CONTRASTIVE LEARNING ACROSS SAMPLES AND FEATURES THROUGH THE LENS OF MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive representation learning typically contrasts corresponding samples across two views, as in InfoNCE. However, the mutual information bound of sample-wise InfoNCE is limited by the number of negative samples and does not explicitly discourage redundancy across features. To address these issues, we study sample-wise and feature-wise contrastive learning through a common mutual information formulation. Our analysis of the feature axis reveals an additional penalty arising from the dependence and non-exchangeability of feature channels and motivates feature standardization before contrastive matching. The two objectives are complementary: the sample-wise objective distinguishes inputs, while the feature-wise objective promotes informative, non-redundant features. We prove that the mutual information captured along either axis lower-bounds the mutual information between the full embedding matrices, providing a principled basis for combining the two objectives. Our extensive experiments across image representation learning and cross-modal retrieval demonstrate that the combined objective improves over either objective alone and achieves superior performance compared with existing self-supervised learning methods.

1 INTRODUCTION

Contrastive representation learning seeks to learn representations by making the representations of two different views of the same input similar, while making them distinguishable from representations of other inputs. A widely used objective is InfoNCE (Oord et al., 2018), where the representation of one view is treated as a query, and the model must identify the representation of its corresponding view among a set of representations from the same batch. Representations corresponding to other inputs act as negative examples, providing alternatives that the model is encouraged to distinguish from. Although effective, this construction imposes two limitations. First, because the set of alternatives is formed from other samples in the batch, the amount of mutual information that the objective can certify is limited by the number of negatives, creating a dependence on batch size (Poole et al., 2019; McAllester & Stratos, 2020; Lee et al., 2025). Second, the objective operates entirely at the level of individual samples: it encourages representations of different inputs to be distinguishable, but places no direct constraint on how information is distributed across the feature channels within each representation. Consequently, different output channels may remain redundant.

Several approaches address these limitations from different directions. BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) avoid explicit negative samples through architectural asymmetry, reducing the dependence on large batches. Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2022), in contrast, explicitly regularize feature channels to encourage non-redundant representations. Feature-wise contrastive objectives have also been explored, including DimCL (Nguyen et al., 2023) and SimCSE++ (Xu et al., 2023), primarily to promote feature diversity or address feature redundancy. These approaches highlight two distinct aspects of representation learning: distinguishing different samples and distributing information across different features. However, the sample-wise and feature-wise contrastive perspectives have not been characterized together through the mutual information formulation underlying InfoNCE, and it remains unclear how they can be combined in a principled manner.

We therefore study the sample and feature axes through a common mutual information formulation, as shown in Figure 1. Let $Z, \hat{Z} \in \mathbb{R}^{B\times D}$ denote the representations produced by two branches,

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.

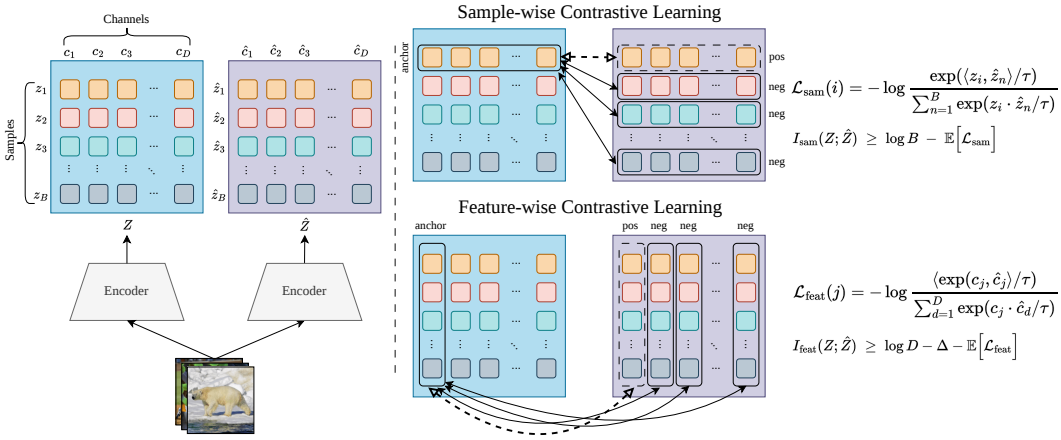


Figure 1: Sample-wise contrastive learning contrasts rows over B samples, while feature-wise contrastive learning contrasts columns over D features, yielding mutual information bounds governed by $\log B$ and $\log D - \Delta$, respectively.

where B indexes samples and D indexes output features. The conventional sample-wise perspective treats the B rows as the population and asks whether an embedding from one branch can identify its corresponding embedding among the samples of the other branch. The feature-wise perspective instead treats the D columns as the population and asks whether a feature channel can identify its corresponding channel among those of the other representation.

Applying the mutual information formulation along the feature axis yields a feature-wise contrastive objective closely related to those used in prior work. Our analysis, however, reveals an additional penalty arising from the dependence and non-exchangeability of feature candidates, a distinction that does not appear in the standard sample-wise formulation. This analysis motivates feature standardization before contrastive matching. It also shows that the feature-wise bound is governed by the feature dimension D rather than the batch size, providing a complementary contrastive signal. Our experiments further highlight the importance of these design choices.

We connect the two axis-wise objectives to the mutual information of the full embedding matrices. We prove that both the sample-wise and feature-wise mutual information quantities lower-bound the joint mutual information between Z and $\hat{Z}$. This establishes that the two axis-wise formulations provide complementary lower bounds on the same underlying quantity. Consequently, we combine them through a convex combination, obtaining a combined objective with a corresponding lower bound on the joint mutual information.

Our formulation also reveals a connection between contrastive learning and redundancy-reduction methods such as Barlow Twins. Although these approaches are typically presented as distinct, both encourage corresponding representations to agree while discouraging redundancy across features. Our analysis shows that they can be viewed within a common matching framework: our feature-wise contrastive objective enforces matching through relative competition among features, whereas Barlow Twins directly penalizes deviations of feature correlations from their desired values.

In summary, our work makes three main contributions: 1) we provide a mutual information characterization of feature-wise contrastive learning; 2) we show that the sample-wise and feature-wise mutual information quantities both lower-bound the mutual information between the full embedding matrices, yielding a principled combined objective and corresponding bound; and 3) we conduct extensive experiments across image representation learning and cross-modal retrieval, showing that the combined objective improves over either objective alone and compares favorably with existing self-supervised learning methods.

2 Problem Formulation

Let $\mathcal{D} = \{(x_n, \hat{x}_n)\}_{n=1}^{M}$ denote a dataset of M paired observations drawn from two branches. The two branches may correspond to two augmented views of the same input, as is standard in self-supervised representation learning (Chen et al., 2020), or to two different modalities (Radford et al.,

2021). Each branch is processed by its respective encoder, ϕ and $\hat{\phi}$ (which may share weights, as is typical in single-network multiview architectures (Chen et al., 2020; Bardes et al., 2022)), mapping inputs into a shared D-dimensional output space. For a minibatch of B paired observations, stacking their representations row-wise yields two matrices

$$Z, \hat{Z} \in \mathbb{R}^{B \times D}, \qquad Z_{i,d} = \phi(x_i)_d, \quad \hat{Z}_{i,d} = \hat{\phi}(\hat{x}_i)_d,$$

with $i \in [B]$ indexing samples and $d \in [D]$ indexing output features.

The matrices Z and $\hat{Z}$ admit two complementary statistical perspectives:

- **Sample-wise:** The statistical population consists of the B samples, with each sample represented by its D output features. We write $z_i := Z_{i,:}$ and $\hat{z}_i := \hat{Z}_{i,:}$ (both in $\mathbb{R}^D$) to denote the two embeddings of sample i. Let $I \sim \mathrm{Unif}[B]$ be independent of $(Z, \hat{Z})$.
- **Feature-wise:** The statistical population consists of the D output features, with each feature represented by its values across the B samples. We write $c_d := Z_{:,d}$ and $\hat{c}_d := \hat{Z}_{:,d}$ (both in $\mathbb{R}^B$) to denote the two representations of feature d. Let $J \sim \mathrm{Unif}[D]$ be independent of $(Z, \hat{Z})$.

The corresponding sample-wise and feature-wise mutual informations are defined as

$$I_{\mathrm{sam}}(Z; \hat{Z}) := I(z_I; \hat{z}_I \mid I) = \frac{1}{B} \sum_{i=1}^{B} I(z_i; \hat{z}_i), \tag{1}$$

$$I_{\mathrm{feat}}(Z; \hat{Z}) := I(c_J; \hat{c}_J \mid J) = \frac{1}{D} \sum_{j=1}^{D} I(c_j; \hat{c}_j). \tag{2}$$

The sample-wise perspective corresponds to the conventional contrastive-learning formulation, where each sample embedding is matched to its corresponding embedding among candidate samples from the other branch. The feature-wise perspective instead treats each output feature as the unit of correspondence, asking whether a feature from one branch can be identified among the D output features of the other branch. In the following sections, we derive contrastive lower bounds for both $I_{\mathrm{sam}}(Z; \hat{Z})$ and $I_{\mathrm{feat}}(Z; \hat{Z})$ and relate them to the joint mutual information $I(Z; \hat{Z})$ between the two embedding matrices.

