RECONSTRUCTION IS NOT PERCEPTION: A SIMPLE PROOF OF NECESSITY BUT INSUFFICIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked autoencoders (MAE) provide a scalable approach to visual representation learning, yet strong classification performance often requires fine-tuning, richer probes, or longer pretraining. We ask whether this mismatch reflects *downstream underdetermination*: models with the same reconstruction loss can nevertheless have substantially different downstream utility. We study this question through reconstruction-loss level sets in the joint encoder–decoder parameter space. Controlled sweeps of ViT-B/16 MAEs on ImageNet-100 reveal substantial variation in linear-probe accuracy among checkpoints at matched reconstruction loss, with the variation generally increasing as reconstruction loss decreases. We derive a first-order local characterization of this phenomenon and adapt a classification-guided predictor–corrector traversal that improves or degrades downstream performance while keeping reconstruction loss nearly fixed. Across nine decoder–loss settings, these walks span 24.4–50.4 percentage points of accuracy, 1.23–$2.05\times$ the matched-loss pretraining spread, and extend beyond the pretraining accuracy range in every setting. The phenomenon also transfers to public ImageNet-1k checkpoints, where walks improve ViT-B/16 and ViT-L/16 by 4.1 and 3.6 percentage points at nearly unchanged reconstruction loss. Mechanistically, reconstruction and classification gradients remain nearly orthogonal throughout pretraining, while walk-induced spectral changes concentrate in a late-learned tail containing only 0.06–0.11% of target variance. These results show that reconstruction loss does not determine downstream representation quality and identify directions that reconstruction weakly constrains yet matter substantially for classification.

1 INTRODUCTION

Learning representations by reconstructing inputs has long been central to unsupervised and self-supervised learning (Hinton & Salakhutdinov, 2006; Vincent et al., 2008; 2010; Devlin et al., 2019; Kingma & Welling, 2013; Van Den Oord et al., 2016; He et al., 2022). Masked autoencoders (MAE) make this approach scalable and transferable for visual representation learning (He et al., 2022).

Yet pretraining loss need not uniquely determine downstream quality. Language models with similar pretraining loss can perform differently downstream (Liu et al., 2023), and the same contrastive objective value can support different downstream behavior (Saunshi et al., 2022). MAE exhibits a related mismatch: strong classification often requires fine-tuning, richer probes, or longer pretraining (He et al., 2022; Balestriero & Lecun, 2024; Przewięźlikowski et al., 2025; Psomas et al., 2026). We ask whether this reflects *downstream underdetermination*: models with the same reconstruction loss but widely different linear-probe accuracies. If so, reconstruction loss alone cannot certify representation quality.

We measure this ambiguity through linear-probe accuracy along reconstruction-loss level sets in the joint encoder–decoder parameter space. In controlled sweeps of ViT-B/16 MAEs on ImageNet-100, accuracy already varies widely among pretraining checkpoints at matched loss, and the variation generally grows as reconstruction loss decreases. We then derive a first-order local characterization of nonlinear level sets and adapt a classification-guided predictor–corrector traversal that jointly updates the encoder and decoder while holding reconstruction loss nearly fixed. The resulting walks span 24.4–50.4 pp in accuracy, 1.23–$2.05\times$ the matched-loss pretraining spread, and exceed the pretraining accuracy range in every setting, including at convergence and across masking ratios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

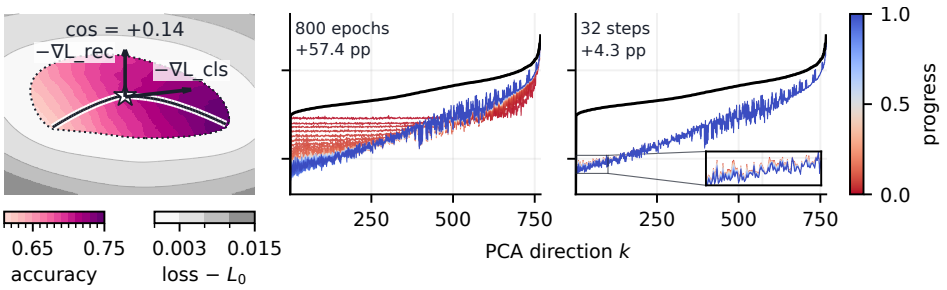

Figure 1: **Level-set geometry and reconstructed-patch PCA spectra**. **(Left)** A local reconstruction-loss level set: magenta shows linear-probe accuracy, gray shows loss deviation from the starting level, and arrows show the negative reconstruction and classification gradients. **(Center)** Prediction spectra over 800 pretraining epochs (+57.4 pp accuracy). **(Right)** Spectra along a 32-step level-set walk (+4.3 pp) at nearly fixed reconstruction loss; the inset magnifies the lowest-variance directions. Colored curves show prediction variance along target-patch PCA directions, ordered by increasing target variance; black shows the target spectrum. **Pretraining fits the dominant spectrum, while walks alter a downstream-sensitive, low-variance tail that reconstruction weakly constrains.**

