIVE EXAMPLE OF THE TRADE-OFF

Latent factors and positive views. We consider a Euclidean illustration of the Gaussian positive-view model in Eq. (1), with two one-dimensional latent blocks: $M = 2$, $d_1 = d_2 = 1$, and $\mathcal{Z} = \mathcal{Z}^{(1)} \times \mathcal{Z}^{(2)} = \mathbb{R}^2$. For this example, let $\mathcal{C} = \{c_1, c_2\}$, $c(i) \sim \mathcal{U}(\mathcal{C})$, and define the class representatives

$$\mathbf{z}_{c_1} = (-\mu, 0)^\top, \qquad \mathbf{z}_{c_2} = (\mu, 0)^\top, \qquad \mu > 0. \tag{12}$$

We take the unaugmented latent instance to be $\mathbf{z}_i = \mathbf{z}_{c(i)}$ and generate positive views as

$$\mathbf{z}_i^+ = \mathbf{z}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \qquad \boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^2, \sigma_2^2), \tag{13}$$

with perturbations independent of the class and across samples. This specifies the anchor distribution for this Gaussian illustration; it does not use the hyperspherical class-conditional distribution. Let $\mathbf{v} = (0, 1)^\top$ and define

$$\eta := \sigma_2 = \sqrt{\mathbf{v}^\top \boldsymbol{\Lambda} \mathbf{v}}, \qquad \eta > 0. \tag{14}$$

Thus η measures augmentation strength along the second, nuisance factor, while $\sigma_1 \geq 0$ is held fixed. Increasing η changes only nuisance variation and leaves the distribution of the first factor unchanged. The notation $z_i^{+,(m)}$ denotes the scalar coordinate in block $\mathcal{Z}^{(m)}$. The second coordinate is independent of the first coordinate and of $c(i)$. If $\sigma_1 = 0$, the class is determined exactly by the sign of $z_i^{+,(1)}$; for $\sigma_1 > 0$, the class-conditional distributions overlap and $c(i)$ remains the generating label.

Here η is a directional Gaussian standard deviation. In the disentanglement benchmarks, the same symbol indexes augmentation configurations; these are distinct parameterizations of augmentation strength and are not quantitatively equated.

Shared representation directions. Let $g : \mathcal{Z} \to \mathcal{X}$ be the invertible generator and set $\mathbf{x}_i^+ = g(\mathbf{z}_i^+)$. Consider a family of encoders $f_\beta : \mathcal{X} \to \mathcal{H}$, with $\mathcal{H} = \mathbb{R}$, defined by

$$\mathbf{h}_i^+ = f_\beta(\mathbf{x}_i^+) = (1 \quad \beta)\, g^{-1}(\mathbf{x}_i^+) = z_i^{+,(1)} + \beta z_i^{+,(2)}, \qquad \beta \in [0, 1]. \tag{15}$$

We retain the main text's bold notation for the representation, even though $\mathbf{h}_i^+$ is one-dimensional here. Its dependence on β is implicit. The encoder uses the augmented observation and does not require its class label. At $\beta = 0$, it preserves the first latent factor and discards the second. Increasing β introduces more nuisance information into the same representation coordinate. Both Φ_{id} and Φ_{cls} act on this shared representation.

Classification probe and accuracy. Take $\mathcal{U} = \mathbb{R}^2$ and define the affine prediction probe

$$\Phi_{\mathrm{cls}}(\mathbf{h}) = A_{\mathrm{cls}}\mathbf{h} + \mathbf{a}, \qquad A_{\mathrm{cls}} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \mathbf{a} = \mathbf{0}. \tag{16}$$

The two outputs are indexed by c_1, c_2, respectively, and the class prediction is

$$\widehat{c}_i = \arg\max_{c \in \mathcal{C}} \big[\Phi_{\mathrm{cls}}(\mathbf{h}_i^+)\big]_c. \tag{17}$$

The corresponding class targets can be chosen as

$$\mathbf{u}_{c_1} = (\mu, -\mu)^\top, \qquad \mathbf{u}_{c_2} = (-\mu, \mu)^\top, \tag{18}$$

so $\Phi_{\mathrm{cls}}(f_\beta(g(\mathbf{z}_c))) = \mathbf{u}_c$ at each class representative. Conditioned on $c(i)$, the representation satisfies

$$\mathbf{h}_i^+ \mid c(i) \sim \mathcal{N}\Big(z_{c(i)}^{(1)}, \sigma_1^2 + \beta^2\eta^2\Big). \tag{19}$$

For positive conditional variance, the log-likelihood ratio is

$$\log \frac{p(h \mid c(i) = c_2)}{p(h \mid c(i) = c_1)} = \frac{2\mu h}{\sigma_1^2 + \beta^2\eta^2}. \tag{20}$$

Thus the decision rule above is Bayes-optimal given $\mathbf{h}_i^+$. For class c_2, correctness requires $\mu + \varepsilon_i^{(1)} + \beta\varepsilon_i^{(2)} > 0$. The noise term has variance $\sigma_1^2 + \beta^2\eta^2$, and the same probability follows for class c_1 by symmetry. Hence

$$\mathrm{Acc}(\beta, \eta) = \Phi_{\mathcal{N}}\left(\frac{\mu}{\sqrt{\sigma_1^2 + \beta^2\eta^2}}\right), \tag{21}$$

where $\Phi_{\mathcal{N}}$ denotes the standard Gaussian CDF and accuracy is defined as one when the denominator is zero. At fixed $\eta > 0$, accuracy decreases as β increases.

Affine recovery probe. For each β, fit the population-optimal affine recovery probe $\Phi_{\mathrm{id}} : \mathcal{H} \to \mathcal{Z}$ to the augmented latent target $\mathbf{z}_i^+$. The dependence of this probe on β is implicit. Write the marginal latent variances as

$$V_1 := \mathrm{Var}(z_i^{+,(1)}) = \mu^2 + \sigma_1^2, \qquad V_2 := \mathrm{Var}(z_i^{+,(2)}) = \eta^2. \tag{22}$$

The balanced classes and centered perturbations imply $\mathbb{E}\mathbf{z}_i^+ = \mathbf{0}$ and $\mathbb{E}\mathbf{h}_i^+ = 0$. Independence of the latent coordinates gives

$$\mathrm{Var}(\mathbf{h}_i^+) = V_1 + \beta^2 V_2, \qquad \mathrm{Cov}(\mathbf{z}_i^+, \mathbf{h}_i^+) = \begin{pmatrix} V_1 \\ \beta V_2 \end{pmatrix}. \tag{23}$$

Consistently with Eq. (3), write $\Phi_{\mathrm{id}}(\mathbf{h}) = \boldsymbol{A}\mathbf{h} + \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{2\times 1}$ and $\boldsymbol{b} \in \mathbb{R}^2$. Centering gives $\boldsymbol{b} = \mathbf{0}$ at the optimum, and

$$\begin{aligned}\mathbb{E}\|\mathbf{z}_i^+ - \boldsymbol{A}\mathbf{h}_i^+\|_2^2 = V_1 + V_2 - 2\boldsymbol{A}^\top \begin{pmatrix} V_1 \\ \beta V_2 \end{pmatrix} \\ + (V_1 + \beta^2 V_2)\|\boldsymbol{A}\|_2^2.\end{aligned} \tag{24}$$

Setting the gradient with respect to $\boldsymbol{A}$ to zero yields

$$\widehat{\mathbf{z}}_i^+ := \Phi_{\mathrm{id}}(\mathbf{h}_i^+) = \frac{1}{V_1 + \beta^2 V_2} \begin{pmatrix} V_1 \\ \beta V_2 \end{pmatrix} \mathbf{h}_i^+. \tag{25}$$

The coordinate-wise reconstruction errors are

$$\mathbb{E}\big(z_i^{+,(1)} - \widehat{z}_i^{+,(1)}\big)^2 = \frac{\beta^2 V_1 V_2}{V_1 + \beta^2 V_2}, \tag{26}$$

$$\mathbb{E}\big(z_i^{+,(2)} - \widehat{z}_i^{+,(2)}\big)^2 = \frac{V_1 V_2}{V_1 + \beta^2 V_2}. \tag{27}$$

Summing gives the minimum affine mean squared recovery error

$$\mathcal{E}_{\mathrm{id}}(\beta) := \mathbb{E}\|\Phi_{\mathrm{id}}(\mathbf{h}_i^+) - \mathbf{z}_i^+\|_2^2 = \frac{V_1 V_2 (1 + \beta^2)}{V_1 + \beta^2 V_2}. \tag{28}$$

This population mean squared error is distinct from the pointwise error tolerance $\varepsilon_{\mathrm{id}}$ in the main theorem. The pooled coefficient of determination is

$$\begin{aligned} R_\beta^2 &:= 1 - \frac{\mathcal{E}_{\mathrm{id}}(\beta)}{\mathbb{E}\|\mathbf{z}_i^+ - \mathbb{E}\mathbf{z}_i^+\|_2^2} \\ &= \frac{V_1^2 + \beta^2 V_2^2}{(V_1 + \beta^2 V_2)(V_1 + V_2)}. \end{aligned} \tag{29}$$

Differentiation gives

$$\frac{\mathrm{d}R_\beta^2}{\mathrm{d}\beta} = \frac{2\beta V_1 V_2 (V_2 - V_1)}{(V_1 + V_2)(V_1 + \beta^2 V_2)^2}. \tag{30}$$

Consequently, if

$$\eta^2 > \mu^2 + \sigma_1^2, \tag{31}$$

increasing β improves pooled affine recovery while reducing classification accuracy. When $V_1 = V_2$, pooled recovery is constant; when $V_2 < V_1$, it decreases with β.

