Information-Theoretic Unsupervised Embedding Quality Evaluation

Mikhail Kuznetsov[1,4], **Ivan Butakov**[1,2,3], **Marina Munkhoeva**[5,4], **Alexey Frolov**[1] **& Ivan Oseledets**[1,3,5]
[1]Applied AI Institute [2]Moscow Independent Research Institute of Artificial Intelligence [3]Institute of Numerical Mathematics, RAS [4]Lomonosov Moscow State University [5]AXXX
Moscow, Russia
mmkuznecov2002@gmail.com, ivan.butakov@applied-ai.ru

Abstract

This study revisits existing unsupervised measures of embedding quality and introduces new metrics rooted in Information Theory. We establish that classical spectral metrics such as rank, effective rank, and NESum form a unified family of Rényi entropies. An extensive evaluation of both existing and new approaches reveals that most failures in the generalization of SSRL models can be explained via linear deficiencies of the embeddings, rather than by more intricate metrics like clustering or entropy. On the other hand, non-linear metrics proved useful for quantifying model alignment.

1 Introduction

In recent years, advances in unsupervised and self-supervised learning have enabled models to achieve performance competitive with supervised methods under standard evaluation protocols (Bardes et al., 2022; Chen et al., 2020). More recent paradigms, including masked modeling and large-scale distillation, have continued to push representation quality (He et al., 2022; Oquab et al., 2023). This progress has proven particularly valuable for representation learning from unlabeled data, creating a correspondingly strong need to quantify the quality of learned representations.

In parallel, embedding models are increasingly trained and deployed as components of multimodal systems, where visual encoders are optimized using weak, noisy, or automatically collected supervision such as image–text pairs (Jia et al., 2021; Radford et al., 2021; Zhai et al., 2022). In this setting, curated labels may be unavailable, domain-specific, or misaligned with the intended deployment, and running a full suite of downstream evaluations for each iteration can be prohibitively expensive. Moreover, multimodal retrieval and alignment rely directly on embedding geometry, issues such as anisotropy or partial collapse can degrade alignment and retrieval even when task-based probes appear competitive. These trends further motivate unsupervised embedding-quality metrics.

Prevailing approaches rely on supervised downstream tasks (primarily linear classification) to serve as a proxy for embedding quality (Bardes et al., 2022; Chen et al., 2020; Perozzi et al., 2014). However, this paradigm is fundamentally limited by its dependence on annotated data and its susceptibility to bias from the arbitrary selection of benchmark tasks. Consequently, there is a clear need for unsupervised evaluation metrics that enable direct assessment and model selection within a label-free framework.

Unfortunately, existing research primarily focuses on simple metrics that typically do not go beyond linear evaluation or the spectral properties of covariance matrices (Tsitsulin et al., 2023). In this context, we believe Information Theory offers a strong alternative by providing quantities and metrics that are interpretable and correspond to actual information content (Polyanskiy and Wu, 2024).

In our work, we explore the application of Information Theory to embedding quality evaluation. In contrast to popular linear metrics, we propose using differential entropy as a robust and universal measure of the information capacity of learned representations. We show that

this measure already accounts for possible linear deficiencies of embeddings, thus incorporating, to some extent, existing approaches, while also capturing non-linear characteristics of representations. We also propose a scalable and practical proxy that provides a cost-effective alternative for estimating entropy while maintaining the aforementioned benefits.

We summarize our key contributions as follows:

- We propose the use of differential entropy as a universal and theoretically grounded measure for the diversity and information capacity of embeddings.
- We provide a decomposition of this entropy into three interpretable components: *scalar* inflation, and *linear* and *non-linear* collapses.
- We review existing spectral metrics and offer a unified perspective on them through the lens of Rényi entropy. This approach connects established linear evaluation methods to the linear entropy term, thereby completing our proposed "taxonomy of collapse."
- We conduct an extensive evaluation across numerous self-supervised representation learning (SSRL) methods to benchmark both existing and novel metrics. Based on this analysis, we identify and recommend optimal measures for assessing the linear and non-linear quality of representations.

The remainder of the paper is structured as follows: in Section 2, the necessary background in information theory is provided; the related works are discussed in Section 3; Section 4 introduces a general taxonomy that interconnects existing and the proposed approach to unsupervised embeddings quality evaluation. A computationally efficient and scalable alternative to direct entropy estimation is proposed in Section 5. We provide the key results of our experimental evaluation in Section 6. Finally, we discuss the results in Section 7.

2 Preliminaries

Elements of Information Theory. For any probability measure $\mathbb{Q} \ll \mathbb{P}$, the Kullback-Leibler (KL) divergence is $\mathsf{KL}[\mathbb{Q} \,\|\, \mathbb{P}] = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right]$, which is non-negative and vanishes if and only if (iff) $\mathbb{P} = \mathbb{Q}$. 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. The mutual information (MI) between X and Y quantifies the divergence between their joint and marginal distributions:

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

When $\mathbb{P}_X$ admits a probability density function (PDF) $p(X)$ with respect to (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

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

Random Slicing. In this work, we denote by μ_{M} the normalized Haar (uniform) probability measure on a compact manifold M, i.e., the unique bi-invariant measure satisfying $\mu_{\mathrm{M}}(\mathrm{M}) = 1$. Hence, to sample uniformly from specific spaces we write $\mathrm{W} \sim \mu_{\mathrm{O}(d)}, \theta \sim \mu_{\mathbb{S}^{d-1}}, \mathrm{A} \sim \mu_{\mathrm{St}(k,d)}$, indicating draws from the Haar measures on orthogonal group $\mathrm{O}(d) = \left\{\mathrm{Q} \in \mathbb{R}^{d\times d} : \mathrm{Q}^{\mathsf{T}}\mathrm{Q} = \mathrm{Q}\mathrm{Q}^{\mathsf{T}} = \mathrm{I}\right\}$, the unit sphere $\mathbb{S}^{d-1} = \left\{X \in \mathbb{R}^d : \|X\|_2 = 1\right\}$, and the Stiefel manifold $\mathrm{St}(k,d) = \left\{\mathrm{Q} \in \mathbb{R}^{d\times k} : \mathrm{Q}^{\mathsf{T}}\mathrm{Q} = \mathrm{I}\right\}$, respectively.

3 Background

Embeddings Quality. The quality of learned representations is traditionally assessed through downstream task performance using supervised probes. Linear probing, where a

linear classifier is trained on frozen embeddings, has become the de facto standard for evaluating representation quality, under the assumption that good representations enable linear separability of downstream task classes. Nearest-neighbor (kNN) probing provides an alternative by evaluating how well embeddings preserve semantic similarity between samples. However, both approaches rely on labeled data, making them unsuitable for model selection and hyperparameter tuning in fully unsupervised settings.

