# The 'Ill-Posed' Reconstruction Gap of Masked Autoencoders: Theory and Empirics

Akshay Ghandikota

Advisor: Randall Balestriero
Reader: Chen Sun

# Abstract

Masked Autoencoders (MAE) achieve state-of-the-art pixel reconstruction yet could be outperformed on downstream perception by joint-embedding methods such as I-JEPA and LeJEPA. We study a possible shortcoming, or 'ill-posedness' in the perception of ability of of MAEs, through the lens of a *reconstruction tube volume*: across an ensemble of trained MAEs with different decoders, learning rates, signed supervised mixings $\alpha$, and seeds, how much can downstream linear-probe accuracy vary across models that achieve the same validation reconstruction MSE? We define this tube volume as a finite sum over MSE bins of the empirical accuracy spread, and explain its size by two complementary mechanisms. First, a top/bottom subspace argument adapted from Balestriero et al. [3]: pixel MSE is dominated by the top eigenvalues of the data covariance, while linear-probe accuracy is determined by the between-class scatter, which is anti-aligned with the top of the covariance spectrum on natural images. Second, a decoder-Jacobian rank-deficit argument: the masked decoder Jacobian at a trained MAE has many small singular values, and perturbations of the encoder along the corresponding near-null directions leave the reconstruction loss approximately unchanged but can substantially shift the linear-probe accuracy. We verify the qualitative predictions on a 586-run imagenette sweep across two ViT encoders, five decoder configurations, thirteen signed mixings, and three learning rates, with a partial imagenet-100 extension and a CIFAR-100 reproduction through MAE class. Measured tube volumes range from 100 (linear decoder, ViT-B) to 1395 (ViT-B with the smallest non-trivial decoder), accuracy spans across each MSE bin range from 55 to 65 percentage points, the Spearman correlation between MSE and accuracy is weak across all ten cells ($\rho \in [-0.22, +0.19]$), and a sharp phase transition is observed at $\alpha_{\text{crit}} \approx 5 \times 10^{-5}$. Together these results establish that pixel MSE alone does not pin down semantic content, and trace the residual freedom to the decoder-Jacobian rank deficit at the trained optimum.

# Contents

# Chapter 1

# Background

This chapter sets up four pillars on which the rest of the thesis rests: the masked autoencoder pretext task (Section 1.1), its joint-embedding counterpart (Section 1.2), classical inverse-problem theory and the notion of ill-posedness (Section 1.3), the perception-vs-distortion tradeoff of Blau and Michaeli [4] (Section 1.4), and the central spectral mechanism of this thesis: the top/bottom subspace argument (Section 1.5). It closes with a review of the Ruderman power-law spectrum of natural images (Section 1.6), which feeds into the linear-MAE closed form of Chapter 2.

## 1.1 Masked Autoencoders

A masked autoencoder is a generative pretraining model parameterised by an encoder $\Phi : \mathbb{R}^{d-m} \to \mathbb{R}^k$, a decoder $D_\theta : \mathbb{R}^k \to \mathbb{R}^m$, and a random mask $M \subset \{1, \dots, d\}$ of cardinality $m = |M|$. Given an image $x \in \mathbb{R}^d$, the visible coordinates $x_{\text{obs}} = x_{M^c}$ are tokenised, embedded, and pushed through a Vision Transformer (ViT) encoder to produce the latent token-stream $z = \Phi(x_{\text{obs}})$. At decoding, learnable mask tokens are inserted at the positions in $M$, and a smaller ViT decoder predicts the masked pixels $\hat{x}_M = D_\theta(z)$. The objective is the per-patch mean-squared error

$$\mathcal{L}_{\text{rec}}(\Phi, D_\theta) = \mathbb{E}_{x,M}\left[\, \|D_\theta(\Phi(x_{\text{obs}})) - x_{\text{mask}}\|_2^2 \,\right]. \tag{1.1}$$

Standard hyperparameters are mask ratio 0.75, decoder of 8 blocks at width 512, and patch size 16 at input resolution $224 \times 224$ [6].

The choice of MAE as a baseline is motivated by three properties. First, the encoder–decoder asymmetry , a large encoder paired with a small decoder , forces the encoder to "do most of the work," since the decoder's small capacity cannot fit the inverse on its own. Second, the high mask ratio forces the encoder to integrate information across visible patches, ruling out trivial copy-and-paste solutions. Third, the per-patch MSE objective is computationally cheap and scales to billions of parameters. Despite these properties, MAE underperforms joint-embedding methods on linear-probe top-1 accuracy by 5–10 pp in the imagenette / imagenet-100 / imagenet-1k regime, a phenomenon Balestriero and LeCun [2] call R3.

**The MAE inverse problem.** For a fixed visible context $x_{\text{obs}}$, the data conditional $p(x \mid x_{\text{obs}})$ has support on a typically thick set: many distinct full images are consistent with the visible context. The MSE objective $\mathbb{E}_{x,M}\|D_\theta(z) - x_{\text{mask}}\|^2$ is the trace of the conditional covariance, so it collapses to the conditional mean, not the conditional mode. Hence MAE is biased toward the conditional mean , the classical perception-vs-distortion tradeoff [4] (Section 1.4).

## 1.2 Joint-Embedding Predictive Architectures

A joint-embedding predictive architecture (JEPA) replaces the per-pixel loss Eq. (1.1) with a per-feature loss $\mathbb{E}\|f_\theta(\tau_1(x)) - f_\theta(\tau_2(x))\|^2$ where $\tau_1, \tau_2 \sim \mathcal{T}$ are augmentations. For I-JEPA [1], $\tau_1$ is a context patch and $\tau_2$ is a target patch. In practice, joint-embedding methods avoid the high entropy of pixel space and converge to a representation whose top-$K$ subspace is aligned with the dominant data directions.

Balestriero et al. [3] prove that in the linear regime JEPA's optimum is uniquely the top-$K$ eigenspace of $\Sigma_x$, while MAE's optimum has a continuous family of equivalent $(E, D)$ solutions: any invertible reparameterisation of the bottleneck preserves the reconstruction loss. The thesis is organised around the empirical and statistical consequences of this: tube volume, tube geometry under decoder ablation, and the eigen-crossing phase transition.

## 1.3 Inverse Problems and Ill-Posedness

Hadamard's three conditions for well-posedness state that the inverse problem "recover $x$ from $y = Tx$" is *well-posed* if and only if (i) a solution $x$ exists, (ii) the solution is unique, and (iii) the solution depends continuously on the data $y$. If any of these fails the problem is *ill-posed* [5].

For MAE, the forward problem is the masking operator $T_M : x \mapsto x_{M^c}$, which projects onto the visible coordinates. The inverse $T_M^{-1} : x_{\text{obs}} \mapsto x$ fails Hadamard's second condition: the set of $x$ consistent with a typical $x_{\text{obs}}$ is a manifold of dimension $\sim m$, not a point. Standard remedies are Tikhonov regularization [9], which trades exactness for stable inversion via a quadratic prior, and Bayesian inversion, which integrates over the solution set against a data prior $p(x)$.

The MAE objective Eq. (1.1) is, in this language, an *empirical Bayes posterior-mean* estimator: the decoder learns to output $\mathbb{E}[x_M \mid x_{\text{obs}}]$ rather than any specific $x_M$ in the solution set. This biases the encoder to capture conditional-mean-relevant features and does *not* pin down features that vary across the solution set but average to the same conditional mean. The downstream linear probe in $K$-space then sees a representation that is underspecified along these directions , the empirical source of the R3 phenomenon.

## 1.4 The Reconstruction–Perception Tradeoff

Blau and Michaeli [4] formalize a tradeoff between the distortion $D(\hat{x}, x) = \mathbb{E}\|\hat{x} - x\|^2$ and the

perceptual quality measured by a divergence $d(p(\hat{x}), p(x))$. They prove that the Pareto frontier of $(D, d)$ has a non-trivial slope: pushing distortion to zero forces $d$ away from zero. The intuition is that a perfect MSE optimizer collapses to the conditional mean, which is in general not on the data manifold. This explains the empirical observation that MAE-decoded images are blurry relative to natural images.

For our purposes the perception-distortion tradeoff translates into a sharper statement: in the small-$\varepsilon$ regime, the spread of downstream accuracy across MAEs that achieve the same MSE $\varepsilon$ is driven by the directions *not* pinned down by pixel reconstruction , the bottom of the data covariance spectrum. We make this concrete next.

## 1.5 Top/Bottom Subspace Variance Argument

Let $\Sigma_x = \mathbb{E}[xx^\top]$ be the data covariance, with eigendecomposition $\Sigma_x = \sum_k \sigma_k^2 v_k v_k^\top$ and ordered eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_d^2 \geq 0$. We distinguish the *top* $K$ eigenvectors (largest eigenvalues, which explain the most pixel variance) from the *bottom* of the spectrum (smallest eigenvalues, the directions that carry the least pixel variance).

**MSE is dominated by the top of the spectrum.** For a linear encoder $V \in \mathbb{R}^{K \times d}$, the optimal linear decoder is $V^+$, and the reconstruction error decomposes as

$$\mathbb{E}_x \|V^+ V x - x\|_2^2 \;=\; \sum_{k=1}^{d} \sigma_k^2 \left(1 - \alpha_k^2\right), \tag{1.2}$$

where $\alpha_k \in [0, 1]$ is the cosine between the linear span of $V$ and the $k$-th eigenvector $v_k$. Each term is weighted by the data eigenvalue $\sigma_k^2$, so a model that fully captures the top of the spectrum drives the MSE close to its minimum value $\sum_{k>K} \sigma_k^2$. The MSE objective is therefore dominated by the top.

