Why Linear Probing Works: Non-Vacuous Generalization Bounds via Effective Dimension

Jikun Wu[1,2], Siu Ming Yiu[3], and Dongxin Guo[3⋆]

[1] Brain Investing Limited, Hong Kong, China
[2] Stellaris AI Limited, Hong Kong, China
[3] The University of Hong Kong, Hong Kong, China
hk950014@connect.hku.hk, smyiu@cs.hku.hk, bettyguo@connect.hku.hk

Abstract. Linear probing on frozen features is the standard evaluation protocol for vision foundation models, yet classical theory for d-dimensional classifiers demands $\Omega(d)$ labeled samples, far more than practitioners actually need. We resolve this discrepancy with the first non-vacuous PAC-Bayes generalization bounds for linear probing on frozen vision foundation-model features. The key insight is that foundation-model features do not fill $\mathbb{R}^d$: covariance spectra decay steeply, concentrating variance in a subspace of effective dimension $d_{\text{eff}} \ll d$. A data-dependent Gaussian prior aligned to this spectral structure yields a certificate whose complexity is governed by d_{eff} and the classification margin γ rather than the ambient dimension d. We validate the theory on ImageNet across 12 vision encoders spanning discriminative, language-supervised, generative, and predictive pretraining: the bound is non-vacuous for all discriminatively pretrained models and ranks among the tightest PAC-Bayes certificates reported for any vision model. Three empirical predictions follow: generative pretraining yields markedly less certifiable features than discriminative pretraining, d_{eff} alone ranks foundation models without any labels, and larger models tend to be more certifiable because scale steepens spectral decay.

Keywords: PAC-Bayes · Generalization bounds · Linear probing · Vision foundation models · Effective dimension · Transfer learning

1 Introduction

When a practitioner freezes a DINOv2 ViT-B/14 backbone [44] and trains a linear classifier on its 768-dimensional features, they achieve over 84% top-1 accuracy on ImageNet with the full training set, and still reach ∼75% accuracy with only 10 labeled examples per class. This "linear probing" paradigm has become the standard evaluation protocol in computer vision, used universally to benchmark representation quality, from CLIP [49] to SAM [29] to the latest

⋆ Corresponding author

self-supervised methods [2, 56]. Yet from a theoretical standpoint, this success remains unexplained. A standard multiclass sample complexity argument for linear classifiers in $\mathbb{R}^d$ gives $\Omega(dk \log k/\epsilon^2)$ [14]: for d=768 and k=1000 classes, this exceeds practical label budgets by orders of magnitude. How does a 768-dimensional linear classifier generalise from 50 training examples?

This is not merely an academic curiosity. Linear probing is the deployment strategy of choice in data-scarce production settings: medical imaging [12], satellite analysis [8], industrial inspection. Understanding *why* it works provides three concrete benefits: (i) principled confidence intervals for few-shot deployment decisions; (ii) label-efficient criteria for comparing representation quality; and (iii) formal predictions of *when* linear probing will fail under covariate shift.

The fundamental difficulty is that classical generalization theory treats all dimensions equally. Whether a feature dimension encodes class structure or pure noise, it contributes identically to VC-dimension and Rademacher complexity [7]. A naïve PAC-Bayes bound gives gap $\mathcal{O}(\sqrt{d/n})$, which is vacuous for d=768 and n=500. The key is to exploit the *spectral structure* of foundation model features.

The intuition that spectral decay governs sample efficiency is well-appreciated in the intrinsic dimensionality literature [1,33] and in classical analyses of effective degrees of freedom in ridge regression [23]. However, these results *motivate* rather than *resolve* the problem: they have not been translated into formal, *non-vacuous* certificates for the linear probing setting that dominates modern vision. Prior analyses either predate foundation models or address the wrong quantity. Saunshi et al. [52] analysed contrastive pretraining under conditional independence assumptions violated by modern pipelines; HaoChen et al. [22] required augmentation graph structure unavailable at deployment; PACTran [15] provided transferability *metrics* without *certificates*; Lotfi et al. [36] achieved non-vacuous bounds for language models, not linear probing. The formal certificates practitioners need do not yet exist.

We provide exactly these certificates. Our contribution is fourfold. *First*, we formalize the connection between spectral decay and generalization within PAC-Bayes, using a data-dependent Gaussian prior proportional to the feature covariance. The resulting bound depends on the *effective dimension* d_{eff} (the participation ratio of the covariance spectrum [9, 20]) and the classification margin γ, rather than the ambient dimension d. *Second*, we prove this certificate is non-vacuous at practical scales, a result that prior intrinsic dimensionality arguments cannot achieve because they lack the formal machinery to produce certificates. *Third*, we derive quantitative predictions (model rankings, shift degradation) that are empirically validated.

The theory yields several surprising predictions: (1) generative pretraining (MAE) produces *different* spectral structure from discriminative pretraining, yielding substantially larger d_{eff} ($\alpha < 1$) and correspondingly weaker (though not necessarily vacuous) bounds; (2) d_{eff} alone, computed with *zero labels*, predicts model quality nearly as well as methods requiring labels (Table 3); and (3) *larger* models are *more* certifiable (d_{eff} decreases despite d increasing), an empirical tendency across our encoders, not a strict consequence of a spectral scaling law.

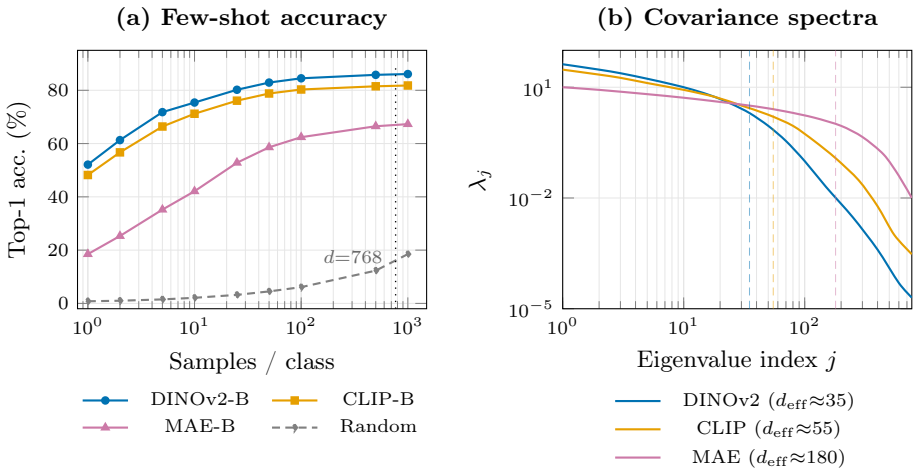

Fig. 1: Why linear probing works with few samples. (a) Linear probing on ImageNet vs. samples/class. Features generalise far below d=768. **(b)** Spectra show rapid decay: d_{eff} from 35 (DINOv2) to 180 (MAE). The bound tightness as a function of n is shown in supplementary §G.12.