Two observations help explain this variation (Figure 1). First, reconstruction and classification gradients remain nearly orthogonal throughout pretraining, leaving a large classification component tangent to the reconstruction level set. Second, the walks preserve the dominant reconstructed-patch spectrum while altering a late-learned tail that contains only 0.06–0.11% of the target variance yet strongly affects classification. The effect also transfers to public ImageNet-1k checkpoints: at nearly unchanged reconstruction loss, walks improve ViT-B/16 by 4.1 pp and ViT-L/16 by 3.6 pp.

Our contributions are as follows:

- We formulate downstream underdetermination as variation in fitted classification performance along reconstruction-loss level sets.
- We locally characterize underdetermination in nonlinear encoder–decoders and adapt a classification-guided predictor–corrector traversal.
- Across controlled MAE sweeps and public ImageNet-1k checkpoints, we expose more downstream variation than conventional pretraining.
- We identify geometric and spectral factors that help explain underdetermination: nearly orthogonal gradients and downstream-sensitive changes in a low-variance tail.

Section 2 formalizes downstream underdetermination and gives a motivating linear-autoencoder example. Section 3 presents the experimental protocol, underdetermination during conventional pretraining, first-order analysis, level-set walks, and spectral results. Finally, Section 4 discusses implications and directions for future work.

2 Background

To study reconstruction vis-à-vis linear probing performance on downstream classification, we assume access to labeled data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^d$ and one-hot encoded targets $\mathbf{y}_i \in \mathbb{R}^c$, which are respectively stacked horizontally into column-major matrices $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d\times n}$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{c\times n}$. Unless otherwise specified, $\mathbf{X}$ is assumed full rank throughout the paper.

Setting. We consider deep networks $\mathbf{f} : \mathbb{R}^d \times \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}^d$ formed by stacking L layers, with parameter $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^L) \in \mathbb{R}^p$ collecting all weights and biases, for $p_1 + p_2 = p$. The model $\mathbf{f}$ is obtained by composing a decoder network $\mathbf{g}_2 : \mathbb{R}^k \times \mathbb{R}^{p_2} \to \mathbb{R}^d$ with an encoder $\mathbf{g}_1 : \mathbb{R}^d \times \mathbb{R}^{p_1} \to \mathbb{R}^k$, where $\mathbf{g}_1$ and $\mathbf{g}_2$ respectively act column-wise over the batched $\mathbf{X}$ and $\mathbf{Z}$. By letting $\boldsymbol{\phi} = \boldsymbol{\theta}^{1:\ell}$ and $\boldsymbol{\psi} = \boldsymbol{\theta}^{\ell+1:L}$, the network computes $\hat{\mathbf{X}} = \mathbf{g}_2(\mathbf{Z}, \boldsymbol{\psi}) \in \mathbb{R}^{d\times n}$, for $\mathbf{Z} = \mathbf{g}_1(\mathbf{X}, \boldsymbol{\phi}) \in \mathbb{R}^{k\times n}$, with the interval $[\ell] := \{1, \ldots, \ell\}$ indexing the encoder parameters,

$1 \leq \ell < L$. For a linear predictor $\mathbf{W} \in \mathbb{R}^{c \times k}$ and $\lambda \geq 0$, we consider the regularized objective

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\psi}, \mathbf{W}) := \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi}) + \lambda \mathcal{L}_{\text{cls}}(\mathbf{W}, \boldsymbol{\phi}) = \frac{1}{2n} \|\hat{\mathbf{X}} - \mathbf{X}\|_F^2 + \frac{\lambda}{2n} \|\mathbf{W}\mathbf{Z} - \mathbf{Y}\|_F^2 \tag{1}$$

whereby setting $\lambda = 0$ recovers the pretraining loss. Finally, we recall that since $\hat{\mathbf{X}} = \mathbf{g}_2(\mathbf{Z}, \boldsymbol{\psi})$, both the decoder and predictor network share the same latent embeddings $\mathbf{Z} = \mathbf{Z}(\boldsymbol{\phi})$.