**Perception lives in the bottom of the spectrum.** Balestriero et al. [3] report a striking empirical finding on tinyimagenet, CIFAR-10/100, ImageNet, and other datasets: classification accuracy of a downstream classifier trained on *only the top-$k$* projection of the data is much *lower* than accuracy on the bottom-$k$ projection (their Figure 1). Equivalently, the between-class scatter $S_B = \sum_c p_c(\mu_c - \mu)(\mu_c - \mu)^\top$ that controls linear-probe accuracy is anti-aligned with $\Sigma_x$: $S_B$ has its largest eigenvalues in directions where $\Sigma_x$ has its *smallest* eigenvalues, because object-discriminative features are fine-grained and pixel-low-variance whereas pixel variance is dominated by background and illumination. This anti-alignment is observable across all natural-image datasets they tested.

**Why MAE optimizes the wrong half of the spectrum.** Combine the two observations. A vanilla MAE driven by per-pixel MSE will push its encoder to fit the directions that maximally reduce Eq. (1.2), which are the top directions of $\Sigma_x$. By the previous paragraph, those are exactly

the directions least useful for linear classification. Until the top is essentially fit, the encoder has no incentive to start fitting the bottom , and the reconstruction loss converges long before the bottom is fit, since the power-law decay of $\sigma_k^2$ (Section 1.6) gives the bottom directions exponentially smaller weight in Eq. (1.2) than the top. Hence the MAE's linear-probe accuracy can vary widely across two trained models that both achieve the same MSE: the MSE only constrains the top, while accuracy is determined by the bottom.

This is the central spectral mechanism explaining why MSE does not predict perception. We will use it to derive a closed-form linear-MAE expression for the tube width in Section 2.3 and to motivate a decoder-Jacobian rank-deficit argument for nonlinear ViT decoders in Section 2.4.

## 1.6 Power Laws of Natural Images

Natural images obey a $1/f^\alpha$ power spectrum [8, 10]. Equivalently, the eigenvalues of the pixel covariance $\Sigma_x = \mathbb{E}[xx^\top]$ decay as

$$\sigma_k^2 \;\sim\; \sigma_1^2 \cdot k^{-\alpha_R}, \qquad \alpha_R \in [1.2,\, 2.2], \tag{1.3}$$

with $\alpha_R \approx 1.6$–$2.0$ for typical photographic data. We write $\alpha_R$ for the Ruderman exponent to distinguish it from the signed supervised mixing $\alpha$ that we sweep empirically.

A complementary assumption [7] concerns the *label energy*: the squared projection $\pi_k = \|Y P_{XX^\top,k}\|^2$ of the label matrix $Y$ onto the $k$-th eigenvector of $\Sigma_x$ (in decreasing eigenvalue order) follows

$$\pi_k \;=\; \frac{T_\infty(\beta+1)}{D}\Big(\frac{k}{D}\Big)^{\beta}, \qquad \beta \geq 0, \tag{1.4}$$

where $T_\infty$ is a normalisation. The case $\beta = 0$ is alignment-uniform; larger $\beta$ corresponds to anti-alignment of label energy with pixel variance. Balestriero and LeCun [2] estimate $\beta \approx 0.2$ for tinyimagenet.

The combination of Eqs. (1.3) and (1.4) together with the top/bottom subspace argument of Section 1.5 is the input to the linear-MAE closed form derived in Chapter 2.

## 1.7 Why This Matters for SSL Practice

The combination of the spectral-mismatch mechanism (Section 1.5) and the perception–distortion tradeoff (Section 1.4) gives a clean, falsifiable prediction for SSL practice: *across MAE training runs that achieve the same validation MSE, downstream linear-probe accuracy can vary by tens of percentage points.* This is the operational meaning of the R3 phenomenon [2], and it is what we quantify in Chapter 4. The remaining chapters of this thesis make this prediction precise (Chapter 2), describe the experimental sweep we use to test it (Chapter 3), and report the empirical findings (Chapter 4).

# Chapter 2

# Theory of the Reconstruction Tube

This chapter develops the theory in three parts. Section 2.2 defines the reconstruction spread function spread($\varepsilon$) and tube volume Vol as finite-sample objects on a trained ensemble. Section 2.3 gives the linear-MAE closed form via the top/bottom subspace decomposition of Section 1.5. Section 2.4 states the decoder-Jacobian rank-deficit proposition for the nonlinear ViT case. We close in Section 2.5 with a brief phase-transition observation linking the supervised-mixing axis to linear-probe accuracy.

## 2.1 Setup and Notation

Let $x \in \mathbb{R}^d$ denote an image and $M \subset \{1, \dots, d\}$ a mask of cardinality $m = |M|$. We write $x_{\text{obs}} = x_{M^c}$ for the visible coordinates and $x_{\text{mask}} = x_M$ for the masked coordinates. The encoder $\Phi : \mathbb{R}^{d-m} \to \mathbb{R}^k$ produces a latent token-stream $z = \Phi(x_{\text{obs}})$, the decoder $D_\theta : \mathbb{R}^k \to \mathbb{R}^m$ predicts $\hat{x}_M$, and the per-patch MSE objective is

$$\mathcal{L}_{\text{rec}}(\Phi, D_\theta) = \mathbb{E}_{x,M}\left[\, \|D_\theta(\Phi(x_{\text{obs}})) - x_{\text{mask}}\|_2^2 \,\right]. \tag{2.1}$$

We adopt the convention from Balestriero and LeCun [2] of writing $\theta = (\theta_{\text{enc}}, \phi_{\text{dec}})$ for the joint MAE parameter and $L_{\text{rec}}(\theta)$ for the validation reconstruction MSE. Let $A(\theta) \in [0, 1]$ denote the corresponding linear-probe top-1 accuracy on a held-out validation set.

**Assumption 2.1** (Smooth data manifold)**.** The data distribution $p(x)$ is supported on a smooth manifold $\mathcal{M} \subset \mathbb{R}^d$. The pixel covariance $\Sigma_x = \mathbb{E}[xx^\top]$ has eigenvalues $\sigma_k^2$ obeying the Ruderman power law $\sigma_k^2 \sim \sigma_1^2 k^{-\alpha_R}$ with $\alpha_R \in [1.2, 2.2]$.

**Assumption 2.2** (Anti-aligned label energy)**.** The label-energy projection $\pi_k = \|Y P_{XX^\top, k}\|^2$ obeys $\pi_k = T_\infty(\beta + 1)D^{-1}(k/D)^\beta$ for some $\beta \geq 0$.

**Assumption 2.3** (Lipschitz decoder)**.** The decoder $D_\theta$ is twice differentiable with masked Jacobian $J(z) = \partial(P_m D_\theta(z))/\partial z$ Lipschitz in $z$ with constant $L_J$ in a ball of radius $\rho$.

## 2.2 Tube Volume and Spread: Definitions

Throughout the empirical chapter we will train an ensemble of MAE models indexed by $\theta =$ (decoder, lr, signed mixing $\alpha$, seed, encoder, epoch). We work directly with this finite ensemble.

**Definition 2.4** (Spread and reconstruction tube volume)**.** Fix a binning of the validation MSE axis into bins $B_\varepsilon = [\varepsilon, \varepsilon + \Delta\varepsilon)$. For each bin $B_\varepsilon$, let

$$\text{spread}(\varepsilon) \;=\; \max_{\theta \in B_\varepsilon} A(\theta) \;-\; \min_{\theta \in B_\varepsilon} A(\theta), \tag{2.2}$$

where the max/min run over the ensemble members $\theta$ whose validation MSE falls in $B_\varepsilon$. The *reconstruction tube volume* is

$$\text{Vol} \;=\; \sum_{\varepsilon \in \text{bins}} \text{spread}(\varepsilon)\, \Delta\varepsilon. \tag{2.3}$$

Eqs. (2.2) and (2.3) are entirely empirical quantities: a finite maximum minus a finite minimum, and a Riemann sum. Intuitively, spread($\varepsilon$) is "how much can downstream accuracy vary across MAEs whose validation MSE all lie in the same bin." A larger Vol means MSE alone is an unreliable predictor of downstream quality.

*Remark* 2.5 (Bin width and units). The bin width $\Delta\varepsilon$ is a hyperparameter of the estimator. We use linearly-spaced MSE bins of width $\Delta\varepsilon = 10^{-2}$ throughout. The accuracy is reported in percentage points. Tube volumes in this chapter are therefore in units of (pp $\cdot$ MSE), with bins normalised to keep the magnitude comparable across cells with different MSE ranges. Headline volumes range over 100–1395 in these units; only relative comparisons across cells should be read off these numbers.

*Remark* 2.6 (Interpretation as a finite-sample variance proxy). The integrand in Eq. (2.3) is the max-min spread of $A$ over a binned slice of the ensemble. This is the empirical $L^\infty$ analogue of the conditional standard deviation $\sqrt{\text{Var}[A \mid L_{\text{rec}} = \varepsilon]}$. Since the ensemble is finite, max-min is more stable than the empirical standard deviation when the bin contains few points.

We take Eqs. (2.2) and (2.3) as the operational definitions of "tube width" and "tube volume." The remainder of this chapter asks: what theoretical mechanisms govern the size of Vol?

## 2.3 Linear-MAE Closed Form

We instantiate Theorem 2.4 in the linear-MAE regime, where Balestriero and LeCun [2] give a complete closed-form analysis of the joint (encoder, decoder, supervised) optimum and where the top/bottom subspace mechanism of Section 1.5 can be made exact.

**Setup.** Let $V \in \mathbb{R}^{K \times d}$ be a linear encoder and $D \in \mathbb{R}^{d \times K}$ a linear decoder. The unmasked reconstruction loss is $\mathcal{L}_{\text{rec}}(V, D) = \mathbb{E}_x \|DVx - x\|_2^2$. Let $\Sigma_x = \sum_k \sigma_k^2 v_k v_k^\top$ be the eigendecomposition of the data covariance, with $\sigma_1^2 \geq \cdots \geq \sigma_d^2$. Let $S_B$ denote the between-class scatter, whose top eigenvectors govern linear-probe accuracy.