3 Mutual Information Objectives

We now derive contrastive variational bounds for the sample-wise and feature-wise mutual information quantities introduced in Section 2, and use their tractable loss terms as training objectives.

3.1 Sample-wise Contrastive Objective

First, we formulate the sample-wise contrastive framework and derive the InfoNCE loss (Oord et al., 2018), which subsequently enables the formulation of the feature-wise contrastive loss. Fix an anchor sample z_i and its corresponding partner $\hat{z}_i$. Let $U \sim \mathrm{Unif}[B]$ be independent of $(z_i, \hat{z}_i)$, and construct a candidate tuple $X = (X_1, \ldots, X_B)$ by placing the true partner $\hat{z}_i$ in slot U and drawing the remaining $B - 1$ candidates independently from the marginal distribution of $\hat{z}$. Since the candidate slots can be re-indexed without changing the softmax objective, we write the resulting bound in the conventional indexing where the positive candidate is $\hat{z}_i$:

$$I(z_i; \hat{z}_i) \;\geq\; I(U; z_i, X) \;\geq\; \log B + \mathbb{E}\left[\log \frac{\exp f^\star(z_i, \hat{z}_i)}{\sum_{n=1}^{B} \exp f^\star(z_i, \hat{z}_n)}\right], \tag{3}$$

where $f^\star(z, \hat{z}) = \log \frac{p(z,\hat{z})}{p(z)\,p(\hat{z})}$ is the true pointwise mutual information critic. The first inequality is proved, via the chain rule for mutual information, in Appendix C.1; the second inequality is

the Barber–Agakov bound (Barber & Agakov, 2003), which more generally holds for any scoring function.

In practice (e.g., InfoNCE), the scoring function is commonly parameterized by the cosine similarity between the embeddings, $\langle z_i, \hat{z}_n \rangle := \left\langle \frac{z_i}{\|z_i\|_2}, \frac{\hat{z}_n}{\|\hat{z}_n\|_2} \right\rangle$, with a temperature parameter $\tau > 0$. Re-indexing the candidate tuple so that the positive partner occupies slot i gives the standard sample-wise contrastive loss:

$$\mathcal{L}_{\text{sam}}(i) = -\log \frac{\exp\big(\langle z_i, \hat{z}_i \rangle / \tau\big)}{\sum_{n=1}^{B} \exp\big(\langle z_i, \hat{z}_n \rangle / \tau\big)}, \tag{4}$$

$$I_{\text{sam}}(Z; \hat{Z}) \;\geq\; \log B - \mathbb{E}\big[\mathcal{L}_{\text{sam}}\big]. \tag{5}$$

Since $\mathcal{L}_{\text{sam}} \geq 0$, the lower bound in Equation 5 cannot exceed $\log B$. Consequently, the batch size imposes a hard ceiling on the mutual information that the sample-wise InfoNCE bound can certify at each step. As demonstrated in our experiments and prior literature (Ash et al., 2022; Chen et al., 2020), the performance of $\mathcal{L}_{\text{sam}}$ can therefore be sensitive to the batch size. Raising this ceiling requires increasing the batch size, which demands substantial computational resources.

3.2 Feature-wise Contrastive Objective

The sample-wise construction treats rows of Z as the population. We now apply the same mutual information identification principle to the columns of Z, treating output features as the population. Fixing an anchor feature c_j, we consider the $D-1$ channels $\hat{c}_d$ ($d \neq j$) as negative candidates. Let $V \in [D]$ denote the index of the true partner among the D candidates after randomly relabeling their positions, and let $Y = (Y_1, \ldots, Y_D)$ denote the resulting candidate tuple, such that $Y_V = \hat{c}_j$. The random relabeling ensures that V is uniform over $[D]$.

The feature-wise construction differs from the sample-wise case because feature channels are jointly generated and are not assumed to be conditionally independent. Specifically, c_j and the candidate channels are produced from the same batch of inputs and are coupled through the shared data and encoder architecture. Consequently, unlike the sample-wise negatives, the negative feature channels cannot generally be treated as conditionally independent of the anchor given its true partner. Moreover, feature channels are not naturally exchangeable; for example, nothing guarantees that channel 7 plays a statistically interchangeable role with channel 12.

As a result, when we formulate the mutual information bound for the feature-wise setting, an additional penalty term survives the derivation (proved in Appendix C.2):

$$I(c_j; \hat{c}_j) \geq I(V; c_j, Y) - \Delta_j \geq \log D - \Delta_j + \mathbb{E}\left[\log \frac{\exp f(c_j, \hat{c}_j)}{\sum_{d=1}^{D} \exp f(c_j, \hat{c}_d)}\right], \tag{6}$$

where the final expression is written using the original feature indexing and the feature-wise dependence penalty is defined as

$$\Delta_j := I\big(c_j; Y_{-V} \mid Y_V, V\big) + I(V; Y) \geq 0, \tag{7}$$

where $Y_{-V} := (Y_d)_{d \neq V}$ denotes the tuple of $D-1$ negative channels.

The first term in Δ_j captures the conditional dependence between the anchor and the negative channels, while the second captures the non-exchangeability of the candidate channels.

When is Δ small? The bound in Equation 6 is informative when Δ is small relative to $\log D$. Its two terms capture two sources of deviation from the ideal feature-wise identification setting: dependence between the anchor and non-corresponding channels, and non-exchangeability among the candidate features. These quantities can be characterized more explicitly under a Gaussian representation model. Suppose the rows $(z_i, \hat{z}_i)$ are i.i.d. jointly Gaussian, and let $\Sigma_{\hat{z}} := \mathrm{Cov}(\hat{z}_i)$ denote the within-branch covariance of a row representation. For standardized features, $\mathrm{tr}(\Sigma_{\hat{z}}) = D$, and we show in Appendix C.4 that

$$\Delta_j \leq -\frac{B}{2} \log(1 - R_j^2) - \frac{B}{2} \log\det \Sigma_{\hat{z}}, \tag{8}$$

where $R_j^2 := 1 - \mathrm{Var}(Z_{i,j} \mid \hat{z}_i)/\,\mathrm{Var}(Z_{i,j} \mid \hat{Z}_{i,j})$ measures the additional dependence between the scalar feature $Z_{i,j}$ and the non-corresponding features $\hat{Z}_{i,-j}$ after conditioning on its matched feature $\hat{Z}_{i,j}$. Thus, Δ_j is small when the non-corresponding features provide little additional information beyond the matched feature and the within-branch covariance is close to isotropic. In particular, $R_j^2 = 0$ and $\Sigma_{\hat{z}} = I_D$ imply $\Delta_j = 0$. These sufficient conditions motivate feature standardization, while Appendix D further examines the behavior of Δ in a controlled setting.

Target objective. Motivated by this bound, we optimize the tractable feature-wise identification term $I(V; c_j, Y)$ through its variational classifier. Before computing the feature-wise objective, we standardize each feature channel across the batch to have zero mean and unit variance. We then parameterize the scoring function using cosine similarity between the standardized feature channels, $\langle c_j, \hat{c}_d \rangle := \left\langle \frac{c_j}{\|c_j\|_2}, \frac{\hat{c}_d}{\|\hat{c}_d\|_2} \right\rangle$, with temperature $\tau > 0$, which yields

$$\mathcal{L}_{\mathrm{feat}}(j) = -\log \frac{\exp\left(\langle c_j, \hat{c}_j \rangle / \tau\right)}{\sum_{d=1}^{D} \exp\left(\langle c_j, \hat{c}_d \rangle / \tau\right)}, \tag{9}$$

$$I_{\mathrm{feat}}(Z; \hat{Z}) \;\geq\; \log D - \Delta - \mathbb{E}\big[\mathcal{L}_{\mathrm{feat}}\big], \tag{10}$$

where $\Delta := \frac{1}{D}\sum_{j=1}^{D} \Delta_j$. Unlike the sample-wise contrastive bound, whose candidate-set size is determined by the minibatch size B, the feature-wise bound is governed by the feature dimension D. Since $\mathcal{L}_{\mathrm{feat}} \geq 0$, the variational certificate in Equation 10 cannot exceed $\log D - \Delta$. Consequently, the information ceiling associated with the number of contrastive candidates is controlled by D instead of B, providing a complementary source of contrastive signal that is not directly limited by the number of samples in the minibatch.

3.3 Combining the Objectives

The sample-wise and feature-wise objectives provide variational lower bounds on two axis-wise mutual information quantities, $I_{\mathrm{sam}}(Z; \hat{Z})$ and $I_{\mathrm{feat}}(Z; \hat{Z})$, respectively. Although these quantities capture different aspects of the information shared between the two representations, both are bounded by the mutual information of the full embedding matrices.