Underdetermination. We now introduce the notion of underdetermination. Given two distinct parameters $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p$, underdetermination arises when the respective models attain the same reconstruction error $\mathcal{L}_{\text{rec}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{rec}}(\boldsymbol{\theta}')$, but different downstream classification performance $\mathcal{L}_{\text{cls}}(\boldsymbol{\phi}, \mathbf{W}^\star(\boldsymbol{\phi})) \neq \mathcal{L}_{\text{cls}}(\boldsymbol{\phi}', \mathbf{W}^\star(\boldsymbol{\phi}'))$, here shown under the optimal linear predictor (Section 3.3).

Importantly, since the aim of pretraining is to guide optimization towards extracting rich features that are useful downstream, underdetermination can be indicative of poor alignment between the pretext task and the downstream one. Motivated by this view, in this work we experimentally explore the problem by asking an important question: *how well does learning by reconstructing the input inform the encoder about retaining features that are useful downstream?* Specifically, if the range of realized downstream accuracy at fixed pretraining error across a cohort of compatible training trajectories is wide, underdetermination provides a strong signal of pretraining suboptimally prioritizing good downstream features. Furthermore, for a well pre-trained encoder $\boldsymbol{\phi}$, if there exist a distinct $\boldsymbol{\phi}'$ in the proximity of $\boldsymbol{\phi}$ showing higher downstream performance, then the pre-training trajectory informed feature selection suboptimally. In the next section, as a motivating example, we draw an intuitive connection between encoder feature selection and the emergence of downstream underdetermination.

2.1 Underdetermination in Linear Autoencoders

As a motivating example, we consider the case of a linear encoder $\mathbf{g}_1(\mathbf{X}, \boldsymbol{\phi}) := \mathbf{V}\mathbf{X}$ and decoder $\mathbf{g}_2(\mathbf{Z}, \boldsymbol{\psi}) := \mathbf{U}\mathbf{Z}$, for $\mathbf{V} \in \mathbb{R}^{k \times d}$, $\mathbf{U} \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\phi} = \text{vec}(\mathbf{V})$, $\boldsymbol{\psi} = \text{vec}(\mathbf{U})$.

In this setting, it is well known that in the vicinity of critical points of the reconstruction error, the encoder features take the form of Equation (2), recovering k eigendirections of the $\mathbf{X}$-whitened subspaces of the task covariance (Balestriero & Lecun, 2024). Furthermore, if $\lambda \to 0$, one instead recovers k eigendirections of the data covariance (Oja, 1982). In Equation (2), this is represented by the index set I and it is the core mechanism driving feature selection in the linear autoencoder.

$$\mathbf{V} \quad \text{spans} \quad \mathbf{P}_{\mathbf{X}\mathbf{X}^\top} \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} (\mathbf{P}_{\mathbf{H}})_{\cdot, I} \quad \text{for } I = \{i_1, \ldots, i_k\} \subset [d] \tag{2}$$

$$\mathbf{H} = \mathbf{D}_{\mathbf{X}\mathbf{X}^\top} + \lambda \mathbf{S}, \quad \mathbf{S} := \left(\mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} \mathbf{P}_{\mathbf{X}\mathbf{X}^\top}^\top \mathbf{X}\mathbf{Y}^\top \mathbf{Y}\mathbf{X}^\top \mathbf{P}_{\mathbf{X}\mathbf{X}^\top} \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} \right) \tag{3}$$

Baldi & Hornik (1989) further note that the non-optimal critical points are strict saddles. In keeping with the saddle-to-saddle dynamics observed in the (deep) linear regime (Pesme & Flammarion, 2023; Saxe et al., 2014) and the kernel regime under small-scale initialization (Jacot et al., 2021), then the saddle points of Equation (2) provide a local model of feature learning in the linear autoencoder. This allows us to describe underdetermination. Specifically, for two index sets $I \neq I'$ of size k realizing the same reconstruction error, one can ask what is the corresponding difference in classification error.

Proposition 1 (Underdetermination in linear autoencoders (informal))**.** *Consider two index sets $I \neq I'$ of size k and assume that the corresponding encoder features admit the same reconstruction error. With abuse of notation, denote $\mathcal{L}_{\text{cls}}(I)$ as the downstream error of the features $\boldsymbol{\phi}$ associated with index set I. Let $s_1 \geq \ldots \geq s_d \geq 0$ denote the eigenvalues of $\mathbf{S}$ (Equation (3)). Then, in the unsupervised case ($\lambda = 0$), the absolute difference in classification performance is bounded by:*

$$0 \leq |\mathcal{L}_{\text{cls}}(I) - \mathcal{L}_{\text{cls}}(I')| \leq \Big| \sum_{j=1}^{k} s_j - \sum_{j=d-k+1}^{d} s_j \Big| \tag{4}$$

Hence, when distinct features at fixed reconstruction error do not capture the same task covariance energy, downstream underdetermination is observed, with range dependent on task-data alignment.