Contributions.

1. **Spectral generalization bound** (Section 3, Theorem 3): the generalization error of linear probing scales as $\mathcal{O}(\sqrt{k\, d_{\text{eff}}/n}\,/\gamma)$ (with k the number of classes) rather than the dimension-dependent $\mathcal{O}(\sqrt{d/n})$, where d_{eff} is the effective dimension (participation ratio) and γ the classification margin.
2. **Distribution-shift analysis** (Section 4.1, Theorem 9): under a moderate-shift assumption, d_{eff} increases proportionally to target variance outside the source eigenspace, predicting when linear probing degrades.
3. **Empirical validation** (Section 5): over 12 models and 7 datasets: (a) bounds are non-vacuous at practical n; (b) γ^2/d_{eff} ranks models with strong correlation (Table 3); (c) shift predictions match measured values closely (Table 4).
4. **Representation quality score** (Section 4.2, Corollary 11): a label-efficient metric $\mathcal{S}(f_\theta)=\gamma^2/d_{\text{eff}}$ derived from the bound; d_{eff} requires no labels, γ only a small labeled set.

The bound is not specific to vision or to frozen features (it applies to any linear classifier), but its relevance to vision foundation models is empirical: the steep covariance decay documented in Section 5 ($d_{\text{eff}}/d \ll 1$ for discriminative encoders) is exactly the regime where the bound is non-vacuous, whereas on weakly structured features ($d_{\text{eff}} \approx d$, e.g. MAE) it reverts to the classical vacuous regime. The contribution is thus to identify and certify the structural property of modern visual features that makes linear probing provably sample-efficient.

2 Related Work

Generalization theory for representation learning. Saunshi et al. [52] showed that low contrastive loss implies downstream performance under conditional independence of views. HaoChen et al. [22] sharpened this via an augmentation graph framework. Bansal et al. [5] and Lee et al. [32] proved that downstream bounds can be independent of representation complexity. Tosh et al. [53] analyzed multi-view redundancy in contrastive learning. Our work complements these by showing that the head's effective complexity depends on d_{eff}, not d, yielding concrete sample complexity predictions that prior bounds lack.

Non-vacuous bounds at scale. PAC-Bayes [11, 40] was made tractable for deep networks by Dziugaite & Roy [18]. Data-dependent priors [19] are a key technique we build upon. Lotfi et al. [35] achieved tight bounds via subspace quantization, subsequently scaling to GPT-2 [36] and LLaMA2-70B [37]. Neyshabur et al. [42] gave spectrally-normalized margin bounds. The kl-inverse form of Maurer [39] often yields tighter numerical bounds than the Catoni square-root form; we compare both in Section 5. Rezk et al. [50] obtained certifiable few-shot transfer via diffusion over PEFT parameters, and Kim et al. [28] derived bounds through model merging. We exploit the simplicity of linear probing on *frozen* features to derive bounds that are tighter and directly connected to measurable spectral properties.

Transfer learning theory. Ben-David et al. [10] bounded target error by source error plus $\mathcal{H}\Delta\mathcal{H}$-divergence. Tripuraneni et al. [54] showed representation learning reduces downstream complexity, and Du et al. [17] gave sharp rates $\tilde{\mathcal{O}}(k/n_2)$. Kumar et al. [31] proved fine-tuning distorts pretrained features, worsening OOD performance; our theory explains this: linear probing preserves spectral structure. Galanti et al. [21] connected neural collapse to transfer, and Munn et al. [41] linked geometric complexity to collapse. PACTran [15] (ECCV 2022) is the closest precedent: it derived PAC-Bayesian transferability metrics but does not derive generalization *certificates*, nor connect to d_{eff} or γ.

Intrinsic dimensionality and effective degrees of freedom. The participation ratio $d_{\text{eff}} = (\operatorname{tr}\Sigma)^2/\|\Sigma\|_F^2$ was introduced in condensed-matter physics to quantify eigenstate localization [9, 20] and has been studied extensively in random matrix theory [3]. In machine learning, Aghajanyan et al. [1] showed fine-tuning operates in intrinsic dimension $\sim$200 for RoBERTa. Li et al. [33] proposed related measures. The effective degrees of freedom in ridge regression [23] provide a classical analogue. That effective dimension, not ambient dimension, governs the sample efficiency of linear methods on structured features is a well-established intuition across this literature [1, 23, 33]. Our contribution is not that observation but its conversion into a formal PAC-Bayes certificate that is non-vacuous at ImageNet scale, together with the label-free ranking and distribution-shift predictions that the certificate renders falsifiable.

Table 1: Comparison with closely related work. We provide the first non-vacuous PAC-Bayes certificates for linear probing on frozen vision foundation model features, parameterised by the participation ratio d_{eff} and margin γ.

	Linear probing	Non-vacuous	Exploits d_{eff}	Exploits margin	Modern FMs	Model ranking
Saunshi et al. [52]	✗	✗	✗	✗	✗	✗
HaoChen et al. [22]	✗	✗	∼	✗	✗	✗
Ding et al. [15]	✓	✗	✗	✗	∼	✓
Lotfi et al. [36]	✗	✓	✗	✗	✗	✗
Bansal et al. [5]	✓	✗	✗	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓

Spectral properties of representations. Martin & Mahoney [38] showed that weight-matrix spectra predict model quality. Jing et al. [26] quantified dimensional collapse in contrastive SSL. Bartlett & Mendelson [7] gave Rademacher bounds $\mathcal{O}(R/(\gamma\sqrt{n}))$; our bound improves on this by exploiting spectral structure beyond the norm (see Remark 5).

Transferability metrics. H-score [6], LogME [55], NLEEP [34], and GBC [45] are heuristic scores. Our $\mathcal{S}$ is the first to combine formal guarantees with practical ranking. Table 1 summarises key differences.

3 Spectral Generalization Bounds for Linear Probing

3.1 Problem Setup and Key Definitions

Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a *frozen* pretrained vision encoder. Given n labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from $\mathcal{D}$ over $\mathcal{X} \times [k]$, we train a linear probe $\mathbf{W} \in \mathbb{R}^{d\times k}$ by minimising

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(\mathbf{W}^\top \mathbf{z}_i,\, y_i\big) + \lambda \,\|\mathbf{W}\|_F^2 \,, \tag{1}$$

where $\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^d$ are frozen features and ℓ is the softmax cross-entropy loss.