The challenge of label-free quality assessment has motivated the development of unsupervised embedding quality metrics. Tsitsulin et al. (2023) identify three complementary perspectives for evaluating representation quality without supervision. The linear classifier perspective examines how easily representations can be linearly transformed to downstream targets and is quantified using metrics such as coherence, which measures alignment of singular vectors with the standard basis (Mohri and Talwalkar, 2011), and the pseudo-condition number, which captures numerical stability and sensitivity of linear systems (Belsley et al., 2005; Ben-Israel, 1966). The numerical linear algebra perspective focuses on the stability and effective dimensionality of the representation space, commonly measured via stable (numerical) rank, which characterizes how evenly variance is distributed across dimensions and governs robustness to subsampling and noise (Roy and Vetterli, 2007; Rudelson and Vershynin, 2007). The high-dimensional probability perspective assesses whether embeddings are distributed uniformly across the available dimensions, with self-clustering metrics quantifying deviations from isotropic distributions on the hypersphere (Assran et al., 2022; Vershynin, 2018).

These perspectives capture distinct failure modes of learned representations: poor alignment with canonical bases, numerical ill-conditioning, or excessive concentration in low-dimensional subspaces. Recent work has further emphasized spectral analysis of representation covariance matrices as a unifying framework for understanding these phenomena. The eigenspectrum encodes fundamental properties of learned representations, including effective dimensionality, isotropy, and susceptibility to various forms of representational collapse (Garrido et al., 2023; He and Ozay, 2022). Spectral metrics can be computed efficiently without labels and provide interpretable diagnostics of representation quality. Moreover, theoretical connections between spectral decay and generalization performance have been established for linear models, most notably in the context of benign overfitting and power-law eigenspectrum (Agrawal et al., 2022; Bartlett et al., 2020), suggesting that spectral analysis may offer principled guidance for model selection beyond empirical correlation.

Spectral Metrics. Spectral metrics analyze the eigenvalue distribution of representation covariance matrices $\Sigma = \mathbb{E}[XX^\mathsf{T}]$, where $X \in \mathbb{R}^d$ denotes the embedding of a data point. For an eigenvalue decomposition $\Sigma = \mathsf{Q}\Lambda\mathsf{Q}^\mathsf{T}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$, the normalized eigenvalue distribution $Q_i = \lambda_i / \sum_j \lambda_j$ characterizes the concentration of variance across the embedding space.

Several metrics quantify different aspects of this distribution. The matrix rank $\mathrm{rank}(\Sigma) = |\{i \mid \lambda_i > 0\}|$ counts non-zero eigenvalues but provides limited information about their relative magnitudes. The effective rank addresses this limitation:

$$\mathrm{erank}(\Sigma) := \exp(\mathsf{H}(Q)) = \exp\left(-\sum_{i=1}^{d} Q_i \log Q_i\right),$$

where $\mathsf{H}(Q) = -\sum_{i=1}^{d} Q_i \log Q_i$ is the Shannon entropy of the normalized eigenvalue distribution. Effective rank captures the number of dimensions that substantially contribute to the representation, with $1 \leq \mathrm{erank}(\Sigma) \leq \mathrm{rank}(\Sigma)$.

NESum (He and Ozay, 2022) measures the ratio of total variance to maximum variance:

$$\mathrm{NESum}(\Sigma) = \sum_{i=1}^{d} \lambda_i / \lambda_1 = \tau / \lambda_1,$$

where $\tau = \mathrm{tr}(\Sigma)$ is the trace. This metric is sensitive to dominance by the leading eigenvalue and can be interpreted as the inverse of the maximum normalized eigenvalue Q_1.

The stable rank, defined as

$$\operatorname{srank}(\Sigma) = \frac{\|\Sigma\|_F^2}{\|\Sigma\|_2^2} = \frac{\sum_{i=1}^{d} \lambda_i^2}{\lambda_1^2},$$

measures the squared Frobenius-to-spectral norm ratio and quantifies the concentration of the spectrum. It can be estimated efficiently using trace and norm computations without full eigendecomposition.

The α-ReQ metric (Agrawal et al., 2022) characterizes eigenspectrum decay by fitting a power law $\lambda_i \propto i^{-\alpha}$ to the sorted eigenvalues. The decay coefficient α has been empirically linked to generalization performance, with $\alpha \approx 1$ corresponding to optimal representations. Values $\alpha \gg 1$ indicate rapid decay (dimensional collapse), while $\alpha \ll 1$ suggests slow decay (inefficient use of dimensions).

Information Theory in SSRL. Contrastive self-supervised representation learning (SSRL) methods, such as Deep InfoMax, are fundamentally linked to maximizing mutual information (MI) between different augmented views of data (Hjelm et al., 2019; Oord et al., 2019). This is typically framed as optimizing

$$\mathsf{I}(f(X'); f(X'')) \to \max,$$

where X is the input data, X', X'' are two independent augmentations, and f is an encoder network. Non-contrastive methods, such as autoencoders (Hinton and Salakhutdinov, 2006) and VICReg (Bardes et al., 2022), also have an InfoMax-like interpretation (Butakov et al., 2025).

However, a conventional contrastive objective alone is insufficient for assessing embedding quality, as it merely measures representation invariance to data augmentations. Since differential entropy quantifies the randomness and diversity of a distribution, it is natural to use $\mathsf{h}(f(X))$ to detect collapses in representations. Butakov et al. (2025) establish a firm connection between embedding entropy and a modified InfoMax-style SSRL objective:

$$\begin{aligned} \mathsf{I}(f(X'); f(X) + Z) &= \mathsf{h}(f(X) + Z) + \underset{\text{invariance term}}{\text{embeddings}} \\ \mathsf{KL}[f(X) \parallel \mathcal{N}(0, \mathrm{I})] &\leq \underset{\text{terms}}{\text{other}} - \mathsf{I}(f(X'); f(X) + Z), \end{aligned}$$

where $f(X)$ is energy-constrained (having unit variance across all components) and Z is an independent Gaussian noise. However, in practice, these expressions are difficult to evaluate due to the high sample complexity of mutual information and entropy estimation. One of our main contributions is providing a *scalable, slicing-based alternative* to the results above.

Sliced Mutual Information. To mitigate the curse of dimensionality, one may average MI over all k-dimensional projections. The k-sliced mutual information (k-SMI) (Goldfeld et al., 2022) between X and Y is defined as

$$\mathsf{SI}_k(X; Y) = \mathsf{I}(\Theta^\mathsf{T} X; \Phi^\mathsf{T} Y \mid \Theta, \Phi),$$

where $\Theta \perp\!\!\!\perp \Phi \perp\!\!\!\perp X, Y$ and $\Theta \sim \mu_{\mathrm{St}(k, d_X)}$, $\Phi \sim \mu_{\mathrm{St}(k, d_Y)}$.

SMI is scalable and cheap to estimate, making it a convenient replacement for MI. However, recent studies suggest that SMI also exhibits severe limitations and inherent biases (Semenenko et al., 2026). One of them — the *redundancy bias* — makes SMI prefer linearly collapsed distributions, rendering it unsuitable for our specific task.

