Your Projection Head Secretly Shapes Representations

Anonymous authors
Paper under double-blind review

Abstract

Projection heads are discarded after contrastive pretraining, yet they strongly affect the representation that remains. We study this effect through the critic induced by the projection head and similarity function, showing that critic form shapes not only how representations are compared, but also which encoder-output distributions are easy to realize. This yields a tradeoff between critic expressiveness, self-supervised loss value, and downstream usability, where restricted critics impose distributional biases that highly expressive non-separable critics can weaken. Experiments support this mechanism: a separable projection-head critic improves linear-probe accuracy over a non-separable MLP critic despite a much lower InfoNCE estimate, and fixed critics steer embeddings toward Gaussian-mixture structures without explicit prior matching.

1 Introduction

Self-supervised representation learning (SSRL) has become a standard approach for learning useful representations from unlabeled data. In contrastive SSRL, an encoder is trained by comparing representations of augmented views of the same input, with classical methods applying this comparison in a projected space (Balestriero et al., 2023; Chen et al., 2020; Ericsson et al., 2022; He et al., 2020; Ozbulak et al., 2023). After pretraining, however, the projection head is discarded, and the pre-projector representation is used for downstream tasks. This makes the projection head's effect on the retained representation nontrivial.

Prior work has mainly studied projection-head design through coarse architectural choices, such as using a projection head at all or choosing a linear versus non-linear projector. These choices have been explained through representation content, including subspace selection, information bottlenecks, and layer-wise feature reweighting (Gupta et al., 2022; Ouyang et al., 2025; Xue et al., 2024), or through representation geometry, including expansion-shrinkage dynamics, dimensional collapse, and uniformity (Gui et al., 2023; Ma et al., 2024; Song et al., 2023). However, these perspectives do not fully explain why more expressive scoring functions can achieve better loss values yet yield worse retained representations (Tschannen et al., 2020), or how critic architecture shapes the distribution of encoder outputs.

Building on the connection between contrastive SSRL and Deep InfoMax (Hjelm et al., 2019; Oord et al., 2019), we interpret objectives such as InfoNCE as variational lower bounds on the mutual information (MI) between representations of augmented views (Chen et al., 2020; Oord et al., 2019). Such bounds are optimized over a variational *critic*, which scores pairs of encoder outputs. In standard projection-head architectures, this critic applies a projection head separately to each output, followed by a simple similarity function (Butakov et al., 2024; Tschannen et al., 2020).

The MI formulation makes the interaction between encoder and critic explicit. Highly expressive non-separable critics can process representation pairs jointly, allowing them to reduce the training loss by absorbing more of the MI-estimation task. This can improve the training loss value without improving the retained encoder representation (Tschannen et al., 2020). In contrast, separable critics score a pair only after transforming each representation independently. They place stronger pressure on the encoder to produce outputs whose distribution is easy for the critic to score.

In InfoNCE-based methods, the projection head together with the similarity function implements this separable form. This places projection heads in an intermediate regime: the critic is expressive enough to capture similarities needed for MI estimation, but restricted enough to pressure the encoder

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.

to adapt its output distribution to the critic family. This suggests a constructive use of critic design. Fixed critics with prescribed structure can be used to steer the resulting embedding distribution.

Our contributions are as follows.

1. **Encoder-critic tradeoff (Section 4).** We formalize how critic expressiveness affects the retained encoder representation, relating critic capacity to a tradeoff between information content and downstream usability.
2. **Distributional bias (Section 5).** We establish distributional bias as the mechanism by which critic architecture governs the encoder output distribution. The variational objective favors distributions matched to the critic: highly expressive non-separable critics can dissipate this bias, whereas restricted or non-learnable critics impose stronger constraints on the representation distribution.
3. **Constructive distribution shaping (Section 6).** We demonstrate the practical consequence of our theory: fixed critics can steer embeddings toward prescribed structures, such as Gaussian mixtures, without an explicit prior-matching loss, adversarial discriminator, or generative model.

2 Related Work

Projection heads in joint-embedding SSRL. The pretraining loss is often imposed in a space different from the representation used downstream. In contrastive methods, projection heads define the space where similarities are computed (Chen et al., 2020; He et al., 2020). In redundancy-reduction and DINO-style methods, related heads define the space where feature statistics are regularized or teacher-student targets are matched (Bardes et al., 2022; Siméoni et al., 2025; Zbontar et al., 2021). Across these settings, the head participates in the pretraining objective, while the pre-head representation is typically retained for downstream use.

Projection-head explanations. Prior work interprets projection heads as non-linear subspace selectors (Gupta et al., 2022), information bottlenecks that remove task-irrelevant factors (Ouyang et al., 2025), or mechanisms that progressively reweight features across layers (Xue et al., 2024). Other studies connect projection heads to expansion and shrinkage dynamics (Gui et al., 2023), sparsity and dimensional collapse (Song et al., 2023), or more uniform feature distributions (Ma et al., 2024). Our work aligns more closely with the latter line of explanation, but attributes the effect to distributional bias induced by the critic family defined during pretraining.

Critic parameterization. Our critic-based framing builds on prior work on MI maximization for representation learning (Tschannen et al., 2020). This work argues that the success of MI objectives cannot be explained by MI maximization alone and emphasizes the inductive bias introduced by the parameterization of MI estimators. This observation motivates our study of critic parameterization as a source of distributional bias in encoder outputs, rather than only as a choice in MI estimation.

Latent priors and distribution matching. Classical latent distribution matching adds an explicit prior-matching objective, using variational regularization, adversarial discriminators, or distribution discrepancies such as MMD and optimal-transport distances (Gretton et al., 2012; Kingma and Welling, 2014; Kolouri et al., 2018; Makhzani et al., 2015; Tolstikhin et al., 2017). Recent work also shows that Gaussian-like structure can arise naturally from contrastive objectives: (Butakov et al., 2024) derive Gaussian and uniform matching under noise-injected Deep InfoMax, while (Betser et al., 2026) show that InfoNCE induces Gaussian structure in learned representations. Our work instead studies how the critic architecture shapes this bias, and how fixed critics can impose other target structures.

3 Background

Notation. For a measurable space $(\Omega, \mathcal{F})$, let us denote the set of all probability measures as $\mathscr{P}(\Omega, \mathcal{F})$ (or $\mathscr{P}$ for short). Let $(\Omega', \mathcal{F}')$ be a second measurable space. By $\mathscr{M}(\mathcal{F}, \mathcal{F}')$ ($\mathscr{M}$ for short) we denote the set of measurable functions $f : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$. For a pair of measures μ, ν defined on the same measurable space, $\mu \ll \nu$ denotes the *absolute continuity* of μ with respect to

(w.r.t.) ν, that is, for any measurable S, $\nu(S) = 0$ implies $\mu(S) = 0$. For $\mu \ll \nu$, $\mathrm{d}\mu/\,\mathrm{d}\nu$ denotes the corresponding *Radon-Nikodym derivative*. Wherever it is needed, we assume the relevant Radon-Nikodym derivatives exist.