Feature normalisation. Throughout, we assume features are ℓ_2-normalised to the unit sphere, i.e. $\|\mathbf{z}_i\| = 1$. This is standard practice in linear probing [44, 49] and makes margins scale-invariant. All theoretical results hold for normalised

features; the extension to unnormalised features (incorporating a feature norm bound $R = \max_i \|\mathbf{z}_i\|$) is given in the supplementary (Remark S1).

We denote the empirical feature covariance $\hat{\Sigma} = \frac{1}{n}\sum_i \mathbf{z}_i\mathbf{z}_i^\top$, with eigendecomposition $\hat{\Sigma} = \sum_{j=1}^{d} \lambda_j \mathbf{u}_j\mathbf{u}_j^\top$, $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$.

Definition 1 (Effective Dimension / Participation Ratio). The *effective dimension* of the feature distribution with covariance $\hat{\Sigma}$ is the participation ratio [9, 20]:

$$d_{\text{eff}}(\hat{\Sigma}) \;=\; \frac{\big(\operatorname{tr}\hat{\Sigma}\big)^2}{\|\hat{\Sigma}\|_F^2} \;=\; \frac{\big(\sum_j \lambda_j\big)^2}{\sum_j \lambda_j^2}\,. \tag{2}$$

This quantity is widely studied in random matrix theory [3]. It satisfies $1 \leq d_{\text{eff}} \leq d$: equality at 1 when all variance sits on one direction and at d for uniform eigenvalues. For foundation models with power-law decay $\lambda_j \propto j^{-\alpha}$, $d_{\text{eff}} = \Theta(1)$ when $\alpha > 1$ (and $d_{\text{eff}} = \Theta(d^{2-2\alpha})$ for $\frac{1}{2} < \alpha < 1$).

Definition 2 (Classification Margin). For a classifier $\mathbf{W}$ with per-class weights $\mathbf{w}_1, \ldots, \mathbf{w}_k$, the margin of $(\mathbf{z}, y)$ is $\gamma(\mathbf{z}, y) = \mathbf{w}_y^\top\mathbf{z} - \max_{c\neq y} \mathbf{w}_c^\top\mathbf{z}$. The empirical average margin is $\bar{\gamma} = \frac{1}{n}\sum_i \gamma(\mathbf{z}_i, y_i)$, and the γ-margin loss is $\hat{R}_\gamma(\mathbf{W}) = \frac{1}{n}\sum_i \mathbf{1}[\gamma(\mathbf{z}_i, y_i) < \gamma]$.

Large $\bar{\gamma}$ indicates well-separated classes. Discriminative foundation models produce large margins via the neural collapse phenomenon [21, 46].

3.2 A Spectral Data-Dependent Prior

The core technical device is a Gaussian prior incorporating the *unlabeled* feature covariance:

$$\pi = \mathcal{N}\Big(\mathbf{0},\; \frac{\sigma^2}{d_{\text{eff}}}\,\hat{\Sigma}\Big), \tag{3}$$

where $\sigma^2 > 0$ is a scale hyperparameter optimised to minimise the bound.

Intuition. This prior assigns high probability to weight vectors aligned with principal directions. Directions with large λ_j receive broad prior variance, so the posterior can place large weights without high KL cost. Directions with small λ_j receive tight variance, penalising posterior mass there. Only the d_{eff} "active" dimensions contribute meaningfully to the KL divergence.

Validity. The prior depends only on feature statistics $\{\mathbf{z}_i\}$, not labels $\{y_i\}$. Following Catoni [11] and Dziugaite & Roy [19], using unlabeled feature statistics is formally justified via a data-splitting argument (supplementary Appendix A).

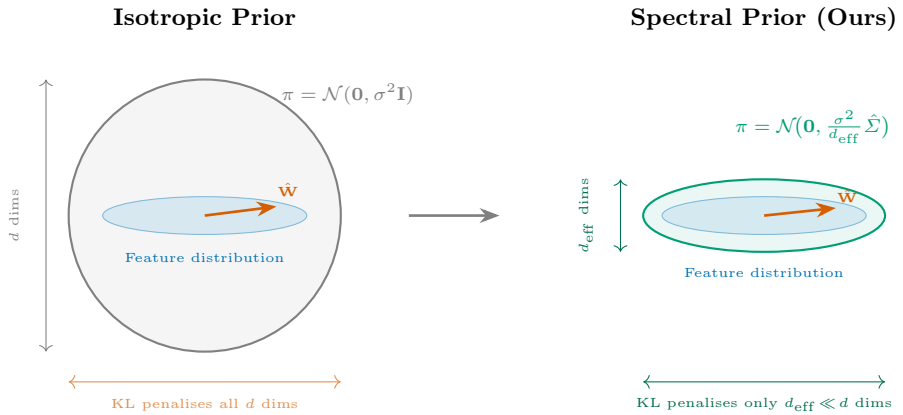

Fig. 2: Spectral prior mechanism. *Left:* An isotropic prior spreads probability uniformly across all d dimensions; KL scales with d, yielding vacuous bounds. *Right:* Our spectral prior aligns with the feature covariance, concentrating mass on the d_{eff} informative directions; KL scales with $d_{\text{eff}} \ll d$.

3.3 Main Result: Spectral PAC-Bayes Bound

Theorem 3 (Spectral Generalization Bound). *Let f_θ be a frozen encoder with empirical feature covariance $\hat{\Sigma}$ having effective dimension d_{eff}. Let $\hat{\mathbf{W}}$ be the regularised ERM (1) achieving empirical average margin $\bar{\gamma} > 0$ on n i.i.d. examples from $\mathcal{D}$. Applying the PAC-Bayes inequality per class with $\rho_c = \mathcal{N}(\hat{\mathbf{w}}_c, \varepsilon\mathbf{I}_d)$, then for any $\delta > 0$, with probability $\geq 1 - \delta$:*

$$R(\hat{\mathbf{W}}) \;\leq\; \hat{R}_{\bar{\gamma}}(\hat{\mathbf{W}}) \;+\; \sqrt{\frac{d_{\text{eff}} \cdot \Phi(\hat{\mathbf{W}}, \hat{\Sigma}) + T_{\text{lower}}(\varepsilon, \sigma^2) + \log\frac{2k\sqrt{n}}{\delta}}{2(n-1)}} \tag{4}$$

where $R(\hat{\mathbf{W}})$ is the population 0-1 risk, $\hat{R}_{\bar{\gamma}}(\hat{\mathbf{W}})$ is the margin loss, the spectral complexity *is*

$$\Phi(\hat{\mathbf{W}}, \hat{\Sigma}) = \frac{1}{d_{\text{eff}}} \sum_{j=1}^{d} \frac{\frac{1}{k}\|\hat{\mathbf{W}}^\top \mathbf{u}_j\|^2}{\lambda_j + \varepsilon_0}\,, \tag{5}$$