Lemma 1 (Axis-wise mutual information lower-bounds joint mutual information)**.**

$$I(Z; \hat{Z}) \geq I_{\mathrm{sam}}(Z; \hat{Z}), \qquad I(Z; \hat{Z}) \geq I_{\mathrm{feat}}(Z; \hat{Z}).$$

The proof is provided in Appendix C.3. Combining Lemma 1 with the sample-wise and feature-wise bounds gives

$$I(Z; \hat{Z}) \geq \max\left\{\log B - \mathbb{E}\big[\mathcal{L}_{\mathrm{sam}}\big],\ \log D - \Delta - \mathbb{E}\big[\mathcal{L}_{\mathrm{feat}}\big]\right\}. \tag{11}$$

Thus, the two axes provide distinct lower-bound certificates of the same joint mutual information, governed by the number of samples B and output features D, respectively.

Equation 11 provides the tighter certificate by taking the maximum of the two bounds. However, it is not directly suitable as a training objective because the feature-wise bound contains the representation-dependent penalty Δ, which is not available as a tractable quantity during optimization. We therefore optimize the tractable loss terms from both objectives through the convex combination

$$\mathcal{L}_{\mathrm{comb}} = \pi \mathcal{L}_{\mathrm{sam}} + (1 - \pi)\mathcal{L}_{\mathrm{feat}}, \qquad \pi \in [0, 1]. \tag{12}$$

For any $\pi \in [0, 1]$, the joint mutual information also satisfies

$$\begin{aligned} I(Z; \hat{Z}) &\geq \pi(\log B - \mathbb{E}[\mathcal{L}_{\mathrm{sam}}]) + (1 - \pi)(\log D - \Delta - \mathbb{E}[\mathcal{L}_{\mathrm{feat}}]) \\ &= \pi \log B + (1 - \pi)(\log D - \Delta) - \mathbb{E}\big[\mathcal{L}_{\mathrm{comb}}\big]. \end{aligned} \tag{13}$$

The practical role of $\mathcal{L}_{\mathrm{comb}}$ is to preserve both tractable training signals: $\mathcal{L}_{\mathrm{sam}}$ distinguishes corresponding samples from other samples, while $\mathcal{L}_{\mathrm{feat}}$ distinguishes corresponding features from other features. The parameter π controls the relative contribution of the two objectives, with $\pi = 1$ and $\pi = 0$ recovering the purely sample-wise and feature-wise objectives, respectively. Our experiments in Section 5.4 show that intermediate values of π outperform either endpoint, supporting the empirical benefit of jointly optimizing the two axes. We analyze the computational and memory costs of the individual and combined objectives in Appendix A.

4 Relation to Barlow Twins

Prior work has shown that sample-contrastive and covariance-based dimension-contrastive objectives are closely related (Garrido et al., 2023). Our formulation provides a complementary mutual-information view by applying the InfoNCE identification principle directly along the feature axis. Specifically, Barlow Twins (Zbontar et al., 2021) encourages agreement between corresponding features while reducing redundancy between distinct features. Given standardized feature channels, it constructs the cross-correlation matrix $C_{dd'} := \langle c_d, \hat{c}_{d'} \rangle$, $C \in \mathbb{R}^{D \times D}$, where $C_{dd'}$ measures the cross-correlation between feature d in one branch and feature d' in the other. Our feature-wise contrastive objective operates on this same cross-branch feature-similarity matrix. In particular, Equation 9 can be written as

$$\mathcal{L}_{\text{feat}}(d) = -\frac{C_{dd}}{\tau} + \log \sum_{d'=1}^{D} \exp\left(\frac{C_{dd'}}{\tau}\right). \tag{14}$$

This imposes a *relative* matching criterion: the positive entry C_{dd} is encouraged to dominate the competing entries $C_{dd'}$, $d' \neq d$, without prescribing their absolute values.

Barlow Twins instead applies an *absolute* matching criterion to the same matrix:

$$\mathcal{L}_{\text{BT}} = \sum_{d=1}^{D} (1 - C_{dd})^2 + \lambda \sum_{d \neq d'} C_{dd'}^2, \tag{15}$$

which directly encourages C to approach the identity matrix by driving corresponding-feature correlations toward one and cross-feature correlations toward zero.

Thus, the two objectives use the same cross-feature matching structure but enforce it differently: our feature-wise contrastive loss uses softmax competition, requiring each feature to be more similar to its counterpart than to alternatives, whereas Barlow Twins directly penalizes deviations from the target correlation structure. They therefore represent relative and absolute forms of cross-feature matching, respectively. More generally, absolute matching is not inherently tied to the feature axis; a Barlow Twins-style objective can analogously be defined over samples. Exploring absolute matching along the sample axis, as well as more general combinations of relative and absolute matching across both axes, is an interesting direction for future work.

5 Experiments

We empirically investigate three questions motivated by our analysis: (i) whether feature-wise contrastive learning provides a useful representation-learning signal on its own; (ii) whether combining the sample-wise and feature-wise objectives improves over optimizing either axis alone; and (iii) whether the distinct dependence of the two objectives on the batch size B and feature dimension D is reflected in their empirical behavior.

5.1 Experimental Setup

Pretraining. We follow the SimCLR framework (Chen et al., 2020) and perform self-supervised pretraining on the ImageNet-1K (ILSVRC2012) training set (Deng et al., 2009). Unless otherwise stated, all methods use a ResNet-50 backbone (He et al., 2016) trained from scratch under the same data-augmentation and optimization pipeline. Our default projector output dimension is $D = 8192$, the pretraining batch size is $B = 256$, and the temperature is $\tau = 0.1$. For the combined objective, we use $\pi = 0.5$ unless otherwise specified. Complete architectural and optimization details are provided in Appendix B.

Evaluation protocols. We evaluate the learned representations using three complementary protocols. First, we report k-nearest-neighbor (k-NN) classification using representations extracted from the frozen backbone. Second, we perform linear evaluation by training a linear classifier on top of the frozen backbone while keeping all backbone parameters fixed. Third, we perform end-to-end fine-tuning of both the backbone and classification head. For ImageNet, we report Top-1 and Top-5 accuracy under all three protocols. The frozen k-NN and linear evaluations measure the quality of the representation produced directly by self-supervised pretraining, whereas full fine-tuning additionally measures how effectively the pretrained representation can be adapted using labeled data.

Table 1: ImageNet benchmark. k-NN, linear-probe, and full-fine-tuning Top-1/Top-5 accuracy (%), together with training time and peak memory. Bold and underline indicate the best and second-best results, respectively.

Method	k-NN		Linear Probe		Full Fine-Tuning		Training Cost		
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Time / epoch	Peak memory	Relative time
SimCLR	44.8	73.9	63.2	85.2	73.9	**91.9**	29.2 min	11.3 GB	1.00×
MoCo	41.8	72.2	61.5	84.1	<u>74.3</u>	**91.9**	24.9 min	6.8 GB	0.85×
BYOL	46.0	74.8	62.5	85.0	**74.5**	**91.9**	41.2 min	11.7 GB	1.41×
TiCo	26.6	53.6	49.7	74.4	72.7	90.9	22.4 min	12.8 GB	0.77×
Barlow Twins	45.6	73.9	62.9	84.3	72.6	90.9	37.9 min	14.0 GB	1.30×
VICReg	46.3	75.2	63.0	85.4	73.7	**91.9**	40.1 min	15.7 GB	1.37×
MACL	46.0	74.4	62.1	83.9	71.8	90.4	30.3 min	11.3 GB	1.03×
SimCLR + DimCL	45.1	74.1	63.4	85.4	73.1	90.9	30.3 min	14.9 GB	1.04×
$\mathcal{L}_{\text{sam}}$	45.5	73.5	64.2	85.8	72.7	91.0	29.2 min	13.6 GB	1.00×
$\mathcal{L}_{\text{feat}}$	<u>50.1</u>	<u>77.7</u>	<u>65.8</u>	<u>86.8</u>	72.6	91.0	30.3 min	14.9 GB	1.04×
$\mathcal{L}_{\text{comb}}$	**50.7**	**78.0**	**67.2**	**87.6**	73.2	<u>91.3</u>	31.1 min	14.9 GB	1.07×

Baselines. We compare against representative sample-wise contrastive, negative-free, and redundancy-reduction methods, including SimCLR (Chen et al., 2020), MoCo (He et al., 2020), BYOL (Grill et al., 2020), TiCo (Zhu et al., 2022), MACL (Huang et al., 2023), DimCL (Nguyen et al., 2023), Barlow Twins (Zbontar et al., 2021), and VICReg (Bardes et al., 2022). We additionally compare the three variants of our formulation: the sample-wise objective $\mathcal{L}_{\text{sam}}$, the feature-wise objective $\mathcal{L}_{\text{feat}}$, and their combination $\mathcal{L}_{\text{comb}}$.