Connection to shared-direction interference. This construction makes the geometry in Corollary 2 explicit. Changing only the nuisance factor by τ changes the representation by $\beta\tau$, while moving between the class representatives changes it by 2μ. For $\beta > 0$ and $\tau \neq 0$, these changes obey

$$\mathbf{h}_r - \mathbf{h}_s = t(\mathbf{h}_j - \mathbf{h}_k), \qquad t = \frac{\beta\tau}{2\mu}, \tag{32}$$

where j, k denote the class-representative inputs and r, s have the same generating class and first latent coordinate. The resulting prediction change has norm

$$\|\Phi_{\mathrm{cls}}(\mathbf{h}_r) - \Phi_{\mathrm{cls}}(\mathbf{h}_s)\|_2 = \sqrt{2}\,\beta|\tau| = |t|\,\|\mathbf{u}_{c_2} - \mathbf{u}_{c_1}\|_2. \tag{33}$$

Thus nuisance variation acts along the same representation direction that the prediction probe uses to distinguish the classes.

Interpretation and scope. The bottleneck $\mathcal{H} = \mathbb{R}$ is an explicit assumption. Retaining both latent factors in separate coordinates would permit exact recovery and classification based solely on the first factor. The encoder family here is prescribed to isolate this mechanism; the example does not assert that optimizing a contrastive objective must produce this family.

The recovery improvement concerns a pooled, variance-weighted score. Nuisance recovery improves with β, while recovery of the first factor deteriorates. Indeed, their coefficients of determination are

$$R^2_{1,\beta} = \frac{V_1}{V_1 + \beta^2 V_2}, \qquad R^2_{2,\beta} = \frac{\beta^2 V_2}{V_1 + \beta^2 V_2}, \tag{34}$$

whose unweighted mean is $1/2$. Thus a better pooled recovery score does not imply improved recovery of every factor or exact identifiability at $\beta = 1$. The example demonstrates that aggregate affine recovery can improve while downstream classification deteriorates when class-relevant and nuisance variations share a representation direction.

Example: anisotropic latent-variable model. Consider a balanced binary classification problem with $Y \in \{-1, +1\}$ and class means

$$\mathbf{c}_+ = (\mu, \mu), \qquad \mathbf{c}_- = (-\mu, -\mu), \qquad \mu > 0.$$

Define the orthonormal directions

$$u = \frac{1}{\sqrt{2}}(1, 1), \qquad v = \frac{1}{\sqrt{2}}(1, -1).$$

The direction u separates the class means, since

$$\langle \mathbf{c}_+, u \rangle = \sqrt{2}\mu, \qquad \langle \mathbf{c}_-, u \rangle = -\sqrt{2}\mu,$$

whereas

$$\langle \mathbf{c}_+, v \rangle = \langle \mathbf{c}_-, v \rangle = 0.$$

Thus, v describes class-internal variation that is irrelevant to the label.

Let the augmented latent variable be

$$\mathbf{z}^+ = Y\sqrt{2}\mu\, u + \varepsilon_u u + \varepsilon_v v, \tag{35}$$

where

$$\varepsilon_u \sim \mathcal{N}(0, \sigma_u^2), \qquad \varepsilon_v \sim \mathcal{N}(0, \sigma_v^2)$$

are mutually independent and independent of Y. We assume $\sigma_u \ll \sigma_v$, so most within-class variation lies along v. The conditional covariance is

$$\mathbf{\Lambda} = \sigma_u^2 u u^\top + \sigma_v^2 v v^\top.$$

Consider the representation

$$\mathbf{h}_\beta := f_\beta(g(\mathbf{z}^+)) = \big(\langle \mathbf{z}^+, u \rangle, \beta \langle \mathbf{z}^+, v \rangle\big) = \Big(Y\sqrt{2}\mu + \varepsilon_u, \beta\varepsilon_v\Big), \qquad \beta \in [0, 1]. \tag{36}$$

Conditioned on the class,

$$\mathbf{h}_\beta = \begin{cases} \big(\sqrt{2}\mu + \varepsilon_u,\ \beta\varepsilon_v\big), & Y = +1, \\ \big(-\sqrt{2}\mu + \varepsilon_u,\ \beta\varepsilon_v\big), & Y = -1. \end{cases} \tag{37}$$

At $\beta = 0$, the class-internal component is discarded, while variability along u is retained. At $\beta = 1$, the representation preserves the complete latent variable, expressed in the orthonormal basis (u, v).

To measure reconstruction in the original latent coordinates, first consider the fixed decoder

$$r(a, b) = a\, u + b\, v.$$

It gives

$$\widetilde{\mathbf{z}}_\beta := r(\mathbf{h}_\beta) = Y\sqrt{2}\mu\, u + \varepsilon_u u + \beta\varepsilon_v v. \tag{38}$$

Consequently,

$$\widetilde{\mathbf{z}}_\beta - \mathbf{z}^+ = -(1-\beta)\varepsilon_v v,$$
$$\left\|\widetilde{\mathbf{z}}_\beta - \mathbf{z}^+\right\|_2 = (1-\beta)|\varepsilon_v|, \tag{39}$$
$$\mathbb{E}\left[\left\|\widetilde{\mathbf{z}}_\beta - \mathbf{z}^+\right\|_2^2\right] = (1-\beta)^2\sigma_v^2. \tag{40}$$

Thus, increasing β reduces the reconstruction error under this fixed decoder.

Since the classes are balanced, $\mathbb{E}[\mathbf{z}^+] = \mathbf{0}$. Using the orthonormality of u and v,

$$\mathbb{E}\left[\left\|\mathbf{z}^+ - \mathbb{E}[\mathbf{z}^+]\right\|_2^2\right] = \mathbb{E}\left[\left(Y\sqrt{2}\mu + \varepsilon_u\right)^2 + \varepsilon_v^2\right]$$
$$= 2\mu^2 + \sigma_u^2 + \sigma_v^2. \tag{41}$$

The population coefficient of determination, defined using total squared error and total latent variance, is therefore

$$R^2_{\beta,\mathrm{dir}} := 1 - \frac{\mathbb{E}\left[\|\widetilde{\mathbf{z}}_\beta - \mathbf{z}^+\|_2^2\right]}{\mathbb{E}\left[\|\mathbf{z}^+ - \mathbb{E}[\mathbf{z}^+]\|_2^2\right]} = 1 - \frac{(1-\beta)^2\sigma_v^2}{2\mu^2 + \sigma_u^2 + \sigma_v^2}. \tag{42}$$

In particular, $R^2_{1,\mathrm{dir}} = 1$, whereas

$$R^2_{0,\mathrm{dir}} = \frac{2\mu^2 + \sigma_u^2}{2\mu^2 + \sigma_u^2 + \sigma_v^2}.$$

Now consider the linear classifier

$$\widehat{Y}(\mathbf{h}) = \begin{cases} +1, & h_1 > 0, \\ -1, & h_1 \le 0. \end{cases} \tag{43}$$

This is also the nearest-class-mean classifier in representation space, whose class means are $(\pm\sqrt{2}\mu, 0)$. For $Y = +1$, the prediction is correct precisely when

$$\sqrt{2}\mu + \varepsilon_u > 0 \quad \Longleftrightarrow \quad \varepsilon_u > -\sqrt{2}\mu. \tag{44}$$

Hence, for $\sigma_u > 0$,

$$\Pr(\widehat{Y} = +1 \mid Y = +1) = \Pr\left(\frac{\varepsilon_u}{\sigma_u} > -\frac{\sqrt{2}\mu}{\sigma_u}\right)$$
$$= 1 - \Phi_{\mathcal{N}}\left(-\frac{\sqrt{2}\mu}{\sigma_u}\right)$$
$$= \Phi_{\mathcal{N}}\left(\frac{\sqrt{2}\mu}{\sigma_u}\right), \tag{45}$$

where $\Phi_{\mathcal{N}}$ is the standard Gaussian cumulative distribution function. For $Y = -1$, correctness requires $\varepsilon_u \le \sqrt{2}\mu$, which gives the same probability. The overall accuracy is therefore

$$\mathrm{Acc}(\beta, \sigma_u, \sigma_v) = \Phi_{\mathcal{N}}\left(\frac{\sqrt{2}\mu}{\sigma_u}\right), \qquad \beta \in [0,1], \quad \sigma_u > 0, \tag{46}$$

with accuracy equal to 1 when $\sigma_u = 0$. Thus, accuracy depends only on the distance $\sqrt{2}\mu$ from either class mean to the decision boundary and the noise standard deviation $\sigma_u = \sqrt{u^\top \mathbf{\Lambda} u}$ along the decision direction. Neither β nor σ_v enters this probability.

The reconstruction score in Eq. (42) must be distinguished from recovery by an unrestricted affine probe. For every $\beta > 0$, the linear decoder

$$r_\beta(a, b) = a\, u + \frac{b}{\beta}\, v \tag{47}$$

satisfies $r_\beta(\mathbf{h}_\beta) = \mathbf{z}^+$ exactly. Thus, the optimal affine-probe reconstruction score is 1 for every $\beta > 0$. However, its Euclidean Lipschitz constant is

$$\mathrm{Lip}(r_\beta) = \max\{1, \beta^{-1}\} = \beta^{-1}, \qquad 0 < \beta \leq 1, \tag{48}$$

so perturbations in the second representation coordinate are amplified as β decreases. At $\beta = 0$, the discarded variable ε_v is independent of the representation; the minimum achievable expected squared reconstruction error is then σ_v^2.