4 Taxonomy of Collapse

Entropy Decomposition. As a universal measure of distribution randomness and diversity, differential entropy already accounts for both linear and non-linear collapses. This is illustrated by the following decomposition:

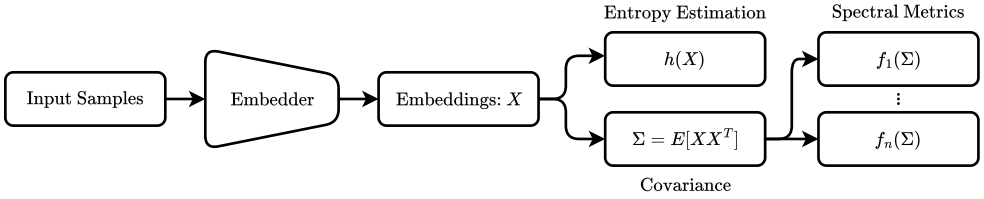

Figure 1: Experimental evaluation pipeline. Input samples are processed through a pre-trained embedder to produce embedding vectors X. The evaluation branches into two parallel pathways: (1) direct entropy estimation $h(X)$ on the embedding distribution, and (2) spectral analysis via the empirical covariance matrix $\Sigma = \mathbb{E}[XX^\mathsf{T}]$, from which we compute various spectral metrics $\{f_1(\Sigma), \cdots, f_n(\Sigma)\}$ including effective rank, NESum, and participation ratio. Both pathways provide complementary perspectives on embedding quality: entropy captures distributional properties directly, while spectral metrics characterize the linear geometry of the representation space.

Proposition 4.1. For a d-dimensional X with covariance matrix Σ and finite $\mathsf{h}(X)$:

$$\mathsf{h}(X) = \underbrace{\frac{d}{2}\log\|\Sigma\|}_{\text{scalar inflation}} + \underbrace{\frac{1}{2}\log\det(\Sigma/\|\Sigma\|)}_{\text{linear collapse}} + \underbrace{\mathsf{h}\left(\Sigma^{-\frac{1}{2}}X\right)}_{\text{non-linear collapse}}$$

where $\|\cdot\|$ is the matrix 2-norm, and $\Sigma^{-\frac{1}{2}}X$ is whitened (decorrelated) X.

Proof of Proposition 4.1. Follows directly from $\mathsf{h}(\mathrm{A}\cdot X) = \log|\det \mathrm{A}| + \mathsf{h}(X)$. □

Spectral Metrics. Linear collapse can be explored through spectral metrics, which characterize the eigenvalue distribution Q of the covariance matrix. We now establish that classical spectral metrics form a unified family rooted in information theory.

Let $\Sigma \in \mathbb{R}^{d\times d}$ be a positive semidefinite covariance matrix with eigenvalue decomposition $\Sigma = \mathrm{Q}\Lambda\mathrm{Q}^\mathsf{T}$, where $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Define the total variance $\tau = \mathrm{tr}(\Sigma) = \sum_{i=1}^d \lambda_i$ and the normalized eigenvalue distribution $Q_i := \lambda_i/\tau$ for $i = 1, \cdots, d$. Note that $Q = (Q_1, \cdots, Q_d)$ forms a probability distribution.

Classical spectral metrics can be defined as follows:

- The matrix rank counts non-zero eigenvalues: $\mathrm{rank}(\Sigma) = |\{i \mid \lambda_i > 0\}|$
- NESum measures the trace-to-max-eigenvalue ratio: $\mathrm{NESum}(\Sigma) = \tau/\lambda_1 = 1/Q_1$
- Effective rank is the exponential of Shannon entropy: $\mathrm{erank}(\Sigma) = \exp(\mathsf{H}(Q))$, where $\mathsf{H}(Q) = -\sum_{i=1}^d Q_i \log Q_i$

These three seemingly disparate metrics are unified through the Rényi entropy framework. For a probability distribution Q and parameter $\alpha \geq 0$ with $\alpha \neq 1$, the Rényi entropy of order α is defined as:

$$\mathsf{H}_\alpha(Q) := \frac{1}{1-\alpha}\log\left(\sum_{i=1}^d Q_i^\alpha\right).$$

Special cases defined by limits include:

$$\mathsf{H}_0(Q) = \lim_{\alpha\to 0}\mathsf{H}_\alpha(Q) = \log|\{i \mid Q_i > 0\}|,$$

$$\mathsf{H}_1(Q) = \lim_{\alpha\to 1}\mathsf{H}_\alpha(Q) = -\sum_{i=1}^d Q_i \log Q_i \quad \left(\begin{smallmatrix}\text{Shannon's}\\ \text{entropy}\end{smallmatrix}\right),$$

$$\mathsf{H}_\infty(Q) = \lim_{\alpha\to\infty}\mathsf{H}_\alpha(Q) = -\log\left(\max_i Q_i\right).$$

The limit $\mathsf{H}_0(Q)$ follows because $Q_i^\alpha \to 1$ for $Q_i > 0$ as $\alpha \to 0$, while $0^\alpha = 0$ for all $\alpha > 0$. For $\mathsf{H}_1(Q)$, we set $f(\alpha) = \sum_{i=1}^d Q_i^\alpha$. At $\alpha = 1$, we have $f(1) = 1$ and $\log f(1) = 0$, giving an indeterminate form $0/0$. Applying L'Hôpital's rule, we derive:

$$\lim_{\alpha\to 1} \frac{\log f(\alpha)}{1-\alpha} = \lim_{\alpha\to 1} \frac{f'(\alpha)/f(\alpha)}{-1} = -\left(\sum_{i=1}^d Q_i \log Q_i\right),$$

where

$$f'(\alpha) = \sum_{i=1}^d Q_i^\alpha \log Q_i \qquad f'(1) = \sum_{i=1}^d Q_i \log Q_i$$

For $\mathsf{H}_\infty(Q)$, factoring out $Q_{\max}^\alpha$ from the sum yields $\lim_{\alpha\to\infty} \sum_{i=1}^d Q_i^\alpha = Q_{\max}^\alpha$, since all terms with $Q_i < Q_{\max}$ vanish exponentially.

Taking exponentials of Rényi entropies recovers the classical linear metrics:

$$\mathrm{rank}(\Sigma) = \exp(\mathsf{H}_0(Q)), \quad \mathrm{erank}(\Sigma) = \exp(\mathsf{H}_1(Q)), \quad \mathrm{NESum}(\Sigma) = \exp(\mathsf{H}_\infty(Q)).$$

This unification is significant because Rényi entropy satisfies the fundamental monotonicity property: $\mathsf{H}_\alpha(Q)$ is non-increasing in α for $\alpha \geq 0$. Since exponentiation is monotonous,

$$\mathrm{NESum}(\Sigma) \leq \mathrm{erank}(\Sigma) \leq \mathrm{rank}(\Sigma).$$

This inequality reveals the different sensitivities of each metric: NESum is most sensitive to the dominant eigenvalue (measuring concentration at the top), effective rank balances all eigenvalues equally through entropy, and rank only distinguishes zero from non-zero modes.