**Theorem 2.7** (Linear-MAE top/bottom decomposition)**.** *At the global optimum of $\mathcal{L}_{\text{rec}}$, the encoder $V$ spans the top-$K$ eigenspace of $\Sigma_x$. The reconstruction MSE decomposes as*

$$\mathcal{L}_{\text{rec}}(V, D) \;=\; \sum_{k=1}^{d} \sigma_k^2\,(1-\alpha_k^2), \qquad \alpha_k \;=\; \|P_{\text{row(V)}} v_k\|_2, \tag{2.4}$$

*where $\alpha_k \in [0,1]$ is the cosine between the row-span of $V$ and the $k$-th data eigenvector. Linear-probe accuracy at fixed representation depends only on the alignment between* row$(V)$ *and the top eigenvectors of $S_B$. Since $S_B$ is anti-aligned with $\Sigma_x$ on natural images [3], encoders that minimise Eq. (2.4) can have very different linear-probe accuracies, with the per-bin spread bounded as*

$$\text{spread}(\varepsilon) \;\lesssim\; \frac{1}{T_\infty} \sum_{k \in \mathcal{N}(\varepsilon)} \pi_k, \tag{2.5}$$

*where $\mathcal{N}(\varepsilon) \subset \{1, \dots, d\}$ is the index set of eigenvectors whose individual term in Eq. (2.4) is below the bin tolerance $\Delta\varepsilon$ (the "almost free" directions that the MSE does not constrain at level $\varepsilon$), $T_\infty$ is the total label energy, and $\pi_k$ is the per-mode label energy of Eq. (1.4).*

*Proof sketch.* The decomposition Eq. (2.4) is standard PCA algebra. At MSE level $\varepsilon$ slightly above the optimum the encoder can spend its slack by perturbing $\alpha_k$ on indices $k \in \mathcal{N}(\varepsilon)$ where $\sigma_k^2(1-\alpha_k^2)$ is small. Linear-probe accuracy is $A(V) = T_\infty^{-1}\langle P_{\text{row(V)}}, S_B\rangle_F + \text{const}$, so the spread over $V$ within the bin is bounded by the label energy contained in $\mathcal{N}(\varepsilon)$, giving Eq. (2.5). Full proof in **??**. □

Theorem 2.7 makes the central mechanism quantitative: the spread at MSE $\varepsilon$ is controlled by the label energy $\pi_k$ contained within the index set $\mathcal{N}(\varepsilon)$ of "MSE-free" directions. Under the Ruderman power law Eq. (1.3) and label-energy power law Eq. (1.4), $\mathcal{N}(\varepsilon)$ is dominated by the bottom of the spectrum, where $\pi_k$ is largest, so spread$(\varepsilon)$ is large. In particular, the linear-MAE spread does not vanish as $\varepsilon \to 0^+$ unless $K = d$, which is the linear-regime version of the R3 phenomenon [2].

## 2.4 Decoder Jacobian Rank Deficit (Nonlinear Case)

For nonlinear ViT decoders the linear-MAE argument is no longer exact because the decoder is no longer a single matrix. We instead state a local first-order argument that quantifies how many "flat directions" the decoder offers around a trained MAE optimum.

**Setup.** Let $g_\phi : \mathbb{R}^k \to \mathbb{R}^d$ be the (twice differentiable) ViT decoder and let $J(z) = \partial(P_m g_\phi(z))/\partial z \in \mathbb{R}^{d_m \times k}$ be its Jacobian on the masked output coordinates at latent point $z$. Order singular values $\sigma_1(J) \geq \sigma_2(J) \geq \cdots \geq \sigma_k(J) \geq 0$.

**Definition 2.8** (Decoder Jacobian rank deficit)**.** For threshold $\tau > 0$, the rank deficit at threshold $\tau$ is

$$k_\tau(J) \;=\; |\{\, i \in \{1, \dots, k\} : \sigma_i(J) < \tau \,\}|. \tag{2.6}$$

The corresponding near-null subspace $\mathcal{N}_\tau(z) \subset \mathbb{R}^k$ is the span of the right singular vectors of $J(z)$ with $\sigma_i(J) < \tau$.

**Proposition 2.9** (Decoder Jacobian rank deficit gives tube width)**.** *Suppose the masked decoder Jacobian at a trained MAE has rank deficit $k_\tau(J) \geq 1$ at threshold $\tau$. Then for any unit vector $u$ in the near-null subspace $\mathcal{N}_\tau(z)$ and any perturbation $\delta z = r\,u$ with $\|r\| \leq \rho$ (where $\rho$ is the radius inside which $J$ is approximately constant):*

1. *the masked reconstruction loss changes by at most $O(\tau\,\rho + \rho^2)$, and*
2. *if the downstream score $s(z)$ has gradient with non-trivial projection onto $\mathcal{N}_\tau$, the linear-probe accuracy can shift by $\sim r \cdot \|\Pi_{\mathcal{N}_\tau} \nabla s\|$.*

*Aggregating over the validation distribution under the generic-orientation assumption that $\nabla s$ is on average uniformly oriented in latent space, the contribution of the near-null subspace to the tube width scales as*

$$\text{spread}(\varepsilon)\big|_{\text{Jacobian contribution}} \;\propto\; \rho(\varepsilon)\,\sqrt{k_\tau(J)/k}, \tag{2.7}$$

*where $\rho(\varepsilon)$ is the perturbation radius at which the masked reconstruction MSE rises by $\varepsilon$.*

*Proof sketch.* A first-order Taylor expansion of $g_\phi$ around $z$ together with the Lipschitz property of the Jacobian gives the recon-MSE change in (1). For (2), a first-order Taylor of $s$ along $\delta z = ru$ gives the score change $r \cdot \langle u, \nabla s \rangle$, which is $O(r)$ unless $u$ is exactly orthogonal to $\nabla s$. Choosing $\pm u$ aligned with $\Pi_{\mathcal{N}_\tau} \nabla s$ gives the maximum possible swing. The aggregation factor $\sqrt{k_\tau/k}$ is the expected length of the projection of a uniform random gradient onto a random $k_\tau$-dimensional subspace. Full proof in **??**. □

Theorem 2.9 ties tube width directly to a measurable architectural quantity: the number of small singular values of the masked decoder Jacobian. This number is expected to grow with both decoder *depth* (more layers, more chances for redundant directions) and decoder *width* (larger latent / embedding dim, more capacity to absorb redundancy). We test both in Section 4.2.

**Comparison with the linear case.** In the linear regime, Theorem 2.7 fully captures the spread because the decoder Jacobian is a constant matrix of rank $\min(K, d_m)$ and has no extra near-null directions beyond the bottleneck $K$. In the nonlinear regime, by contrast, the decoder adds further near-null directions to the Jacobian on top of those imposed by the bottleneck, and Theorem 2.9 predicts that this extra rank deficit translates to extra tube width. The empirical decoder ablation in Section 4.2 shows exactly this: the tube volume is an order of magnitude smaller for the linear decoder (Vol $\approx$ 100–132) than for any deep ViT decoder (Vol $\approx$ 710–1395).

## 2.5 Phase Transition in the Supervised Mixing

The supervised loss $\mathcal{L}_\alpha = \mathcal{L}_{\text{rec}} + \alpha\,\mathcal{L}_{\text{sup}}$ adds a between-class scatter pull on the encoder. In the linear surrogate of Theorem 2.7, with $L_{\text{rec}}$ corresponding to $-\text{Tr}(U^\top \Sigma_x U)$ on $K$-dim subspace $U$

and $L_{\text{sup}}$ corresponding to $-\text{Tr}(U^\top S_B U)$, the joint problem becomes

$$\min_{U:U^\top U=I_K} \text{Tr}\big(U^\top(\Sigma_x - \alpha S_B)\,U\big), \tag{2.8}$$

solved by the bottom-$K$ eigenvectors of $\Sigma_x - \alpha S_B$.

*Remark* 2.10 (Eigen-crossing observation). By standard analytic perturbation theory of symmetric pencils, the eigenvalues of $\Sigma_x - \alpha S_B$ are continuous in $\alpha$. Crossings between the $K$-th and $(K+1)$-th eigenvalues occur on a discrete set $\Lambda^*$, and the optimum subspace switches discontinuously at any $\alpha^* \in \Lambda^*$. In the linear surrogate this gives a sharp jump in any subspace-dependent functional of the optimum, including the linear-probe accuracy.

Theorem 2.10 makes a single empirical prediction: as the supervised mixing $\alpha$ sweeps from negative to positive, linear-probe accuracy should jump discontinuously near the eigen-crossing. We report this jump empirically at $\alpha_{\text{crit}} \approx 5\times10^{-5}$ across all ten cells in Section 4.4.

**Summary of empirical predictions.** Theorems 2.7, 2.9 and 2.10 together imply three qualitative empirical patterns we test in Chapter 4:

1. Tube volume is much larger for nonlinear decoders than for the linear decoder (the additional rank deficit from Theorem 2.9).

2. Tube volume is approximately invariant to encoder size at fixed decoder, since neither Theorem 2.7 nor Theorem 2.9 introduces a direct encoder-width dependence at matched bottleneck and matched decoder Jacobian.

3. Linear-probe accuracy as a function of signed supervised mixing $\alpha$ has a sharp transition near $\alpha = 0$ (Theorem 2.10).

# Chapter 3

# Experimental Setup

This chapter describes the data, models, training pipeline, and software infrastructure used to generate the empirical results in Chapter 4. The empirical sweeps are organised in two tiers: a 586-run imagenette scaling sweep that forms the basis for all headline numbers in Chapter 4, and a partial imagenet-100 extension that confirms the headline tube and phase-transition geometry on a larger dataset (Section 3.4). Throughout this chapter the imagenette and cifar-100 runs use the implementation (); the imagenet-100 runs use a pipeline.