D EXPERIMENTAL SETUP

D.1 NUMERICAL SIMULATIONS

For the numerical simulations presented in the main paper, we instantiate the isotropic von Mises–Fisher latent variable model using the settings that yielded the highest identifiability (i.e., invertibility) scores in Reizinger et al. (2025b). Namely, we set $d = 5$, $|\mathcal{I}| = 1000$, and $|\mathcal{C}| = 100$. The generator function g is an invertible multilayer perceptron (MLP) with five layers and leaky ReLU activation functions. Its weight matrices $\boldsymbol{W}_i \in \mathbb{R}^{d\times d}$ are random but well-conditioned, with ℓ_2-normalized rows.

We train an MLP encoder f with layer output dimensions $(d \times 10, d \times 50, d \times 50, d \times 50, d \times 50, d \times 10, d)$. Batch normalization is applied to each hidden layer.

At each of the 150 training epochs, we generate $|\mathcal{I}|$ i.i.d. observations. We use the DIET framework to train the encoder with the training settings prescribed by Gan et al. (2025). Specifically, we use a linear projector $\boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{|\mathcal{I}|\times d}$ that is ℓ_2-normalized at each training step. We use the AdamW optimizer with a mini-batch size of 32 and a learning rate of 1×10^{-4}, with 15 epochs of warmup using a linear learning rate scheduler.

For evaluation, we draw 5096 i.i.d. samples, of which 1000 are used for validation. For the classification task, we train a linear probe for 100 epochs or until convergence, using 10% of the training data as a held-out set. Robustness is evaluated on the same fixed samples selected from the validation set, perturbed with random ℓ_∞ noise. Noise magnitude was set to $8/255$ for RGB images, and $4/255$, otherwise.

To measure invertibility, we fit a linear least-squares probe to predict $\mathbf{z}$ from $f(g(\mathbf{z}))$ and compute the coefficient of determination, R^2, between the predicted and true latents.

D.2 DISENTANGLEMENT EXPERIMENTS

We use the same training settings, but replace the MLP encoder f with a ResNet-18. During validation on each dataset, we compute the mean correlation coefficient (MCC) between the representations and the latent factors. We use a projected gradient descent attack conditioned on the categorical latent factor of each dataset to generate adversarial samples.

E EXTENDED TABLE ON DISLIB

Table 3: Identifiability and category mean correlation coefficient (MCC) on clean, ℓ_∞ and category adversarial noise holdout validation sets for disLib Locatello et al. (2019). Results are reported as mean $\pm$ standard deviation by selecting distinct train/validation splits using three random seeds. The augmentation strength $\eta = i$ denotes the first i transformations from the ordered set $(\{\mathrm{id}_X\}, \{\text{crop}\}, \{\text{crop}, \mathbf{x}\text{-flip}\}, \{\text{crop}, \mathbf{x}\text{-}\mathbf{y}\text{-flip}\}, \{\text{crop}, \mathbf{x}\text{-}\mathbf{y}\text{-flip}, \{\text{color-jitter}\}\})$. **Green** indicates the best performance; **red** indicates the worst.

Dataset	Latent MCC	$\eta = 0$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
dSprites	scale	0.70 ± 0.03	0.94 ± 0.00	0.94 ± 0.01	0.95 ± 0.00	-
	orientation	0.10 ± 0.02	0.46 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	-
	x coordinate	0.85 ± 0.00	0.94 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	-
	y coordinate	0.84 ± 0.01	0.94 ± 0.01	0.94 ± 0.00	0.05 ± 0.00	-
	category (clean)	0.34 ± 0.00	0.93 ± 0.00	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.96 \pm 0.00}$	-
	category (noisy)	0.34 ± 0.00	$\mathbf{0.83 \pm 0.04}$	0.85 ± 0.05	$\mathbf{0.92 \pm 0.01}$	-
	category (adversarial)	0.34 ± 0.00	$\mathbf{0.83 \pm 0.04}$	0.85 ± 0.05	$\mathbf{0.92 \pm 0.01}$	-
shapes3D	floor color	0.84 ± 0.05	0.86 ± 0.02	0.88 ± 0.03	0.91 ± 0.01	0.79 ± 0.02
	wall color	0.81 ± 0.00	0.87 ± 0.02	0.89 ± 0.00	0.91 ± 0.01	0.92 ± 0.01
	object color	0.75 ± 0.03	0.83 ± 0.02	0.83 ± 0.03	0.83 ± 0.07	0.84 ± 0.00
	size	0.81 ± 0.02	0.93 ± 0.00	0.92 ± 0.00	0.94 ± 0.01	0.98 ± 0.00
	azimuth	0.82 ± 0.03	0.91 ± 0.02	0.88 ± 0.02	0.91 ± 0.01	0.62 ± 0.06
	category (clean)	0.79 ± 0.01	$\mathbf{0.90 \pm 0.02}$	0.92 ± 0.01	0.93 ± 0.00	$\mathbf{0.99 \pm 0.00}$
	category (noisy)	0.78 ± 0.01	$\mathbf{0.88 \pm 0.01}$	0.89 ± 0.04	0.91 ± 0.01	$\mathbf{0.95 \pm 0.01}$
	category (adversarial)	0.72 ± 0.02	$\mathbf{0.82 \pm 0.03}$	0.85 ± 0.03	0.85 ± 0.02	$\mathbf{0.93 \pm 0.03}$
smallNORB	instance	0.33 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.49 ± 0.01	0.58 ± 0.01
	elevation	0.68 ± 0.02	0.85 ± 0.01	0.85 ± 0.01	0.83 ± 0.01	0.76 ± 0.01
	rotation	0.31 ± 0.01	0.42 ± 0.02	0.28 ± 0.01	0.25 ± 0.01	0.35 ± 0.01
	lighting	0.83 ± 0.02	0.92 ± 0.00	0.92 ± 0.00	0.92 ± 0.00	0.72 ± 0.01
	category (clean)	0.62 ± 0.01	$\mathbf{0.80 \pm 0.01}$	0.82 ± 0.01	0.82 ± 0.01	$\mathbf{0.89 \pm 0.00}$
	category (noisy)	0.62 ± 0.01	0.32 ± 0.10	0.33 ± 0.09	$\mathbf{0.29 \pm 0.13}$	$\mathbf{0.66 \pm 0.03}$
	category (adversarial)	0.50 ± 0.02	0.18 ± 0.10	$\mathbf{0.10 \pm 0.06}$	0.10 ± 0.08	$\mathbf{0.57 \pm 0.09}$
cars3D*	elevation	0.74 ± 0.01	0.87 ± 0.01	0.84 ± 0.02	0.77 ± 0.02	0.89 ± 0.01
	azimuth	0.76 ± 0.00	0.79 ± 0.01	0.11 ± 0.02	0.11 ± 0.03	0.14 ± 0.02
	category (clean)	0.37 ± 0.01	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.66 \pm 0.00}$	0.64 ± 0.03	0.59 ± 0.02
	category (noisy)	0.36 ± 0.01	0.57 ± 0.02	$\mathbf{0.65 \pm 0.00}$	0.59 ± 0.04	$\mathbf{0.27 \pm 0.08}$
	category (adversarial)	0.33 ± 0.03	0.36 ± 0.08	$\mathbf{0.43 \pm 0.05}$	0.28 ± 0.06	$\mathbf{0.01 \pm 0.03}$

F THE CONE EFFECT OF EXISTING DGPS FOR GENERAL SSL

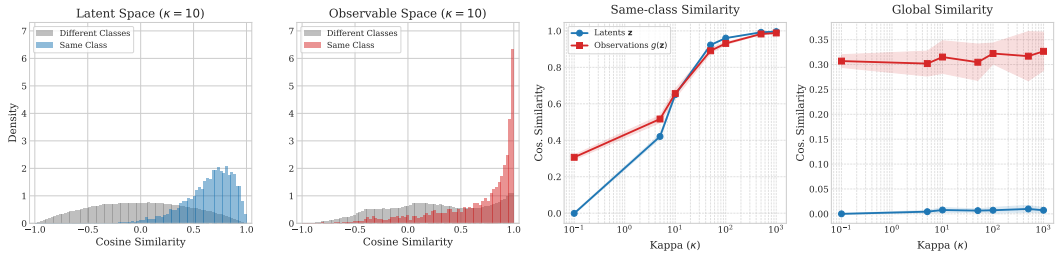


Figure 6: Even though the generators are well-conditioned, invertible, and relatively shallow MLPs, the cone-effect is still prevasive enough to distort the distances in observable spaces. Not only are classes from the same class artificially concentrated in observation space, but also the overall similarity between pairs generated from diferent class anchors is increased. For lower concentration values, we measure that the expected distance between observables of the same class to be greatly increased with respect to the expected marginal similarity between any two pairs, which likely drives classification performance on probes learned on top of $\tilde{f}$.