5.2 ImageNet Representation Quality

We first evaluate whether feature-wise contrastive learning provides a useful representation-learning signal on its own and whether combining the sample-wise and feature-wise objectives is beneficial. As shown in Table 1, $\mathcal{L}_{\text{feat}}$ performs strongly under both k-NN and linear evaluation, demonstrating that contrasting feature channels can learn effective representations independently of a conventional sample-wise objective. Combining the two axes further improves frozen-representation performance and achieves the strongest results across these evaluation protocols, supporting the complementarity of the sample-wise and feature-wise signals. The differences are less pronounced after full fine-tuning, where supervised optimization can substantially reshape the pretrained representation and reduce the influence of the pretraining objective. The additional feature-wise computation introduces only a modest increase in training time and memory usage, with a detailed complexity analysis provided in Appendix A.

5.3 Cross-Modal Retrieval

We evaluate whether the sample-wise/feature-wise formulation extends beyond two augmented views of the same modality. On Flickr30K (Young et al., 2014), we instantiate the two branches using a frozen ViT-L/14 image encoder (Dosovitskiy et al., 2021) and a frozen BERT-base text encoder (Devlin et al., 2019), training only linear projection heads to isolate the effect of the alignment objective from representation learning within the modality-specific encoders. We evaluate both image-to-text and text-to-image retrieval using standard recall metrics. As shown in Table 2, the feature-wise objective alone is less effective than in the single-modality setting, but combining the sample-wise and feature-wise objectives provides the strongest retrieval performance, suggesting that the two axes remain complementary even when the paired representations are produced by heterogeneous modalities and encoder architectures.

5.4 Analysis and Ablations

We next examine the mechanisms suggested by our mutual-information analysis by varying the batch size B, feature dimension D, interpolation weight π, and temperature τ.

Effect of feature dimension. The feature-wise bound in Equation 10 depends explicitly on the number of feature candidates D, whereas the sample-wise bound in Equation 5 has no corresponding dependence on D. We therefore vary the projector output dimension while holding the batch size and

Table 2: Cross-modal retrieval on Flickr30K (Young et al., 2014). Bold marks the best entry in each column.

Method	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
VICReg	9.63	25.29	36.52	9.00	24.82	34.79
$\mathcal{L}_{\text{sam}}$	8.02	23.47	33.66	7.82	23.53	34.10
$\mathcal{L}_{\text{feat}}$	7.77	22.71	31.93	6.07	19.94	29.16
$\mathcal{L}_{\text{comb}}$	**9.75**	**28.09**	**39.32**	**10.47**	**27.12**	**38.85**

all other training settings fixed. As shown in Table 3, the sample-wise objective is largely insensitive to the feature dimension, while the feature-wise and combined objectives benefit from increasing D. This behavior is consistent with the different roles that the feature dimension plays in the two contrastive formulations.

Table 3: k-NN, linear-probe, and full fine-tuning Top-1 accuracy (%) vs. projector output dimension D.

D	k-NN			Linear probe			Full Fine-Tuning		
	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$
512	45.4	49.2	50.2	64.2	64.3	66.1	72.5	74.2	73.2
2048	45.5	50.1	50.4	64.2	65.5	66.7	72.6	73.1	73.3
8192	45.5	50.1	50.7	64.2	65.8	67.2	72.7	72.6	73.2

Complementarity of the two objectives. The combined objective in Equation 12 interpolates between purely feature-wise and purely sample-wise learning through the weight π. We vary this weight while keeping all other training settings fixed to determine whether the two objectives provide complementary signals. Table 4 shows that intermediate mixtures outperform either single-axis objective, indicating that the benefit of the combined formulation arises from jointly exploiting the sample and feature axes rather than simply selecting the stronger individual objective.

Table 4: ImageNet Top-1 accuracy (%) as a function of the combination weight π in $\mathcal{L}_{\text{comb}}$.

π	0.0	0.25	0.5	0.75	1.0
k-NN	50.1	50.7	50.7	48.6	45.5
Linear	65.8	66.9	67.2	66.1	64.2
Full fine-tune	72.6	73.3	73.2	72.9	72.7

Sensitivity to batch size. The sample-wise bound in Equation 5 is directly governed by the number of samples in the minibatch, whereas the candidate-set term of the feature-wise bound is controlled by the feature dimension. We therefore vary the pretraining batch size while fixing D and the remaining training configuration. As shown in Table 5, the sample-wise objective improves substantially as the batch size increases, while the feature-wise objective is less sensitive to B and performs comparatively better at smaller batch sizes. The combined objective also remains strong across batch sizes. This behavior is consistent with the different roles that the batch size plays in the two contrastive formulations.

Table 5: k-NN, linear-probe, and full-fine-tuning Top-1 accuracy (%) vs. pretraining batch size B.

B	k-NN			Linear Probe			Full Fine-Tuning		
	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$
64	37.2	46.4	46.8	60.4	63.8	64.1	69.9	71.8	72.1
128	41.3	48.2	48.6	62.3	64.9	65.0	71.3	72.3	72.8
256	45.5	50.1	50.7	64.2	65.8	67.2	72.7	72.6	73.2

Temperature sensitivity. Finally, we study the sensitivity of the objectives to the softmax temperature by jointly varying τ for both axes. As shown in Table 6, all three objectives are sensitive to the choice of temperature, with the strongest k-NN and linear-probe performance obtained at the default setting $\tau = 0.1$. The feature-wise objective is particularly sensitive to a very small temperature, while the combined objective remains stronger than either individual objective across the tested values. The differences are again less pronounced after full fine-tuning. This behavior highlights the importance of appropriately scaling the relative similarities in both contrastive formulations.

Table 6: k-NN, linear-probe, and full-fine-tuning Top-1 accuracy (%) vs. temperature τ.

τ	k-NN			Linear Probe			Full Fine-Tuning		
	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$	$\mathcal{L}_{\text{sam}}$	$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{comb}}$
0.01	30.1	9.2	33.7	48.3	16.1	51.9	65.2	65.6	66.8
0.1	45.5	50.1	50.7	64.2	65.8	67.2	72.7	72.6	73.2
1.0	38.7	42.9	45.0	56.2	58.3	59.4	72.9	73.1	74.5

6 Related Work

Contrastive Learning and Mutual Information. Contrastive learning typically operates along the sample axis, with InfoNCE (Oord et al., 2018) underlying methods such as MoCo (He et al., 2020) and SimCLR (Chen et al., 2020). Its mutual-information interpretation and dependence on negative candidates have been extensively studied (Poole et al., 2019; McAllester & Stratos, 2020). Recent work has further analyzed InfoNCE through its minibatch geometry (Lee et al., 2025), identifiability (Rusak et al., 2025), projection heads (Ouyang et al., 2025), probabilistic formulations (Wang et al., 2025), and relation to mutual-information estimation (Ryu et al., 2026), while related work has derived non-contrastive objectives from mutual-information principles (Guo et al., 2026). Our work instead asks whether the InfoNCE identification principle can be applied along the feature axis.

Redundancy Reduction and Feature Structure. A complementary family of methods regulates the structure of learned features. Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2022), and whitening-based methods (Ermolov et al., 2021) reduce redundancy through correlation, covariance, or whitening constraints. More recent work has studied dimensional collapse and decorrelation (Sansone et al., 2025), entropy-based feature diversification (Chakraborty et al., 2025), spectral regularization of feature covariance (Ghanooni et al., 2025), and theoretical properties of whitening (Han et al., 2026). In contrast, our feature-wise objective promotes informative and non-redundant features through relative contrastive competition rather than explicit covariance, entropy, whitening, or spectral penalties.

Feature-wise Contrastive Learning. Feature-level objectives have been explored in Zero-CL (Zhang et al., 2022b), ARB (Zhang et al., 2022a), DimCL (Nguyen et al., 2023), and SimCSE++ (Xu et al., 2023). In particular, DimCL and SimCSE++ apply InfoNCE along the feature axis, yielding objectives closely related to $\mathcal{L}_{\text{feat}}$. Our contribution is instead the mutual-information analysis: we show that feature-wise identification lower-bounds feature-wise mutual information up to a dependence and non-exchangeability penalty Δ, relate both axis-wise quantities to $I(Z; \hat{Z})$, and derive the combined bound. We further show that the feature-wise objective is competitive on its own and extends beyond single-modality representation learning. Relatedly, Garrido et al. (2023) connect sample- and dimension-contrastive learning through covariance-based criteria, whereas our analysis concerns softmax-based identification.