Elements of Information Theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with sample space Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$ defined on $\mathcal{F}$. Consider random vectors $X : \Omega \to \mathbb{R}^{d_x}$ and $Y : \Omega \to \mathbb{R}^{d_y}$ with joint distribution $\mathbb{P}_{X,Y}$ and marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. For a probability measure $\mathbb{Q} \ll \mathbb{P}$, the Kullback-Leibler (KL) divergence from $\mathbb{P}$ to $\mathbb{Q}$ is $\mathsf{KL}[\mathbb{Q} \parallel \mathbb{P}] = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right]$, which is non-negative and vanishes if and only if (iff) $\mathbb{P} = \mathbb{Q}$. The mutual information (MI) between X and Y quantifies the divergence between the joint distribution and the product of marginals:

$$\mathsf{I}(X;Y) = \mathbb{E} \log \frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y} = \mathsf{KL}\left[\mathbb{P}_{X,Y} \,\middle\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\right].$$

Here $\mathrm{PMI}(x,y) \stackrel{\text{def}}{=} \frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y}(x,y)$ is the *point-wise mutual information* between X and Y.

When $\mathbb{P}_X$ admits a probability density function (PDF) $p(X)$ w.r.t. the Lebesgue measure, the differential entropy is defined as $\mathsf{h}(X) = -\mathbb{E}[\log p(X)]$, where $\log(\,\cdot\,)$ denotes the natural logarithm. Likewise, the joint entropy $\mathsf{h}(X,Y)$ is defined via the joint density $p(X,Y)$, and conditional entropy is $\mathsf{h}(X \mid Y) = -\mathbb{E}[\log p(X \mid Y)]$. Under the existence of PDFs, MI satisfies the identities

$$\mathsf{I}(X;Y) = \mathsf{h}(X) - \mathsf{h}(X \mid Y) = \mathsf{h}(Y) - \mathsf{h}(Y \mid X) = \mathsf{h}(X) + \mathsf{h}(Y) - \mathsf{h}(X,Y). \tag{1}$$

Data Processing Inequality. A fundamental property of MI is that it cannot be increased by deterministic processing or, more generally, by Markov kernels. Furthermore, MI is preserved under invertible transformations. This is formalized by the *data processing inequality* (DPI) (Polyanskiy and Wu, 2024, Theorem 3.7): for a Markov chain $X \to Y \to Z$, $\mathsf{I}(X;Y) \geq \mathsf{I}(X;Z)$, with the equality holding for $Z = f(Y)$, where f is measurably invertible.

DPI is often applied together with the Markov chain reversal property (Polyanskiy and Wu, 2024, Section 3.4), which states that two conditionally independent Markov chains sharing a common source variable actually form a single Markov chain:

$$W \leftarrow V \leftarrow X \to Y \to Z \qquad \Longleftrightarrow \qquad W \to V \to X \to Y \to Z$$

Variational lower bounds. In practice, KL and MI are intractable, necessitating alternative representations of these quantities. Among these representations, the most widely adopted are the *Donsker-Varadhan* (DV) (Donsker and Varadhan, 1983) and *Nguyen-Wainwright-Jordan* (NWJ) (Nguyen et al., 2010) variational lower bounds.

***Definition.* 3.1.** Let $\mathbb{P}, \mathbb{Q} \in \mathscr{P}(\Omega, \mathcal{F})$ be a pair of probability measures and $T \in \mathscr{M}(\mathcal{F}, \mathcal{B}_{\mathbb{R}})$ be a measurable function from Ω to $\mathbb{R}$. Define

$$\mathcal{L}_{\mathrm{DV}}[\mathbb{P}, \mathbb{Q}; T] = \mathbb{E}_{\mathbb{P}}\, T - \log \mathbb{E}_{\mathbb{Q}}\, e^{T} \qquad \text{(Donsker-Varadhan (Donsker and Varadhan, 1983))}$$

$$\mathcal{L}_{\mathrm{NWJ}}[\mathbb{P}, \mathbb{Q}; T] = \mathbb{E}_{\mathbb{P}}\, T - \mathbb{E}_{\mathbb{Q}}\, e^{T-1} \qquad \text{(Nguyen-Wainwright-Jordan (Nguyen et al., 2010))}$$

Lemma 3.2. If $\mathbb{P} \ll \mathbb{Q}$, then $\mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] = \sup_{T \in \mathscr{M}} \mathcal{L}_{\mathrm{DV}} = \sup_{T \in \mathscr{M}} \mathcal{L}_{\mathrm{NWJ}}$, and the supremum is attained at

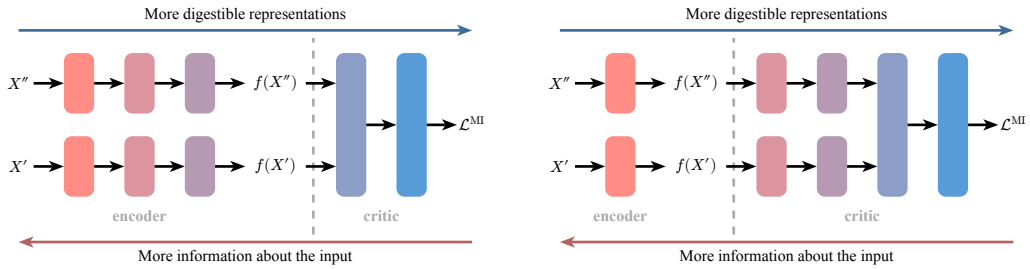
$$\arg\sup_{T \in \mathscr{M}} \mathcal{L}_{\mathrm{DV}} = \left\{ \log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}} + \alpha \;\middle|\; \alpha \in \mathbb{R} \right\} \qquad \arg\sup_{T \in \mathscr{M}} \mathcal{L}_{\mathrm{NWJ}} = \left\{ \log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}} + 1 \right\}$$

***Remark.* 3.3.** Since $\mathsf{I}(X;Y) = \mathsf{KL}\left[\mathbb{P}_{X,Y} \,\middle\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\right]$, the same bounds also hold for MI (Belghazi et al., 2018). In this particular case, we will use the notation $\mathcal{L}\left[\mathbb{P}_{X,Y}, \mathbb{P}_X \otimes \mathbb{P}_Y; T\right] \stackrel{\text{def}}{=} \mathcal{L}^{\mathrm{MI}}[X, Y; T]$.

Contrastive SSRL setup. For the analysis below, it is useful to view the projection head and similarity function as a critic. Given encoder outputs $Z' = f(X')$ and $Z'' = f(X'')$, a projection head g and similarity s define a separable critic

$$T(Z', Z'') = s(g(Z'), g(Z'')).$$

Thus, changing the projection head changes the critic family used by the contrastive objective.



(a) Strong encoder, weak critic; representations are more refined, but contain less information.

(b) Weak encoder, strong critic; representations are less refined, but contain more information.

Figure 1: The tradeoff between encoder and critic expressivity: while the joint critic remains the same, assigning more layers to the encoder leads to higher information content, but lower refinement.

4 Deep InfoMax Revisited

We start our analysis by revisiting the Deep InfoMax paradigm (Hjelm et al., 2019; Oord et al., 2019), which suggests maximizing $\mathsf{I}(f(X'); f(X''))$ — mutual information between representations of augmented views of the same input X. Since many contrastive SSRL objectives can be interpreted as MI lower-bound optimization, this view captures the common structure of projection-head-based contrastive methods (Butakov et al., 2024).