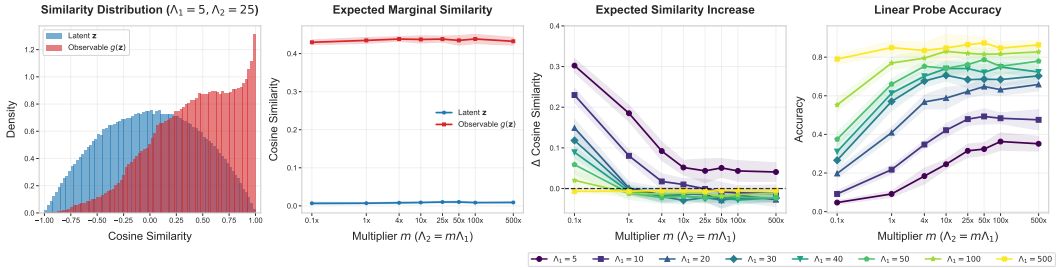


Figure 7: Cosine-similarity of $\mathbf{z}$ vs. $g(\mathbf{z})$ using the anisotropic latent model proposed by Rusak et al. (2025) with covariance matrix $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_1, \Lambda_2, \ldots, \Lambda_2)$. (Left-most) The unconditional similarity distribution is still biased towards higher similarities by the forward pass on g. (Middle) The effect is prevasive for various possible combinations of Λ_1 with multiple degrees of anisotropy controlled by $\Lambda_2 = m\Lambda_1$. (Right) Linear probe accuracy depends on Λ_1 and m on the latent space.

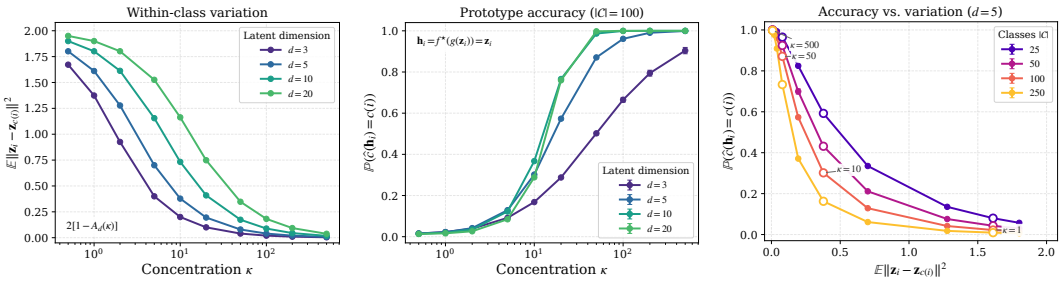


Figure 8: Analysis of the vMF model, showing that increasing κ reduces within-class variation and improves nearest-prototype classification accuracy under exact latent recovery, $\mathbf{h}_i = f^\star(g(\mathbf{z}_i)) = \mathbf{z}_i$.

The cone-effect is a phenomenon first described in multimodal contrastive learning Liang et al. (2022), where the cosine similarity of any pair of vectors increases as they pass each non-linear layer of an arbitrary neural network. Here, we extend the theoretical results of Liang et al. (2022) to include Leaky Relu activation functions to show that this phenomena is also pervasive in the MLP used for g. In Figures 6 and 7 we visualize the impact of this effect on both the isotropic and anisotropic latent variable models. Additionally, in Figure 8 we see that the cone effect is persistent across multiple dimensions d and number of classes $C = |\mathcal{C}|$.

Theorem 3 (Cone effect on leaky-ReLU layers)**.** *Let $\phi_a(t) = \max\{t, at\}$ act coordinate-wise, with $a \in (0,1)$, and let $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ have independent entries $W_{ij} \sim \mathcal{N}(0, 1/d)$. For any fixed nonzero inputs $\boldsymbol{x}, \boldsymbol{y}$ with $\rho := \cos(\boldsymbol{x}, \boldsymbol{y}) < 1$,*

$$d_{\cos}\big(\phi_a(\boldsymbol{W}\boldsymbol{x}), \phi_a(\boldsymbol{W}\boldsymbol{y})\big) < d_{\cos}(\boldsymbol{x}, \boldsymbol{y})$$

with probability at least $1 - C(a,\rho)/d$, where $C(a,\rho)$ is finite and independent of d.

Proof. By positive homogeneity, assume $\|\boldsymbol{x}\| = \|\boldsymbol{y}\| = 1$. The pairs

$$(U_i, V_i) := \sqrt{d}\,\big((\boldsymbol{W}\boldsymbol{x})_i, (\boldsymbol{W}\boldsymbol{y})_i\big)$$

are i.i.d. standard Gaussian pairs with correlation ρ. Write

$$\phi_a(t) = bt + c|t|, \qquad b = \frac{1+a}{2}, \quad c = \frac{1-a}{2}.$$

Joint sign symmetry makes the mixed terms vanish, giving

$$m := \mathbb{E}[\phi_a(U_i)\phi_a(V_i)] = b^2\rho + c^2\mathbb{E}[|U_iV_i|], \qquad q := \mathbb{E}[\phi_a(U_i)^2] = b^2 + c^2.$$

Since $\rho < 1$, the product U_iV_i is negative with positive probability. Hence

$$\mathbb{E}[|U_iV_i|] > \mathbb{E}[U_iV_i] = \rho, \qquad \text{so} \qquad \frac{m}{q} > \rho.$$

The output cosine is

$$\frac{d^{-1}\sum_i \phi_a(U_i)\phi_a(V_i)}{\sqrt{\big(d^{-1}\sum_i \phi_a(U_i)^2\big)\big(d^{-1}\sum_i \phi_a(V_i)^2\big)}}.$$

Each empirical average has variance of order d^{-1}. By Chebyshev's inequality and a union bound, these averages lie within any fixed positive tolerance of m, q, q with probability $1 - O(d^{-1})$. Since $q > 0$ and $m/q > \rho$, a sufficiently small tolerance ensures that the output cosine exceeds ρ. The claimed decrease in cosine distance follows. □

In Figure 7 we also include the same experiment on learning the linear probe directly on $\mathbf{z}$. Indeed, for both cases, we observed that the cosine similarity of any two latents to be skewed to increase substantially. Interesting, although both increase on average despite the constraints of the latent variable model (modeled by κ on Figure 8 and $\Lambda_2 = m\Lambda_1$ on Figure 7), the expected cosine similarity of pairs of the same class seemed to surpass that of expected marginal similarity in the observation space, which should elucidate the deceptively high accuracy on noiseless samples.

G ADDITIONAL RESULTS

G.1 VISUALIZATION OF THE TRADE-OFF IN DSPRITES

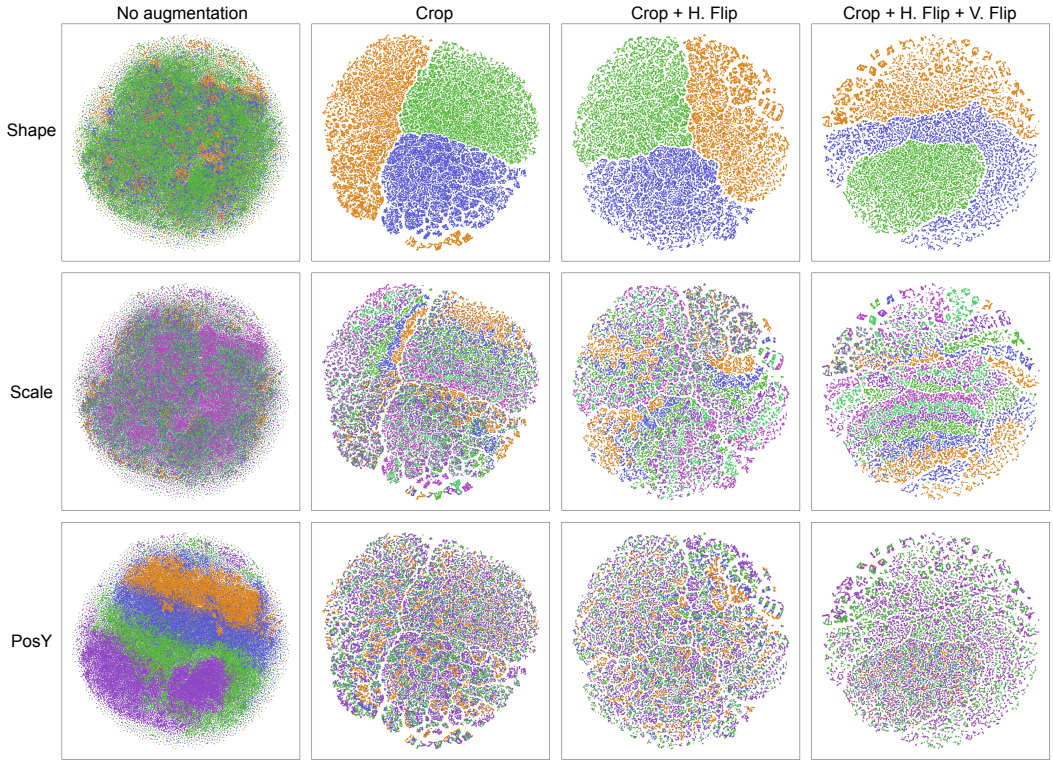


Figure 9: t-SNE van der Maaten and Hinton (2008) for the validation set of one of the seeds for increasingly stronger augmentations (i.e., higher η) in the dSprites dataset.