7 Conclusion

We studied sample-wise and feature-wise contrastive learning through a common mutual information formulation. Our analysis characterizes the feature-wise objective in terms of a mutual-information bound that accounts for dependence and non-exchangeability among feature channels, and connects both axis-wise objectives to the mutual information between the full embedding matrices. This provides a principled basis for combining the two objectives. Experiments across representation learning and cross-modal retrieval show that combining both axes improves performance.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing portions of the manuscript, improving clarity and presentation, and providing feedback on mathematical derivations and proof presentation. All AI-assisted content was reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies general-purpose representation learning objectives and does not involve the collection of new human-subject data or the deployment of models in high-stakes applications. Our experiments use established publicly available benchmark datasets under their standard research protocols, and we do not introduce or release new personal or sensitive data. Beyond general considerations associated with large-scale representation learning, we are not aware of additional ethical concerns specific to the proposed method.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce both the theoretical and empirical results of this work. The formulation of the sample-wise, feature-wise, and combined objectives is provided in the main text, together with the experimental setup and evaluation protocols in Section 5. Appendix B gives the complete implementation details, including architecture, data augmentation, optimization, pretraining, and evaluation settings. Our experiments use standard publicly available benchmark datasets and established evaluation protocols, with all method-specific hyperparameters and training configurations reported in the paper and appendix. We additionally report the relevant computational requirements and efficiency measurements to facilitate comparison and replication of the experiments.

REFERENCES

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

David Barber and Felix V. Agakov. The im algorithm: A variational approach to information maximization. In *NIPS*, pp. 201–208, 2003.

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.

Deep Chakraborty, Yann LeCun, Tim G. J. Rudner, and Erik Learned-Miller. Improving pre-trained self-supervised embeddings through effective entropy maximization. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 433–441. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/chakraborty25a.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pp. 3015–3024. PMLR, 2021.

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

Naghmeh Ghanooni, Waleed Mustafa, Dennis Wagner, Sophie Fellenz, Anthony Lin, and Marius Kloft. Mitigating spurious features in contrastive learning with spectral regularization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 107856–107888. Curran Associates, Inc., 2025. doi: 10.52202/085713-3597. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9aeabac8f13e3ea49eef15cf1154b3c4-Paper-Conference.pdf.

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

Zhaohan Daniel Guo, Bernardo Avila Pires, Khimya Khetarpal, Dale Schuurmans, and Bo Dai. Representation learning via non-contrastive mutual information. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=GoLd6k51zK.

Zhiwei Han, Stefan Matthes, and Hao Shen. Provable affine identifiability of nonlinear cca under latent distributional priors. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3835–3843. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/han26b.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Zizheng Huang, Haoxing Chen, Ziqi Wen, Chao Zhang, Huaxiong Li, Bo Wang, and Chunlin Chen. Model-aware contrastive learning: Towards escaping the dilemmas. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, Jul. 2023.

Chungpa Lee, Sehee Lim, Kibok Lee, and Jy yong Sohn. On the similarities of embeddings in contrastive learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ChmJZ9V2o1.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Thanh Nguyen, Trung Xuan Pham, Chaoning Zhang, Tung M Luu, Thang Vu, and Chang D Yoo. Dimcl: dimensional contrastive learning for improving self-supervised learning. *IEEE Access*, 11: 21534–21545, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *International Conference on Learning Representations*, volume 2025, pp. 76941–76960, 2025.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pp. 5171–5180. PMLR, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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 (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4159–4167. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/rusak25a.html.

Jongha Ryu, Pavan Yeddanapudi, Xiangxiang Xu, and Gregory Wornell. Contrastive predictive coding done right for mutual information estimation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 28546–28576, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/304e0ae755d5d7e9016e3d05700c35b2-Paper-Conference.pdf.

Emanuele Sansone, Tim Lebailly, and Tinne Tuytelaars. Collapse-proof non-contrastive self-supervised learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52833–52860. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/sansone25a.html.

Bokun Wang, Yunwen Lei, Yiming Ying, and Tianbao Yang. On discriminative probabilistic modeling for self-supervised representation learning. In *International Conference on Learning Representations*, volume 2025, pp. 38217–38247, 2025.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. Simcse++: Improving contrastive learning for sentence embeddings from two perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12028–12040, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16579–16588. IEEE, 2022a.

Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-CL: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RAW9tCdVxLj.

Jiachen Zhu, Rafael M Moraes, Serkan Karakulak, Vlad Sobol, Alfredo Canziani, and Yann LeCun. Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. *arXiv preprint arXiv:2206.10698*, 2022.

We provide additional analysis, implementation details, and theoretical proofs in the appendices. The appendices are organized as follows:

- **Appendix A: Computational and Memory Complexity.** Analysis of the computational and memory costs of the sample-wise, feature-wise, and combined objectives.
- **Appendix B: Implementation Details.** Complete architectural, optimization, pretraining, and evaluation details for the experiments.
- **Appendix C: Proofs.** Proofs of the sample-wise and feature-wise mutual information bounds, as well as the relation between the axis-wise and joint mutual information, and feature-wise dependence penalty analysis.
- **Appendix D: Synthetic Analysis of the Feature-Wise Dependence Penalty.** Controlled Gaussian experiments illustrating the behavior of Δ under feature dependence and non-exchangeability.

A Computational and Memory Complexity

We analyze the computational and memory costs of the sample-wise, feature-wise, and combined objectives. We focus on the cost of computing the contrastive objectives themselves, excluding the shared encoder and projection network.

Sample-wise objective. The sample-wise objective computes pairwise similarities among the B sample representations across the two branches. Since each similarity is computed between D-dimensional vectors, evaluating all B^2 pairwise similarities requires $\mathcal{O}(B^2D)$ operations, while explicitly materializing the corresponding logits requires $\mathcal{O}(B^2)$ memory. The row-wise normalization of the representations requires only $\mathcal{O}(BD)$ computation and therefore does not affect the leading-order complexity.

Feature-wise objective. The feature-wise objective instead computes pairwise similarities among the D feature channels across the two branches. Since each similarity is computed between vectors of length B, evaluating all D^2 pairwise similarities requires $\mathcal{O}(BD^2)$ operations, while explicitly materializing the corresponding logits requires $\mathcal{O}(D^2)$ memory. The feature standardization and normalization performed before computing the similarities require only $\mathcal{O}(BD)$ computation and therefore do not affect the leading-order complexity.

Combined objective. The combined objective evaluates both the sample-wise and feature-wise losses, resulting in a total similarity-computation cost of

$$\mathcal{O}(B^2D + BD^2) = \mathcal{O}\big(BD(B + D)\big). \tag{16}$$

If the two logit matrices are materialized simultaneously, their memory cost is $\mathcal{O}(B^2 + D^2)$. Alternatively, the two losses can be evaluated sequentially, in which case the two logit matrices need not coexist in memory.

Table 7 summarizes these costs.

Table 7: Computational and memory complexity of the contrastive objectives. The memory cost refers to the similarity logits and excludes the shared encoder and projector activations.

Objective	Similarity computation	Logit memory
$\mathcal{L}_{\text{sam}}$	$\mathcal{O}(B^2D)$	$\mathcal{O}(B^2)$
$\mathcal{L}_{\text{feat}}$	$\mathcal{O}(BD^2)$	$\mathcal{O}(D^2)$
$\mathcal{L}_{\text{comb}}$	$\mathcal{O}(B^2D + BD^2)$	$\mathcal{O}(B^2 + D^2)$

Under our default configuration, $B = 256$ and $D = 8192$. The sample-wise objective therefore produces $B^2 = 65{,}536$ logits, whereas the feature-wise objective produces $D^2 = 67{,}108{,}864$ logits. When stored densely, these require approximately 0.125 MiB and 128 MiB, respectively, in FP16/BF16, or 0.25 MiB and 256 MiB in FP32.

The relative arithmetic cost of the feature-wise and sample-wise similarity computations is $BD^2/(B^2D) = D/B$, which equals 32 under our default configuration. This factor applies only to the computation of the contrastive similarities and not to the total training cost, since the encoder and projection network are shared across objectives and account for a substantial fraction of the end-to-end computation. We therefore additionally report measured training time and peak device memory in Table 1.

B Implementation Details

B.1 ImageNet Pretraining

Architecture. We use a ResNet-50 backbone with the final classification layer removed, yielding a 2048-dimensional representation after global average pooling. The backbone is initialized from scratch. The projection head consists of two linear layers, with a 2048-dimensional hidden layer followed by batch normalization and ReLU. For our sample-wise, feature-wise, and combined objectives, the output dimension of the projection head is set to $D = 8192$. Batch normalization is also applied to the output of the projection head, without a subsequent nonlinearity.

Data augmentation. For each image, we independently generate two augmented views following the standard SimCLR augmentation pipeline (Chen et al., 2020). Each view is obtained using a random resized crop to 224×224 pixels with crop scale in $[0.08, 1.0]$, followed by random horizontal flipping with probability 0.5. Color jittering is applied with probability 0.8, using brightness, contrast, saturation, and hue strengths of 0.8, 0.8, 0.8, and 0.2, respectively. Images are converted to grayscale with probability 0.2, and Gaussian blur is applied with probability 0.5 using a standard deviation sampled uniformly from $[0.1, 2.0]$. Finally, images are normalized using the ImageNet channel-wise mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$.

Optimization. We pretrain all models for 100 epochs with a global batch size of $B = 256$. Optimization is performed using LARS with momentum 0.9 and weight decay 10^{-6}. Batch-normalization parameters and biases are excluded from weight decay. We use square-root learning-rate scaling, with learning rate

$$\eta = 0.075\sqrt{B},$$

which gives $\eta = 1.2$ for $B = 256$. The learning rate is linearly warmed up during the first 10 epochs and subsequently decayed to zero using a cosine schedule. Training is performed using mixed-precision computation, and synchronized batch normalization is used for distributed training.