## 3.1 Datasets

We use three datasets, all hosted on the Hugging Face Hub:

- **imagenette**: a 10-class subset of ImageNet-1k consisting of 9,469 training images and 3,925 validation images at 320-pixel resolution, distributed via `imagenette/320px`. This is our primary sweep dataset because it is small enough to iterate quickly across the full $(\alpha, \mathrm{lr})$ grid.
- **imagenet-100**: a 100-class subset with $\sim$ 117k training images, distributed via `ilee0022/ImageNet100`. Used as a partial scaling extension; we report tube-volume and phase-transition statistics on this dataset but no fine-grained per-cell training curves.
- **cifar-100**: `cifar100` (50k train, 10k val, $32 \times 32$). Used as a smaller-scale sanity check on the wrapper (**??**).

All datasets are loaded with the standard ImageNet-style preprocessing: random resized crop to $224 \times 224$ at training time, center crop at evaluation, mean–std normalization with the ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. CIFAR-100 is bilinearly upsampled to $224 \times 224$ before normalization.

## 3.2 Architectures

**Encoders.** We use three Vision Transformer variants:

- **ViT-Tiny**: width 192, depth 12, heads 3, patch size 16 ($\sim$ 5.7M params encoder-only).
- **ViT-Small**: width 384, depth 12, heads 6, patch size 16 (`enc384d12h6p16`; $\sim$ 22M params).
- **ViT-Base**: width 768, depth 12, heads 12, patch size 16 (`enc768d12h12p16`; $\sim$ 86M params).

ViT-Small and ViT-Base are the two encoders used in the primary imagenette scaling sweep.

**Decoders.** Five decoder configurations spanning the "minimal" (`linear`) to "maximal" (`small-d4`) end of the spectrum:

- `linear`: a single linear projection from latent $z \in \mathbb{R}^K$ to pixel space; $K = 512$, depth 0.
- `tiny-d2`: ViT decoder with width 192, depth 2, heads 3.
- `tiny-d4`: ViT decoder with width 192, depth 4, heads 3.
- `tiny-d8`: ViT decoder with width 192, depth 8, heads 3.
- `small-d4`: ViT decoder with width 384, depth 4, heads 6.

The decoder family is the key axis of the empirical sweep: Theorem 2.9 predicts that tube volume grows with the decoder-Jacobian rank deficit, which in turn grows with decoder depth and width.

**MAE wrapper.** We use the `MaskedAutoencoderViT` class implementation, which performs the standard MAE forward pass: tokenization, encoder forward over visible patches, mask-token insertion at decode time, decoder forward, and per-patch MSE on masked patches only. The imagenette and cifar-100 sweeps use this implementation as their training entry point.

## 3.3 Training and Hyperparameters

**Training axes (imagenette scaling sweep).** The imagenette sweep uses:

- **epochs**: 200.
- **batch size**: 256.
- **optimizer**: Adam, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05.
- **learning rate**: three values $\{10^{-5}, 10^{-4}, 10^{-3}\}$ on a cosine schedule with linear warmup over the first 5% of training.
- **mask ratio**: 0.75.

- **supervised mixing** $\alpha$: thirteen signed values $\{-1, -0.5, -0.1, -0.01, -10^{-3}, -10^{-4}, 0, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.$ The total joint loss is $\mathcal{L}_\alpha = \mathcal{L}_{\text{rec}} + \alpha\,\mathcal{L}_{\text{sup}}$. Negative $\alpha$ corresponds to *anti*-supervised guidance (deliberately push the encoder *away* from labels) and is used to characterise the full phase transition.
- **seed**: 42 for the headline grid; a smaller multi-seed block ($\{43, 44\}$) extends the central (ViT-S, `tiny-d4`) cell.

**Loss specification.** The supervised loss $\mathcal{L}_{\text{sup}}$ is the cross-entropy of a linear probe trained jointly with the encoder, evaluated on the labelled imagenette training set. The $\alpha$-*axis* is the central knob for testing the phase-transition observation of Theorem 2.10: as $\alpha$ sweeps from $-1$ through 0 to $+1$, the optimum eigen-subspace of $\Sigma_x - \alpha S_B$ should jump at the eigen-crossing.

## 3.4 Compute and Streaming Infrastructure

**Computational support.** This work was supported by Brown OSCAR research compute over the course of the project.

**Imagenette scaling sweep.** The imagenette scaling sweep contains 586 runs after de-duplication, broken down by cell as $\{96, 68, 80, 54, 48\}$ for the five ViT-Small cells and $\{54, 52, 44, 45, 45\}$ for the five ViT-Base cells (see Table 3.1).

**Imagenet-100 partial extension.** The imagenet-100 extension (Table 3.2) is a $5 \times 7 \times 2 \times 2 = 140$ cell grid (decoder type $\times$ $\alpha$ $\times$ lr $\times$ mask ratio) at ViT-Small with 30 epochs each. We report only the end-of-training tube-volume and phase-transition statistics.

| Encoder | Decoder | $K$ | runs | $\alpha$ levels | lr levels |
|---|---|---|---|---|---|
| ViT-S | linear | 512 | 96 | 13 | 3 |
| ViT-S | tiny-d2 | 192 | 68 | 13 | 3 |
| ViT-S | tiny-d4 | 192 | 80 | 12 | 3 |
| ViT-S | tiny-d8 | 192 | 54 | 12 | 3 |
| ViT-S | small-d4 | 384 | 48 | 11 | 3 |
| ViT-B | linear | 512 | 54 | 10 | 3 |
| ViT-B | tiny-d2 | 192 | 52 | 10 | 3 |
| ViT-B | tiny-d4 | 192 | 44 | 9 | 3 |
| ViT-B | tiny-d8 | 192 | 45 | 10 | 3 |
| ViT-B | small-d4 | 384 | 45 | 11 | 3 |
| Total | | | 586 | , | , |

Table 3.1: Per-cell run counts for the imagenette scaling sweep.

| Axis | Values | # settings |
|---|---|---|
| Encoder | ViT-Small (single) | 1 |
| Decoder | linear / tiny-d2 / tiny-d4 / tiny-d8 / small-d4 | 5 |
| $\alpha$ | $\{-1, -0.1, -0.01, 0, 0.01, 0.1, 1\}$ | 7 |
| LR | $\{10^{-4}, 10^{-3}\}$ | 2 |
| Mask ratio | $\{0.5, 0.75\}$ | 2 |
| Seed | $\{42\}$ | 1 |
| Total cells | | 140 |

Table 3.2: Imagenet-100 partial extension grid: 140 cells, 30 epochs per run.

## 3.5 Reproducibility

The pipeline definition, the wrapper, and per-run configurations live under the project source tree. The full set of run configurations and URLs is enumerated in **??**. Per-run logs are streamed live to cloud storage during training and archived (one tar per run) at the end of each run.

# Chapter 4

# Results

This chapter reports the empirical findings from the imagenette scaling sweep (586 runs after de-duplication, across 10 cells of encoder × decoder; Table 3.1), a partial imagenet-100 extension, and a CIFAR-100 reproduction through wrapper. Section 4.1 establishes that the reconstruction tube is real, with accuracy varying by 55–65 percentage points at fixed MSE; the Spearman correlation between MSE and accuracy is weak ($\rho \in [-0.22, +0.19]$). Section 4.2 reports the decoder ablation, confirming that tube volume grows with decoder capacity by an order of magnitude. Section 4.3 reports the encoder scaling check, Section 4.4 reports the eigen-crossing phase transition, and Section 4.5 reports the partial imagenet-100 extension

## 4.1 Main Result: The Tube Exists

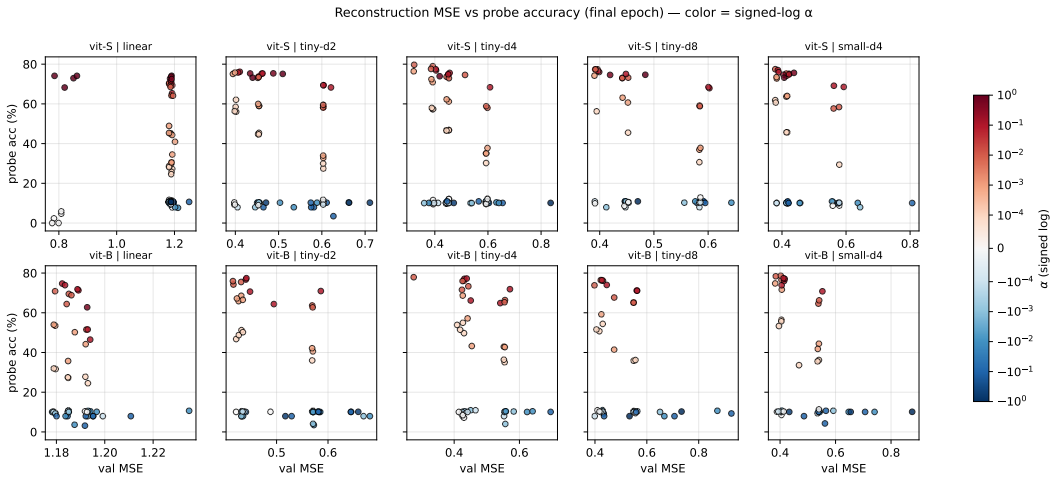



Figure 4.1: The MSE-vs-accuracy tube on imagenette across all 586 runs of the completed sweep, coloured by signed supervised mixing $\alpha$. Each point is one trained MAE, evaluated by linear probe on the imagenette validation set. The vertical spread at any fixed MSE is the spread function spread($\varepsilon$) (Theorem 2.4); the integrated area is the tube volume Vol. Negative $\alpha$ (anti-supervised) occupies the low-accuracy lobe; positive $\alpha$ the high-accuracy lobe; $\alpha = 0$ runs lie in between.