To illustrate how factors become entangled we additionally perform a t-SNE analysis of the embeddings subject to increasingly stronger augmentations (denoted as η in the main text) with respect to several latent factors on the dSprites dataset. By looking at the rightmost column of Figure 9, it is possible to see that the categorical latents become well separated, whereas such structure is not preserved on the scale and y-position factors. The trend is reversed by looking at the y-position factor. Structure over this latent only emerges in the embedding visualization when no augmentation is performed. We conjecture that this is due to the convolutional inductive bias of the ResNet-18. As soon as one enforces scale symmetries by enforcing random crops, this structure is immediately lost.

G.2 VISUALIZATION OF THE TRADE-OFFS IN DISLIB

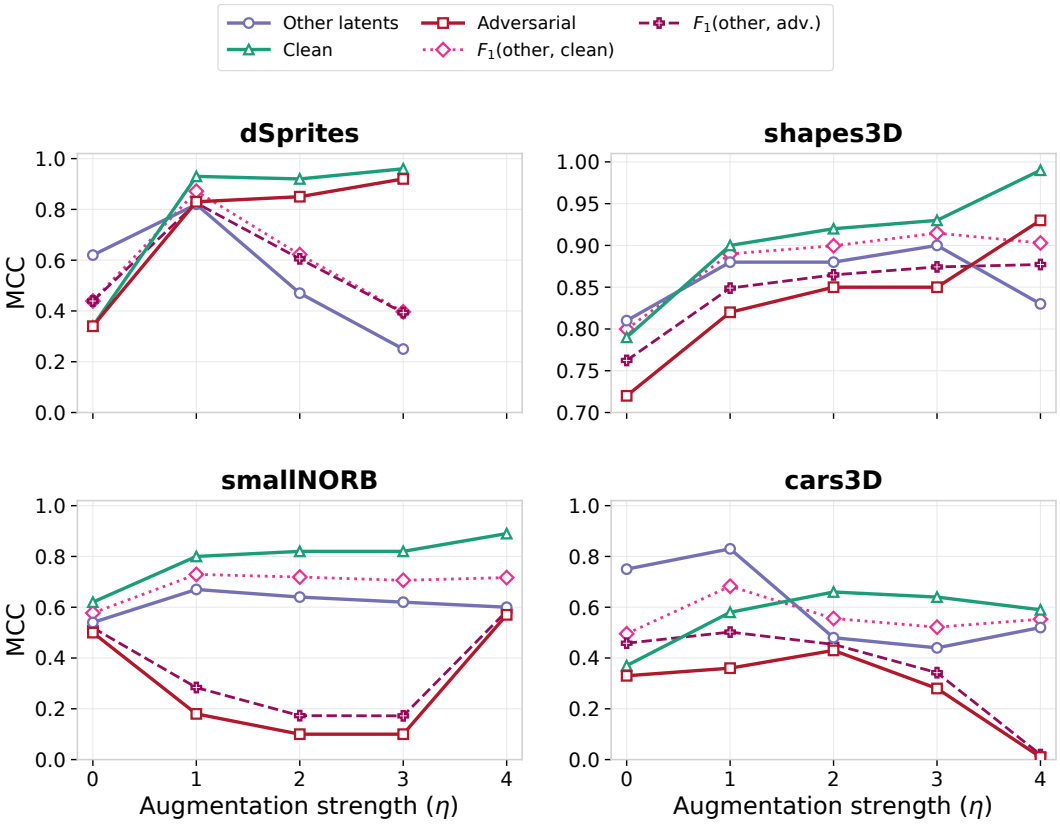


Figure 10: F1 measure between the MCC of the latents in the clean and noisy datasets versus the MCC of the remainder of the latent factors.

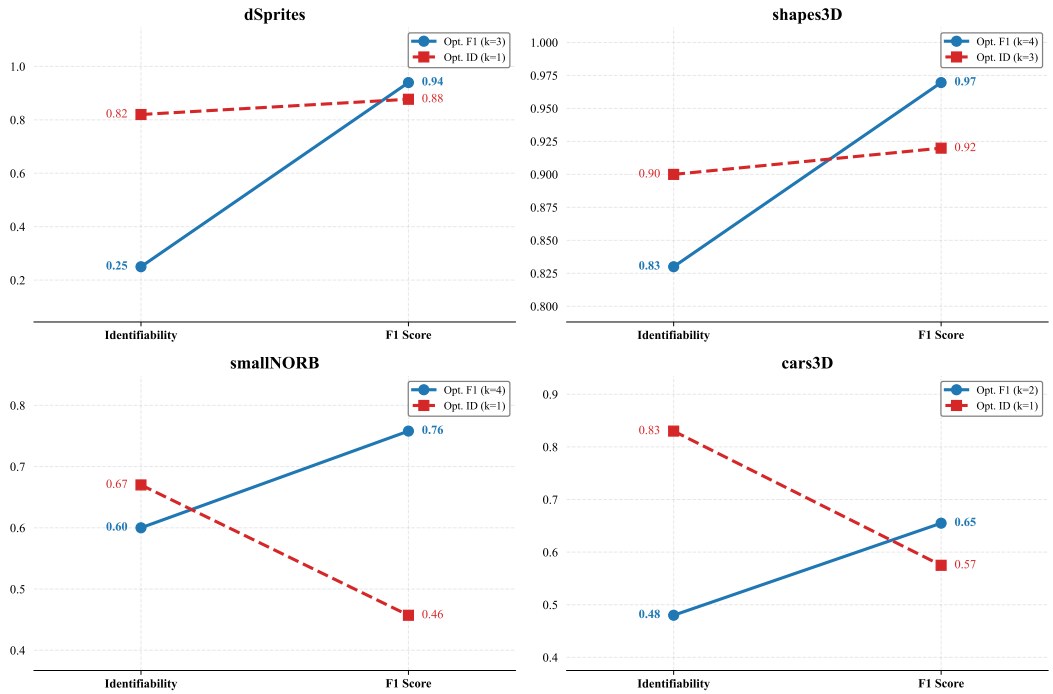


Figure 11: Invertibility/identifiability versus stability (downstream) trade-off cross across the disLib benchmark.

G.3 Label-changing augmentations and noise-induced regularization

Table 4: Effect of label-flip probability on regression performance and identifiability. A precision of 4 decimal places was extraordenarly used to make the trade-off between robustness and identifiability clearer. Note that κ is the same downstream.

	Classification R^2				Invertibility R^2	
p_{flip}	$\epsilon = 0$		$\epsilon = 0.03$		SL	SSL
	SL	SSL	SL	SSL		
0.0	0.9256	**0.9244**	**0.2289**	**0.2605**	**0.9931**	**0.9824**
0.1	0.9285	0.9152	0.2572	0.7325	0.9920	0.9602
0.2	**0.9308**	0.9013	0.3873	0.8076	0.9910	0.9455
0.3	0.9291	0.8894	0.5094	0.8024	0.9869	0.9248
0.4	0.9234	0.8885	0.7188	0.8267	0.9783	0.9252
0.5	**0.9179**	**0.8628**	**0.8008**	**0.8575**	**0.9642**	**0.8819**

Our analysis has focused on augmentations that preserve the downstream label. In practice, however, transformations such as random crops and cutouts may remove class-relevant information and therefore violate this assumption. Nevertheless, cropping can support strong self-supervised representations, even without domain-specific augmentations when training at sufficient scale Moutakanni et al. (2024). This raises the question of whether occasional violations of label preservation can themselves be beneficial. We investigate one possible mechanism: a regularizing effect associated with noisy training targets, motivated by analyses of label-noise SGD Damian et al. (2021).

Label corruption in instance discrimination. Consider the instance-discrimination formulation of CL. Let $i \in \mathcal{I}$ denote the instance index and define

$$q_u(j \mid \boldsymbol{T}(\mathbf{x}_i)) := \frac{\exp\big(u(\boldsymbol{T}(\mathbf{x}_i), j)\big)}{\sum_{k\in\mathcal{I}} \exp\big(u(\boldsymbol{T}(\mathbf{x}_i), k)\big)}.$$

The uncorrupted objective is

$$\mathcal{L}_0(u) = \mathbb{E}_{i,\boldsymbol{T}}\left[-\log q_u(i \mid \boldsymbol{T}(\mathbf{x}_i))\right].$$

To model target corruption, replace i with a random index $\widetilde{i}$, where

$$\mathbb{P}(\widetilde{i} = j \mid i) = (1-p)\mathbf{1}_{\{j=i\}} + p\, r(j \mid i), \qquad r(i \mid i) = 0.$$

Here, p is the flip probability and $r(\cdot \mid i)$ specifies the distribution over replacement indices. The expected objective becomes

$$\mathcal{L}_p(u) = (1-p)\mathcal{L}_0(u) + p\,\mathbb{E}_{i,\boldsymbol{T}}\left[-\sum_{j\neq i} r(j \mid i)\log q_u(j \mid \boldsymbol{T}(\mathbf{x}_i))\right]. \tag{49}$$

Thus, label corruption changes the expected training objective. When corruption is resampled during training, it also introduces an additional source of stochasticity into the gradient estimates. These effects motivate a connection with noise-induced regularization, but do not establish equivalence to adding zero-mean gradient noise to the original objective. Moreover, target corruption is only a simplified proxy for label-changing augmentations, which also modify the input distribution.

Empirical robustness–invertibility trade-off. We investigate this connection by independently corrupting instance targets with probability p_{flip} during self-supervised training. We then evaluate downstream regression from $f(\mathbf{x}_i)$ to $\mathbf{z}_{c_i}$. For comparison, we train a supervised model with the same corruption probability, replacing the regression target $\mathbf{z}_{c_i}$ with a different class target. Table 4 reports downstream performance on clean and perturbed inputs, together with latent identifiability.