Objectives. Unless otherwise specified, the temperature is set to $\tau = 0.1$ for both the sample-wise and feature-wise objectives. Before computing the sample-wise objective, sample representations are ℓ_2-normalized. For the feature-wise objective, each feature channel is first standardized across the batch to have zero mean and unit variance and is subsequently ℓ_2-normalized. For the combined objective, we use $\pi = 0.5$, such that the two losses receive equal weight. The default feature dimension is $D = 8192$; the experiments studying feature dimensionality vary only D while keeping the remaining training configuration fixed.

B.2 ImageNet Evaluation

All evaluation protocols use the representation produced by the ResNet-50 backbone before the projection head.

k-NN evaluation. For k-NN evaluation, the pretrained backbone is kept fixed. Images from both the training and validation sets are resized to 256 pixels, center-cropped to 224×224, and normalized using the ImageNet statistics. We construct the feature bank using the ImageNet training set and classify each validation representation using its $k = 200$ nearest neighbors. Neighbor votes are weighted according to their cosine similarity using a temperature of 0.1. We report Top-1 and Top-5 accuracy on the ImageNet validation set.

Linear evaluation. For linear evaluation, the pretrained backbone is frozen and only a linear classifier is optimized. During training, images are randomly resized and cropped to 224×224 and randomly flipped horizontally. Validation images are resized to 256 pixels and center-cropped to 224×224. The classifier is trained for 90 epochs using SGD with momentum 0.9, zero weight decay, and a learning rate of 0.1 for a global batch size of 256. The learning rate is scaled linearly with the batch size and decayed using a cosine schedule without warmup. We report the best Top-1 and Top-5 validation accuracy over training.

Full fine-tuning. For full fine-tuning, both the pretrained backbone and a newly initialized linear classification head are optimized using ImageNet labels. We use the same training and validation transformations as in linear evaluation. Fine-tuning is performed for 30 epochs using SGD with momentum 0.9, zero weight decay, and an initial learning rate of 0.05 for a global batch size of 256. The learning rate is scaled linearly with batch size and follows a cosine schedule without warmup. We report the best Top-1 and Top-5 validation accuracy.

B.3 Cross-Modal Retrieval

For Flickr30K, we use a frozen ViT-L/14 image encoder and a frozen BERT-base text encoder. Only linear projection heads are trained, mapping the two modalities to a common $D = 8192$-dimensional representation space.

The projection heads are trained for 30 epochs with a batch size of 256 using an optimizer and learning rate similar to those used for the ImageNet benchmark. We use $\tau = 0.1$ and $\pi = 0.5$ for the contrastive objectives. At test time, image and text embeddings are ranked using cosine similarity. We evaluate using the standard Flickr30K test split and report Recall@1, Recall@5, and Recall@10 for both image-to-text and text-to-image retrieval.

B.4 Implementation and Hardware

Our implementation is based on PyTorch and uses distributed data-parallel training. ImageNet experiments use 16-bit mixed precision and synchronized batch normalization across devices. The global batch size is 256. Experiments are conducted using 2 NVIDIA V100 GPUs with 32 GB of memory each.

C Proofs

This section contains the proofs of the theoretical results used in the main text:

- Proof of the sample-wise mutual information inequality in Equation 3.
- Proof of the feature-wise mutual information inequality in Equation 6.
- Proof of Lemma 1, relating the axis-wise mutual information quantities to the joint mutual information.
- Gaussian characterization of the feature-wise dependence penalty Δ and sufficient conditions under which it is small.

C.1 Proof of the left inequality in equation 3

We show that

$$I(z_i; \hat{z}_i) \;\geq\; I(U; z_i, X)$$

holds as a direct consequence of the chain rule for mutual information. Fix an anchor z_i, jointly distributed with its true partner $\hat{z}_i$ as $(z_i, \hat{z}_i) \sim p(z, \hat{z})$, and let $U \in [B]$ be a uniform index, $p(U = u) = 1/B$, drawn independently of $(z_i, \hat{z}_i)$. Write $X = (X_1, \ldots, X_B)$ for the candidate tuple, generated conditionally on $(z_i, U = u)$ by placing the true partner in slot u,

$$X_u \sim p(\hat{z} \mid z_i),$$

and drawing the remaining candidates independently from the marginal,

$$X_n \sim p(\hat{z}), \qquad n \neq u.$$

The $B-1$ negative candidates are drawn independently of z_i, of X_u, and of each other.

Expansion A (condition on U first). By the chain rule,

$$I(z_i; X, U) = I(z_i; U) + I(z_i; X \mid U). \tag{17}$$

Since U is drawn independently of $(z_i, \hat{z}_i)$, $I(z_i; U) = 0$. Conditioned on $U = u$, the tuple X consists of the true partner X_u together with $B-1$ negatives that are independent of z_i. Therefore,

$$I(z_i; X \mid U) = I(z_i; X_U \mid U) + I(z_i; X_{-U} \mid X_U, U) \tag{18}$$

$$= I(z_i; \hat{z}_i), \tag{19}$$

where the second term vanishes by the independence of the negative candidates. Substituting into Equation 17 gives

$$I(z_i; X, U) = I(z_i; \hat{z}_i). \tag{20}$$

Expansion B (condition on X first). By the chain rule again,

$$I(z_i; X, U) = I(z_i; X) + I(U; z_i \mid X), \tag{21}$$

where we use the symmetry of conditional mutual information.

Identifying $I(U; z_i, X)$. Applying the chain rule to $I(U; z_i, X)$ gives

$$I(U; z_i, X) = I(U; X) + I(U; z_i \mid X). \tag{22}$$

Marginalizing over z_i, the candidate in the true-partner slot also follows the marginal distribution $p(\hat{z})$. Therefore,

$$p(X \mid U = u) = \prod_{n=1}^{B} p(X_n),$$

which does not depend on u. Hence, U and X are marginally independent and $I(U; X) = 0$. Equation 22 therefore gives

$$I(U; z_i, X) = I(U; z_i \mid X),$$

and Equation 21 becomes

$$I(z_i; X, U) = I(z_i; X) + I(U; z_i, X). \tag{23}$$

Combining. Equating Equations 20 and 23 gives

$$I(z_i; \hat{z}_i) = I(z_i; X) + I(U; z_i, X). \tag{24}$$

Since mutual information is non-negative, $I(z_i; X) \geq 0$, and therefore

$$I(z_i; \hat{z}_i) \;\geq\; I(U; z_i, X),$$

which proves the left inequality in Equation 3.

C.2 Proof of the left inequality in Equation 6

We show that

$$I(c_j; \hat{c}_j) \;\geq\; I(V; c_j, Y) - \Delta_j,$$

where the feature-wise dependence penalty is

$$\Delta_j = I\big(c_j; Y_{-V} \mid Y_V, V\big) + I(V; Y),$$

as a consequence of the chain rule for mutual information and the non-negativity of mutual information.

Fix an anchor feature c_j and its corresponding partner $\hat{c}_j$. Let $V \in [D]$ denote the position of the true partner after independently and uniformly relabeling the D candidate channels, and let $Y = (Y_1, \ldots, Y_D)$ denote the resulting candidate tuple. By construction, $Y_V = \hat{c}_j$, $V \sim \mathrm{Unif}[D]$, and V is independent of the underlying feature representations prior to relabeling.

For convenience, define

$$A_j := I\big(c_j; Y_{-V} \mid Y_V, V\big), \tag{25}$$

so that

$$\Delta_j = A_j + I(V;Y). \tag{26}$$

Expansion A (condition on V first). By the chain rule,

$$I(c_j;Y,V) = I(c_j;V) + I(c_j;Y \mid V). \tag{27}$$

Since the relabeling is independent of the feature representations, $I(c_j;V) = 0$. Applying the chain rule to the second term gives

$$I(c_j;Y \mid V) = I(c_j;Y_V \mid V) + I(c_j;Y_{-V} \mid Y_V, V) \tag{28}$$

$$= I(c_j;\hat{c}_j) + A_j, \tag{29}$$

where the second equality follows because Y_V is the true partner $\hat{c}_j$ and the random relabeling is independent of $(c_j, \hat{c}_j)$. Therefore,

$$I(c_j;Y,V) = I(c_j;\hat{c}_j) + A_j. \tag{30}$$

Expansion B (condition on Y first). By the chain rule,

$$I(c_j;Y,V) = I(c_j;Y) + I(V;c_j \mid Y), \tag{31}$$

where we use the symmetry of conditional mutual information. Applying the chain rule to $I(V;c_j,Y)$ gives

$$I(V;c_j,Y) = I(V;Y) + I(V;c_j \mid Y). \tag{32}$$

Rearranging,

$$I(V;c_j \mid Y) = I(V;c_j,Y) - I(V;Y). \tag{33}$$

Substituting Equation 33 into Equation 31 gives

$$I(c_j;Y,V) = I(c_j;Y) + I(V;c_j,Y) - I(V;Y). \tag{34}$$

Combining. Equating Equations 30 and 34 gives

$$I(c_j;\hat{c}_j) + A_j = I(c_j;Y) + I(V;c_j,Y) - I(V;Y). \tag{35}$$

Rearranging,

$$I(c_j;\hat{c}_j) = I(c_j;Y) + I(V;c_j,Y) - A_j - I(V;Y). \tag{36}$$

Using Equation 26,

$$I(c_j;\hat{c}_j) = I(c_j;Y) + I(V;c_j,Y) - \Delta_j. \tag{37}$$

Since $I(c_j;Y) \geq 0$, we obtain

$$I(c_j;\hat{c}_j) \;\geq\; I(V;c_j,Y) - \Delta_j,$$

which proves the left inequality in Equation 6.