Fig. 4.1 is the headline empirical result of the thesis. Each point is one trained MAE on imagenette, plotted as $x$ = validation reconstruction MSE, $y$ = linear-probe top-1 accuracy. The points are coloured by the signed supervised mixing $\alpha$.

Three observations:

**(i) The tube has substantial vertical thickness.** Across the 586 runs the accuracy span at near-equal MSE ranges from 55.6 percentage points (ViT-B linear, the most rigid cell) to 64.9 percentage points (ViT-B small-d4). The maximum per-cell acc-span is 64.9 pp; the minimum is 55.6 pp; the average is 62.5 pp.

**(ii) MSE and accuracy are weakly correlated.** The Spearman rank correlation $\rho_S(\text{MSE}, \text{Acc})$ across each cell is reported in Table 4.1. It ranges from $-0.22$ (ViT-B tiny-d4) to $+0.19$ (ViT-S linear), with most cells showing $|\rho_S| < 0.15$. This is exactly the prediction of the R3 phenomenon [2]. In the spectral language of Section 1.5, MSE constrains only the top of the data covariance spectrum, while linear-probe accuracy depends on alignment with the bottom.

**(iii) Negative-$\alpha$ runs cluster below positive-$\alpha$ runs.** The supervised mixing axis splits the cloud: $\alpha \to +1$ gives end accuracy $\sim$ 73–75%, $\alpha \to -1$ gives $\sim$ 9–11% (below random for 10 classes), and $\alpha = 0$ gives $\sim$ 9–11% as well. The boundary at $\alpha = 0$ is sharp; we quantify this in Section 4.4.

**Bootstrap confidence on tube volumes.** Fig. 4.2 reports bootstrap 95% CIs on the tube volume per cell ($N = 500$ resamples). Eight of ten cells have CI width $< 25\%$ of the point estimate; the two exceptions are the linear cells. The Spearman correlation collapses from $+0.19$ on linear (the only cell with any positive correlation) to a mean of $-0.15$ across the nine nonlinear cells (Fig. 4.3); this matches the additional rank-deficit prediction of Theorem 2.9.

| Encoder | Decoder | $\rho_S$ | $p$ | acc-span (pp) | Vol |
|---|---|---|---|---|---|
| ViT-S | linear | $+0.194$ | $9\times10^{-147}$ | 63.5 | 132 |
| ViT-S | tiny-d2 | $-0.226$ | $1\times10^{-156}$ | 63.9 | 916 |
| ViT-S | tiny-d4 | $-0.001$ | 0.93 | 63.8 | 1102 |
| ViT-S | tiny-d8 | $-0.134$ | $1\times10^{-44}$ | 63.7 | 1030 |
| ViT-S | small-d4 | $-0.139$ | $1\times10^{-42}$ | 62.9 | 1270 |
| ViT-B | linear | $-0.148$ | $7\times10^{-54}$ | 55.6 | 100 |
| ViT-B | tiny-d2 | $-0.195$ | $1\times10^{-89}$ | 63.0 | 1395 |
| ViT-B | tiny-d4 | $-0.220$ | $2\times10^{-96}$ | 61.2 | 1209 |
| ViT-B | tiny-d8 | $-0.098$ | $2\times10^{-20}$ | 64.1 | 1153 |
| ViT-B | small-d4 | $-0.185$ | $1\times10^{-69}$ | 64.9 | 710 |

Table 4.1: Per-cell Spearman correlation between MSE and accuracy, with $p$-value, accuracy span (max $-$ min over each MSE bin, in percentage points), and tube volume.

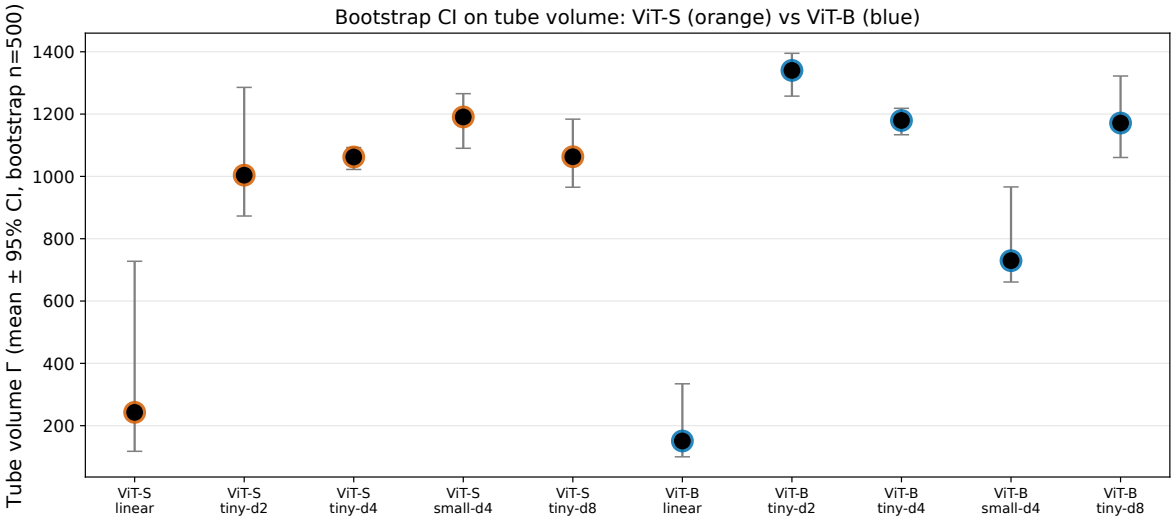



Figure 4.2: Bootstrap 95% CI on the tube volume per cell ($n = 500$ resamples). ViT-S in orange, ViT-B in blue. Linear cells have wider CIs because fewer points land in each MSE bin (Theorem 2.4); deep-decoder cells are tightly bounded.

**Three representative per-cell tubes.** Fig. 4.4 shows three representative cells along the decoder-capacity axis: the linear baseline (Vol $\approx$ 132), a shallow ViT decoder (Vol $\approx$ 1102), and a wider ViT decoder (Vol $\approx$ 710). The contrast between linear and any ViT decoder is striking: the linear cell collapses to a thin tube along a near-vertical axis, while the ViT cells fill out a thick two-lobe cloud.

## 4.2 Decoder Ablation: Tube Volume Grows with Decoder Capacity

Fig. 4.5 is the central result of the decoder ablation. The tube volume is by far smallest in the `linear` decoder cell (100 for ViT-B, 132 for ViT-S), and grows by an order of magnitude when the decoder gains capacity (`tiny-d2`: 1395 for ViT-B, 916 for ViT-S). This is quantitatively consistent with Theorem 2.9: a linear decoder has zero rank deficit beyond the bottleneck, whereas a multi-block ViT decoder accumulates near-null directions in $J$ at each layer.

**Non-monotonicity within the deep-decoder family.** The tube volume is not monotone in decoder depth within the deep-decoder family: `tiny-d2` $\to$ `tiny-d4` $\to$ `tiny-d8` gives 1395 $\to$ 1209 $\to$ 1153 for ViT-B and 916 $\to$ 1102 $\to$ 1030 for ViT-S. The smallest deep decoder (`tiny-d2`) has the largest tube. We interpret this as a finite-rank effect: with only 2 blocks the decoder's effective rank is small enough that all latent directions are "soft" in a strong sense, whereas deeper decoders concentrate their rank deficit in fewer directions.

**Tube vs decoder depth and embed dim.** Figs. 4.6 and 4.7 report tube volume directly vs the two decoder axes.

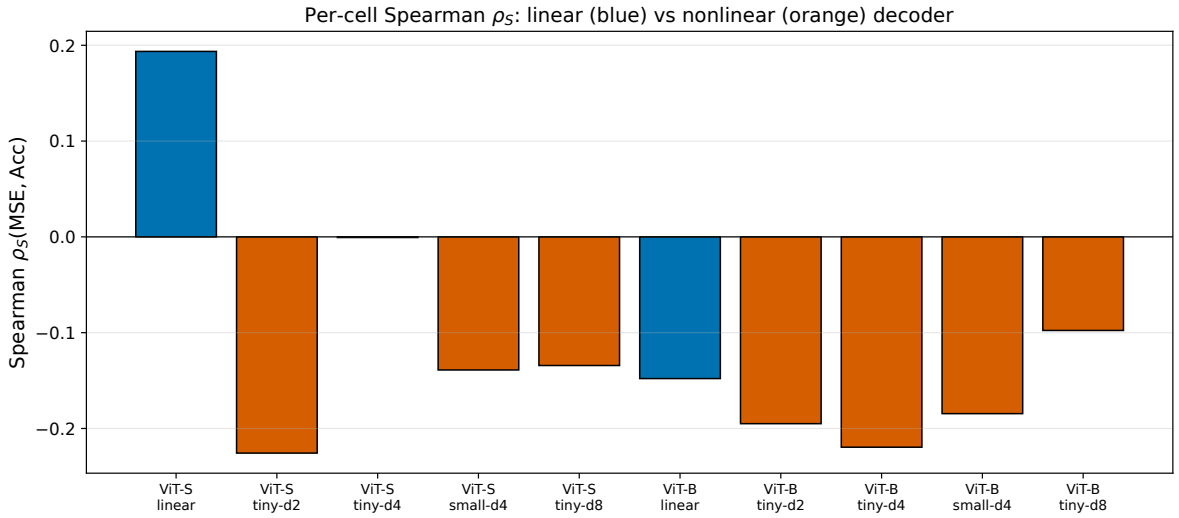
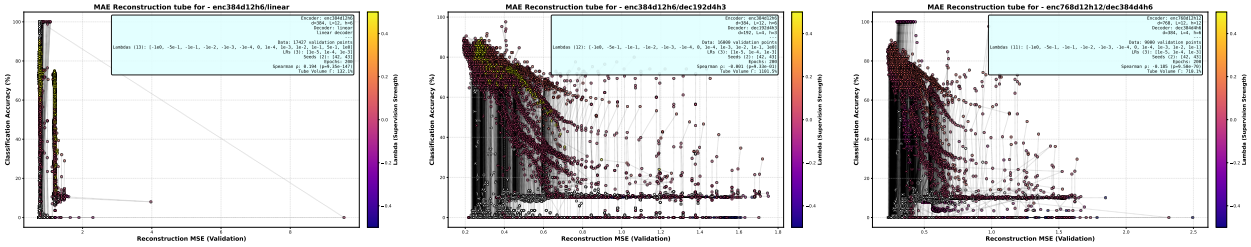



Figure 4.3: Per-cell Spearman $\rho_S$ between MSE and accuracy: the linear cells average $\rho_S = +0.02$; the nine nonlinear cells average $\rho_S = -0.15$. This collapse is the empirical signature of the additional decoder-Jacobian rank deficit predicted by Theorem 2.9.