Connection to Power-Law Spectra. The Rényi framework enables precise characterization when eigenvalues follow a power law $\lambda_i \propto i^{-\alpha}$. Under this model, the total variance can be expressed using the generalized harmonic number $H_{d,\alpha} = \sum_{i=1}^d i^{-\alpha}$. Specifically, if $\lambda_i = \lambda_1 \cdot i^{-\alpha}$, then:

$$\tau = \sum_{i=1}^d \lambda_i = \lambda_1 \sum_{i=1}^d i^{-\alpha} = \lambda_1 H_{d,\alpha} \quad \Longrightarrow \quad \mathrm{NESum}(\Sigma) = \tau/\lambda_1 = H_{d,\alpha}.$$

Asymptotic behavior of $H_{d,\alpha}$ characterizes a spectral regime:

- For $\alpha \approx 1$: $H_{d,\alpha} \approx \log(d)$, yielding balanced eigenvalue distribution
- For $\alpha \gg 1$: $H_{d,\alpha} = O(1)$, indicating rapid decay and dimensional collapse
- For $\alpha \ll 1$: $H_{d,\alpha} = \Theta(d^{1-\alpha})$, indicating slow decay and diffuse representations

This connection enables estimation of the decay coefficient α directly from the scalar ratio $\mathrm{NESum}(\Sigma)$ without requiring full eigendecomposition or log-log regression, providing a computationally stable alternative to the original α-ReQ approach (Agrawal et al., 2022).

Participation Ratio. To obtain tighter bounds on α, we introduce the participation ratio (Rényi-2 effective dimension):

$$d_2(\Sigma) = \exp(\mathsf{H}_2(Q)) = \left(\sum_{i=1}^d Q_i^2\right)^{-1} = \frac{\tau^2}{\|\Sigma\|_F^2} = \left(\sum_i \lambda_i\right)^2 / \left(\sum_i \lambda_i^2\right),$$

where $\|\Sigma\|_F = \sqrt{\sum_{i=1}^d \lambda_i^2}$ is the Frobenius norm. The participation ratio offers an intermediate measure between NESum (which depends only on λ_1) and effective rank (which weights all eigenvalues equally), and can be computed efficiently from trace and Frobenius norm without full eigendecomposition.

Under the power-law model $\lambda_i = \lambda_1 \cdot i^{-\alpha}$, the participation ratio satisfies:

$$d_2(\Sigma) = \left(H_{d,\alpha}\right)^2 / H_{d,2\alpha},$$

where $H_{d,2\alpha} = \sum_{i=1}^{d} i^{-2\alpha}$. This provides a second independent constraint on α: given observed values $\widehat{\mathrm{NESum}}(\Sigma) = H_{d,\alpha}$ and $d_2(\Sigma) = \left(H_{d,\alpha}\right)^2 / H_{d,2\alpha}$, we can solve for α more precisely than from NESum alone. Specifically, $H_{d,2\alpha} = (\widehat{\mathrm{NESum}}(\Sigma))^2 / d_2(\Sigma)$, which bounds 2α and thus refines the estimate of α.

This two-scalar approach (NESum + participation ratio) provides tighter bounds on the decay coefficient while remaining computationally tractable, as both quantities can be estimated using randomized linear algebra methods such as stochastic trace estimation.

5 Aligned SMI

This section proposes a novel surrogate that can be used as a scalable alternative to Mutual Information and differential entropy under certain circumstances, including our specific task. Inspired by Sliced Mutual Information, we define *Aligned* k-SMI (aSMI):

$$\mathsf{SI}_k^{\parallel}(X;Y) = \mathsf{I}(\mathrm{A}^\mathsf{T}X; \mathrm{A}^\mathsf{T}Y \mid \mathrm{A}), \qquad \begin{matrix} \mathrm{A} \perp\!\!\!\perp X, Y \\ \mathrm{A} \sim \mu_{\mathrm{St}(k,d)} \end{matrix} \tag{2}$$

Here, as opposed to conventional SMI, we use a single matrix A to project both X and Y. This entails the following merits and demerits:

- \+ Whilst independent projectors bias traditional SMI toward low-rank (linearly collapsed) distributions (Semenenko et al., 2026), identical projection matrices eliminate this issue entirely.
- — Using a single projector restricts us to X and Y of the same dimensionality.
- — Contrary to SMI, which nullifies iff $X \perp\!\!\!\perp Y$, $\mathsf{SI}_k^{\parallel}(X;Y) = 0$ **does not imply** $\boldsymbol{X \perp\!\!\!\perp Y}$**.**

Despite the severity of the latter issue, which prohibits aSMI from being adopted as a replacement for SMI in most settings, the following result establishes the former as a scalable proxy for measuring entropy:

Theorem 5.1. Let X be a d-dimensional random vector such that $\mathrm{var}[X_i] \leq 1$ for any i. Let $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$. Then

$$\mathsf{SI}_k^{\parallel}(X; X+Z) \leq \frac{k}{2} \log\left(1 + \frac{1}{\sigma^2}\right),$$

with the equality iff $X \sim \mathcal{N}(0, \mathrm{I})$.

This theorem mirrors a distribution matching result from (Butakov et al., 2025, Theorem 3.4), but leverages scalable sliced metric instead of hard-to-estimate MI.

Since the Gaussian distribution maximizes entropy under energy constraints, and Theorem 5.1 establishes aSMI as a surrogate measure of normality, we employ $\mathsf{SI}_k^{\parallel}(X; X+Z)$ as a proxy for the differential entropy estimation.

6 Experiments

Experimental Setup. We evaluate embedding quality metrics in two settings.

The first setting employs linear probing on eight diverse image classification datasets to measure downstream task performance across varied visual domains. Each dataset represents a distinct classification challenge —from fine-grained recognition (Stanford Cars, Flowers-102) to scene understanding (SUN397) and texture discrimination (DTD) — enabling robust evaluation of how well representations perform on downstream tasks. For each dataset, a frozen pre-trained encoder is augmented with a single trainable linear projection layer optimized via cross-entropy loss.

The second setting performs zero-shot analysis on the ImageNet100 validation set to compute intrinsic embedding properties without task-specific tuning. We utilize the

ImageNet100 validation split specifically because: (1) it provides diverse visual content, (2) these validation images were not seen during pre-training of evaluated models. All spectral metrics, covariance analyses, and information-theoretic measures are computed on embeddings extracted from this set.

This dual-setting design enables correlation analysis between unsupervised embedding diagnostics (computed on ImageNet100) and downstream task performance (measured across eight datasets), revealing which intrinsic properties best predict transfer learning success.