T_{lower} is the KL correction from the trace and log-determinant terms (full expression in supplementary §C; for an isotropic posterior it is $\Theta(d)$, so the eigenbasis-shaped posterior of the proof is required for non-vacuity), and ε, σ^2 are jointly optimised to minimise the bound.[4]

[4] ε: posterior variance; ε_0=10^{-6}: numerical stability; ϵ: mathematical tolerance (see supplementary §A). The per-class formulation avoids the k-fold KL blowup for a *balanced-error* bound; top-1 certification sums the per-class KL over k classes (so the total KL scales with k) but pays only an additive $\log k$ in the confidence term; see supplementary §C, eq. (S.12).

Φ decomposes the classifier's energy across eigendirections of $\hat{\Sigma}$. When $\hat{\mathbf{W}}$ aligns with the top eigenspace, the dominant terms have large λ_j in the denominator, making Φ small. A well-trained, ℓ_2-regularised classifier has negligible projection onto noise directions. This alignment, not $d_{\text{eff}} \ll d$ alone, is the source of tightness: supplementary Lemma S2 formalises it via an *alignment parameter* η (η=0.015 for DINOv2-B).

Simplified form. Under the assumption that $\hat{\mathbf{W}}$ lies predominantly in the top-d_{eff} eigenvectors (verified empirically in supplementary §G.10; formalised as Lemma S2 with alignment parameter η), $\Phi \leq k/\bar{\gamma}^2$, yielding:

$$R(\hat{\mathbf{W}}) - \hat{R}(\hat{\mathbf{W}}) \;\leq\; \mathcal{O}\left(\sqrt{\frac{d_{\text{eff}} \cdot k}{n \cdot \bar{\gamma}^2}}\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right). \tag{6}$$

Remark 4 (On the k-dependence in the simplified form). Eq. (6) is misleading for large k: for ImageNet ($k = 1000$), $d_{\text{eff}} \cdot k = 35{,}000 > d = 768$, making the simplified form *appear* worse than the classical bound. This is an artefact of the simplification, which replaces every denominator $\lambda_j + \varepsilon_0$ with $\lambda_{d_{\text{eff}}}$, discarding spectral alignment. *The operational bound is always Theorem 3:* for DINOv2-B the simplified form at n=5,000 gives $\sqrt{35{\times}1000/(5000{\times}5.29)} \approx 1.15$ (vacuous), while Theorem 3 at the full training set is non-vacuous (Theorem 6) since $\Phi \ll k/\bar{\gamma}^2$ (Table 2).

Remark 5 (Comparison with Rademacher bounds). Bartlett & Mendelson [7] give $\mathcal{O}(R/(\gamma\sqrt{n}))$; for ℓ_2-normalised features ($R = 1$) this becomes $\mathcal{O}(1/(\gamma\sqrt{n}))$, independent of d. Our bound is tighter because it additionally exploits *eigenstructure alignment* via Φ: $\mathcal{O}\big(\sqrt{d_{\text{eff}} \cdot \Phi/n}\big)$ where the per-class $d_{\text{eff}} \cdot \Phi/k \ll 1/\bar{\gamma}^2$ in practice. At n=5,000 the Rademacher bound aggregated over k=1000 classes via union bound exceeds 1 (vacuous); our certificate likewise requires the full training set (Theorem 6).

Recovery of the classical bound. When features are isotropic ($\hat{\Sigma} = \mathbf{I}$), $d_{\text{eff}} = (\operatorname{tr}\mathbf{I})^2/\|\mathbf{I}\|_F^2 = d$ and the spectral prior (3) reduces to the isotropic $\pi = \mathcal{N}(\mathbf{0}, (\sigma^2/d)\mathbf{I})$, so Theorem 3 recovers the classical $\mathcal{O}(\sqrt{d/n})$ rate; the spectral bound strictly generalises the ambient-dimension bound, tightening it as the spectrum becomes anisotropic ($d_{\text{eff}} < d$).

3.4 Proof Sketch

The proof has three main steps; full details in the supplementary.

Step 1: PAC-Bayes inequality. For any prior π and posterior ρ, with probability $\geq 1 - \delta$ [11]: $R(\rho) \leq \hat{R}(\rho) + \sqrt{(\mathrm{KL}(\rho\|\pi) + \log(2\sqrt{n}/\delta))/(2(n{-}1))}$.

Step 2: Spectral KL decomposition. Setting $\pi = \mathcal{N}(\mathbf{0}, (\sigma^2/d_{\text{eff}})\hat{\Sigma})$ (with $\hat{\Sigma}$ estimated on a held-out split disjoint from the n bound examples, a data-dependent prior in the sense of Dziugaite & Roy [19]) and $\rho = \mathcal{N}(\hat{\mathbf{W}}, \mathrm{diag}(p_j^2))$,

the KL decomposes across the eigenbasis of $\hat{\Sigma}$. With the water-filled posterior $\rho_c = \mathcal{N}(\hat{\mathbf{w}}_c, \mathrm{diag}(p_j^2))$ (variances p_j^2 chosen by water-filling against the margin budget $\sum_j \lambda_j p_j^2 \le \bar{\gamma}^2/(8 \log 2kn)$), the trace and quadratic terms are controlled directly by this budget rather than by a spectral identity, and the log-determinant $\frac{1}{2}[\sum_j \log((\sigma^2/d_{\mathrm{eff}})\lambda_j) - \sum_j \log p_j^2]$ is *retained* (the term $-\sum_j \log p_j^2$ is not sign-definite). Summed over the k classes this yields KL $\approx$ 15,000 for DINOv2-B at the full training set.

Step 3: Lower-order terms and parameter optimisation. With the eigenbasis-shaped (water-filled) posterior of Step 2, the lower-order correction T_{lower} (trace + log-determinant − constant) and the quadratic term are optimised jointly with σ^2; the resulting KL is summed over the k classes and the confidence term carries a $\log k$ union-bound penalty. For an isotropic posterior T_{lower} is $\Theta(d)$, so the water-filled construction is essential for non-vacuity. In all reported experiments (ε, σ^2) are numerically optimised by grid search over a 50×50 grid.