Because MI is intractable in practical scenarios, variational lower bounds from Definition 3.1 (or relaxed, less tight counterparts) are maximized instead, with InfoNCE loss (Oord et al., 2019) (a special case of NWJ (Poole et al., 2019)) being the most widely adopted choice. This leads to the following problem:

$$\mathcal{L}^{\text{MI}}[f(X'), f(X''); T] \longrightarrow \max_{T,f}, \tag{2}$$

where $\mathcal{L}^{\text{MI}}$ is any variational lower bound on MI, f is the encoder model, T is a learnable similarity measure (*critic*). Note, however, that (2) can be reframed as the following problem:

$$\mathcal{L}^{\text{MI}}[X', X''; T \circ (f \times f)] \longrightarrow \max_{T,f}, \tag{3}$$

where $h \circ g$ is a composition, and $h \times g$ is a Cartesian product of two functions. Since the distributions in (3) are now fixed, this is plain MI estimation between X' and X'', where $T \circ (f \times f) \stackrel{\text{def}}{=} T_{\text{joint}}$ serves as a *joint critic* trained to discern between augmentations coming from the same input and augmentations acquired from different inputs. After learning the joint critic, we remove the common part T, thus acquiring f that outputs features useful for downstream tasks.

Here we make our *first key observation*: the decomposition of the joint critic T_{joint} into encoder f and critic T is not unique. One can assign the last layers of f to T, making the critic more expressive and changing which representation is retained. Conversely, choosing stronger T (e.g., non-linear instead of bilinear) is essentially equivalent to removing last layers from f, thus yielding less processed representations.

This provides an intuition for the "sweet spot" on the critic complexity spectrum: when T is weak, outputs of f are over-processed and useful only for discerning between $\mathbb{P}_{X',X''}$ and $\mathbb{P}_{X'} \otimes \mathbb{P}_{X''}$. On the other hand, extremely expressive T performs all the "heavy lifting" for MI estimation and, therefore, sends almost no optimization signal to f; see Figure 1. We formalize our intuition below:

Proposition 4.1. Consider the setup from (3). Let $f = g_n \circ \ldots \circ g_1$ be a feed-forward network of n layers, and ${L_k}'/{L_k}''$ be outputs of the k-th layer on X'/X'' respectively. Then

$$\textbf{(a)}\ \mathsf{I}({L_k}'; X) \geq \mathsf{I}({L_k}'; {L_k}'') \qquad \textbf{(b)}\ \mathsf{I}({L_1}'; {L_1}'') \geq \ldots \geq \mathsf{I}({L_n}'; {L_n}'') \geq \mathcal{L}^{\text{MI}}\left[X', X''; T_{\text{joint}}\right]$$

Let $\mathscr{M}_k$ be a set of extended critics T_k that incorporate the last $n-k$ layers of f into T. Then

$$\sup_{T_k \in \mathscr{M}_k} \mathcal{L}^{\mathrm{MI}}[L_k', L_k''; T_k] \geq \mathcal{L}^{\mathrm{MI}}\left[X', X''; T_{\mathrm{joint}}\right]$$

This proposition establishes a natural information ordering among encoder layers: subsequent internal representations cannot contain more information than the preceding ones. On the other hand, later layers can yield representations that are more accessible to simpler critics, since they allow the same lower bound on MI to be attained with shallower critic architectures.

5 Distribution shaping

Although intuitive, the surface-level explanation from Section 4 does not describe the *precise mechanisms* that increase the digestibility of subsequent representations. Therefore, the role of projection heads still remains somewhat ambiguous, and some practical questions (e.g., the optimal choice of critic architecture) remain unanswered. In this section, we dive deeper into the aforementioned tradeoff and show that distribution biases imposed by weaker/shallower (or even fixed) critics are the ones who drive the refinement of representations.

To provide initial intuition, we consider a limit case, where critic T is non-learnable (fixed). The lemma below provides a result dual to Lemma 3.2. In particular, we show that fixing the critic function and the second distribution $\mathbb{Q}$ yields a uniquely defined optimal distribution while maximizing $\mathcal{L}[\mathbb{P}, \mathbb{Q}; T]$ over $\mathbb{P}$ under certain regularization constraints.

Lemma 5.1. (Distribution matching duality) Consider a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ and $T \in \mathscr{M}(\mathcal{F}, \mathcal{B}_{\mathbb{R}})$. Define $\mathbb{G}$ such that $\frac{\mathrm{d}\,\mathbb{G}}{\mathrm{d}\,\mathbb{Q}} \sim e^T$. Let $\mathscr{R} = \{\mathbb{P} \in \mathscr{P} \mid \mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] \leq \mathsf{KL}[\mathbb{G} \parallel \mathbb{Q}]\}$. Then

$$\sup_{\mathbb{P} \in \mathscr{R}} \mathcal{L}_{\mathrm{DV}} = \sup_{\mathbb{P} \in \mathscr{R}} \mathcal{L}_{\mathrm{NWJ}} = \mathsf{KL}[\mathbb{G} \parallel \mathbb{Q}],$$

where the supremum is attained at $\mathbb{P} = \mathbb{G}$.

The lemma shows that, once the critic is fixed, the variational objective is no longer neutral with respect to the optimized distribution. It favors the Gibbs distribution induced by the critic. Thus, critic form can be read not only as an estimator parameterization, but also as an implicit preference over representation distributions.

While useful for providing initial formal grounds, this result does not directly cover DIM-style SSRL. First, we constrain $\mathbb{P}$ by considering only the distributions from $\mathscr{R}$. In reality, such constraints are not imposed, at least not explicitly. While some SSRL setups regularize and smooth the representations' distribution by adding small noise (Butakov et al., 2024), which effectively limits $\mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}]$, this is not a common technique.

Secondly, and most importantly, in the DIM setup, $\mathbb{Q}$ is not fixed, since $\mathbb{P} = \mathbb{P}_{Z',Z''}$ and $\mathbb{Q} = \mathbb{P}_{Z'} \otimes \mathbb{P}_{Z''}$; therefore, both distributions depend on the encoder f. Unfortunately, this implies that it is not possible to derive strict distribution matching results for this case, since many distributions can correspond to one $\log T = \frac{\mathrm{d}\,\mathbb{P}_{Z',Z''}}{\mathrm{d}\,\mathbb{P}_{Z'} \otimes \mathbb{P}_{Z''}}$.[1] However, below we show that it is still possible to derive one-way results: for some fixed critics, it is possible to show that some predefined distribution is a global maximizer of a SSRL loss (but not necessarily a unique one).