Linear Probing Datasets. We evaluate downstream performance on eight image classification benchmarks spanning diverse visual domains. Food-101 (Bossard et al., 2014) contains 101 food categories with 101,000 images. CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) comprise 60,000 32×32 natural images across 10 and 100 classes respectively. SUN397 (Xiao et al., 2010) provides 108,754 images across 397 scene categories. Stanford Cars (Krause et al., 2013) contains 16,185 images of 196 car models. Describable Textures Dataset (DTD) (Cimpoi et al., 2014) includes 5,640 texture images across 47 categories. Oxford-IIIT Pets (Parkhi et al., 2012) contains 7,349 images of 37 pet breeds. Flowers-102 (Nilsback and Zisserman, 2008) comprises 8,189 images of 102 flower categories. This diversity ensures that metric correlations are not artifacts of specific domain characteristics.

ImageNet100 Subset. ImageNet100 is a stratified subset of the ImageNet-1K dataset (Deng et al., 2009), containing 100 classes selected to represent diverse object categories. We utilize the validation split, comprising 5,000 images (50 per class), for zero-shot embedding analysis. This subset provides sufficient statistical power for covariance matrix estimation while maintaining computational tractability for high-dimensional operations. The class distribution spans animal species, vehicles, household objects, and natural scenes, preserving the semantic diversity of the full ImageNet dataset. All images are resized to 224×224 pixels and normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]).

Model Architectures. We evaluate 16 pre-trained models spanning four architecture families and three training paradigms. Vision Transformers include CLIP-ViT-B/32 (Radford et al., 2021) (512-dim), CLIP-RN50 (1024-dim), timm-vit-base and timm-vit-small (Wightman, 2019) (768-dim and 384-dim), DINO-ViT-B/8, DINO-ViT-B/16, DINO-ViT-S/8, DINO-ViT-S/16 (Caron et al., 2021) (768-dim and 384-dim), DINOv2-ViT-B/14 and DINOv2-ViT-B/14-reg (Oquab et al., 2023) (768-dim), MoCoV3-ViT-B (Chen et al., 2021) (768-dim), and MAE-base (He et al., 2022) (768-dim). Convolutional architectures include ResNet-50 (He et al., 2016) (2048-dim), VICReg-ResNet-50 and VICReg-ResNet-50×2 (Bardes et al., 2022) (2048-dim and 4096-dim), and SimCLR-ResNet-1x (Chen et al., 2020) (2048-dim). All models are used with official pre-trained weights without modification.

Implementation Details. For linear probing, we train a single fully-connected layer with cross-entropy loss for 10 epochs using the Adam optimizer (Kingma and Ba, 2017) with learning rate 0.001 and batch size 512. The encoder remains frozen throughout training. We report test accuracy alongside k-NN classification accuracy (k=5) using cosine similarity in the embedding space. For zero-shot analysis, we extract embeddings for all 5,000 ImageNet100 validation images in a single forward pass. Covariance matrices are computed using unbiased estimation: $\Sigma = (n-1)^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^\top$, where z_i are centered embeddings. Sliced mutual information estimates use projection dimension $k = 2$, Gaussian noise scales $\sigma \in \{0.01, 0.1, 1.0\}$, and $M = 50$ random projections per noise scale. The KSG estimator employs $k = 1$ nearest neighbor for MI computation. PCA filtering retains dimensions with eigenvalues exceeding the noise scale threshold: $\lambda_i > \sigma$.

To identify the optimal preprocessing methodology for information-theoretic metrics, we evaluate multiple configurations. We test four vector preprocessing variants: raw embeddings, L2-normalized vectors (unit norm), standardized features (zero mean and unit variance per dimension), and ZCA-whitened embeddings. For each variant, we compute both sliced mutual information (SMI) and aligned sliced mutual information (ASMI) with three Gaussian noise scales ($\sigma \in \{0.01, 0.1, 1.0\}$). Additionally, we apply PCA-based dimensionality filtering with and without subsequent whitening, retaining only dimensions with eigenvalues exceeding the noise scale threshold matching the original noise scales. This

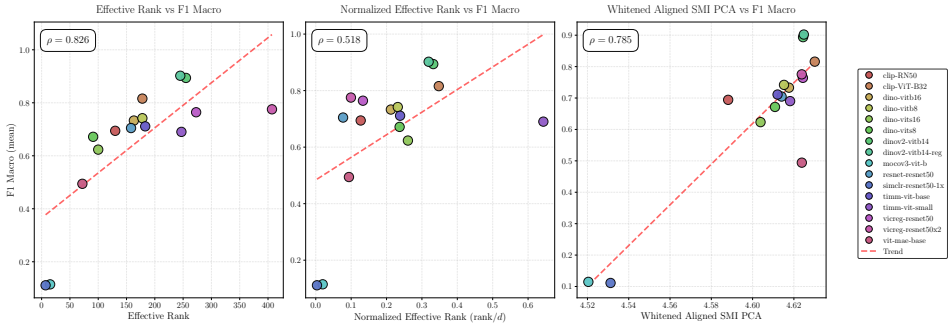

Figure 2: Correlation between unsupervised metrics and mean F1 macro score across eight classification benchmarks for 16 pre-trained models. Each point represents one model architecture. Effective rank **(left)** achieves Spearman $\rho = 0.826$ ($p = 8.2 \times 10^{-5}$) with $R^2 = 0.570$. Normalized effective rank (effective rank divided by embedding dimension) **(middle)** shows reduced correlation ($\rho = 0.518$, $p = 0.040$) with $R^2 = 0.296$. Whitened aligned sliced mutual information with PCA filtering **(right)** achieves competitive rank correlation ($\rho = 0.785$, $p = 3.1 \times 10^{-4}$) but superior linear fit with $R^2 = 0.830$. All correlations are statistically significant ($p < 0.05$) and computed with $n = 16$ models. Dashed lines indicate linear trends (used for R^2 computation).

ablation study yields multiple candidate metrics, from which we select the best-performing configuration based on correlation with downstream task performance. The correlation analysis in Figure 2 presents results from the optimal configuration: whitened aligned sliced mutual information with PCA filtering.

Correlation Analysis. We evaluate the correlation between unsupervised metrics and downstream task performance using Spearman rank correlation coefficient. Figure 2 presents the relationship between effective rank and mean macro F1 score across eight classification tasks. The information-theoretic metric explains 83% of variance in downstream performance compared to 57% for effective rank, demonstrating substantially reduced scatter and more consistent predictive behavior across different model architectures.

6.1 Representational Alignment Analysis

Beyond evaluating individual embedding quality, we investigate whether SMI captures representational similarity between models. We compute pairwise SMI between all 16 models using standardized embeddings and convert the resulting similarity matrix to distances (distance = normalized(1/SMI)) for hierarchical clustering with Ward's linkage.

Figure 3 illustrates clustering by model family and training methods. DINO models (dino-vits16, dino-vits8, dino-vitb16, dino-vitb8) form a tight cluster at the bottom-right, reflecting their shared self-distillation approach. The timm models (timm-vit-small, timm-vit-base) cluster together, as do VICReg variants (vicreg-resnet50, vicreg-resnet50x2). CLIP models occupy separate branches — clip-ViT-B32 and clip-RN50 form distinct groups, consistent with their different architectures despite shared multimodal training. The two collapsed models (mocov3-vit-b, simclr-resnet-1x) cluster together at the top, separated from well-performing models.