The two components of Δ_j capture the two departures from the conventional sample-wise construction. The term $I(c_j;Y_{-V} \mid Y_V, V)$ vanishes when the negative channels are conditionally independent of the anchor given its true partner and index, while $I(V;Y)$ vanishes when the candidate tuple contains no information about the location of the true partner. Thus, both terms vanish under the corresponding conditional-independence and exchangeability conditions.

C.3 Proof of Lemma 1

We first prove the sample-wise inequality. For every $i \in [B]$, $z_i = Z_{i,:}$ is a deterministic function of Z, while $\hat{z}_i = \hat{Z}_{i,:}$ is a deterministic function of $\hat{Z}$. Therefore, by the data-processing inequality,

$$I(z_i;\hat{z}_i) \;\leq\; I(Z;\hat{Z}), \qquad \text{for every } i \in [B]. \tag{38}$$

Averaging over the sample indices and using the definition of $I_{\mathrm{sam}}(Z;\hat{Z})$ gives

$$I_{\mathrm{sam}}(Z;\hat{Z}) = \frac{1}{B}\sum_{i=1}^{B} I(z_i;\hat{z}_i) \tag{39}$$

$$\leq \frac{1}{B}\sum_{i=1}^{B} I(Z;\hat{Z}) \tag{40}$$

$$= I(Z;\hat{Z}). \tag{41}$$

The feature-wise inequality follows analogously. For every $j \in [D]$, $c_j = Z_{:,j}$ is a deterministic function of Z, while $\hat{c}_j = \hat{Z}_{:,j}$ is a deterministic function of $\hat{Z}$. Therefore,

$$I(c_j; \hat{c}_j) \;\le\; I(Z; \hat{Z}), \qquad \text{for every } j \in [D]. \tag{42}$$

Averaging over the feature indices and using the definition of $I_{\text{feat}}(Z; \hat{Z})$ gives

$$I_{\text{feat}}(Z; \hat{Z}) = \frac{1}{D} \sum_{j=1}^{D} I(c_j; \hat{c}_j) \tag{43}$$

$$\le \frac{1}{D} \sum_{j=1}^{D} I(Z; \hat{Z}) \tag{44}$$

$$= I(Z; \hat{Z}). \tag{45}$$

Therefore,

$$I(Z; \hat{Z}) \ge I_{\text{sam}}(Z; \hat{Z}), \qquad I(Z; \hat{Z}) \ge I_{\text{feat}}(Z; \hat{Z}). \qquad \blacksquare$$

C.4 Gaussian Characterization of the Feature-Wise Dependence Penalty

We provide a sufficient-condition analysis for the feature-wise dependence penalty

$$\Delta_j = I(c_j; Y_{-V} \mid Y_V, V) + I(V; Y)$$

under a Gaussian representation model. Suppose that the rows $(z_i, \hat{z}_i)$ of $(Z, \hat{Z})$ are i.i.d. jointly Gaussian, and let

$$\Sigma_{\hat{z}} := \mathrm{Cov}(\hat{z}_i)$$

denote the within-branch covariance of a row representation. Since the rows are identically distributed, $\Sigma_{\hat{z}}$ does not depend on i.

Conditional-dependence term. By construction, $Y_V = \hat{c}_j$, while conditional on V, Y_{-V} is a re-indexing of the $D - 1$ non-corresponding feature channels. Define

$$\hat{C}_{-j} := (\hat{c}_d)_{d \ne j}.$$

Since mutual information is invariant to bijective re-indexing,

$$I(c_j; Y_{-V} \mid Y_V, V) = I(c_j; \hat{C}_{-j} \mid \hat{c}_j). \tag{46}$$

For any sample index i, define

$$R_j^2 := 1 - \frac{\mathrm{Var}(Z_{i,j} \mid \hat{z}_i)}{\mathrm{Var}(Z_{i,j} \mid \hat{Z}_{i,j})}. \tag{47}$$

Because the rows are identically distributed, this quantity does not depend on i. It measures the reduction in the residual variance of $Z_{i,j}$ obtained by conditioning on the non-corresponding features $\hat{Z}_{i,-j}$ in addition to its matched feature $\hat{Z}_{i,j}$.

Since the B rows are independent, conditional mutual information is additive across samples:

$$I(c_j; \hat{C}_{-j} \mid \hat{c}_j) = \sum_{i=1}^{B} I(Z_{i,j}; \hat{Z}_{i,-j} \mid \hat{Z}_{i,j}) \tag{48}$$

$$= B\, I(Z_{i,j}; \hat{Z}_{i,-j} \mid \hat{Z}_{i,j}). \tag{49}$$

For jointly Gaussian variables,

$$I(Z_{i,j}; \hat{Z}_{i,-j} \mid \hat{Z}_{i,j}) = \frac{1}{2} \log \frac{\mathrm{Var}(Z_{i,j} \mid \hat{Z}_{i,j})}{\mathrm{Var}(Z_{i,j} \mid \hat{z}_i)} \tag{50}$$

$$= -\frac{1}{2} \log(1 - R_j^2). \tag{51}$$

Combining the above expressions gives

$$I(c_j; Y_{-V} \mid Y_V, V) = -\frac{B}{2} \log(1 - R_j^2). \tag{52}$$

Non-exchangeability term. Let

$$P_{\hat{z}} = \mathcal{N}(0, \Sigma_{\hat{z}})$$

denote the Gaussian distribution of a row representation in the second branch, and consider the isotropic reference distribution

$$Q_{\hat{z}} = \mathcal{N}(0, I_D).$$

The latter is exchangeable across feature coordinates. For the B independent rows,

$$D_{\mathrm{KL}}\left(P_{\hat{Z}} \,\|\, Q_{\hat{Z}}\right) = B\, D_{\mathrm{KL}}\left(P_{\hat{z}} \,\|\, Q_{\hat{z}}\right). \tag{53}$$

The random relabeling used to construct (V, Y) from the feature channels defines a stochastic mapping applied identically under the two distributions. By the data-processing inequality for KL divergence,

$$D_{\mathrm{KL}}\left(P_{V,Y} \,\|\, Q_{V,Y}\right) \le D_{\mathrm{KL}}\left(P_{\hat{Z}} \,\|\, Q_{\hat{Z}}\right). \tag{54}$$

Under the isotropic reference distribution, the feature coordinates are exchangeable, so the distribution of Y does not depend on the position V of the matched feature. Hence,

$$Q_{V,Y} = P_V Q_Y.$$

Using the chain rule for KL divergence,

$$D_{\mathrm{KL}}\left(P_{V,Y} \,\|\, P_V Q_Y\right) = I(V;Y) + D_{\mathrm{KL}}(P_Y \| Q_Y) \tag{55}$$

$$\ge I(V;Y). \tag{56}$$

Combining this inequality with Equation 54 gives

$$I(V;Y) \le D_{\mathrm{KL}}\left(P_{\hat{Z}} \,\|\, Q_{\hat{Z}}\right). \tag{57}$$

For zero-mean Gaussian distributions,

$$D_{\mathrm{KL}}\left(\mathcal{N}(0, \Sigma_{\hat{z}}) \,\|\, \mathcal{N}(0, I_D)\right) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_{\hat{z}}) - \log\det\Sigma_{\hat{z}} - D\right).$$

Therefore,

$$I(V;Y) \le \frac{B}{2}\left(\mathrm{tr}(\Sigma_{\hat{z}}) - \log\det\Sigma_{\hat{z}} - D\right). \tag{58}$$

Combining Equations 52 and 58 yields

$$\Delta_j \le -\frac{B}{2}\log(1 - R_j^2) + \frac{B}{2}\left(\mathrm{tr}(\Sigma_{\hat{z}}) - \log\det\Sigma_{\hat{z}} - D\right). \tag{59}$$

If the feature coordinates are standardized to have unit marginal variance, then

$$\mathrm{diag}(\Sigma_{\hat{z}}) = \mathbf{1}, \qquad \mathrm{tr}(\Sigma_{\hat{z}}) = D.$$

Equation 59 therefore reduces to

$$\Delta_j \le -\frac{B}{2}\log(1 - R_j^2) - \frac{B}{2}\log\det\Sigma_{\hat{z}}. \tag{60}$$

Consequently, $R_j^2 = 0$ and $\Sigma_{\hat{z}} = I_D$ are sufficient for $\Delta_j = 0$. The first condition means that the non-corresponding features provide no additional information about $Z_{i,j}$ once its matched feature $\hat{Z}_{i,j}$ is known, while the second makes the Gaussian feature distribution isotropic and therefore exchangeable across coordinates.