Proposition 5.2. (Gaussian mixture distribution) Let $Z \in \mathbb{R}^d$ be the embedding before adding noise, and $\tilde{Z} = Z + \varepsilon$ be the embedding after the injection of Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\mathrm{noise}} I)$. Let the target distribution $p_{\mathrm{target}(\tilde{Z})}$ (Gaussian Mixture) be defined as a symmetric equiprobable mixture of K multivariate normal distributions, whose centers $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$, and whose covariance matrices are equal to $\sigma^2_{\mathrm{class}} I$. Then the optimal fixed critic $T^*\left(Z, \tilde{Z}\right)$ used to maximize InfoNCE, up to an additive constant, has the form:

$$T\left(Z, \tilde{Z}\right) = -\frac{\left\|Z - \tilde{Z}\right\|^2}{2\sigma^2_{\mathrm{noise}}} - \log \sum_{k=1}^{K} \exp\left(-\frac{\left\|\tilde{Z} - \boldsymbol{\mu}_k\right\|^2}{2\sigma^2_{\mathrm{class}}}\right)$$

Proof. The optimal critic (similarity function) in the limit for InfoNCE is equivalent to the log-likelihood ratio:

[1] For example, consider $Z' \perp\!\!\!\perp Z''$, which corresponds to PMI $= 0$ regardless of the marginal distributions.

$$T^*(Z, \tilde{Z}) = \log p(\tilde{Z}|Z) - \log p_{\text{target}}(\tilde{Z}) + \text{const}.$$

Since $\tilde{Z} = Z + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}} I)$, the conditional distribution is normal. The logarithm of its density is:

$$\log p(\tilde{Z}|Z) = -\frac{\|\tilde{Z} - Z\|^2}{2\sigma^2_{\text{noise}}} - \frac{d}{2}\log(2\pi\sigma^2_{\text{noise}}).$$

The target marginal distribution $p_{\text{target}(\tilde{Z})}$ is a mixture of K Gaussians:

$$p_{\text{target}(\tilde{Z})} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{(2\pi\sigma^2_{\text{class}})^{\frac{d}{2}}} \exp\left(-\frac{\|\tilde{Z} - \boldsymbol{\mu}_k\|^2}{2\sigma^2_{\text{class}}}\right)$$

Taking the logarithm of the target distribution density, we obtain:

$$\log p_{\text{target}(\tilde{Z})} = \log\left(\sum_{k=1}^{K} \exp\left(-\frac{\|\tilde{Z} - \boldsymbol{\mu}_k\|^2}{2\sigma^2_{\text{class}}}\right)\right) - \log K - \frac{d}{2}\log(2\pi\sigma^2_{\text{class}}).$$

Substituting $\log p(\tilde{Z}|Z)$ and $\log p_{\text{target}(\tilde{Z})}$ into the initial expression for the critic $T^*(Z, \tilde{Z})$ and absorbing all terms independent of $\tilde{Z}$ into a single constant, we arrive at the required expression.

□

The duality in Lemma 5.1 gives a formal version of the distributional view. When the critic is fixed, the variational objective favors a Gibbs distribution determined by the critic. This does not mean that standard contrastive SSRL performs exact distribution matching, because in SSRL both $\mathbb{P}_{Z',Z''}$ and $\mathbb{P}_{Z'} \otimes \mathbb{P}_{Z''}$ depend on the encoder. It does show, however, that critic form can act as an implicit preference over representation distributions. The fixed-critic construction above makes this preference explicit.

This perspective also clarifies the role of projection-head critics. A fully separable critic using an inner product, $T(Z', Z'') = g(Z')^{\mathsf{T}} g(Z'')$, cannot universally approximate arbitrary functions of two variables. Therefore, the variational bound cannot be optimized equally well for all encoder-output distributions under this critic family. To improve the objective, the encoder must produce representations whose pairwise structure is accessible through the restricted critic. In this sense, the projection head does not merely post-process representations before the loss. Together with the similarity function, it defines a critic family that biases the encoder-output distribution.

The strength of this bias depends on critic expressiveness. If the critic is too restricted, the encoder may be pushed toward distributions that satisfy the critic but discard useful information. If the critic is too expressive, it can model pairwise relations directly and reduce the pressure on the encoder to expose useful structure. Projection-head MLPs occupy an intermediate regime: they increase the capacity of the separable critic, but still constrain comparison to pass through independently transformed representations. This provides the mechanism behind the critic-complexity tradeoff studied in Section 4 and motivates the empirical comparisons in Section 6.

6 Experiments

We use experiments to test two consequences of the preceding analysis. First, we compare critic architectures to study whether a better InfoNCE value necessarily leads to a better retained encoder representation. Second, we use fixed critics to test whether critic form can steer the embedding distribution toward a prescribed structure.

Shared setup The models are trained with a contrastive objective based on InfoNCE. The critic comparison uses CIFAR-10 images at 32×32 resolution. The fixed-critic visualizations also use MNIST, selected CIFAR-100 coarse superclasses, and selected ImageNet classes, with two-dimensional retained embeddings. During training, two augmented views are generated for each input image using standard data augmentations, and the chosen critic scores their encoder outputs.

6.1 Critic Comparison: Separable vs MLP

We compare two critic architectures under the same contrastive training pipeline. The **separable critic** applies a shared projection head to each encoder output independently, normalizes the projected representations, and scores them by a temperature-scaled dot product. The **non-separable MLP** critic concatenates the two encoder outputs and scores the pair jointly with a multilayer perceptron. The separable critic has the form $T_{\text{sep}(z',z'')} = \frac{\langle g(z'), g(z'') \rangle}{\tau}$, where g is shared across both views. The MLP critic instead uses $T_{\text{MLP}(z',z'')} = \frac{\text{MLP}([z';z''])}{\tau}$, and can therefore model pairwise interactions after concatenation. This comparison isolates the effect of critic parameterization while keeping the retained representation before the critic fixed.

We evaluate each encoder using the InfoNCE estimate $-\mathcal{L}_{\text{NCE}}$, linear-probe accuracy on frozen encoder features, and spectral statistics of the embedding covariance. Effective rank is defined as $\exp\left(H\left(\frac{\lambda}{\sum_i \lambda_i}\right)\right)$, where λ_i are covariance eigenvalues. Stable rank is $\sum_i \frac{\lambda_i}{\lambda_{\max}}$. Together with the normalized eigenspectrum and cumulative explained variance, these metrics describe how variance is distributed across embedding directions.

The MLP critic achieves a substantially higher InfoNCE estimate than the separable critic (6.7370 ± 0.2470 versus 1.8614 ± 0.0117 at the final epoch). It also yields higher effective rank (14.12 ± 1.02 versus 11.01 ± 1.80) and higher stable rank (3.92 ± 0.29 versus 3.43 ± 0.54). However, this does not translate into better downstream performance: the separable critic achieves higher final linear-probe accuracy (0.6886 ± 0.0138 versus 0.6511 ± 0.0073).

A depth- and parameter-matched comparison on a seven-class ImageNet subset shows the same loss–downstream mismatch: the depth-2 separable critic reaches 91.15% linear-probe accuracy versus 87.93% for the joint MLP, despite a lower InfoNCE estimate (1.579 versus 4.289); see Table 7.