Provided hierarchy demonstrates that SMI methods also can be used to measure the representational alignment, capturing both architectural similarities and training paradigm effects across different models.

7 Discussion

In this work, we revisit the problem of unsupervised embeddings quality evaluation. Based on an extensive benchmarking of linear metrics for assessing the quality and diversity of compressed representations, we selected effective rank as the most robust and informative

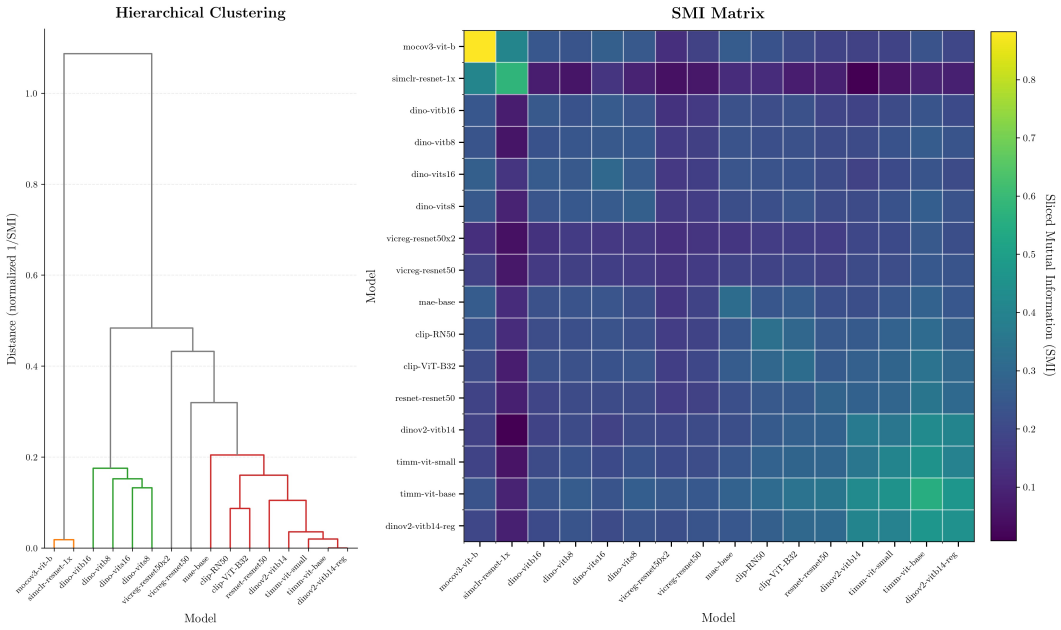

Figure 3: Representational alignment via hierarchical clustering of 16 pre-trained models. **Left**: Dendrogram based on pairwise aligned SMI (standardized preprocessing, distance = normalized(1/SMI)). **Right**: SMI matrix. Model families cluster together: DINO variants (bottom-right), timm models, VICReg variants, and CLIP models in separate branches. Collapsed models (mocov3-vit-b, simclr-resnet-1x) separate distinctly at top.

measure. Effective rank, by quantifying the intrinsic dimensionality and the distribution of variance within the embedding space, consistently correlated with downstream task performance.

Our theoretical analysis shows that classical spectral metrics — rank, effective rank, and NESum — can be viewed as different orders ($\alpha = 0, 1, \infty$) of Rényi entropy applied to the normalized eigenvalue distribution. This connection helps explain their different sensitivities to spectral concentration and their monotonicity ordering $\text{NESum} \leq \text{erank} \leq \text{rank}$. Under power-law spectra, this framework enables direct estimation of the decay coefficient α from spectral scalars. Additionally, the participation ratio (Rényi-2 dimension) provides a computationally efficient second constraint for refining α estimates.

To address the limitations of linear analysis, we developed a novel non-linear evaluation metric — aligned Sliced Mutual Information (aSMI). This metric was enhanced with PCA-based filtering and whitening procedures to (a) improve its stability and discriminative power and (b) eliminate linear term interference. The results demonstrate that the aligned SMI metric possesses predictive power for downstream task performance that is similar to, or even surpassing, that of the best linear metrics.

Crucially, the proposed aSMI metric offers a significant advantage by capturing non-linear deficiencies in representations that linear metrics like effective rank inherently miss. This capability allows for the detection of subtler flaws in how semantic information is organized. Consequently, the combination of effective rank for linear assessment and aligned SMI for non-linear analysis provides a more well-rounded framework for evaluation of embedding quality.

Reproducibility statement. To ensure reproducibility of our results, we provide complete proofs in Section B and experimental details in Section 6.

References

Agrawal, K. K., Mondal, A. K., Ghosh, A., and Richards, B. α-ReQ: Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, *35*, 17626–17638, 2022.

Assran, M., Balestriero, R., Duval, Q., Bordes, F., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., and Ballas, N. The hidden uniform cluster prior in self-supervised learning. *Arxiv Preprint Arxiv:2210.07277*, 2022.

Bardes, A., Ponce, J., and LeCun, Y. Variance-invariance-covariance regularization for self-supervised learning. *ICLR, Vicreg*, *1*(2), 2022.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070, 2020.

Belsley, D. A., Kuh, E., and Welsch, R. E. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

Ben-Israel, A. On error bounds for generalized inverses. *SIAM Journal on Numerical Analysis*, *3*(4), 585–592, 1966.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. *European Conference on Computer Vision*, 446–461, 2014.

Butakov, I., Semenenko, A., Tolmachev, A., Gladkov, A., Munkhoeva, M., and Frolov, A. Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax. *The Thirteenth International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=mAmCdASmJ5

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660, 2021.

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

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649, 2021.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3606–3613, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, 2009.

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. *International Conference on Machine Learning*, 10929–10974, 2023.

Goldfeld, Z., Greenewald, K., Nuradha, T., and Reeves, G. k-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. https://openreview.net/forum?id=L-ceBdl2DPb

He, B., and Ozay, M. Exploring the gap between collapsed & whitened features in self-supervised learning. *International Conference on Machine Learning*, 8613–8634, 2022.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked Autoencoders Are Scalable Vision Learners. *CVPR*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, 2016.

Hinton, G. E., and Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507, 2006. https://doi.org/10.1126/science.1127647

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

Jia, C., Yang, Y., Xia, Y., and others. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Arxiv Preprint Arxiv:2102.05918*, 2021.

Kingma, D. P., and Ba, J. *Adam: A Method for Stochastic Optimization*, 2017.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, *69*(6), 66138, 2004. https://doi.org/10.1103/PhysRevE.69.066138

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561, 2013.

Krizhevsky, A., Hinton, G., and others. *Learning multiple layers of features from tiny images. (2009)*, 2009.