Increasing the flip probability substantially improves performance under perturbations in both settings. Between $p_{\text{flip}} = 0$ and 0.5, downstream R^2 at $\epsilon = 0.03$ increases from 0.23 to 0.80 for SL and from 0.26 to 0.86 for SSL. Over the same range, identifiability decreases from 0.99 to 0.96 and from 0.98 to 0.88, respectively. Clean downstream performance also decreases at the highest corruption level, although moderate corruption slightly improves it for SL.

These results are consistent with a trade-off between robust downstream prediction and latent recovery, as suggested by our theoretical analysis. They also support the hypothesis that noisy training targets can promote robustness. Furthermore, it builds towards a mechanistic argument to empirical claims over the robustness of SSL over SL Navarro et al. (2021); Zhong et al. (2022); Çağatan et al. (2025).

G.4 CONNECTIONS WITH EQUIVARIANT SSL

Table 5: CIFAR-10 replication of Equivariant SSL using equivariant ResNet-18 backbone Wang et al. (2024). Results are reported for baseline ($\eta = 0$), random + crop ($\eta = 1$), and SimCLR ($\eta = 2$). Invertibility relates to the accuracy of the parameters of the data transformation inferred that can be linearly inferred by f. The % improvement represents the relative gain in F1 score over the $\eta = 0$ baseline.

Metrics	$\eta = 0$	$\eta = 1$	$\eta = 2$
invertibility	**1.00**	1.00	**0.93**
category (clean)	**0.72**	0.79	**0.83**
category (noisy)	**0.65**	0.73	**0.79**
robustness (F1)	**0.68**	0.76	**0.81**
% improve in F1	-	10.85%	**18.10%**

The interplay between the degrees of freedom one can infer from the observations and downstream classification has been recently examined in equivariant SSL Wang et al. (2024). Indeed, Wang et al. (2024) shows a connection between equivariant SSL and instance discrimination. The idea behind their theory is that each augmentation $\boldsymbol{T}$ should add uncertainty with respect to the underlying data generating process so that the solution learned during gradient descent does not degenerate to trivial solutions. We use the code provided in their publication and observe that a trade-off in terms of invertibility and downstream accuracy emerges in the CIFAR-10 dataset (see Table 5).

H INSTANCE DISCRIMINATION.

Augmentations and identifiability. Many joint-embedding methods promote agreement across views while preventing collapse, although their objectives and guarantees need not be equivalent (Garrido et al., 2023; Jha et al., 2024). Identifiability analyses often use an auxiliary observable variable. In the augmentation-based interpretation considered here, this role is assigned to transformations from

$$\mathcal{T} = \{\boldsymbol{T} : \mathcal{X} \to \mathcal{X}\}.$$

Under the conditional block structure used in generalized contrastive learning (Hyvarinen et al., 2019), the latent log-density has to factorize given $\boldsymbol{T}$,

$$\log p(\mathbf{z} \mid \boldsymbol{T}) \propto \sum_{m=1}^{M} q_m(\mathbf{z}^{(m)}, \boldsymbol{T}), \qquad \mathrm{Im}(q_m) = \mathbb{R}. \tag{50}$$

Since the latents are unobserved, an encoder from a sufficiently rich class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{H}\}$ instead supports observable-side functions satisfying

$$\widetilde{q}_m(f(g(\mathbf{z})), \boldsymbol{T}) \approx q_m(\mathbf{z}^{(m)}, \boldsymbol{T}), \qquad m = 1, \ldots, M.$$

Note that dispite this strong assumption, effective modelling necessitates a priori knowledge of $\mathcal{T}$ and $\mathcal{F}$.

Instance discrimination loss For convenience, we use the instance-discrimination formulation of contrastive learning Wu et al. (2018); Balestriero (2023); Reizinger et al. (2025b); Rusak et al. (2025); Ibrahim et al. (2024). Let $J \in \mathcal{I}$ denote the instance label, with $J \sim \mathcal{U}(\mathcal{I})$, and let $C = c(J)$. Write

$$\mathcal{I}_c := \{j \in \mathcal{I} : c(j) = c\}.$$

To match the uniform class prior in (2), we assume that the sets $\mathcal{I}_c$ have equal cardinality.

The observable-side functions $\{\widetilde{q}_m(f(\mathbf{x}), \boldsymbol{T})\}_{m=1}^{M}$ are representation-level quantities; they do not themselves form a categorical distribution over instances. A contrastive loss uses the induced representation to score candidate instances and normalizes these scores into $\widetilde{q}_f(\cdot \mid \mathbf{x}) \in \Delta^{|\mathcal{I}|}$.

The population instance-discrimination objective is

$$f^\star \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{J,\mathbf{X}} \left[-\log \widetilde{q}_f(J \mid \mathbf{X}) \right]. \tag{51}$$

Here, the expectation includes the randomness of the observation conditional on the instance label. This objective admits the decomposition

$$\begin{aligned} \mathbb{E}_{J,\mathbf{X}} \left[-\log \widetilde{q}_f(J \mid \mathbf{X}) \right] = H(J \mid \mathbf{X}) \\ + \mathbb{E}_{\mathbf{X}} \left[D_{\mathrm{KL}} \left(p(\cdot \mid \mathbf{X}) \,\|\, \widetilde{q}_f(\cdot \mid \mathbf{X}) \right) \right]. \end{aligned} \tag{52}$$

Consequently, if the model family can represent the true instance posterior and a population minimum is attained, then, almost surely,

$$\widetilde{q}_{f^\star}(j \mid \mathbf{x}) = p(J = j \mid \mathbf{X} = \mathbf{x}) = p(J = j \mid \mathbf{Z} = \mathbf{z}), \qquad \mathbf{z} = g^{-1}(\mathbf{x}). \tag{53}$$

The second equality follows from the invertibility of g.

I INSTANCE DISCRIMINATION AND OTHER CONTRASTIVE LOSSES

The objective in (51) defines a general instance-discrimination loss in terms of a categorical distribution

$$\pi_\theta(\cdot \mid \mathbf{x}_i) \in \Delta^{|\mathcal{I}|}.$$

Different contrastive objectives correspond to different parameterizations of this categorical distribution through the encoder f_θ. In practice, the distributions below are usually evaluated conditional on sampled augmentations, and the corresponding losses take expectation over the augmentation distribution.

In the notation of the main text, positive and negative candidates are defined at the instance level. For an anchor $\mathbf{x}_i$, the positive candidate is another augmented view of the same indexed instance, whereas negatives are indexed instances distinct from i. The formulas below therefore retain the standard summation over the candidate index set $\mathcal{I}$. At the same time, the data-generation process may induce a coarser class structure over instances. That is, there may exist a class assignment $c(i)$ that groups instances which are, on average, more alike than instances assigned to different classes. If this structure is reflected in the geometry of the augmented observations, then an instance-discrimination objective may implicitly learn a categorical distribution aligned with these classes.

InfoNCE. For InfoNCE, the categorical distribution is parameterized by dot-product logits between augmented representations. Let $f_\theta \in \mathcal{F}$ be an encoder,

$$f_\theta : \mathcal{X} \to \mathbb{R}^d,$$

and let $\tau > 0$ be a temperature parameter. Given two independently sampled augmentations $\boldsymbol{T}_a, \boldsymbol{T}_b \sim p_{\mathcal{T}}$, the anchor is $\boldsymbol{T}_a(\mathbf{x}_i)$, the positive candidate is $\boldsymbol{T}_b(\mathbf{x}_i)$, and the negative candidates are

$$\{\boldsymbol{T}_b(\mathbf{x}_\ell) : \ell \neq i\}.$$

The InfoNCE categorical distribution is

$$\pi_{\mathrm{NCE}}(j \mid \mathbf{x}_i; \boldsymbol{T}_a, \boldsymbol{T}_b) = \frac{\exp\left(f_\theta(\boldsymbol{T}_a(\mathbf{x}_i))^\top f_\theta(\boldsymbol{T}_b(\mathbf{x}_j))/\tau\right)}{\sum_{\ell \in \mathcal{I}} \exp\left(f_\theta(\boldsymbol{T}_a(\mathbf{x}_i))^\top f_\theta(\boldsymbol{T}_b(\mathbf{x}_\ell))/\tau\right)}. \tag{54}$$

Substituting π_{NCE} into (51), with expectation over sampled augmentations, gives

$$\mathcal{L}_{\mathrm{InfoNCE}}(\theta) = -\mathbb{E}_{i \sim \mathcal{U}(\mathcal{I})} \mathbb{E}_{\boldsymbol{T}_a, \boldsymbol{T}_b \sim p_{\mathcal{T}}} \left[\log \pi_{\mathrm{NCE}}(i \mid \mathbf{x}_i; \boldsymbol{T}_a, \boldsymbol{T}_b)\right]. \tag{55}$$

Equivalently, written in the usual positive-versus-negative form,

$$\mathcal{L}_{\mathrm{InfoNCE}}(\theta) = -\mathbb{E}_{i \sim \mathcal{U}(\mathcal{I})} \mathbb{E}_{\boldsymbol{T}_a, \boldsymbol{T}_b \sim p_{\mathcal{T}}} \left[\log \frac{\exp\left(f_\theta(\boldsymbol{T}_a(\mathbf{x}_i))^\top f_\theta(\boldsymbol{T}_b(\mathbf{x}_i))/\tau\right)}{\sum_{\ell \in \mathcal{I}} \exp\left(f_\theta(\boldsymbol{T}_a(\mathbf{x}_i))^\top f_\theta(\boldsymbol{T}_b(\mathbf{x}_\ell))/\tau\right)}\right]. \tag{56}$$

Thus, InfoNCE is an instance of (51) where class probabilities are induced by dot-product scores between augmented representations, provided the negative indices are understood as ranging over the available instance set rather than only over a minibatch Reizinger et al. (2025b).