Our construction illustrates underdetermination arising from a simple mechanism of feature selection. At the same time, modern reconstruction-based algorithms, most prominently Masked AutoEncoders

(MAE) (He et al., 2022), indirectly shape the feature selection mechanism in several important ways, namely by employing mask-based augmentation (Devlin et al., 2019) and by introducing asymmetry between the encoder and decoder. For an input image $\bar{\mathbf{X}} \in \mathbb{R}^{t \times s}$ tokenized into t patches of size s(Dosovitskiy et al., 2021), a binary mask $\mathbf{m} \in \{0,1\}^t \overset{\text{iid}}{\sim} \text{Ber}(\rho)$ is drawn to generate two views $\mathbf{X}_1 = \bar{\mathbf{X}}_{\mathbf{m},\cdot}$, $\mathbf{x}_2 = \bar{\mathbf{X}}_{\mathbf{1}-\mathbf{m},\cdot}$ which are used to compute the loss (Zhang et al., 2022)

$$\mathcal{L}_{\text{MAE}}(\boldsymbol{\theta}, \bar{\mathbf{X}}) := \mathbb{E}_{\bar{\mathbf{X}}} \mathbb{E}_{\mathbf{X}_1, \mathbf{X}_2 | \bar{\mathbf{X}}, \rho} \left\| \mathbf{g}_2(\mathbf{g}_1(\mathbf{X}_1, \boldsymbol{\phi}), \boldsymbol{\psi}) - \mathbf{X}_2 \right\|_F^2 \tag{5}$$

Our simple linear construction then begs the question: *is underdetermination a prevalent problem in masked autoencoders? How wide is the downstream underdetermination range typically observed in practice?* To investigate these questions, in Section 3 we take a level set view of the problem, by exploring the range of downstream classification accuracy observed at fixed reconstruction errors representative of trained models, by controlling different model design choices.

2.2 Related Work

Reconstruction and downstream representations. Reconstruction has long been used for representation learning, from vanilla and denoising autoencoders to masked modeling (Hinton & Salakhutdinov, 2006; Vincent et al., 2008; 2010; Devlin et al., 2019; He et al., 2022; Xie et al., 2022). Despite strong transfer, reconstruction-trained features can remain poorly aligned with downstream classification under linear evaluation, motivating longer pretraining, fine-tuning, or richer probing strategies(He et al., 2022; Przewięźlikowski et al., 2025; Psomas et al., 2026). In particular,Balestriero & Lecun (2024) show that for standard autoencoders, reconstruction preferentially falls into high-variance directions of the representation, while features useful downstream only emerge later in training. These results highlight a potential mismatch between reconstruction quality and downstream representation quality, which our work extends to masked objectives. Importantly, we propose a novel algorithm to probe the representation space in search of classification-aligned directions.

Pretraining loss. Similar pretraining loss need not imply similar downstream performance. Liu et al. (2023) study the problem in language modeling and relate the gap to implicit bias, while Saunshi et al. (2022) show that contrastive objectives can yield different downstream behavior across learning algorithms. In this work, we expose directions in which a local representation can be significantly improved, highlighting the insufficiency of reconstruction for selecting good classification features.

Identifiability. Nelson et al. (2026) formally study the problem of whether optimal representations are equivalent to one another, or to a latent representation, up to specified transformations. We instead ask how much downstream variation is permitted at a fixed reconstruction-loss value, in the setting where a pretext loss is optimized in order to indirectly maximize downstream accuracy.