Mohri, M., and Talwalkar, A. Can matrix coherence be efficiently and accurately estimated?. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 534–542, 2011.

Nilsback, M.-E., and Zisserman, A. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729, 2008.

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

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., and others. Dinov2: Learning robust visual features without supervision. *Arxiv Preprint Arxiv:2304.07193*, 2023.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3498–3505, 2012.

Perozzi, B., Al-Rfou, R., and Skiena, S. DeepWalk: online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710, 2014. https://doi.org/10.1145/2623330.2623732

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., and others. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763, 2021.

Roy, O., and Vetterli, M. The effective rank: A measure of effective dimensionality. *2007 15th European Signal Processing Conference*, 606–610, 2007.

Rudelson, M., and Vershynin, R. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, *54*(4), 21–es, 2007.

Semenenko, A., Butakov, I., Frolov, A., and Oseledets, I. Curse of Slicing: Why Sliced Mutual Information is a Deceptive Measure of Statistical Dependence. *The Fourteenth International Conference on Learning Representations*, 2026. https://openreview.net/forum?id=KxeBgh1zWr

Tsitsulin, A., Munkhoeva, M., and Perozzi, B. Unsupervised embedding quality evaluation. *Topological, Algebraic and Geometric Learning Workshops 2023*, 169–188, 2023.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press, 2018.

Wightman, R. *PyTorch Image Models*. GitHub, 2019. https://doi.org/10.5281/zenodo.4414861

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492, 2010.

Zhai, X., Kolesnikov, A., Beyer, L., and others. LiT: Zero-Shot Transfer with Locked-image Text Tuning. *CVPR*, 2022.

A Supplementary Theory

Lemma A.1. (Polyanskiy and Wu (2024, Example 2.4)) $\mathsf{h}(\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}\log\big((2\pi e)^d \det\Sigma\big)$.

Lemma A.2. Let X be a d-dimensional absolutely continuous vector with mean vector m and covariance matrix Σ. Then

$$\mathsf{h}(X) = \mathsf{h}(\mathcal{N}(m,\Sigma)) - \mathsf{KL}[X \parallel \mathcal{N}(m,\Sigma)]$$

Proof of Lemma A.2. Let p be a PDF of X, and φ — a PDF of $\mathcal{N}(m,\Sigma)$. Then

$$\mathsf{KL}[X \parallel \mathcal{N}(m,\Sigma)] = \mathbb{E}_{\mathbb{P}_X} \log\frac{p(X)}{\varphi(X)} = -\mathsf{h}(X) - \mathbb{E}_{\mathbb{P}_X}\log\varphi(X) \overset{\star}{=} -\mathsf{h}(X) - \mathbb{E}_{\mathcal{N}(m,\Sigma)}\log\varphi(X) = \mathsf{h}(\mathcal{N}(m,\Sigma)) - \mathsf{h}(X),$$

where the $(\star)$ transition is due to $\log\varphi(X)$ being a quadratic form of X, and X and $\mathcal{N}(m,\Sigma)$ having the same first- and second-order moments. □

B Complete Proofs

Theorem 5.1. Let X be a d-dimensional random vector such that $\mathrm{var}[X_i] \le 1$ for any i. Let $Z \sim \mathcal{N}(0,\sigma^2\mathrm{I})$. Then

$$\mathsf{SI}_k^{\parallel}(X; X+Z) \le \frac{k}{2}\log\left(1+\frac{1}{\sigma^2}\right),$$

with the equality iff $X \sim \mathcal{N}(0,\mathrm{I})$.

Proof of Theorem 5.1. First, note that the equality holds if and only if all the projections $(\mathrm{A}^\mathsf{T}X; \mathrm{A}^\mathsf{T}(X+Z))$ are jointly Gaussian. Indeed, as $\mathrm{A}^\mathsf{T}\mathrm{A} = \mathrm{I}_k$ and

$$\mathsf{I}(\mathrm{A}^\mathsf{T}X; \mathrm{A}^\mathsf{T}(X+Z) \mid \mathrm{A}) = \mathsf{I}(\mathrm{A}^\mathsf{T}X; \mathrm{A}^\mathsf{T}X + \mathrm{A}^\mathsf{T}Z \mid \mathrm{A}) = \mathsf{I}(X'; X'+Z \mid \mathrm{A}),$$

where X' is k-dimensional, $\mathrm{var}[X_{i'}] = 1$, $Z' \sim \mathcal{N}(0,\sigma^2\mathrm{I}_k)$, and $X' \perp\!\!\!\perp Z'$, we can apply Lemma A.2:

$$\begin{aligned}\mathsf{I}(X'; X'+Z \mid \mathrm{A}) &\le \mathsf{h}(X'+Z' \mid \mathrm{A}) - \mathsf{h}(X'+Z' \mid X', \mathrm{A}) = \mathsf{h}(X'+Z' \mid \mathrm{A}) - \mathsf{h}(Z' \mid \mathrm{A}) \\ &= \mathsf{h}(X'+Z' \mid \mathrm{A}) - \frac{k}{2}\log(2\pi e\sigma^2) \le \frac{k}{2}\log(2\pi e(1+\sigma^2)) - \frac{k}{2}\log(2\pi e\sigma^2) = \frac{k}{2}\log\left(1+\frac{1}{\sigma^2}\right),\end{aligned}$$

with the equality iff $X'+Z'$ is Gaussian for every A. However, because Z' is normal and independent, $X'+Z'$ is Gaussian iff X' is Gaussian. A fundamental property of Gaussian distribution is that X is normal iff all its projections are normal. Therefore, the equality holds iff X is Gaussian. □

Table 1: Entropy decomposition components computed on ImageNet100 validation set embeddings for all 16 pre-trained models. Models are sorted by effective rank. The three components (scalar inflation, linear collapse, nonlinear collapse) form an additive decomposition of differential entropy.

Model	Eff Rank	Scalar Infl	Linear Collapse	Nonlinear Collapse
vicreg-resnet50x2	545	4280.52	−17388.92	1795.20
dinov2-vitb14	524	1739.09	−1843.59	1085.27
dinov2-vitb14-reg	397	1247.07	−1884.12	1084.07
vicreg-resnet50	363	2722.82	−7410.23	1848.43
timm-vit-small	292	861.99	−849.73	540.45
dino-vitb16	248	1051.33	−2175.40	1077.46
timm-vit-base	226	1721.22	−2248.33	1084.98
dino-vitb8	217	961.97	−2125.38	1085.43
dino-vits8	184	591.11	−980.66	541.10
resnet-resnet50	174	3336.85	−8869.61	1869.00
clip-RN50	120	−965.40	−4902.79	1441.54
mae-base	80	−42.34	−2854.34	1085.66
mocov3-vit-b	14	2036.90	−4059.36	1084.48
simclr-resnet-1x	5	8529.08	−19290.44	1314.90
clip-ViT-B32	178	318.40	−1412.06	726.49
dino-vits16	100	681.24	−1010.22	540.94

C Experimental Results Tables

ImageNet100 Zero-Shot Analysis. Table 1 presents the entropy decomposition results for all 16 evaluated models computed on ImageNet100 training set embeddings. Effective rank quantifies the spectral dimensionality of learned representations. Scalar inflation captures the overall variance scale of the embedding distribution. Linear collapse measures spectral anisotropy through the determinant of the normalized covariance matrix. Nonlinear collapse quantifies deviation from Gaussianity in the whitened representation space.