DIET. For DIET-style parametric instance discrimination, the categorical distribution is induced by a learnable classifier over instance indices. Let

$$W = \{\mathbf{w}_j\}_{j \in \mathcal{I}}, \qquad \mathbf{w}_j \in \mathbb{R}^d,$$

denote classifier weights. Conditional on a sampled augmentation $\boldsymbol{T} \sim p_{\mathcal{T}}$, define

$$\pi_{\mathrm{DIET}}(j \mid \mathbf{x}_i; \boldsymbol{T}) = \frac{\exp\left(\mathbf{w}_j^\top f_\theta(\boldsymbol{T}(\mathbf{x}_i))\right)}{\sum_{\ell \in \mathcal{I}} \exp\left(\mathbf{w}_\ell^\top f_\theta(\boldsymbol{T}(\mathbf{x}_i))\right)}. \tag{57}$$

Substituting π_{DIET} into (51), with expectation over sampled augmentations, gives

$$\mathcal{L}_{\mathrm{DIET}}(\theta, W) = -\mathbb{E}_{i \sim \mathcal{U}(\mathcal{I})} \mathbb{E}_{\boldsymbol{T} \sim p_{\mathcal{T}}} \left[\log \pi_{\mathrm{DIET}}(i \mid \mathbf{x}_i; \boldsymbol{T})\right]. \tag{58}$$

Equivalently,

$$\mathcal{L}_{\mathrm{DIET}}(\theta, W) = -\mathbb{E}_{i \sim \mathcal{U}(\mathcal{I})} \mathbb{E}_{\boldsymbol{T} \sim p_{\mathcal{T}}} \left[\log \frac{\exp\left(\mathbf{w}_i^\top f_\theta(\boldsymbol{T}(\mathbf{x}_i))\right)}{\sum_{\ell \in \mathcal{I}} \exp\left(\mathbf{w}_\ell^\top f_\theta(\boldsymbol{T}(\mathbf{x}_i))\right)}\right]. \tag{59}$$

Thus, InfoNCE and DIET share the same likelihood and cross-entropy backbone in (51). They differ in how the categorical distribution over instance labels is parameterized: InfoNCE uses dot-product scores between augmented representations, whereas DIET uses a parametric classifier whose weights represent instance classes. In both cases, positives and negatives are defined by instance identity. Nevertheless, if there exists a class assignment $c : \mathcal{I} \to \mathcal{C}$ such that same-class instances are systematically more alike than different-class instances, the learned categorical distribution may implicitly reflect this coarser class structure Wang et al. (2023); Reizinger et al. (2025b).

J HIGHER CONCENTRATION LEADS TO LOWER OVERLAP AND MORE ROBUST CLASSIFERS

Class-conditional distributions, concentration, and overlap. Our latent variable model needs to embed some class structure in order for the invertibility hypothesis to make sense. We study latent distrubtions that are concentrated on a set of classes $\mathcal{C} = \{c\}$ with minimal overlap to ensure robust discriminator exists. Specifically, for each class c, fix $\beta_c \in (0, 1)$ and define its concentration radius around $\mathbf{z}_c$ by

$$\alpha_c := \inf \{r \geq 0 : \mathbb{P}(\|\mathbf{Z} - \mathbf{z}_c\|_{\mathcal{Z}} \leq r \mid C = c) \geq 1 - \beta_c\}. \tag{60}$$

Thus, the Borel set

$$B_c := \{\mathbf{z} \in \mathcal{Z} : \|\mathbf{z} - \mathbf{z}_c\|_{\mathcal{Z}} \leq \alpha_c\}$$

satisfies $\mu_c(B_c) \geq 1 - \beta_c$.

For $c \neq c'$, define the signed separation gap

$$s_{cc'} := \|\mathbf{z}_c - \mathbf{z}_{c'}\|_{\mathcal{Z}} - \alpha_c - \alpha_{c'}. \tag{61}$$

Holding the representatives and tail probabilities fixed, smaller concentration radii increase $s_{cc'}$ and shrink the intersection of the concentration balls. In particular, $s_{cc'} > 0$ guarantees that B_c and $B_{c'}$ are disjoint, although the distributions may still overlap.

We quantify distributional overlap by

$$\omega_{cc'} := \int_{\mathcal{Z}} \min\{p_c(\mathbf{z}), p_{c'}(\mathbf{z})\}\, d\nu(\mathbf{z}). \tag{62}$$

If $s_{cc'} > 0$, then

$$\omega_{cc'} \leq \min\{1, \beta_c + \beta_{c'}\}. \tag{63}$$

Indeed, disjointness implies $\mu_{c'}(B_c) \leq \beta_{c'}$, and hence

$$\omega_{cc'} \leq \int_{B_c} p_{c'}\, d\nu + \int_{B_c^c} p_c\, d\nu \leq \beta_{c'} + \beta_c.$$

These assumptions therefore control distributional overlap whenever the concentration balls are disjoint; smaller radii alone do not imply a general monotonic decrease in $\omega_{cc'}$.

Gaussian example. Suppose that $\mathcal{Z} = \mathbb{R}^d$, equipped with the Euclidean norm, and

$$\mathbf{Z} \mid C = c \sim \mathcal{N}(\mathbf{z}_c, \sigma^2 I_d).$$

Then

$$\alpha_c(\beta_c) = \sigma \sqrt{F^{-1}_{\chi^2_d}(1 - \beta_c)},$$

where $F^{-1}_{\chi^2_d}$ denotes the quantile function of a chi-squared distribution with d degrees of freedom. Consequently, decreasing σ reduces the concentration radii at fixed tail probabilities.

For two classes with the same σ, their overlap is

$$\omega_{cc'} = 2\Phi_{\mathrm{N}}\left(-\frac{\|\mathbf{z}_c - \mathbf{z}_{c'}\|_2}{2\sigma}\right), \tag{64}$$

where Φ_{N} is the standard normal distribution function. For fixed distinct representatives, the overlap decreases to zero as $\sigma \to 0$.

Von Mises–Fisher example. Suppose that $\mathbf{Z}$ takes values on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and

$$p_c(\mathbf{z}) = C_d(\kappa) \exp(\kappa \mathbf{z}_c^\top \mathbf{z}), \qquad \|\mathbf{z}_c\|_2 = 1,$$

where $\kappa \geq 0$ is a common concentration parameter and $C_d(\kappa)$ is the normalizing constant. Here, $\mathbf{z}_c$ is the mean direction. Increasing κ decreases the Euclidean concentration radii at fixed tail probabilities.

Let

$$\theta_{cc'} := \arccos(\mathbf{z}_c^\top \mathbf{z}_{c'})$$

denote the angular separation. Since $\min\{a, b\} \leq \sqrt{ab}$,

$$\omega_{cc'} \leq \int_{\mathbb{S}^{d-1}} \sqrt{p_c p_{c'}}\, dS = \frac{C_d(\kappa)}{C_d(\kappa \cos(\theta_{cc'}/2))}. \tag{65}$$

For fixed distinct mean directions, this upper bound decreases to zero as $\kappa \to \infty$.

The induced class posterior by Instance Discrimination. Because $C = c(j)$, aggregating the Bayes-optimal instance posterior yields

$$\begin{aligned}\pi(c \mid \mathbf{x}) &:= \sum_{c(j)=c} \widetilde{q}_{f^\star}(j \mid \mathbf{x}) \\ &= p(C = c \mid \mathbf{Z} = \mathbf{z}) \\ &= \frac{p_c(\mathbf{z})}{\sum_{c' \in \mathcal{C}} p_{c'}(\mathbf{z})}, \qquad \mathbf{z} = g^{-1}(\mathbf{x}).\end{aligned} \tag{66}$$

The concentration and overlap assumptions characterize the ambiguity remaining in this posterior. For two classes c, c', considered with equal prior probabilities, the Bayes classification error is

$$R^\star_{cc'} = \frac{1}{2}\omega_{cc'}. \tag{67}$$

Therefore, whenever $s_{cc'} > 0$,

$$R^\star_{cc'} \leq \frac{1}{2}\min\{1, \beta_c + \beta_{c'}\}. \tag{68}$$

More generally, if all concentration balls are pairwise disjoint, the multiclass Bayes classifier $h^\star(\mathbf{x}) \in \arg\max_c \pi(c \mid \mathbf{x})$ satisfies

$$\mathbb{P}\left(h^\star(\mathbf{X}) \neq C\right) \leq \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} \beta_c. \tag{69}$$

Indeed, a classifier assigning class c on B_c can err only outside the true class's concentration ball, and the Bayes classifier cannot have greater risk.

Thus, population-optimal instance discrimination recovers the class posterior through aggregation, while concentration around separated representatives bounds its classification error.