3 A Level-Set View of Underdetermination

3.1 Experimental Setup

We pretrain ViT-B/16 encoders (Dosovitskiy et al., 2021) on ImageNet-100 (Deng et al., 2009), using masked autoencoding (He et al., 2022). At the standard masking ratio of 0.75, we consider three decoder sizes and sweep three learning rates and six weight-decay values, yielding 54 runs. Furthermore, we train the default decoder at four further masking ratios $0.3, 0.5, 0.6, 0.9$, sweeping the same six weight decays at the fixed learning rate 3×10^{-4}, adding 24 runs. At checkpoints throughout training, we measure fixed-mask validation reconstruction loss $\widehat{\mathcal{L}}_{\text{rec}} := \mathcal{L}_{\text{MAE}}$ using Equation (5) at each run's own masking ratio, together with validation linear probe accuracy, fitting each probe to the frozen encoder. The full experimental setup is described in Section B, with pretraining curves shown in Figures 5 and 6. Finally, let $\text{Acc}(\mathbf{W}, \boldsymbol{\phi})$ denote the validation accuracy of encoder $\boldsymbol{\phi}$ with linear head $\mathbf{W}$, and let $\mathbf{W}^\star_{\text{CE}}(\boldsymbol{\phi})$ denote a head fitted by minimizing the training cross-entropy while holding $\boldsymbol{\phi}$ fixed. For a nonempty level set $S_r \subset \mathbb{R}^p$ of $\widehat{\mathcal{L}}_{\text{rec}}$ at error $r > 0$, we quantify classification underdetermination by its accuracy spread, defined as follows:

$$\Delta_{\text{Acc}}(r) := \max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in S_r} \left| \text{Acc}(\mathbf{W}^\star_{\text{CE}}(\boldsymbol{\phi}_1), \boldsymbol{\phi}_1) - \text{Acc}(\mathbf{W}^\star_{\text{CE}}(\boldsymbol{\phi}_2), \boldsymbol{\phi}_2) \right|. \tag{6}$$

In practice, we approximate this quantity from below using finitely many training trajectories.

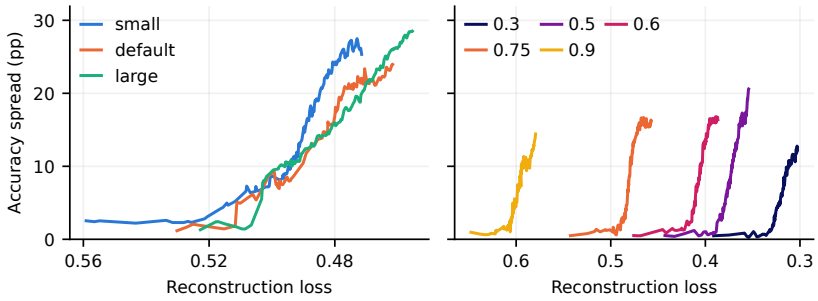

Figure 2: **Spread of linear probe accuracy vs reconstruction loss during pretraining**, with the x-axis inverted to match the direction of training. **(Left)** Three decoder sizes at masking ratio 0.75; each curve represents the maximum range across learning rates and weight decays. **(Right)** Five masking ratios at fixed decoder size; showing the maximum range across learning rates. **Across all settings, underdetermination increases over training as the reconstruction loss decreases.**

3.2 Downstream Underdetermination in MAE Pretraining

We investigate the extent to which classification underdetermination arises during standard MAE pretraining. In keeping with Equation (6), we construct level sets by training models that share the same parameter space and reconstruction objective $\widehat{\mathcal{L}}_{\mathrm{MAE}}$. We thus collect training trajectories of fixed-decoder configurations and fixed masking ratios, two central design choices in MAE, and vary learning rate and weight decay to elicit distinct trajectories and linear-probe accuracies. To test whether underdetermination persists under different conditions, we repeat the analysis across three decoder configurations at masking ratio 0.75 and across five masking ratios with the default decoder.

Because checkpoints are observed at discrete loss values, we interpolate each run's accuracy trajectory onto a grid of 600 reconstruction-loss values over the range shared by all runs in its configuration. At each grid point, we compute the difference between the maximum and minimum interpolated accuracies. Figure 2 plots this spread with inverted x-axis, matching the training direction.

In every setting, the accuracy spread generally increases as the reconstruction loss decreases. At matched reconstruction losses, small decoders generally exhibit larger spread, which is particularly pronounced in the low-loss regime, although this ordering does not hold at every loss. At masking ratios 0.3, 0.5, 0.6, 0.75, and 0.9, the spread reaches 12.8, 20.6, 16.8, 16.7, and 14.4 pp, respectively.

Together with the individual training curves (Figures 5 and 6), the results indicate that for MAE pretraining, achieving low reconstruction loss is necessary but not sufficient for learning strong classification features. Along each trajectory, lower reconstruction loss accompanies higher linear-probe accuracy. Yet low loss does not determine feature quality: models at the same low loss can show widely different accuracies, and this ambiguity generally grows as the loss decreases.