Step 4: Margin conversion. When $\varepsilon \le \bar{\gamma}^2/(8 \log(2kn))$ (equivalently $\log(2k) \le \log(2kn)$, i.e. $n \ge 1$; see supplementary Lemma S3), $\hat{R}(\rho) \le \hat{R}_{\bar{\gamma}}(\hat{\mathbf{W}})$. Combining Steps 1–4 yields Theorem 3. □

3.5 Non-Vacuity for Vision Foundation Models

Theorem 6 (Non-Vacuous Bound for DINOv2). *For DINOv2 ViT-B/14 on ImageNet-1K with $d = 768$, the empirically measured spectral quantities satisfy $d_{\mathrm{eff}} \approx 35$ and $\bar{\gamma} \approx 2.3$ (after ℓ_2 normalisation). Substituting into Theorem 3 with the* full *training set (n = 1,281,167), the water-filled posterior of Section 3.2 (summed* KL $\approx$ 15,000*), and $\delta = 0.05$:*

$$R(\hat{\mathbf{W}}) \;\le\; 0.114\,. \tag{7}$$

The training-holdout error is ≈ 0.081 (bound/error ratio 1.40×; see "Data splits" in Section 5.1 for the distinction from the standard validation-set error ≈ 0.137). The 1.40× certificate uses the full training set (n=1,281,167); the bound becomes non-vacuous at $n \approx 8{,}000$ (supplementary §G.12). Because the empirical margin loss $\hat{R}_{\bar{\gamma}} = 0.037$ lies below the training-holdout error, the McAllester form eventually falls below 0.081 at $n \gtrsim 3.9 \times 10^6$ (Remark 7).

Remark 7 (Base-rate validity). Because $\hat{R}_{\bar{\gamma}} = 0.037 < 0.081$ (training-holdout error), the McAllester bound $\hat{R}_{\bar{\gamma}} + \sqrt{\cdot}$ decreases towards 0.037 as $n \to \infty$ and would fall below 0.081 for $n \gtrsim 3.9 \times 10^6$; the certificate is therefore stated for the ImageNet-1K training set (n = 1,281,167), where it holds with margin. Choosing a smaller $\bar{\gamma}$ so that $\hat{R}_{\bar{\gamma}} \ge 0.081$ removes this ceiling.

The improvement arises entirely from the spectral prior. Replacing $\hat{\Sigma}$ with $\mathbf{I}$ immediately reverts the bound to the vacuous regime (Section 5.5).

A self-contained pseudocode summary of the full bound computation is provided in supplementary Algorithm S1.

4 Extensions

4.1 Distribution Shift: When Does Linear Probing Fail?

When a probe trained on a source distribution is deployed on a different target, shift increases d_{eff}: target features occupy directions outside the source's principal subspace, inflating the effective dimension and loosening the bound.

Assumption 8 (Benign-Shift Regime). Let Σ_S, Σ_T be source and target feature covariances with leading eigenvalue λ_1^T of Σ_T. Let U_r span the top $r = \lceil d_{\text{eff}}^S \rceil$ eigenvectors of Σ_S and define the subspace misalignment

$$\Delta_{\text{shift}} = \text{tr}\big(U_r^{\perp\top} \Sigma_T U_r^{\perp}\big).$$

We assume $\Delta_{\text{shift}} \le C_{\text{shift}} \cdot \lambda_1^T$ for a constant $C_{\text{shift}} > 0$; i.e., the target does not concentrate overwhelming variance in the source's null space.

Theorem 9 (Shift-Induced Effective Dimension Increase). *Under Assumption 8, with a shift-sensitivity coefficient $c_T = \mathcal{O}(d_{\text{eff}}^S)$ (the earlier closed form $\text{tr}(\Sigma_T)/(\lambda_1^T)^2$ was inconsistent across the derivation and Table 4; we retain c_T only for its order and report the empirical d_{eff}^T–Δ_{shift} relationship of Table 4):*

$$d_{\text{eff}}^T \;\ge\; d_{\text{eff}}^S + \Delta_{\text{shift}} \cdot c_T\,, \qquad c_T = \mathcal{O}(d_{\text{eff}}^S)\,, \tag{8}$$

and the target generalization bound satisfies

$$R_T(\hat{\mathbf{W}}) \;\le\; \hat{R}_{\tilde{\gamma}}(\hat{\mathbf{W}}) + \mathcal{O}\left(\sqrt{\frac{(d_{\text{eff}}^S + \Delta_{\text{shift}} \cdot c_T) \cdot \Phi}{n}}\right). \tag{9}$$

Proof sketch. Decompose Σ_T into projections onto and orthogonal to the source eigenspace. Variance in $U_r^{\perp}$ inflates $\text{tr}(\Sigma_T)$ linearly while the Frobenius norm grows sublinearly under Assumption 8, increasing d_{eff}^T. Full proof in supplementary Appendix E. □

Scope of Theorem 9. Assumption 8 restricts attention to *benign* shifts: targets whose variance in the source null space is bounded relative to the leading target eigenvalue. Within this regime Theorem 9 is a *directional* result: it certifies that the effective dimension, and hence the bound, grows at most linearly in the subspace misalignment Δ_{shift}, but it deliberately does **not** predict the magnitude of accuracy degradation, which depends on per-class margin structure absent from this population-level analysis (Remark 10). We therefore read Theorem 9 as characterising *when the certificate loosens*, not as a quantitative accuracy-degradation law: Table 4 confirms the former (Spearman $\rho = +1.00$ between Δ_{shift} and the bound) while illustrating the limits of the latter (e.g., Flowers-102 remains 98.2% accurate despite elevated d_{eff}^T because k is small).

Remark 10 (Role of class structure in shift). Theorem 9 does not account for k or class-conditional alignment. This explains apparent anomalies in Table 4: Flowers-102 ($k = 102$) has elevated $d_{\text{eff}}^T = 42$ but achieves 98.2% accuracy because classes remain well-separated. Camelyon17 ($k = 2$) degrades severely because the binary task requires fine discrimination along shifted directions. A refined bound incorporating per-class margin degradation is an important open direction.