The spectral structure helps explain this discrepancy. Although the non-separable MLP critic uses more effective dimensions, it produces an uneven covariance spectrum with a sharp drop near the leading components, close to reported intrinsic-dimension estimates for CIFAR-10 (Pope et al., 2021). This suggests that the MLP critic preserves more of the dominant geometric structure of the input distribution. In contrast, the separable projection-head critic transforms the covariance spec-

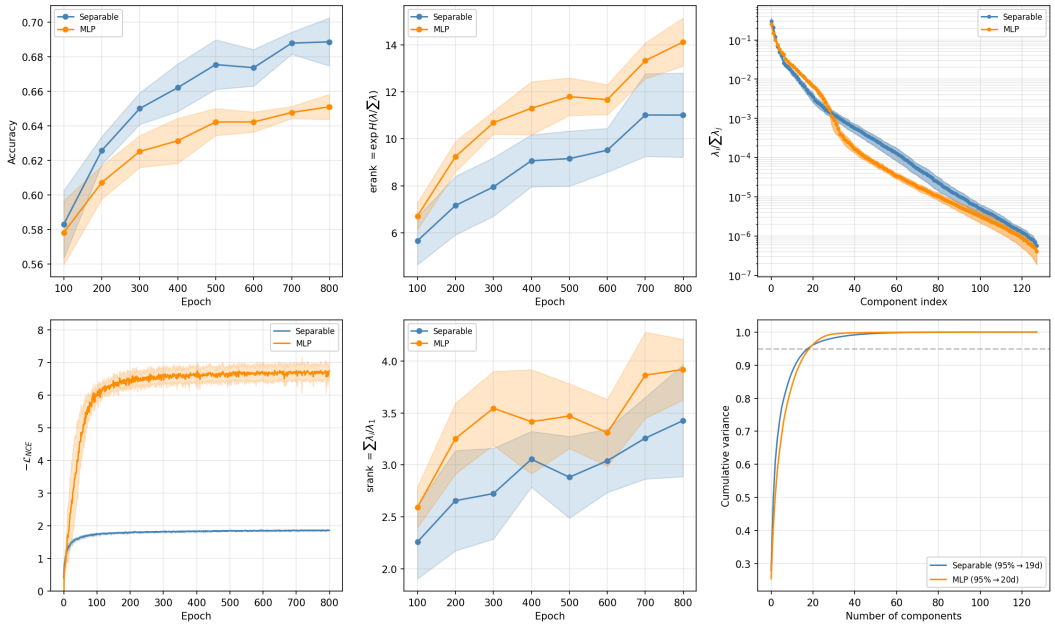


Figure 3: Comparison of separable and non-separable MLP critics on CIFAR-10. Top row: linear-probe accuracy, effective rank, and normalized covariance eigenspectrum. Bottom row: InfoNCE estimate, stable rank, and cumulative variance explained. Curves show mean over six seeds with shading denoting one standard deviation.

trum into a near-perfect power law with no comparable drop. This indicates more active reshaping of the encoder distribution: the separable critic spreads information into weaker spectral directions, making the representation more digestible for a linear probe. Thus, spectral rank metrics are useful for detecting collapse, but they do not by themselves characterize downstream usability.

6.2 Fixed-Critic Gaussian Mixtures

We next test whether a fixed critic can impose a prescribed embedding geometry. For this purpose, we use the Gaussian-mixture critic derived in Proposition 5.2. The critic contains an alignment term that keeps paired embeddings close and a prototype-mixture term that favors concentration around a fixed set of centers. No learnable projection head, adversarial discriminator, generative model, or explicit prior-matching loss is used.

We consider three two-dimensional prototype configurations: an equilateral triangle, an elongated triangle, and four prototypes arranged as a rotated square. On CIFAR-10, we use the class subsets (airplane, automobile, bird), (cat, deer, dog), and (frog, horse, ship, truck). For ImageNet, the selected class subsets for each configuration are listed in Figure 5. The prototypes are fixed throughout training, and only the encoder is optimized.

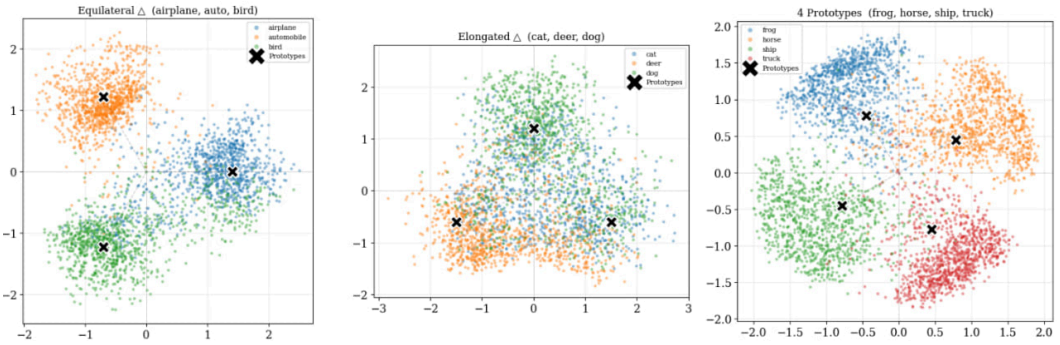


Figure 4: Visualization of learned representations and prototypes on CIFAR-10. Left: three-class subset (airplane, automobile, bird) forming an approximately equilateral configuration. Middle: three-class subset (cat, deer, dog) with an elongated geometry. Right: four-class subset (frog, horse, ship, truck).

On CIFAR-10, the embeddings exhibit prototype-oriented structure, with visible class overlap in the two-dimensional space (Figure 4). The equilateral and four-prototype settings produce clear prototype-aligned structure. The elongated triangle setting is more difficult and exhibits visible class overlap, but the global geometry still follows the prescribed critic structure.

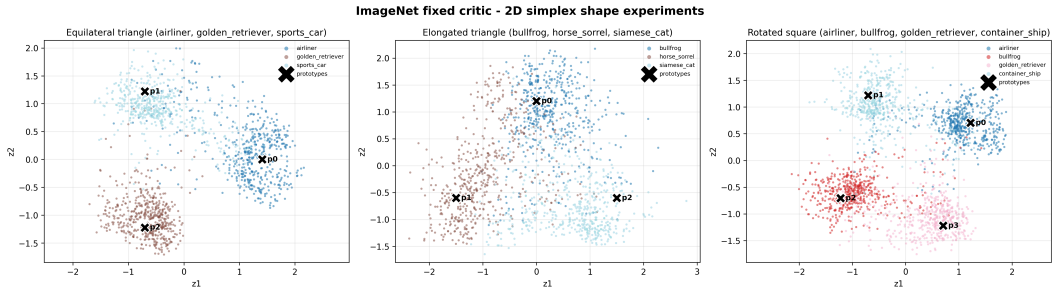


Figure 5: Two-dimensional fixed-critic embeddings for selected ImageNet classes. Left: equilateral triangle (airliner, golden retriever, sports car). Middle: elongated triangle (bullfrog, horse sorrel, Siamese cat). Right: four-prototype configuration (airliner, bullfrog, golden retriever, container ship). Colored points show encoder embeddings; black crosses mark fixed prototypes.

On the selected ImageNet classes, embeddings concentrate around the prescribed prototypes in all three configurations, with visible dispersion and overlap between classes (Figure 5). These plots demonstrate prototype-oriented organization on controlled class subsets. Aggregate downstream and

clustering comparisons with learned critics on three controlled ImageNet class-subset configurations appear in Table 6.

Additional MNIST and CIFAR-100 experiments, including their class configurations and results, appear in Appendix C (Figure 6, Figure 7, and Table 5).

Together, these experiments support the distribution-shaping interpretation. They show that fixed critic form can steer the retained encoder representation toward selected mixture-like structures. This occurs without adding a separate distribution-matching mechanism, which distinguishes the construction from explicit prior-matching or adversarial approaches.