3.3 Locally Characterizing Underdetermination

We now provide conditions for which underdetermination is observable and which will inform our experimental investigation. Let $S_r := \{(\boldsymbol{\phi}, \boldsymbol{\psi}) \in \mathbb{R}^p : \mathcal{L}_{\mathrm{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi}) = r\}$ denote a non-empty level set of the reconstruction loss, for $r > 0$. For a parameter $(\boldsymbol{\phi}, \boldsymbol{\psi}) \in S_r$, we ask whether there exist local directions that improve downstream classification performance without altering the reconstruction loss. If found, such directions provide evidence of the pre-training loss suboptimally exploring the local representation space in terms of selecting features that are useful for downstream classification.

To make our analysis well posed, r is assumed to be regular according to Sard's Lemma (Sard, 1942), so that $\nabla \mathcal{L}_{\mathrm{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi}) \neq \mathbf{0}$ a. e. in S_r (Lemma 2, Section A). Next, by recalling that the linear probe $\mathbf{W}$ and the decoder $\mathbf{g}_2$ in Equation (1) share only the latent codes $\mathbf{Z} = \mathbf{Z}(\boldsymbol{\phi})$, we proceed by writing the downstream classification loss in terms of the encoder's parameter. This is done by taking the ordinary least-squares solution $\mathbf{W}^\star \in \operatorname{argmin}_{\mathbf{W}} \mathcal{L}_{\mathrm{cls}}(\mathbf{W}, \boldsymbol{\phi})$ and setting

$$\bar{\mathcal{L}}_{\mathrm{cls}}(\boldsymbol{\phi}) := \mathcal{L}_{\mathrm{cls}}(\boldsymbol{\phi}, \mathbf{W}^\star) \quad \text{for } \mathbf{W}^\star \in \operatorname*{argmin}_{\mathbf{W}} \mathcal{L}_{\mathrm{cls}}(\boldsymbol{\phi}, \mathbf{W}) \tag{7}$$

which can be solved[1] in closed form in $\mathbf{W}$ (Section A.2.2). Finally, by letting $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$, we take $\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) := (\bar{\mathcal{L}}_{\text{cls}} \circ \pi)(\boldsymbol{\theta})$, for $\pi(\boldsymbol{\theta}) := \boldsymbol{\phi}$, so that $\bar{\mathcal{L}}_{\text{cls}}$ can be represented in the level set S_r.

For a starting parameter $\boldsymbol{\theta} \in S_r$, the constrained optimization problem

$$\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) + \mu(\mathcal{L}_{\text{rec}}(\boldsymbol{\theta}) - r) = 0 \tag{8}$$

asks when, at first order, it is possible to vary the classification loss without leaving the level set.

Proposition 2 (First-order stationarity conditions)**.** *Let* $\mathcal{T}_{\boldsymbol{\theta}} S_r = \{\mathbf{u} \in \mathbb{R}^p : \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rec}}, \mathbf{u} \rangle = 0\}$. *Solving Equation* (8) *is equivalent at first order to the projected gradient flow*

$$\dot{\boldsymbol{\theta}} = -\mathbf{P}_{\mathcal{T}} \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) \qquad \textit{for} \qquad \mathbf{P}_{\mathcal{T}} := \mathbf{I}_p - \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rec}} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rec}}^\top}{\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{rec}}\|_2^2} \tag{9}$$

Particularly, the flow is stationary for $\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) = 0$ *or* $\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) \in \text{span}(\mathcal{L}_{\text{rec}}(\boldsymbol{\theta}))$ *and* $\nabla_{\psi} \mathcal{L}_{\text{rec}} = 0$.

Starting from a base point $\boldsymbol{\theta} \in S_r$, Proposition 2 provides a necessary halting condition for the classification gradient flow within the level set. Until the condition is met, there exist a direction for the encoder to move locally within S_r along the classification gradient.

We now turn our attention to understanding what factors in the autoencoder architecture affect the existence of directions that are locally unexplored by the pretraining optimization trajectory. Importantly, for an encoder parameter $\boldsymbol{\phi}$ in the level set, we consider the optimal local decoder $\boldsymbol{\psi}^\star = \boldsymbol{\psi}^\star(\boldsymbol{\phi}) := \text{argmin}_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi})$ and define $\bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) = \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi})\big|_{\boldsymbol{\psi}^\star}$ (Section A.2.4). By restricting the level set to $S_r^{\text{enc}} = \{\boldsymbol{\phi} : \bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) = r\} \subset S_r$, we can isolate directions in the encoder level set that are *unexplored* under the locally optimal decoder. Crucially, if classification-aligned features are found among those free directions, then the pretraining optimization trajectory is locally suboptimal for downstream classification. Hence, for a direction $\mathbf{u} \in \mathcal{T}_{\boldsymbol{\phi}} S_r^{\text{enc}}$, $\mathbf{u} \neq \mathbf{0}$, we proceed by asking what conditions satisfy level set tangency $\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}, \mathbf{u} \rangle = 0$ (proof in Section A.2.5).