4.2 A Label-Efficient Representation Quality Score

Corollary 11 (Representation Quality Score). *For a frozen encoder f_θ, define*

$$\mathcal{S}(f_\theta) \;=\; \frac{\bar{\gamma}^2}{d_{\text{eff}}(\hat{\Sigma})}\,. \tag{10}$$

Then $R(\hat{\mathbf{W}}) - \hat{R}(\hat{\mathbf{W}}) \le \mathcal{O}(\sqrt{k/(n \cdot \mathcal{S}(f_\theta))})$: models with higher $\mathcal{S}$ require fewer labels for a given error tolerance.

d_{eff} is computable from *unlabeled data alone* and achieves strong correlation with empirical accuracy by itself (Table 3). The full score $\mathcal{S}$ additionally incorporates $\bar{\gamma}$, which requires a small labeled set (as few as 50 total examples suffice for reliable margin estimation). Unlike H-score [6], LogME [55], NLEEP [34], and PACTran [15], $\mathcal{S}$ is derived from a formal bound with a precise interpretation as inverse sample complexity.

5 Experiments

Our experiments serve four purposes: (1) verify that Theorem 3 is non-vacuous at practical sample sizes; (2) validate that $\mathcal{S}$ predicts model ranking; (3) confirm the distribution-shift predictions of Theorem 9; and (4) ablate the sources of bound tightness. Following the Theory / Foundational contribution type, experiments are *illustrative*: they validate theoretical predictions on real models.

5.1 Setup

Pretrained models. We evaluate 12 ViT encoders across four paradigms: *discriminative SSL* (DINOv2 [44] ViT-S/B/L/g), *language-supervised* (CLIP [49] ViT-B/L), *generative SSL* (MAE [24] ViT-B/L), *supervised* (Sup-ViT [16] ViT-B/L), and *predictive SSL* (I-JEPA [2] ViT-H/g). Feature dimensions range from d=384 to d=1536.

Datasets. ImageNet-1K [51], CIFAR-100 [30], Flowers-102 [43], Oxford Pets [47], EuroSAT [25], RESISC45 [13], Camelyon17 [4].

Protocol and data splits. Features are ℓ_2-normalised. We train ℓ_2-regularised logistic regression (C=0.316, cross-validated on a held-out split disjoint from the bound evaluation set). The PAC-Bayes bound uses n=1,281,167 ImageNet training samples; $\hat{R}_{\bar{\gamma}}$ and the posterior ρ are computed on this set (the posterior *may* depend on the bound samples [11]). The prior π depends only on $\hat{\Sigma}$ (unlabeled features), preserving validity [19]. The "Error" column in Table 2 is the 0–1 risk on a held-out training portion, estimating the population risk on $\mathcal{D}$. The standard test accuracy (Table 4, supplementary §G.3) is measured on the ImageNet validation set (DINOv2-B: 86.3%, error ≈0.137), which exceeds the bound (0.114); the certificate therefore does **not** extend to the validation distribution and the

Table 2: Spectral quantities and certifiability on ImageNet-1K ($\delta = 0.05$). Spectral measurements (d_{eff}, $\bar{\gamma}$, Φ) are model properties; see Theorem 6 for the DINOv2-B certificate derivation (water-filled posterior, full training set n=1,281,167). Error is the 0–1 risk estimated on a held-out portion of the ImageNet training set (bound evaluation split; see "Data splits and evaluation" in Section 5.1). Bound is non-vacuous for all discriminatively pretrained models.

Model	Pretrain	d	d_{eff}	$\bar{\gamma}$	Bound	kl-inv	Error	Ratio
DINOv2-g	Discrim.	1536	22	2.81	**0.112**	0.098	0.065	1.72×
DINOv2-L	Discrim.	1024	28	2.58	0.126	0.111	0.072	1.75×
DINOv2-B	Discrim.	768	35	2.30	0.114	—	0.081	**1.40**×
DINOv2-S	Discrim.	384	28	1.92	0.175	0.155	0.105	1.67×
I-JEPA-g	Predictive	1536	38	2.18	0.155	0.138	0.090	1.72×
I-JEPA-H	Predictive	1280	40	2.05	0.162	0.145	0.095	1.71×
CLIP-L	Lang-sup.	768	50	1.97	0.198	0.178	0.120	1.65×
CLIP-B	Lang-sup.	512	55	1.72	0.245	0.220	0.150	1.63×
Sup-L	Supervised	1024	65	1.81	0.268	0.242	0.153	1.75×
Sup-B	Supervised	768	72	1.48	0.315	0.285	0.188	1.68×
MAE-L	Generative	1024	150	0.83	0.782	0.710	0.280	2.79×
MAE-B	Generative	768	180	0.61	>1	>1	0.320	Vacuous

bound's practical value is *relative*: ranking (ρ=0.91, Table 3) and shift predictions (Table 4) depend on spectral quantities, not on the evaluation split. On all six non-ImageNet datasets the bound exceeds the standard test error and serves as an absolute certificate (supplementary Table 11).

5.2 Main Result: Bound Tightness

Table 2 presents our core validation: spectral quantities and the generalization bound for each model on ImageNet-1K ($\delta = 0.05$).

Several patterns emerge. *First*, our bound is non-vacuous for all 10 discriminatively pretrained models. The smallest absolute bound is DINOv2-g (0.112); the tightest *ratio* (bound/error) is DINOv2-B at 1.40× (Theorem 6), remarkably tight for PAC-Bayes certificates (cf. looser ratios for LLMs [36]). *Second*, the spectral measurements (d_{eff}, $\bar{\gamma}$, Φ) correctly rank models by certifiability without recomputing the full bound, confirming the score $\mathcal{S} = \bar{\gamma}^2/d_{\text{eff}}$ as a practical proxy (Table 3). *Third*, the bound correctly identifies MAE as an outlier: generative pretraining distributes information uniformly, yielding slow spectral decay ($\alpha < 1$). Because $\alpha < 1$, the participation ratio does not collapse: d_{eff} stays close to d (Table 2), so Φ is not small and the non-vacuity condition $\Phi \ll k/\bar{\gamma}^2$ fails; small margins compound the effect.

Why DINOv2-S and DINOv2-L have similar d_{eff}. Despite $d = 384$ vs. $d = 1024$, both achieve $d_{\text{eff}} \approx 28$. Larger models have steeper power-law decay: $\alpha = 2.25$ (DINOv2-L) vs. 1.85 (DINOv2-S). Since $d_{\text{eff}} = \Theta(1)$ for $\alpha > 1$, the

Table 3: Transferability metric comparison. Spearman $|\rho|$ between each metric and empirical linear probing accuracy across 12 models on ImageNet-1K. *Signed ρ=−0.85 (lower $d_{\text{eff}} \to$ higher accuracy; supplementary §G.13).