7 Limitations

Our analysis studies critic-induced distributional bias through idealized variational objectives and selected critic families. While this captures the mechanisms relevant to projection-head-based contrastive learning, the conclusions may depend on optimization quality, architecture choices, augmentations, and finite-sample effects. Empirically, our experiments focus on standard vision benchmarks and selected critic forms, including separable, non-separable, and fixed critics. The fixed-critic constructions demonstrate controllable distribution shaping for prescribed structures such as Gaussian mixtures, but do not cover all possible target distributions or downstream tasks.

8 Discussion

In this work, we revisited the role of projection heads in contrastive self-supervised learning through the lens of the Deep InfoMax equivalence (Hjelm et al., 2019; Oord et al., 2019). By viewing contrastive losses as variational lower bounds on mutual information, we revealed the information-theoretic and distribution matching trade-off controlled by the architectural separation line between the encoder and the variational critic. If the critic is highly expressive, it absorbs the burden of MI estimation, extracting task-relevant signals directly and leaving the encoder's representations unrefined. Conversely, restricting the critic shifts the representational burden to the encoder. This dynamic explains why pure information content is an unreliable predictor of downstream linear probing accuracy (Tschannen et al., 2020), and highlights the fundamental trade-off between representational informativeness and amenability to simple downstream models.

Crucially, we attribute this refinement to strong distributional biases imposed by the critic's architectural constraints. For instance, a fully separable critic relying on an inner product cannot universally approximate all functions of two variables. Because the maximum MI cannot be achieved for arbitrary output distributions under this bottleneck, the encoder is forced to adapt its outputs into specific distributions – such as Gaussians – that minimize the loss for that restricted family. This perspective not only explains the inferior downstream performance of dense critics, but also mathematically establishes that a critic's architecture inherently dictates the distribution of the learned representations.

To empirically validate these findings, we craft critics that automatically impose a target distribution, such as a Gaussian mixture, on the learned representations. The results align with our theoretical predictions and hold across datasets of varying complexity. Notably, the induced clusters often correspond to the true underlying classes. We consider this property to be especially useful, since it potentially allows for simple unsupervised classification and clustering.

References

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., and others. A cookbook of self-supervised learning. *Arxiv Preprint Arxiv:2304.12210*, 2023.

Bardes, A., Ponce, J., and LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=xm6YD62D1Ub

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual Information Neural Estimation. In J. Dy & A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, pp. 531–540. PMLR, 2018. https://proceedings.mlr.press/v80/belghazi18a.html

Betser, R., Gofer, E., Levi, M. Y., and Gilboa, G. InfoNCE Induces Gaussian Distribution. *Arxiv Preprint Arxiv:2602.24012*, 2026.

Butakov, I., Semenenko, A., and others. Noise-Injected Deep InfoMax. *Arxiv Preprint Arxiv:2410.06993*, 2024.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *ICML*, 2020.

Donsker, M. D., and Varadhan, S. R. Asymptotic evaluation of certain markov process expectations for large time. IV. *Communications on Pure and Applied Mathematics*, *36*(2), 183–212, 1983. https://doi.org/10.1002/cpa.3160360204

Ericsson, L., Gouk, H., Loy, C. C., and Hospedales, T. M. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42–62, 2022. https://doi.org/10.1109/MSP.2021.3134634

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723–773, 2012.

Gui, Y., Ma, C., and Zhong, Y. *Unraveling Projection Heads in Contrastive Learning: Insights from Expansion and Shrinkage*, 2023. https://arxiv.org/abs/2306.03335

Gupta, K., Ajanthan, T., Hengel, A. van den, and Gould, S. *Understanding and Improving the Role of Projection Head in Self-Supervised Learning*, 2022. https://arxiv.org/abs/2212.11491

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *CVPR*, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *International Conference on Learning Representations*, 2019. https://openreview.net/forum?id=Bklr3j0cKX

Kingma, D. P., and Welling, M. Auto-Encoding Variational Bayes. *International Conference on Learning Representations*, 2014. https://openreview.net/forum?id=33X9fd2-9FyZd

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced Wasserstein auto-encoders. *International Conference on Learning Representations*, 2018.

Ma, J., Hu, T., and Wang, W. Deciphering the Projection Head: Representation Evaluation Self-supervised Learning. In K. Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 4724–4732. International Joint Conferences on Artificial Intelligence Organization, 2024. https://doi.org/10.24963/ijcai.2024/522

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. *Adversarial Autoencoders*, 2015. https://arxiv.org/abs/1511.05644

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theor.*, *56*(11), 5847–5861, 2010. https://doi.org/10.1109/TIT.2010.2068870

Oord, A. van den, Li, Y., and Vinyals, O. *Representation Learning with Contrastive Predictive Coding*, 2019. https://arxiv.org/abs/1807.03748

Ouyang, Z., Hu, K., Zhang, Q., Wang, Y., and Wang, Y. Projection Head is Secretly an Information Bottleneck. *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=L0evcuybH5

Ozbulak, U., Lee, H. J., Boga, B., Anzaku, E. T., Park, H.-min, Messem, A. V., Neve, W. D., and Vankerschaver, J. Know Your Self-supervised Learning: A Survey on Image-based Generative and Discriminative Training. *Transactions on Machine Learning Research*, 2023. https://openreview.net/forum?id=Ma25S4ludQ

Polyanskiy, Y., and Wu, Y. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024. https://books.google.ru/books?id=CySo0AEACAAJ

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On Variational Bounds of Mutual Information. In K. Chaudhuri & R. Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, pp. 5171–5180. PMLR, 2019. https://proceedings.mlr.press/v97/poole19a.html

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The Intrinsic Dimension of Images and Its Impact on Learning. *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=XJk19XzGq2J

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., and others. Dinov3. *Arxiv Preprint Arxiv:2508.10104*, 2025.

Song, Z., Su, X., Wang, J., Qiang, W., Zheng, C., and Sun, F. *Towards the Sparseness of Projection Head in Self-Supervised Learning*, 2023. https://arxiv.org/abs/2307.08913

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *Arxiv Preprint Arxiv:1711.01558*, 2017.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On Mutual Information Maximization for Representation Learning. *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=rkxoh24FPH

Xue, Y., Gan, E., Ni, J., Joshi, S., and Mirzasoleiman, B. Investigating the Benefits of Projection Head for Representation Learning. *The Twelfth International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=GgEAdqYPNA

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In M. Meila & T. Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, pp. 12310–12320. PMLR, 2021. https://proceedings.mlr.press/v139/zbontar21a.html

A Complete proofs

Proposition 4.1. Consider the setup from (3). Let $f = g_n \circ ... \circ g_1$ be a feed-forward network of n layers, and ${L_k}'/{L_k}''$ be outputs of the k-th layer on X'/X'' respectively. Then

$$\textbf{(a)}\ \mathsf{I}({L_k}'; X) \geq \mathsf{I}({L_k}'; {L_k}'') \qquad \textbf{(b)}\ \mathsf{I}({L_1}'; {L_1}'') \geq ... \geq \mathsf{I}({L_n}'; {L_n}'') \geq \mathcal{L}^{\mathrm{MI}}\left[X', X''; T_{\mathrm{joint}}\right]$$

Let $\mathscr{M}_k$ be a set of extended critics T_k that incorporate the last $n-k$ layers of f into T. Then

$$\sup_{T_k \in \mathscr{M}_k} \mathcal{L}^{\mathrm{MI}}[{L_k}', {L_k}''; T_k] \geq \mathcal{L}^{\mathrm{MI}}\left[X', X''; T_{\mathrm{joint}}\right]$$

Proof of Proposition 4.1. Since the augmentations are conditionally independent for any given X, we have the following directed acyclic causality graph:

$${L_n}' \leftarrow ... \leftarrow {L_1}' \longleftarrow X' \leftarrow X \rightarrow X'' \longrightarrow {L_1}'' \rightarrow ... \rightarrow {L_n}'',$$

Due to the Markov chain reversal property (Polyanskiy and Wu, 2024), it can be rearranged into the following Markov chain:

$${L_n}' \rightarrow ... \rightarrow {L_1}' \longrightarrow X' \rightarrow X \rightarrow X'' \longrightarrow {L_1}'' \rightarrow ... \rightarrow {L_n}''$$

Therefore, due to DPI, we have the first part of the Proposition's statement.