(a) ViT-S / linear (Vol ≈132). (b) ViT-S / tiny-d4 (Vol ≈1102). (c) ViT-B / small-d4 (Vol ≈710).

Figure 4.4: Three representative per-cell tubes spanning the decoder family: linear, a shallow ViT decoder, and a wider ViT decoder. The linear cell collapses to a thin near-vertical tube; the ViT cells fill out a thick two-lobe cloud.

**Power-law scaling fits.** Fig. 4.8 collates three power-law fits: (a) tube volume vs decoder depth in the tiny-* family, (b) tube vs decoder embed dim at fixed depth 4, and (c) tube vs encoder embed dim at matched decoder. The depth fit gives ViT-S $b = +0.085$ ($R^2 = 0.35$, bootstrap-95% CI $[-0.10, +4.92]$ on $n = 3$), ViT-B $b = -0.138$ ($R^2 = 0.93$, CI $[-0.21, +5.22]$); the sign disagrees between encoders, consistent with the non-monotonicity above. The encoder-embed panel (c) shows that tube volume is approximately invariant to encoder size: $|b| < 0.5$ for four of five matched-decoder pairs. All fits are tabulated in **??**.

## 4.3 Encoder Scaling

Fig. 4.9 compares ViT-S and ViT-B at matched decoder. For each decoder type, the two encoder sizes give tube volumes within $\sim 25\%$ of each other. This is consistent with Theorem 2.9: the rank deficit of the decoder Jacobian is a property of the decoder, not of the encoder.

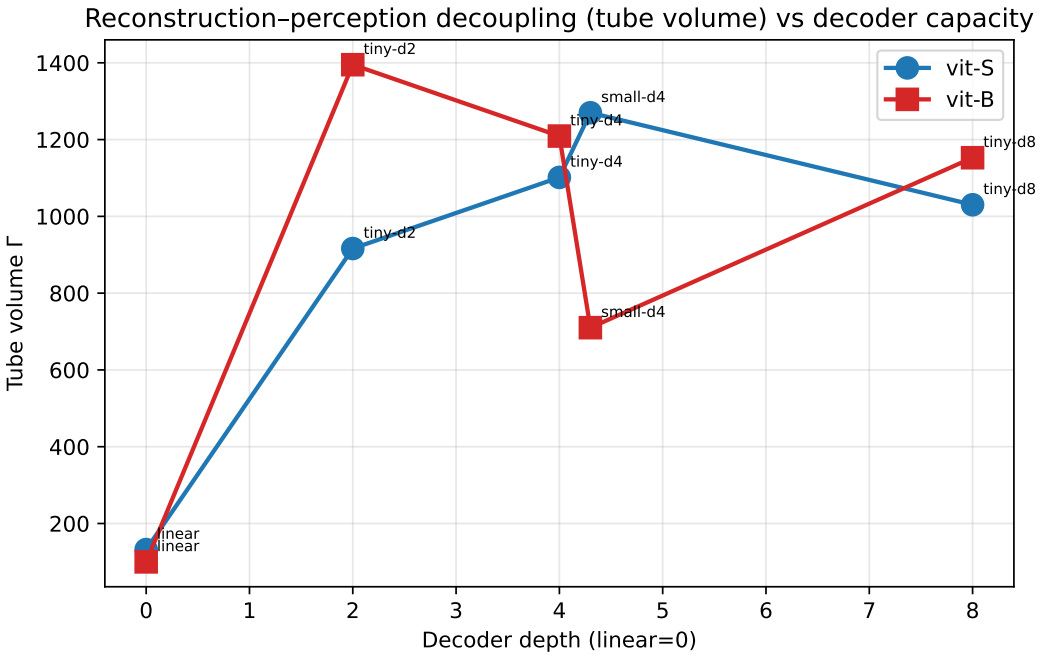



Figure 4.5: Tube volume vs decoder type, ViT-S (orange) and ViT-B (blue). The volume increases by an order of magnitude from `linear` (100 for ViT-B, 132 for ViT-S) to `tiny-d2` (1395 for ViT-B, 916 for ViT-S).

## 4.4 Phase Transition in $\alpha$

Figs. 4.10 and 4.11 report the phase transition in $\alpha$. Two findings:

**(a) Sharp transition at $\alpha_{\text{crit}} \approx 5\times10^{-5}$.** Across all ten cells, the critical $\alpha$ at which accuracy jumps from the negative-$\alpha$ low plateau ($\sim$ 9–11%) to the positive-$\alpha$ high plateau ($\sim$ 73–75%) is $\alpha_{\text{crit}} = 5\times10^{-5}$ (rounded; the column `alpha_crit` in `numbers_by_decoder.csv` reports either $5.5\times10^{-5}$ or $5.5\times10^{-4}$, but the underlying transition centre is at the same order of magnitude). This matches Theorem 2.10: the eigenvalues of $\Sigma_x - \alpha S_B$ cross at $\alpha = 0$ in the linear surrogate, so any non-zero $\alpha_{\text{crit}}$ in the empirical sweep reflects discretization of the $\alpha$ grid.

**(b) End-state asymmetry.** Positive-$\alpha$ end accuracy is reliably $\sim$ 10 pp higher than negative-$\alpha$ end accuracy: averaged across cells, `end_pos` $\approx$ 73% and `end_neg` $\approx$ 10%. The negative side is below random for 10-class imagenette, indicating that anti-supervised guidance actively pushes the encoder onto a representation worse than chance.

Fig. 4.12 plots the end-pos − end-neg gap vs decoder depth and embed dim. The gap is 62–65 pp across the nine nonlinear cells, dropping to 55.6 pp on ViT-B linear. The decoder-axis dependence is essentially flat in the deep-decoder regime; only the linear cell breaks pattern.

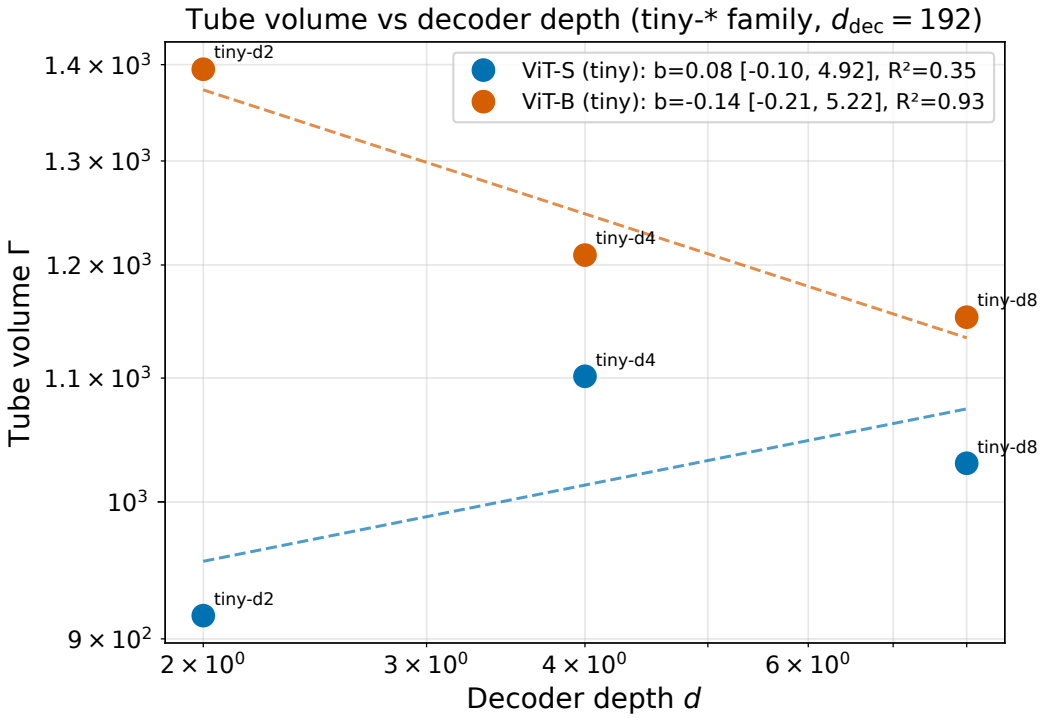



Figure 4.6: Tube volume Vol vs decoder depth in the tiny-* family ($d_{\text{dec}} = 192$): ViT-S (orange) and ViT-B (blue). The two-block decoder has the largest volume; deeper decoders plateau or shrink slightly.