K FINITE-PRECISION CONSEQUENCES OF VMF CONCENTRATION

This appendix provides the derivation supporting the finite-precision argument used in the main text. We assume that latent representations are normalized and lie on the unit hypersphere,

$$\mathbf{z}_i \in S^{d-1} = \left\{ \mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 = 1 \right\}. \tag{70}$$

Each class c is represented by a normalized prototype

$$\mathbf{z}_c \in S^{d-1}, \qquad \|\mathbf{z}_c\|_2 = 1. \tag{71}$$

Conditioned on class membership, we model instance representations with a von Mises–Fisher distribution,

$$\mathbf{z}_i \mid c(i) = c \sim \mathrm{vMF}(\mathbf{z}_c, \kappa), \tag{72}$$

with density

$$p(\mathbf{z}_i \mid c) = C_d(\kappa) \exp\left(\kappa \mathbf{z}_c^\top \mathbf{z}_i\right), \qquad \mathbf{z}_i \in S^{d-1}. \tag{73}$$

Here $C_d(\kappa)$ is the normalizing constant,

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}, \tag{74}$$

where I_ν denotes the modified Bessel function of the first kind. The vMF distribution is the canonical analogue of an isotropic Gaussian distribution on the unit hypersphere, with $\mathbf{z}_c$ acting as the mean direction and κ controlling concentration around that direction Sra (2016).

Let θ_i denote the geodesic angle between the instance representation $\mathbf{z}_i$ and its class prototype $\mathbf{z}_c$:

$$\theta_i = \arccos\left(\mathbf{z}_c^\top \mathbf{z}_i\right). \tag{75}$$

Since both vectors have unit norm,

$$\mathbf{z}_c^\top \mathbf{z}_i = \cos\theta_i. \tag{76}$$

Substituting this identity into Eq. (73) gives

$$p(\mathbf{z}_i \mid c) \propto \exp\left(\kappa \cos\theta_i\right). \tag{77}$$

Thus, increasing κ increases the relative likelihood of points with small angular distance to the class prototype.

K.1 LOCAL TANGENT-SPACE APPROXIMATION

For large κ, most of the probability mass of the vMF distribution lies near $\mathbf{z}_c$. We therefore use normal coordinates in the tangent space at $\mathbf{z}_c$,

$$T_{\mathbf{z}_c} S^{d-1} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{z}_c^\top \mathbf{u} = 0 \right\}. \tag{78}$$

This tangent space has dimension $d-1$. Let

$$r = \|\mathbf{u}\|_2. \tag{79}$$

The exponential map from the tangent space to the sphere gives

$$\mathbf{z}_i = \exp_{\mathbf{z}_c}(\mathbf{u}) = \cos r\, \mathbf{z}_c + \sin r\, \frac{\mathbf{u}}{r}. \tag{80}$$

Taking the inner product with $\mathbf{z}_c$ gives

$$\mathbf{z}_c^\top \mathbf{z}_i = \cos r. \tag{81}$$

Since r is the geodesic distance from $\mathbf{z}_c$ to $\mathbf{z}_i$, we identify

$$r = \theta_i. \tag{82}$$

For small angles, the cosine admits the Taylor expansion

$$\cos\theta_i = 1 - \frac{\theta_i^2}{2} + \frac{\theta_i^4}{24} + O(\theta_i^6). \tag{83}$$

Substituting Eq. (83) into Eq. (77) gives

$$p(\mathbf{z}_i \mid c) \propto \exp\left[\kappa\left(1 - \frac{\theta_i^2}{2} + \frac{\theta_i^4}{24} + O(\theta_i^6)\right)\right] \tag{84}$$

$$\propto \exp\left(-\frac{\kappa\theta_i^2}{2} + O(\kappa\theta_i^4)\right). \tag{85}$$

Ignoring the constant factor $\exp(\kappa)$, the leading-order term is Gaussian in the tangent-space radius:

$$p(\mathbf{u} \mid c) \approx \left(\frac{\kappa}{2\pi}\right)^{(d-1)/2} \exp\left(-\frac{\kappa}{2}\|\mathbf{u}\|_2^2\right). \tag{86}$$

Therefore, locally,

$$\mathbf{u} \approx \mathcal{N}\left(\mathbf{0}, \kappa^{-1}\mathbf{I}_{d-1}\right). \tag{87}$$

This is the standard large-concentration approximation of the vMF distribution: near its mode, the hypersphere is locally flat, and the vMF behaves like an isotropic Gaussian in the $(d-1)$-dimensional tangent space **?**Sra (2012).

K.2 Typical angular scale

From Eq. (87), the squared tangent-space norm has expectation

$$\mathbb{E}[\|\mathbf{u}\|_2^2] = \operatorname{tr}\left(\kappa^{-1}\mathbf{I}_{d-1}\right) = \frac{d-1}{\kappa}. \tag{88}$$

Since $\|\mathbf{u}\|_2 = \theta_i$ locally, this gives

$$\boxed{\mathbb{E}[\theta_i^2] \approx \frac{d-1}{\kappa}.} \tag{89}$$

Thus, the typical angular deviation of an instance from its class prototype is

$$\boxed{\theta_{\text{typ}} \approx \sqrt{\frac{d-1}{\kappa}}.} \tag{90}$$

Therefore, the typical angular distance decreases as $O(\kappa^{-1/2})$ for fixed d.

K.3 Finite-precision and effective-resolution threshold

Let ε denote the minimum cosine-scale difference that the full computational pipeline can reliably resolve. In the idealized hardware-level case, ε may be taken as the floating-point machine precision $\varepsilon_{\text{mach}}$. More generally, however, the relevant threshold is an effective resolution ε_{eff}, which may be much larger than machine epsilon.

If dot-product differences smaller than ε cannot be reliably resolved, then same-class deviations from the prototype become effectively indistinguishable once

$$\boxed{\kappa \gtrsim \frac{d-1}{2\varepsilon}.} \tag{91}$$

When $\varepsilon = \varepsilon_{\text{mach}}$, Eq. (91) describes an extreme hardware-level indistinguishability regime. However, in practical learning systems the effective resolution may be dominated by other sources of variation before machine precision becomes relevant. These include stochastic optimization noise, finite sample effects, mini-batch variability, normalization effects, estimator variance, regularization, model sensitivity, limited embedding dimension, and the resolution of the downstream decision rule.

As κ increases, their deviations from the prototype collapse as $O(\kappa^{-1})$. Once the expected cosine deviation is comparable to the ϵ_{mach}, further increases in concentration no longer produce meaningfully resolvable within-class structure under dot-product-based computations.

K.4 Angular-resolution threshold

If the relevant operation requires explicitly resolving angular or Euclidean displacements rather than dot products, then the appropriate condition is instead

$$\theta_{\mathrm{typ}} \lesssim \varepsilon. \tag{92}$$

Using Eq. (90), this gives

$$\sqrt{\frac{d-1}{\kappa}} \lesssim \varepsilon. \tag{93}$$

Equivalently,

$$\boxed{\kappa \gtrsim \frac{d-1}{\varepsilon^2}.} \tag{94}$$

This angular threshold is much larger than the dot-product threshold because angles shrink as $O(\kappa^{-1/2})$, whereas cosine deviations shrink as $O(\kappa^{-1})$.

K.5 Numerical example for $d = 5$

For the setting used in the experiments with $d = 5$, we have

$$d - 1 = 4. \tag{95}$$

The prototype dot-product threshold in Eq. (91) becomes

$$\kappa \gtrsim \frac{4}{2\varepsilon} = \frac{2}{\varepsilon}. \tag{96}$$

If ε is interpreted as machine precision, then using the standard IEEE-754 machine epsilons

$$\varepsilon_{\mathrm{float32}} = 2^{-23} \approx 1.19 \times 10^{-7}, \tag{97}$$

and

$$\varepsilon_{\mathrm{float64}} = 2^{-52} \approx 2.22 \times 10^{-16}, \tag{98}$$

we obtain

$$\kappa_{\mathrm{dot,float32}} \gtrsim \frac{2}{1.19 \times 10^{-7}} \approx 1.68 \times 10^{7}, \tag{99}$$

$$\kappa_{\mathrm{dot,float64}} \gtrsim \frac{2}{2.22 \times 10^{-16}} \approx 9.0 \times 10^{15}. \tag{100}$$

For explicit angular resolution, Eq. (94) instead gives

$$\kappa_{\mathrm{angle,float32}} \gtrsim \frac{4}{(1.19 \times 10^{-7})^2} \approx 2.8 \times 10^{14}, \tag{101}$$

$$\kappa_{\mathrm{angle,float64}} \gtrsim \frac{4}{(2.22 \times 10^{-16})^2} \approx 8.1 \times 10^{31}. \tag{102}$$

These values describe an extreme hardware-level resolution limit.

However, if ε is interpreted as an effective resolution $\varepsilon_{\mathrm{eff}}$, the threshold can occur much earlier. For $d = 5$, Eq. (96) gives

$$\kappa \gtrsim \frac{2}{\varepsilon_{\mathrm{eff}}}. \tag{103}$$

For example, if the effective cosine resolution is $\varepsilon_{\mathrm{eff}} = 0.04$, then

$$\kappa \gtrsim 50. \tag{104}$$

If $\varepsilon_{\mathrm{eff}} = 0.02$, then

$$\kappa \gtrsim 100. \tag{105}$$

Thus, for $d = 5$, an effective cosine resolution of only a few percent naturally predicts degradation of instance-level identifiability in the range $\kappa = 50$–100, even though this regime is far from the hard representational limits of float32 or float64.