The second part follows from the equality

$$\mathcal{L}^{\mathrm{MI}}\left[X', X''; T_{\mathrm{joint}}\right] = \mathcal{L}^{\mathrm{MI}}[{L_k}', {L_k}''; T_k],$$

where $T_k = T \circ \left(g_n \circ ... \circ g_{k+1} \times g_n \circ ... \circ g_{k+1}\right)$. Taking the supremum over all possible T_k on the right-hand side, while keeping T and f fixed on the left concludes the proof. □

Lemma 5.1. (Distribution matching duality) Consider a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ and $T \in \mathscr{M}(\mathcal{F}, \mathcal{B}_{\mathbb{R}})$. Define $\mathbb{G}$ such that $\frac{\mathrm{d}\,\mathbb{G}}{\mathrm{d}\,\mathbb{Q}} \sim e^T$. Let $\mathscr{R} = \{\mathbb{P} \in \mathscr{P} \mid \mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] \leq \mathsf{KL}[\mathbb{G} \parallel \mathbb{Q}]\}$. Then

$$\sup_{\mathbb{P} \in \mathscr{R}} \mathcal{L}_{\mathrm{DV}} = \sup_{\mathbb{P} \in \mathscr{R}} \mathcal{L}_{\mathrm{NWJ}} = \mathsf{KL}[\mathbb{G} \parallel \mathbb{Q}],$$

where the supremum is attained at $\mathbb{P} = \mathbb{G}$.

Proof of Lemma 5.1. Note that

$$\mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] - \mathcal{L}_{\mathrm{DV}}[T; \mathbb{P}, \mathbb{Q}] = \mathbb{E}_{\mathbb{P}} \log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}} - \mathbb{E}_{\mathbb{P}} \log \frac{\mathrm{d}\,\mathbb{G}}{\mathrm{d}\,\mathbb{Q}} = \mathbb{E}_{\mathbb{P}} \log \frac{\mathrm{d}\,\mathbb{P}}{\mathrm{d}\,\mathbb{Q}} \frac{\mathrm{d}\,\mathbb{Q}}{\mathrm{d}\,\mathbb{G}} = \mathsf{KL}[\mathbb{P} \parallel \mathbb{G}]$$

Therefore,

$$\mathcal{L}_{\mathrm{DV}}[T; \mathbb{P}, \mathbb{Q}] = \mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] - \mathsf{KL}[\mathbb{P} \parallel \mathbb{G}]$$

Note that $\mathsf{KL}[\mathbb{P} \parallel \mathbb{Q}] \leq \mathsf{KL}[\mathbb{G} \parallel \mathbb{Q}]$ by the definition of $\mathscr{R}$, and $-\mathsf{KL}[\mathbb{P} \parallel \mathbb{G}] \leq 0$. The latter term attains its maximal value iff $\mathbb{P} = \mathbb{G}$, which also delivers maximum to the first term.

The result for $\mathcal{L}_{\mathrm{NWJ}}$ follows from $\mathcal{L}_{\mathrm{NWJ}} \leq \mathcal{L}_{\mathrm{DV}}$ (Belghazi et al., 2018), with the equality at $\mathbb{P} = \mathbb{G}$. □

B Details of implementation

B.1 Backbone Architecture

For all experiments on MNIST, CIFAR-10, and CIFAR-100, we employ a modified ResNet-18 backbone. The first convolutional layer is replaced by a 3×3 convolution with stride 1 and padding 1 (no maxpool), adapted to the input channel count (1 for MNIST, 3 for CIFAR). The classification head is replaced by a three-layer MLP projection head. Full architecture details are provided in Table 1.

Table 1: Neural network architectures used in simplex-shape experiments.

Module	**Architecture**
ResNet-18 backbone (shared)	conv1: Conv2d(C_{in}, 64, ks=3, stride=1, pad=1, bias=False) maxpool: Identity() layer1–layer4: standard ResNet-18 residual blocks
Projection head (appended to backbone)	Linear(512, 512, bias=False) → BatchNorm1d(512) → ReLU Linear(512, 128, bias=False) → BatchNorm1d(128) → ReLU Linear(128, d, bias=True)

B.2 Data Augmentation

For CIFAR-10 and CIFAR-100 experiments, each training sample is presented as a pair $(\boldsymbol{x}_{\text{aug}}, \boldsymbol{x}_{\text{clean}})$, where the clean view is the normalized original image and the augmented view is produced by the following transform:

Table 2: CIFAR augmentation pipeline (PairTransform).

View	**Transforms**
Augmented	RandomResizedCrop(32, scale=(0.2, 1.0), interp=BILINEAR) RandomApply([ColorJitter(0.4, 0.4, 0.4, 0.1)], p=0.8) RandomGrayscale(p=0.2) RandomHorizontalFlip(p=0.5) ToTensor() · Normalize(mean, std)
Clean	ToTensor() · Normalize(mean, std)

B.3 Training Hyperparameters

Training hyperparameters are summarised in Table 3. For all CIFAR experiments, Gaussian noise with standard deviation σ_{inject} is added to the clean embedding $f(\boldsymbol{x}_{\text{clean}}))$ during training only.

Table 3: Training hyperparameters for each dataset.

Hyperparameter	**MNIST**	**CIFAR-10**	**CIFAR-100**
Embedding dim d	2	2	2 / 3
Batch size	512	1024	1024
Epochs	200	300	600
Optimizer	Adam	Adam	Adam
Learning rate	10^{-3}	10^{-3}	10^{-3}
LR schedule	CosineAnnealingLR	CosineAnnealingLR	CosineAnnealingLR
Min LR ($\eta_{\min}$)	10^{-5}	10^{-5}	10^{-5}
$T_{\max}$	200	300	600
σ_{noise}	0.6	0.4	0.4
σ_{class}	0.0	0.2	0.2
Noise std σ_{inject}	0.6	0.4	0.4
Prototype radius r	1.41	1.41	1.41 / 1.0

B.4 CIFAR-100 Superclass Partitioning

CIFAR-100 experiments use coarse (superclass) labels. The three plotted configurations use the following class subsets:

Table 4: CIFAR-100 superclass subsets used in 2D shape experiments.