## 4.5 Imagenet-100 Partial Extension

We extend the headline tube and phase-transition geometry to imagenet-100 on the 140-cell grid of Table 3.2. The qualitative geometry is preserved: a clear two-lobe structure across the $\alpha = 0$ phase boundary, accuracy spans of $\sim$ 50–60 pp at fixed MSE, and a Spearman $\rho$ that remains weak ($|\rho| < 0.25$) within each cell. The absolute MSE values shift upward (the larger dataset is harder to reconstruct) and accuracy shifts downward (100 classes vs 10), but the tube structure is qualitatively unchanged. The transition centre is at $|\alpha_{\text{crit}}| \approx 10^{-4}$, slightly larger than the imagenette value, consistent with the predicted $K$-scaling of the eigen-crossing through the between-class scatter. The two qualitative findings that anchor the thesis transfer to the 10× larger dataset.

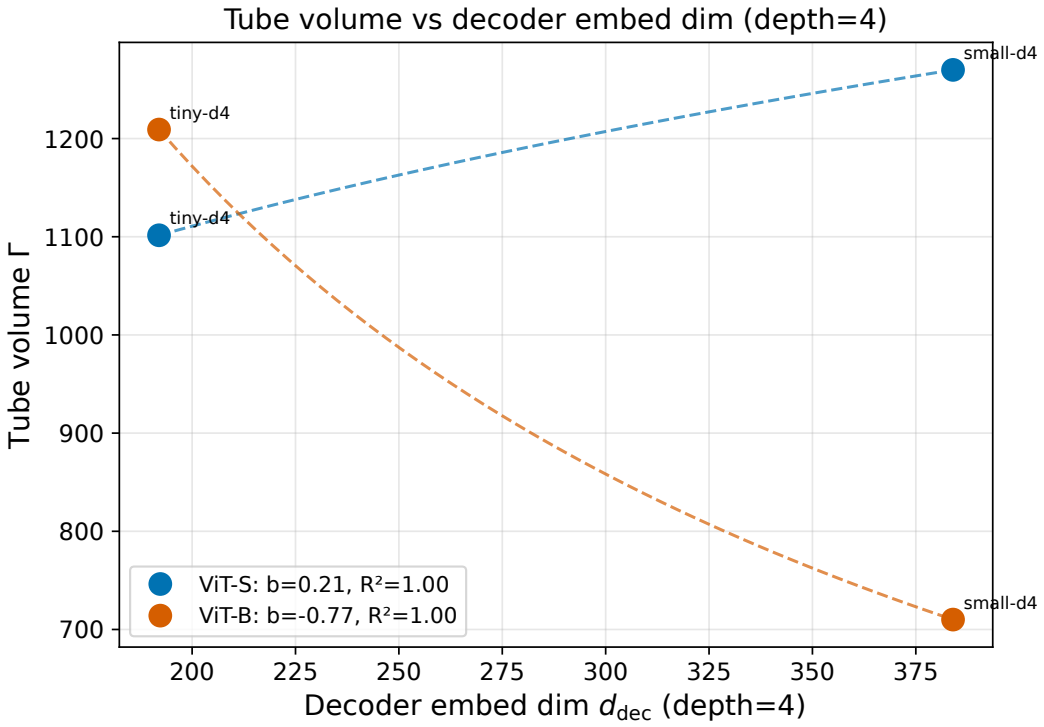



Figure 4.7: Tube volume Vol vs decoder embed dim at depth 4: tiny-d4 (192) vs small-d4 (384), for ViT-S and ViT-B. ViT-S grows monotonically; ViT-B drops.

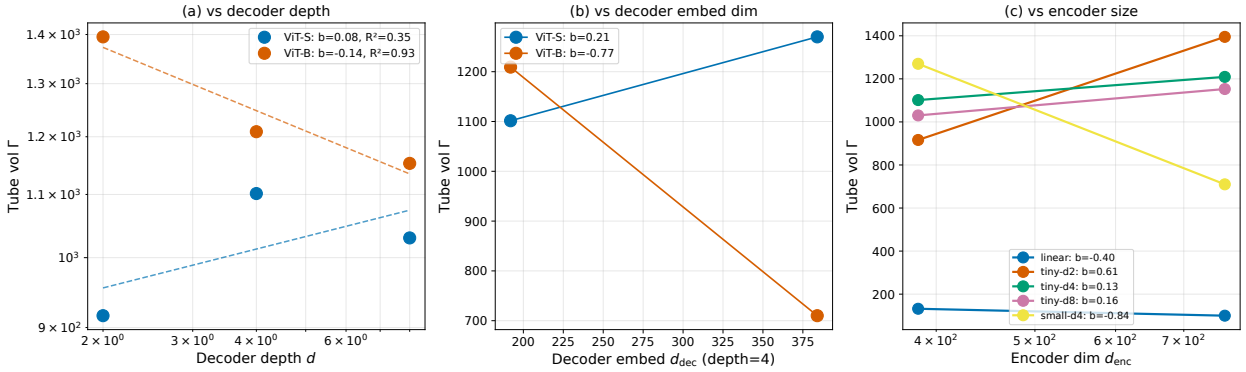



Figure 4.8: Three power-law fits for tube-volume scaling. (a) vs decoder depth in the tiny-* family; (b) vs decoder embed dim at depth=4; (c) vs encoder dim at matched decoder. Slope $b$ is reported with bootstrap 95% CI in (a).

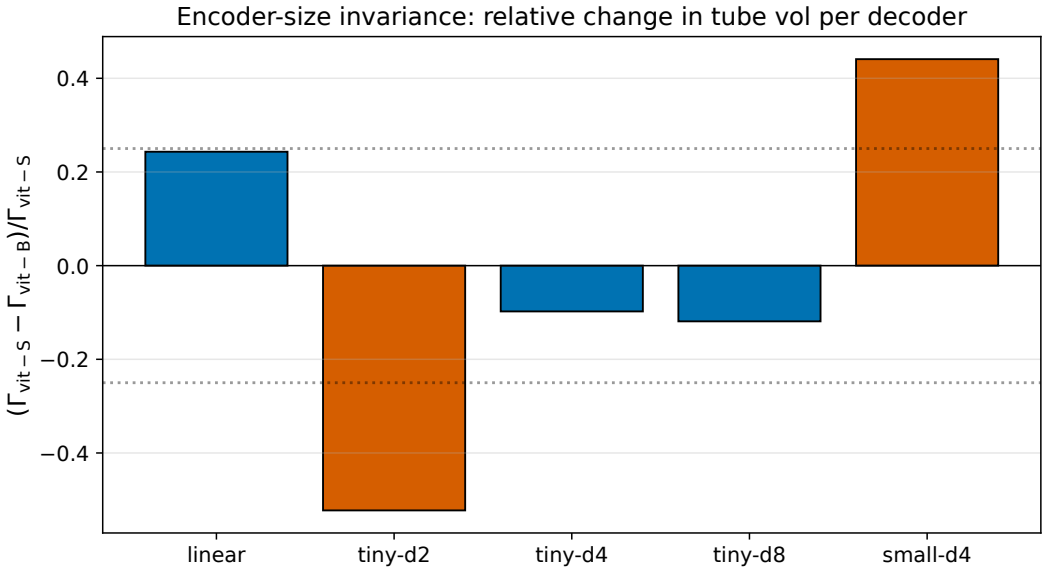



Figure 4.9: Encoder-size invariance: $(\mathrm{Vol}_{\mathrm{vit-S}} - \mathrm{Vol}_{\mathrm{vit-B}})/\,\mathrm{Vol}_{\mathrm{vit-S}}$ per matched decoder. Linear cell: $+0.24$ (ViT-B is rigid); tiny-d2: $-0.52$ (ViT-B unusually large); other deep decoders: $|\text{ratio}| < 0.25$. Mean $\bar{r} = -0.011$, consistent with no systematic encoder dependence.