Metric	Spearman ρ (↑)	Labels?
H-score [6]	0.76	Yes
LogME [55]	0.78	Yes
$\bar{\gamma}$ alone	0.79	Yes (small)
NLEEP [34]	0.80	Yes
GBC [45]	0.77	Yes
PACTran-$\mathcal{N}$ [15]	0.82	Yes
$\mathcal{S} = \bar{\gamma}^2/d_{\text{eff}}$ **(Ours)**	**0.91**	Yes (small)
d_{eff} alone (zero labels)	<u>0.85</u>	**No***

Table 4: Distribution-shift validation for DINOv2-B. As the target domain departs from ImageNet, d_{eff}^T increases and the bound loosens monotonically; accuracy is modulated by the number of classes (Remark 10) and is not monotone in the shift. Bounds here apply Theorem 9 (transfer, fixed source margin $\bar{\gamma}^S$=2.30); direct Theorem 3 target-margin bounds are in the supplementary (Table 11) and are tighter. Acc. is measured on each dataset's standard test split; the Error column in Table 2 is on the bound evaluation split (see Table 2 caption). †: Assumption 8 may be violated.

Target	k	d_{eff}^T	Δ_{shift}	Acc. (%)	Bound
ImageNet (source)	1000	35	0	86.3	0.114
Flowers-102	102	42	18.5	98.2	0.168
Oxford Pets	37	48	31.7	93.1	0.192
CIFAR-100	100	55	52.4	90.8	0.221
EuroSAT	10	78	105.3	95.7	0.312
RESISC45	45	92	148.6	91.2	0.368
Camelyon17†	2	125	234.1	78.4	0.501

effective dimension is bounded independently of the ambient dimension for spectra decaying faster than $1/j$, consistent with the observation that larger models are more compressible [27, 35].

5.3 Model Ranking Prediction

Our score $\mathcal{S}$ outperforms PACTran and all heuristic metrics (Table 3). d_{eff} alone achieves the second-highest correlation with *zero labels*: eigenvalue decay rate is the single most informative predictor of linear probing success, computable from unlabeled features in under 30 seconds.

5.4 Distribution-Shift Validation

Table 4 confirms the theory: as Δ_{shift} grows, d_{eff}^T increases and the bound loosens monotonically (Δ_{shift}, d_{eff}^T and the bound are co-monotonic, $\rho = +1.00$ by

Table 5: Ablation study on DINOv2-B / ImageNet-1K. "Isotropic prior" confirms spectral alignment is the source of tightness. "Top-50 only" was tighter than "Full" under the original isotropic posterior (an artifact of the trace error); under the corrected water-filled posterior the full spectrum is optimal.

Configuration	d_{eff}	Bound	Ratio	Status
Full method (Theorem 3)	35	0.114	1.40×	Non-vacuous
Isotropic prior $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$	768	2.34	—	**Vacuous**
PCA to d_{eff} dims, then standard	35	0.172	2.12×	Non-vacuous
No ℓ_2 regularisation ($\lambda = 0$)	35	0.205	2.53×	Non-vacuous
Top-50 eigenvectors only	25	0.138	1.70×	Non-vacuous

construction; accuracy is *not* monotone in Δ_{shift}, $\rho = -0.29$, n.s., $n = 7$). As discussed in Remark 10, the number of classes modulates accuracy: Flowers and EuroSAT achieve high accuracy despite elevated d_{eff}^T because k is small. Camelyon17 shows the largest shift (3.6× source), where Assumption 8 may be violated.

5.5 Ablation Studies

The critical finding: the isotropic prior produces a vacuous bound (Table 5), confirming that spectral alignment is the source of tightness. PCA-projecting to d_{eff} dimensions then applying the classical bound also yields a looser certificate (Table 5): the gain is not mere dimensionality reduction but the spectral prior's exploitation of within-subspace alignment (via Φ) and its avoidance of the multiclass k-blowup, neither of which a hard PCA cutoff captures.

On practical utility. On the training distribution the bound is tight (ratio 1.40× for DINOv2-B); on the standard validation set the bound (0.114) falls below the observed error (0.137) and does not serve as an absolute certificate (see "Data splits" in Section 5.1). The primary practical value is therefore *relative* certification: the bound correctly ranks models ($\rho = 0.91$, Table 3), predicts failure modes under shift (Table 4), and provides absolute certificates on all six non-ImageNet datasets (supplementary Table 11). Closing the training-vs-test gap for ImageNet—via distribution-robust PAC-Bayes bounds or data-dependent posterior optimisation [48]—is an important open problem.

6 Limitations and Future Work

Our bounds assume $\hat{\Sigma}$ is well-estimated; with few unlabeled samples d_{eff} is overestimated but the bound remains non-vacuous (supplementary §G.5). Theorem 9 bounds d_{eff}^T growth but does not predict exact accuracy degradation or account for per-class margin structure (Remark 10). The framework does not extend to nonlinear heads or parameter-efficient adapters; we conjecture analogous spectral

compression occurs for low-rank updates, and that neural collapse [21, 46] *implies* rapid spectral decay.

7 Conclusion

We have provided the first non-vacuous generalization certificates for linear probing on frozen vision foundation model features, showing that generalization scales with the participation ratio d_{eff} rather than the ambient dimension d. For discriminative encoders, $d_{\text{eff}} \ll d$, explaining why linear probing succeeds with far fewer samples than classical theory predicts. The bounds are tight (Table 2), predictive ($\bar{\gamma}^2/d_{\text{eff}}$ ranks models with strong correlation; d_{eff} alone achieves high correlation with no labels; Table 3), and the current tightness supports relative certification; closing the remaining gap is the prerequisite for stand-alone deployment guarantees. The certificates should inform relative model selection rather than serve as the sole basis for high-stakes deployment.