D Synthetic Study of the Feature-Wise Dependence Penalty

The feature-wise bound in Equation 10 contains the dependence penalty Δ, which is generally not tractable for learned representations. We therefore study synthetic Gaussian representations for which $I(c_j; \hat{c}_j)$ and the Gaussian characterization of Δ_j can be computed analytically, while the feature-wise loss can be estimated by Monte Carlo sampling. This allows us to directly compare the certificate $\log D - \mathbb{E}[\mathcal{L}_{\mathrm{feat}}(j)]$ with the corresponding feature-wise mutual information and examine when the penalty Δ_j is necessary.

Representation model. We draw the B rows of $(Z, \hat{Z})$ i.i.d. from a one-factor Gaussian model. For each sample $i \in [B]$ and feature $d \in [D]$,

$$\hat{Z}_{i,d} = \sqrt{\rho_d}\, g_i + \sqrt{1-\rho_d}\, e_{i,d}, \qquad Z_{i,d} = \alpha\, e_{i,d} + \gamma\, g_i + \sigma\, \varepsilon_{i,d}, \tag{61}$$

where g_i, $e_{i,d}$, and $\varepsilon_{i,d}$ are independent standard Gaussian variables and $\sigma = 1$. The shared factor g_i induces dependence among the features of $\hat{Z}$, with correlation $\sqrt{\rho_d \rho_{d'}}$ between features $d \neq d'$, while $e_{i,d}$ is specific to feature d. Each feature of $\hat{Z}$ has unit variance, so the representation is standardized by construction. Since the rows are independent,

$$I(c_j; \hat{c}_j) = -\frac{B}{2} \log\left(1 - r_j^2\right), \qquad r_j := \mathrm{Corr}(Z_{i,j}, \hat{Z}_{i,j}), \tag{62}$$

and the dependence term in Equation 52 can similarly be evaluated from the conditional variances in Equation 47.

Evaluation. For each configuration, we sample minibatches with $B = 256$, standardize each feature channel across the minibatch, and compute $\mathcal{L}_{\mathrm{feat}}(j)$ as in Equation 9. When all features are statistically identical, we average over 40 minibatches and 256 anchor features; when a single anchor feature is considered, we average over 400 minibatches. Since the variational bound holds for any scoring function, we report the largest certificate over $\tau \in \{0.05, 0.1, 0.2, 0.5\}$. The Monte Carlo standard errors are below 0.1 nats in all configurations. Figure 2 summarizes the three settings considered below.

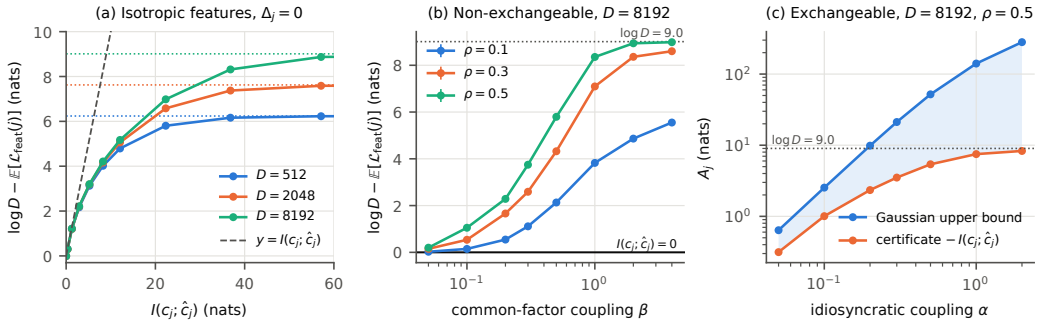


Figure 2: Synthetic study of the feature-wise dependence penalty with $B = 256$. (a) Isotropic features with $\Delta_j = 0$: the certificate $\log D - \mathbb{E}[\mathcal{L}_{\mathrm{feat}}(j)]$ remains below the feature-wise mutual information and saturates at $\log D$ (dotted lines). (b) Non-exchangeable candidates: although $I(c_j; \hat{c}_j) = 0$, the certificate approaches $\log D$. (c) Exchangeable candidates with $\Delta_j = A_j$: the excess of the certificate over the mutual information and the Gaussian characterization bracket A_j as in Equation 63.

Isotropic features. We first set $\rho_d = 0$ and $\gamma = 0$, such that the features are independent across channels and $Z_{i,j}$ is a noisy copy of its matched feature $\hat{Z}_{i,j}$. In this setting, $R_j^2 = 0$ and $\Sigma_{\hat{z}} = I_D$, and therefore $\Delta_j = 0$ by Equation 60. We vary α to control the matched correlation r_j and hence $I(c_j; \hat{c}_j)$. As shown in Figure 2(a), the certificate remains below the feature-wise mutual information across feature dimensions and approaches the $\log D$ ceiling as the mutual information increases. The same behavior is observed for $B \in \{64, 256, 1024\}$. This verifies the expected behavior of the feature-wise bound when $\Delta_j = 0$.

Non-exchangeable candidates. We next consider non-exchangeable feature candidates. We set $\rho_j = 0$ for the matched feature and $\rho_d = \rho$ for $d \neq j$, such that only the negative candidates depend on the common factor. The anchor is defined as $Z_{i,j} = -\beta g_i + \sigma \varepsilon_{i,j}$, corresponding to $\alpha = 0$ and $\gamma = -\beta$. It is therefore independent of $\hat{Z}_{i,j}$, giving $I(c_j; \hat{c}_j) = 0$ for every β, while remaining negatively correlated with the negative candidates through g_i. The matched feature can consequently receive the largest relative similarity despite carrying no information about the anchor.

As shown in Figure 2(b), the certificate increases with both β and ρ and can approach the $\log D$ ceiling. For $D = 8192$, it reaches 8.99 nats compared with $\log D = 9.01$; at the default temperature $\tau = 0.1$,

the corresponding certificate is 6.70 nats. Thus, the variational term alone can substantially exceed the feature-wise mutual information, and the penalty Δ_j is required for the bound in Equation 6. Both components of Δ_j are present in this construction: the negative candidates contain information about the anchor through the common factor, and the matched candidate is statistically distinguishable from the remaining candidates without observing the anchor. Importantly, this occurs despite all features being standardized, showing that standardization alone does not ensure a small Δ_j.

Exchangeable candidates. Finally, we isolate the dependence component of the penalty. We set $\rho_d = \rho$ for every d, so that the candidate features are exchangeable and $I(V;Y) = 0$. Equation 7 then gives $\Delta_j = A_j$, where $A_j = I(c_j; Y_{-V} \mid Y_V, V)$ is defined in Equation 25. We choose $\gamma = -\alpha\sqrt{(1-\rho)/\rho}$, which gives $\mathrm{Cov}(Z_{i,j}, \hat{Z}_{i,j}) = 0$ and therefore $I(c_j; \hat{c}_j) = 0$ for every α. Although the anchor shares the feature-specific component $e_{i,j}$ with its matched feature, this covariance is cancelled by its dependence on the common factor. The matched feature can nevertheless remain identifiable relative to the negative candidates because the softmax in Equation 9 is invariant to a common shift of the similarities.

In this setting,

$$\log D - \mathbb{E}\big[\mathcal{L}_{\mathrm{feat}}(j)\big] - I(c_j; \hat{c}_j) \;\le\; A_j \;\le\; -\frac{B}{2}\log\big(1 - R_j^2\big), \tag{63}$$

where the first inequality follows from Equation 6 with $I(V;Y) = 0$, and the second follows from Equation 52. Figure 2(c) shows both quantities for $D = 8192$ and $\rho = 0.5$. The excess of the certificate over the mutual information increases with α and reaches 8.30 nats, approaching $\log D$ despite $I(c_j; \hat{c}_j) = 0$. Since $I(V;Y) = 0$, this excess arises entirely from the dependence between the anchor and the negative candidates. The Gaussian characterization upper-bounds the excess in all configurations and remains within a factor of 2.0–2.5 for weak coupling ($\alpha \le 0.1$).

Summary. These experiments illustrate the role of the feature-wise dependence penalty. When $\Delta_j = 0$, the certificate remains below the feature-wise mutual information and exhibits the expected $\log D$ ceiling. Subtracting the Gaussian characterization of Δ_j in Equation 60 restores a valid lower bound in all configurations. These results support the interpretation of Δ as an intrinsic component of the feature-wise bound and are consistent with the conditions identified following Equation 8: the penalty is small when non-corresponding features provide little additional information about the anchor and the within-branch covariance is close to isotropic.