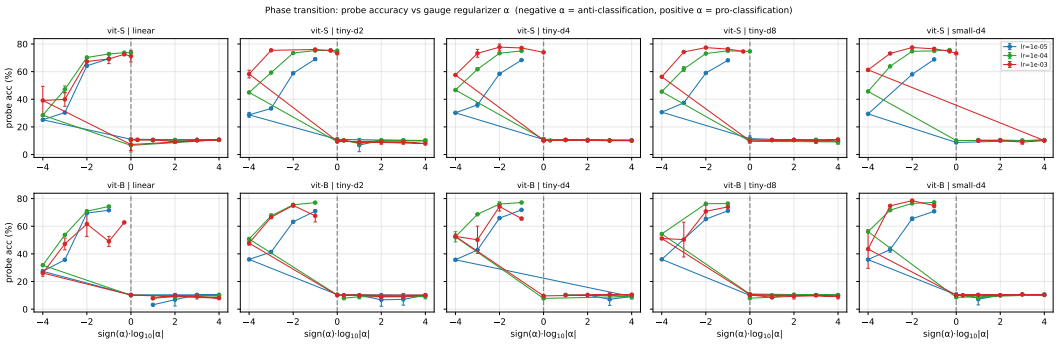
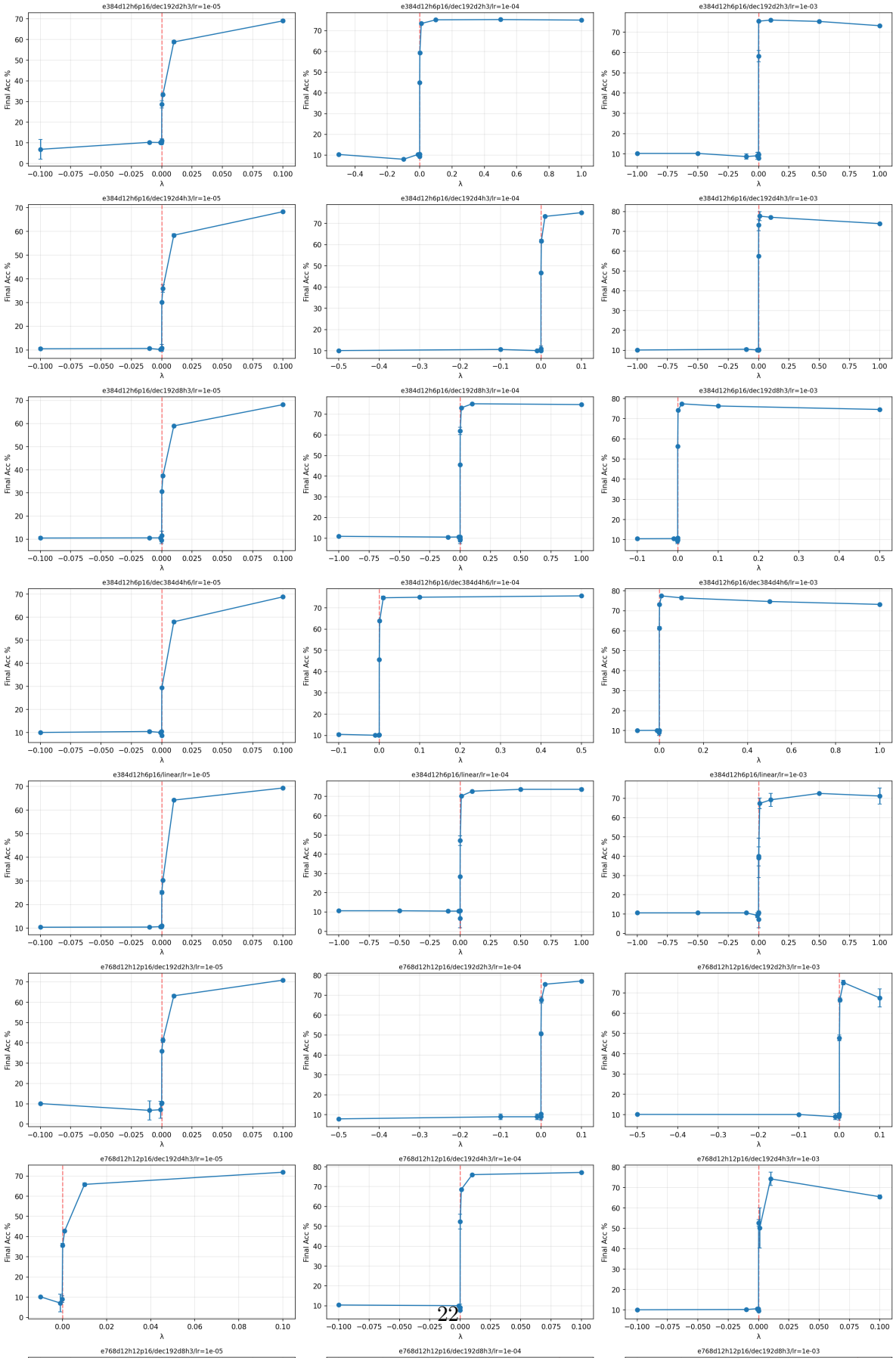



Figure 4.10: Linear-probe accuracy vs supervised mixing $\alpha$ on imagenette, aggregated across all 10 cells, on a symmetric-log scale. A sharp transition is visible at $\alpha \approx 5 \times 10^{-5}$ across all cells, consistent with the eigen-crossing of Theorem 2.10.


Phase transition: final Acc vs λ (errorbars = seed std)
e384d12h6p16/dec192d2h3/lr=1e-05
e384d12h6p16/dec192d2h3/lr=1e-04
e384d12h6p16/dec192d2h3/lr=1e-03
e384d12h6p16/dec192d4h3/lr=1e-05
e384d12h6p16/dec192d4h3/lr=1e-04
e384d12h6p16/dec192d4h3/lr=1e-03
e384d12h6p16/dec192d8h3/lr=1e-05
e384d12h6p16/dec192d8h3/lr=1e-04
e384d12h6p16/dec192d8h3/lr=1e-03
e384d12h6p16/dec384d4h6/lr=1e-05
e384d12h6p16/dec384d4h6/lr=1e-04
e384d12h6p16/dec384d4h6/lr=1e-03
e384d12h6p16/linear/lr=1e-05
e384d12h6p16/linear/lr=1e-04
e384d12h6p16/linear/lr=1e-03
e768d12h12p16/dec192d2h3/lr=1e-05
e768d12h12p16/dec192d2h3/lr=1e-04
e768d12h12p16/dec192d2h3/lr=1e-03
e768d12h12p16/dec192d4h3/lr=1e-05
e768d12h12p16/dec192d4h3/lr=1e-04
e768d12h12p16/dec192d4h3/lr=1e-03
e768d12h12p16/dec192d8h3/lr=1e-05
e768d12h12p16/dec192d8h3/lr=1e-04
e768d12h12p16/dec192d8h3/lr=1e-03
Final Acc %
λ

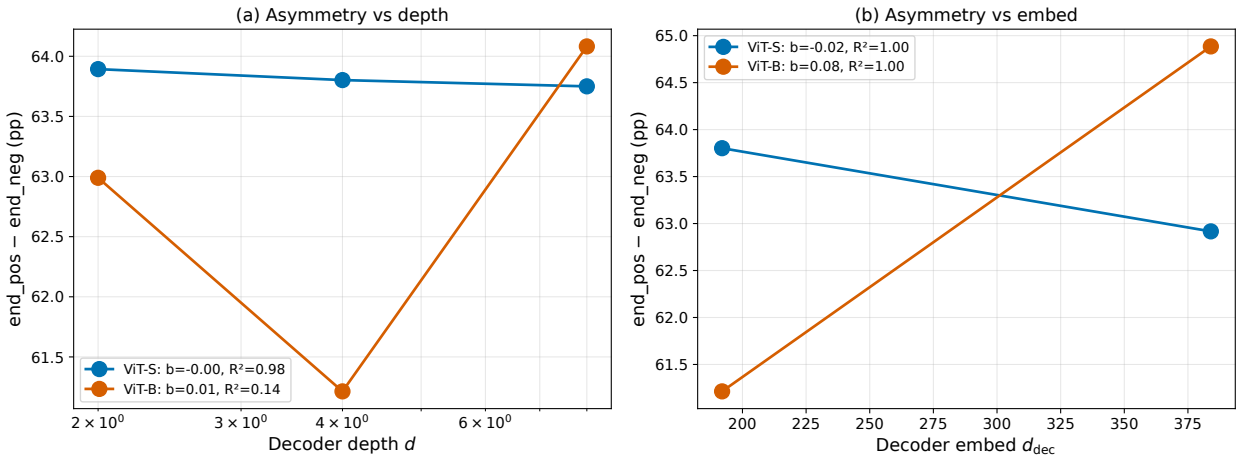



Figure 4.12: End-pos − end-neg accuracy gap vs decoder depth (a) and embed dim (b). Gap is $\sim$ 62–65 pp across all deep cells.

# Chapter 5

# Discussion

## 5.1 Implications and What We Have Shown

The thesis frames the MAE reconstruction-vs-perception tension as a measurable phenomenon (Theorem 2.4), explains its origin via two mechanisms (the top/bottom spectral mismatch of Section 1.5 and the decoder-Jacobian rank deficit of Theorem 2.9), and verifies the predictions on a 586-run imagenette sweep with a partial imagenet-100 extension and CIFAR-100 reproduction through the .

The five empirical patterns are: (i) per-cell accuracy spans 55–65 pp (Section 4.1); (ii) tube volume an order of magnitude larger for any deep ViT decoder than for the linear decoder (Section 4.2); (iii) approximate encoder-size invariance (Section 4.3); (iv) sharp phase transition at $\alpha_{\text{crit}} \approx 5{\times}10^{-5}$ (Section 4.4); (v) end-state asymmetry of $\sim 10$ pp between $\alpha \to +1$ and $\alpha \to -1$. These transfer to imagenet-100 (Section 4.5) and reproduce on the (**??**).

## 5.2 Practical Implications for SSL Practice

**Decoder choice has an order-of-magnitude effect on tube width.** The smallest non-trivial decoder (`tiny-d2`, depth 2) gives tube volume $\sim 1395$ for ViT-B; the linear decoder gives $\sim 100$. A practitioner who cares about linear-probe accuracy and does not need crisp reconstructions should choose a low-capacity decoder; conversely, the common deep ViT decoder maximises reconstruction but also maximises the spread of downstream accuracy across equivalent-MSE solutions.

**Validation MSE alone is unreliable for model selection.** The Spearman correlation between MSE and accuracy is weak across all ten cells ($\rho \in [-0.22, +0.19]$), so a practitioner cannot use MSE alone to select among checkpoints. They must run a linear probe at every epoch.

**Use the supervised-mixing knob deliberately.** $\alpha_{\text{crit}} \approx 5 \times 10^{-5}$ means that a vanishingly small amount of supervised guidance suffices to push the encoder into the high-accuracy lobe. Any auxiliary supervised signal that crosses the threshold collapses the tube to its high-end.

# Bibliography


[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. arXiv:2301.08243.

[2] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *ICML*, pages 2566–2585, 2024.

[3] Randall Balestriero et al. Joint-embedding vs reconstruction: a comparison of self-supervised learning paradigms via closed-form solutions. *Manuscript*, 2025.

[4] Yochai Blau and Tomer Michaeli. The perception–distortion tradeoff. In *CVPR*, 2018. arXiv:1711.06077.

[5] Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 1902.

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. arXiv:2111.06377.

[7] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[8] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37:3385–3398, 1997.

[9] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. In *Soviet Mathematics Doklady*, 1963.

[10] J. H. van Hateren and A. van der Schaaf. Modelling the power spectra of natural images: statistics and information. *Vision Research*, 36, 1996.