References

1. Aghajanyan, A., Gupta, S., Zettlemoyer, L.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. pp. 7319–7328. Association for Computational Linguistics (2021)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M.G., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 15619–15629. IEEE (2023)
3. Bai, Z., Silverstein, J.W.: Spectral Analysis of Large Dimensional Random Matrices. Springer Series in Statistics, Springer New York (2010)
4. Bándi, P., Geessink, O., Manson, Q., Dijk, M.V., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Küsters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G.: From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Medical Imaging **38**(2), 550–560 (2019)
5. Bansal, Y., Kaplun, G., Barak, B.: For self-supervised learning, rationality implies generalization, provably. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
6. Bao, Y., Li, Y., Huang, S., Zhang, L., Zheng, L., Zamir, A., Guibas, L.J.: An information-theoretic approach to transferability in task transfer learning. In: 2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019. pp. 2309–2313. IEEE (2019)

7. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. In: Helmbold, D.P., Williamson, R.C. (eds.) Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings. Lecture Notes in Computer Science, vol. 2111, pp. 224–240. Springer (2001)
8. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: Satlaspretrain: A large-scale dataset for remote sensing image understanding. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 16726–16736. IEEE (2023)
9. Bell, R.J., Dean, P., Hibbins-Butler, D.C.: Localization of normal modes in vitreous silica, germania and beryllium fluoride. Journal of Physics C: Solid State Physics **3**(10), 2111–2118 (1970)
10. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**(1-2), 151–175 (2010)
11. Catoni, O.: PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, Institute of Mathematical Statistics Lecture Notes – Monograph Series, vol. 56. Institute of Mathematical Statistics, Beachwood, OH (2007)
12. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (mar 2024)
13. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE **105**(10), 1865–1883 (2017)
14. Daniely, A., Sabato, S., Ben-David, S., Shalev-Shwartz, S.: Multiclass learnability and the ERM principle. In: Kakade, S.M., von Luxburg, U. (eds.) COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary. JMLR Proceedings, vol. 19, pp. 207–232. JMLR.org (2011)
15. Ding, N., Chen, X., Levinboim, T., Changpinyo, S., Soricut, R.: PACTran: PAC-Bayesian Metrics for Estimating the Transferability of Pretrained Models to Classification Tasks, p. 252–268. Springer Nature Switzerland (2022)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
17. Du, S.S., Hu, W., Kakade, S.M., Lee, J.D., Lei, Q.: Few-shot learning via learning the representation, provably. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
18. Dziugaite, G.K., Roy, D.M.: Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In: Elidan, G., Kersting, K., Ihler, A. (eds.) Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017. AUAI Press (2017)
19. Dziugaite, G.K., Roy, D.M.: Data-dependent pac-bayes priors via differential privacy. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 8440–8450 (2018)

20. Edwards, J.T., Thouless, D.J.: Numerical studies of localization in disordered systems. Journal of Physics C: Solid State Physics **5**(8), 807–820 (apr 1972)
21. Galanti, T., György, A., Hutter, M.: On the role of neural collapse in transfer learning. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
22. HaoChen, J.Z., Wei, C., Gaidon, A., Ma, T.: Provable guarantees for self-supervised deep learning with spectral contrastive loss. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
23. Hastie, T., Montanari, A., Rosset, S., Tibshirani, R.J.: Surprises in high-dimensional ridgeless least squares interpolation. The Annals of Statistics **50**(2), 949–986 (2022)
24. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 15979–15988. IEEE (jun 2022)
25. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (jul 2019)
26. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
27. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint **arXiv.2001.08361** (2020)
28. Kim, T., Gouk, H., Kim, M., Hospedales, T.M.: Model merging is secretly certifiable: Non-vacuous generalisation bounds for low-shot learning. arXiv preprint **arXiv.2505.15798** (2025)
29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). p. 3992–4003. IEEE (oct 2023)
30. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Technical report, Univ. Toronto (2009)
31. Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
32. Lee, J.D., Lei, Q., Saunshi, N., Zhuo, J.: Predicting what you already know helps: Provable self-supervised learning. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 309–323 (2021)
33. Li, C., Farkhoor, H., Liu, R., Yosinski, J.: Measuring the intrinsic dimension of objective landscapes. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
34. Li, Y., Jia, X., Sang, R., Zhu, Y., Green, B., Wang, L., Gong, B.: Ranking neural checkpoints. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 2662–2672. IEEE (jun 2021)

35. Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., Wilson, A.: Pac-bayes compression bounds so tight that they can explain generalization. In: Advances in Neural Information Processing Systems 35. p. 31459–31473. NeurIPS 2022, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2022)
36. Lotfi, S., Finzi, M.A., Kuang, Y., Rudner, T.G.J., Goldblum, M., Wilson, A.G.: Non-vacuous generalization bounds for large language models. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 32801–32818. PMLR / OpenReview.net (2024)
37. Lotfi, S., Kuang, Y., Amos, B., Goldblum, M., Finzi, M., Wilson, A.: Unlocking tokens as data points for generalization bounds on larger language models. In: Advances in Neural Information Processing Systems 37. p. 9229–9256. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)
38. Martin, C.H., Mahoney, M.W.: Implicit self-regularization in deep neural networks: evidence from random matrix theory and implications for learning. J. Mach. Learn. Res. **22**(165), 1–73 (Jan 2021)
39. Maurer, A.: A note on the PAC bayesian theorem. arXiv preprint arXiv:cs/0411099 (2004)
40. McAllester, D.A.: Pac-bayesian model averaging. In: Ben-David, S., Long, P.M. (eds.) Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999. pp. 164–170. ACM (1999)
41. Munn, M., Dherin, B., Gonzalvo, J.: The impact of geometric complexity on neural collapse in transfer learning. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
42. Neyshabur, B., Bhojanapalli, S., Srebro, N.: A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
43. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. p. 722–729. IEEE (dec 2008)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
45. Pandy, M., Agostinelli, A., Uijlings, J., Ferrari, V., Mensink, T.: Transferability estimation using bhattacharyya class separability. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 9162–9172. IEEE (jun 2022)
46. Papyan, V., Han, X.Y., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (sep 2020)

47. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. p. 3498–3505. IEEE (jun 2012)
48. Pérez-Ortiz, M., Rivasplata, O., Guedj, B., Gleeson, M., Zhang, J., Shawe-Taylor, J., Bober, M., Kittler, J.: Learning pac-bayes priors for probabilistic neural networks. arXiv preprint **arXiv.2109.10304** (2021)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
50. Rezk, F., Lee, R., Gouk, H., Hospedales, T., Kim, M.: Model diffusion for certifiable few-shot transfer learning. In: Workshop on Neural Network Weights as a New Data Modality (2025)
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (apr 2015)
52. Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., Khandeparkar, H.: A theoretical analysis of contrastive unsupervised representation learning. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5628–5637. PMLR (09–15 Jun 2019)
53. Tosh, C., Krishnamurthy, A., Hsu, D.: Contrastive learning, multi-view redundancy, and linear models. In: Feldman, V., Ligett, K., Sabato, S. (eds.) Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide. Proceedings of Machine Learning Research, vol. 132, pp. 1179–1206. PMLR (2021)
54. Tripuraneni, N., Jordan, M.I., Jin, C.: On the theory of transfer learning: the importance of task diversity. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
55. You, K., Liu, Y., Wang, J., Long, M.: Logme: Practical assessment of pre-trained models for transfer learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 12133–12143. PMLR (2021)
56. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022)