Experiment	**Prototype config**	**Superclasses**
Equilateral △	$K = 3$, equilateral	flowers, people, vehicles 1
Elongated △	$K = 3$, elongated	trees, fish, insects
Rotated square	$K = 4$, rotated	flowers, people, vehicles 1, trees

C Additional fixed-critic visualizations

The following plots extend the two-dimensional prototype experiments in the main text. Colored points show retained encoder embeddings; black crosses mark the fixed prototypes. Each panel uses only the selected classes named in its caption.

On MNIST, the equilateral triangle, elongated triangle, and four-prototype configurations use digits $(0, 1, 2)$, $(3, 4, 5)$, and $(6, 7, 8, 9)$, respectively.

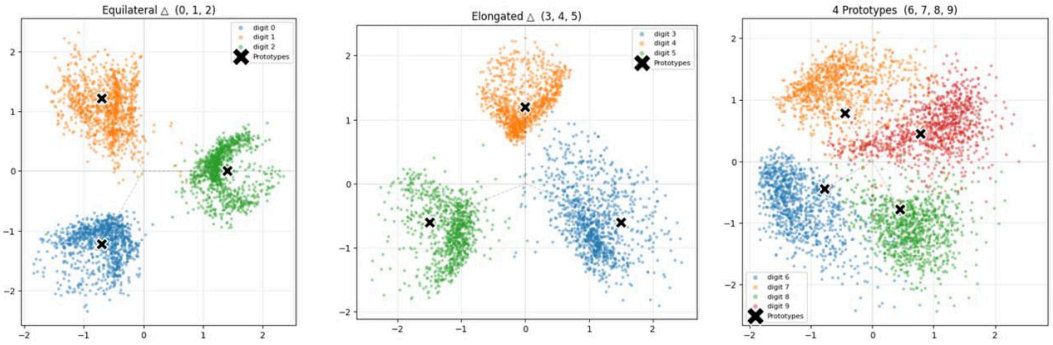


Figure 6: Visualization of learned representations and prototypes on MNIST. Left: three-class subset (digits 0, 1, 2) forming an approximately equilateral configuration. Middle: three-class subset (digits 3, 4, 5) with an elongated geometry. Right: four-class subset (digits 6–9).

On MNIST, the learned embeddings closely follow the prescribed prototype configurations. The class clusters are compact and well separated across all three settings, including the elongated triangle and the four-prototype geometry. This provides a controlled example in which changing only the fixed critic changes the geometry of the retained encoder representation.

The CIFAR-100 configurations show prototype-oriented structure with varying degrees of class overlap; these plots use selected coarse classes rather than all 100 fine classes.

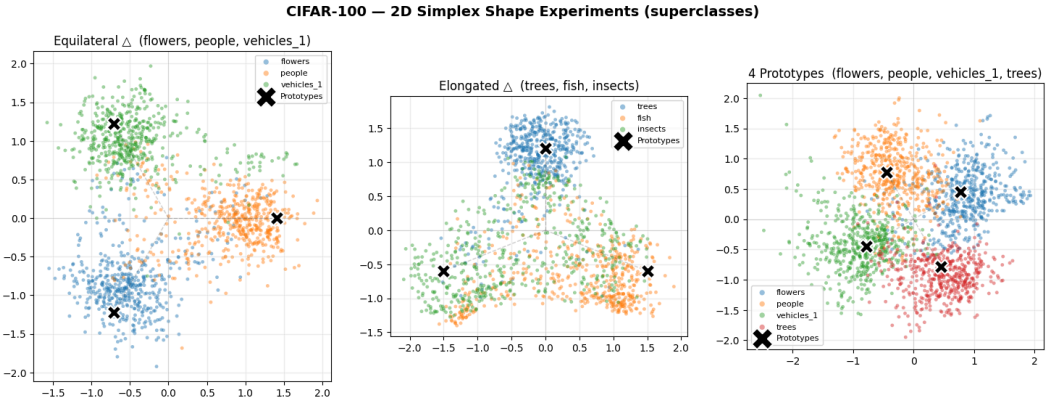


Figure 7: Fixed-critic embeddings for selected CIFAR-100 coarse superclasses. Left: equilateral triangle (flowers, people, vehicles 1). Middle: elongated triangle (trees, fish, insects). Right: four-prototype configuration (flowers, people, vehicles 1, trees).

Table 5: CIFAR-100 coarse-superclass results for the three fixed-critic geometries in Figure 7. Higher is better except for Davies–Bouldin (DB).

Geometry	NMI	ARI	k-means	Proto. acc.	Silhouette	DB
Equilateral	0.518	0.577	83.6%	83.4%	0.391	0.875
Elongated	0.404	0.394	69.2%	68.7%	0.214	2.127
Square	0.757	0.802	92.1%	92.0%	0.513	0.623

The elongated configuration shows the most overlap, while the four-prototype configuration gives the strongest class recovery among these selected subsets.

C.1 Fixed versus learned critics

The fixed Gaussian-mixture critic is also compared with learned separable and joint MLP critics on three controlled ImageNet class-subset configurations. Table Table 6 reports aggregate downstream and clustering measures.

Table 6: Aggregate downstream and clustering results over three controlled ImageNet class-subset configurations. Higher is better except for Davies–Bouldin (DB).

Critic	Linear	5-NN	NMI	ARI	Clust. acc.	Sil.	DB
Fixed Gaussian mixture	89.69%	88.39%	0.655	0.678	87.02%	0.434	0.825
Learned separable	86.14%	85.70%	0.567	0.537	81.21%	0.298	1.094
Learned joint MLP	87.00%	86.25%	0.599	0.619	84.87%	0.302	1.099

The fixed critic has the best aggregate score in every reported measure, although the configuration-level ranking can vary.

D Matched critic comparison on an ImageNet subset

This experiment uses a seven-class ImageNet subset, a 128-dimensional retained representation, a ResNet-18 encoder, identical splits and training, and three seeds. Separable critics use a shared projector, L2 normalization, and a temperature-0.5 dot product; joint MLPs score concatenated pairs. Critics are matched by parameter count within each depth, and none uses internal BatchNorm. This 128-dimensional critic comparison is distinct from the two-dimensional fixed-critic visualization in Figure 5.

Table 7: Matched-capacity critic comparison on a seven-class ImageNet subset. Numbers in parentheses give critic depth. Linear-probe and 5-NN accuracies are percentages; values after ± are across three seeds.

Critic	Architecture	Params	InfoNCE	Linear	5-NN	Eff. rank
Sep. (2)	128→512→128	131,712	1.579±0.004	91.15±0.46	89.44±0.86	8.924
MLP (2)	256→511→1	131,839	4.289±0.064	87.93±1.13	83.71±0.91	7.805
Sep. (3)	128→256→256→128	131,712	1.562±0.004	90.58±0.36	87.50±0.28	7.144
MLP (3)	256→256→256→1	131,841	4.233±0.053	86.55±0.16	83.24±0.25	7.986
Original MLP	256→512→512→1	394,753	4.341±0.013	87.45±1.40	83.76±0.46	8.580

At both matched depths, the separable critic gives higher linear-probe and 5-NN accuracy despite a lower InfoNCE estimate. Effective-rank ordering reverses between depths 2 and 3, so effective rank alone does not explain the downstream gap.