Theorem 1 (Free encoder directions (informal))**.** *Let* $\bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) = r$ *and* $\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) \neq \mathbf{0}$*, so that*

$$\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}} = \frac{1}{n} \left(D_{\boldsymbol{\phi}} \mathbf{Z}\right)^\top \left(D_{\mathbf{Z}} \mathbf{g}_2\right)^\top \text{vec}\big(\hat{\mathbf{X}} - \mathbf{X}\big)\big|_{\boldsymbol{\psi}^\star}. \tag{10}$$

Let $\mathbf{u} \in \mathbb{R}^{p_1}$ *with* $\mathbf{u} \notin \ker(D_{\boldsymbol{\phi}} \mathbf{Z})$*, so that* $\mathbf{u}$ *moves the latent code* $\mathbf{Z}$*. Then* $\mathbf{u} \in \mathcal{T}_{\boldsymbol{\phi}} S_r^{\text{enc}}$*, i.e.* $\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}, \mathbf{u} \rangle = 0$*, if and only if*

$$D_{\mathbf{Z}} \mathbf{g}_2 \, D_{\boldsymbol{\phi}} \mathbf{Z} \, \mathbf{u} \;\perp\; \text{vec}(\hat{\mathbf{X}} - \mathbf{X}) \quad \text{(error redistribution)}; \tag{11}$$

and each of the following implies Equation (11)*:*

(i) $D_{\boldsymbol{\phi}} \mathbf{Z} \, \mathbf{u} \in \ker\left(D_{\mathbf{Z}} \mathbf{g}_2\right)$ (decoder invariant).

(ii) $\exists$ *a sequence* $\delta_k \downarrow 0$ *with* $\bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi} + \delta_k \mathbf{u}) = r$ *for all* k (decoder expressible).

Theorem 1 provides sufficient conditions for the existence of free directions within the level set. Condition (i) represents encoder parameter perturbations to which the decoder is invariant. Condition (ii) captures perturbations whose changes in the reconstruction output can be absorbed by the decoder. Importantly, Equation (11) captures changes in the decoder output that are reabsorbed by redistributing the reconstruction error over the training data while keeping the average error fixed. This includes output perturbations moving orthogonally to the support of the data distribution.

Theorem 1, has implications for understanding implicit model selection during pretraining, highlighting one mechanism of underdetermination. Specifically, if the tangential perturbation $\mathbf{u} \in \text{span}(\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{cls}})$, then there exist encoder parameter directions that are well-aligned with downstream classification while being orthogonal to the pretraining trajectory. Due to the decoder's local optimality, such directions correspond to genuine classification-aligned representations that the pretraining trajectory locally missed. The above results suggest an experimental protocol for exploring underdetermination: test the stationarity condition of Proposition 2 and then follow the gradient flow while adapting the decoder to ensure its representation does not degrade.

Remark. The analysis provided in this section holds to the first order. Second order effects–such as those arising from discrete step size, mini-batch optimization and preconditioning–may push the pretraining trajectory towards exploring the free directions that are not seen by gradient flow. In the sequel, our experiments account for second-order effects appearing in realistic training settings.

[1] If $\mathcal{L}_{\text{cls}} \in \mathcal{C}^1(\mathbb{R}^{p_1}, \mathbb{R}^{c \times k})$, then since $\mathcal{L}_{\text{cls}}(\boldsymbol{\phi}, \mathbf{W})$ is convex in $\mathbf{W}$, the construction is well posed.

3.4 Exploiting Downstream Underdetermination with Level-Set Walks

Algorithm 1 Level-set walk (one direction); full details in Algorithms 2 and 3.