Linear Probing Performance. Table 2 reports downstream task performance via linear probing across image classification benchmarks. Mean accuracy and F1 macro scores are computed by averaging across all eight datasets. The k-NN column reports 5-nearest-neighbor classification accuracy using cosine similarity in the embedding space, providing a parameter-free baseline. Effective rank and embedding dimensionality are reported for each model, with their ratio (EffRank/Dim) providing a normalized measure of spectral dimensionality utilization. Condition number quantifies the numerical stability of the representation covariance matrix.

Aligned Sliced Mutual Information. In our experiments, we sample 128 random projectors A and use the KSG k-nearest neighbors mutual information estimator (Kraskov et al., 2004) to approximate $\mathsf{I}(\mathrm{A}X;\mathrm{A}Y)$. We use $k_{\mathrm{NN}} = 1$, as it proved to be the most robust choice. The projection dimensionality is set to $k = 2$.

Table 3 presents aSMI estimates computed on embeddings extracted from the eight downstream task datasets with standardized preprocessing: mean centering and unit variance scaling per dimension. For each model, aSMI is computed on embeddings from each dataset and then averaged across all eight datasets.

PCA-Filtered Aligned Sliced Mutual Information.

Table 2: Linear probing and k-NN results for all 16 pre-trained models evaluated on eight image classification datasets: Food-101, CIFAR-10, CIFAR-100, SUN397, Stanford Cars, DTD, Oxford-IIIT Pets, and Flowers-102. Mean accuracy and F1 macro scores are averaged across all eight test sets. Models are sorted by descending mean accuracy.

Backbone	**Acc**	**F1**	**k-NN**	**Eff Rank**	**Emb Dim**	**EffRank/ Dim**	**Condition Number**
dinov2-vitb14-reg	0.903	0.902	0.876	245	768	0.319	2.0×10^9
dinov2-vitb14	0.894	0.894	0.844	255	768	0.332	3.5×10^{11}
clip-ViT-B32	0.819	0.816	0.732	178	512	0.348	4.0×10^8
vicreg-resnet50x2	0.777	0.775	0.631	407	4096	0.099	3.5×10^6
vicreg-resnet50	0.766	0.764	0.621	273	2048	0.133	8.8×10^4
dino-vitb8	0.746	0.742	0.590	178	768	0.232	3.1×10^{10}
dino-vitb16	0.736	0.733	0.560	163	768	0.212	1.2×10^{11}
timm-vit-base	0.713	0.711	0.592	183	768	0.238	1.1×10^{11}
resnet-resnet50	0.707	0.705	0.621	158	2048	0.077	5.1×10^4
clip-RN50	0.703	0.694	0.603	130	1024	0.127	1.8×10^7
timm-vit-small	0.691	0.690	0.606	247	384	0.643	2.1×10^{10}
dino-vits8	0.678	0.672	0.534	91	384	0.237	7.4×10^8
dino-vits16	0.631	0.623	0.478	100	384	0.260	2.8×10^8
mae-base	0.506	0.494	0.382	72	768	0.094	2.4×10^9
simclr-resnet-1x	0.123	0.111	0.075	7	2048	0.003	5.4×10^{12}
mocov3-vit-b	0.133	0.115	0.091	15	768	0.020	8.4×10^9

Table 4 reports aSMI estimates computed on embeddings extracted from the eight downstream task datasets with PCA filtering followed by whitening (ZCA preprocessing). For each model and dataset, dimensions with eigenvalues below the noise scale threshold are discarded prior to mutual information estimation, removing noise-dominated components. The aSMI values are then averaged across all eight datasets. The preprocessing pipeline filters low-variance dimensions and subsequently whitens the retained subspace to unit variance per dimension.

Table 3: Aligned sliced mutual information with standardized preprocessing across all 16 models. aSMI values show limited differentiation between well-performing models when computed in the full embedding space without dimensionality filtering.

Backbone	**Accuracy**	**ASMI (standardized)**			**Eff Rank**
		$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 1.0$	
dinov2-vitb14-reg	0.903	8.463	4.622	0.699	245
dinov2-vitb14	0.894	8.467	4.626	0.700	255
clip-ViT-B32	0.819	8.455	4.607	0.696	178
vicreg-resnet50x2	0.777	8.456	4.617	0.690	407
vicreg-resnet50	0.766	8.454	4.612	0.691	273
dino-vitb8	0.746	8.448	4.611	0.695	178
dino-vitb16	0.736	8.457	4.615	0.696	163
timm-vit-base	0.713	8.451	4.613	0.699	183
resnet-resnet50	0.707	8.465	4.623	0.698	158
clip-RN50	0.703	8.438	4.592	0.691	130
timm-vit-small	0.691	8.465	4.625	0.699	247
dino-vits8	0.678	8.448	4.611	0.694	91
dino-vits16	0.631	8.447	4.605	0.696	100
mae-base	0.506	8.453	4.569	0.691	72
simclr-resnet-1x	0.123	7.456	3.664	0.460	7
mocov3-vit-b	0.133	8.295	4.454	0.690	15

Table 4: Aligned sliced mutual information with PCA filtering and whitening across all 16 models. The PCA-whitening pipeline provides consistent estimates by removing noise-dominated dimensions before applying the information-theoretic metric.

Backbone	**Accuracy**	**ASMI-PCA-whiten (ZCA)**			**Eff Rank**
		$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 1.0$	
dinov2-vitb14-reg	0.903	8.463	4.622	0.700	245
dinov2-vitb14	0.894	8.467	4.625	0.700	255
clip-ViT-B32	0.819	8.455	4.607	0.696	178
vicreg-resnet50x2	0.777	8.456	4.615	0.690	407
vicreg-resnet50	0.766	8.453	4.612	0.691	273
dino-vitb8	0.746	8.458	4.616	0.698	178
dino-vitb16	0.736	8.460	4.618	0.697	163
timm-vit-base	0.713	8.451	4.613	0.699	183
resnet-resnet50	0.707	8.465	4.615	0.698	158
clip-RN50	0.703	8.440	4.595	0.691	130
timm-vit-small	0.691	8.460	4.618	0.699	247
dino-vits8	0.678	8.448	4.611	0.694	91
dino-vits16	0.631	8.447	4.609	0.696	100
mae-base	0.506	8.473	4.628	0.699	72
simclr-resnet-1x	0.123	7.839	3.906	0.460	7
mocov3-vit-b	0.133	8.311	4.505	0.690	15