Require: start $\boldsymbol{\theta}_0$, direction $\sigma \in \{-1, +1\}$, step s, tolerances τ, ε, convergence window M and threshold γ

$r_{\mathrm{tr}} \leftarrow \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}_0), \quad r \leftarrow \widehat{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\theta}_0), \quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0, \quad \mathbf{W} \leftarrow \mathbf{W}^{\star}_{\mathrm{CE}}(\boldsymbol{\phi}_0), \quad \mathrm{Acc}^{\mathrm{val}}_0 \leftarrow \mathrm{Acc}(\mathbf{W}, \boldsymbol{\phi}_0)$

for $t = 1, 2, \ldots$ **do**

  $\mathbf{g}_R \leftarrow \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}), \quad \mathbf{g}_C \leftarrow \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{cls}}(\mathbf{W}, \boldsymbol{\phi}), \quad \mathbf{v} \leftarrow \mathbf{g}_C - \dfrac{\mathbf{g}_C^{\top}\mathbf{g}_R}{\|\mathbf{g}_R\|_2^2}\,\mathbf{g}_R$

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \sigma\, s\, \mathbf{v}/\|\mathbf{v}\|_2$ ▷ predictor: tangent step

  while $|\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r_{\mathrm{tr}}| > \tau$ **do** ▷ corrector: damped Newton, $\mathbf{g}_R$ recomputed each iteration

    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\, \dfrac{\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r_{\mathrm{tr}}}{\|\mathbf{g}_R\|_2^2}\,\mathbf{g}_R$ ▷ $\eta \le 1$ by backtracking

  end while

  $\mathbf{W} \leftarrow \mathbf{W}^{\star}_{\mathrm{CE}}(\boldsymbol{\phi})$

  $\mathrm{Acc}^{\mathrm{val}}_t \leftarrow \mathrm{Acc}(\mathbf{W}, \boldsymbol{\phi})$

  break if $|\widehat{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r| > \varepsilon$ **or** $\big[t \ge M$ **and** $-\sigma(\mathrm{Acc}^{\mathrm{val}}_t - \mathrm{Acc}^{\mathrm{val}}_{t-M}) < \gamma\big]$

end for

The projected gradient flow in Proposition 2 identifies a local direction along which the reconstruction loss remains unchanged to first order while the fitted classification loss changes. We use this direction to probe classification variation within a reconstruction level set. By letting, $\mathbf{g}_R := \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{rec}}(\boldsymbol{\theta})$ and $\mathbf{g}_C := \nabla_{\boldsymbol{\theta}} \bar{\mathcal{L}}_{\mathrm{cls}}(\boldsymbol{\theta})$, we denote the projected classification gradient by $\mathbf{v} := \mathbf{P}_{\mathcal{T}} \mathbf{g}_C$ (Equation (9)). The gradient flow follows $-\mathbf{v}$, which decreases the fitted classification loss, whereas $+\mathbf{v}$ increases it.

To expose local variation empirically, we turn these infinitesimal directions into finite predictor–corrector walks along reconstruction level sets. Two issues arise experimentally. First, tangency preserves reconstruction loss only to first order: a step of length s incurs an $O(s^2)$ deviation from S_r, which can accumulate over repeated steps. We therefore follow every tangent predictor step with a correction back toward the level set. Second, an exact loss level cannot generally be maintained numerically, so we work with an ε-thickened level set $S_r^{\varepsilon} := \big\{\boldsymbol{\theta} : |\widehat{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r| \le \varepsilon\big\}$

Starting from $\boldsymbol{\theta}_0$, we construct each walk with the iterative predictor–corrector scheme in Algorithm 1. A fixed subset of the training set supplies the update gradients and local reconstruction constraint. Let $\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}$ denote its fixed-mask reconstruction loss and $r_{\mathrm{tr}} := \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}_0)$ be its value at the start. At each iteration, we compute $\mathbf{g}_R$ from this loss and $\mathbf{g}_C$ on unmasked inputs from the same subset through the fitted head $\mathbf{W}^{\star}_{\mathrm{CE}}(\boldsymbol{\phi})$, which we re-fit after each accepted encoder update. The predictor advances a distance s along $\sigma\mathbf{v}/\|\mathbf{v}\|_2$: $\sigma = -1$ decreases the fitted classification loss, whereas $\sigma = +1$ increases it. The corrector then applies damped Newton updates until $|\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r_{\mathrm{tr}}| \le \tau$, with $\tau \ll \varepsilon$.

After every accepted update, we evaluate validation reconstruction loss and linear-probe accuracy. We stop a trajectory when its validation loss leaves S_r^{ε} or when the change in validation linear-probe accuracy in the intended direction falls below a prescribed threshold. Importantly, the convergence of validation linear-probe accuracy is used only as a halting criterion to avoid spending computation after progress in the predetermined direction becomes negligible. Because the reported spread is the difference between the maximum and minimum accuracy over all visited checkpoints, the stopping condition cannot increase it; it can only omit later checkpoints and thereby make the observed spread a more conservative lower bound. Complete pseudocode is provided in Algorithms 2 and 3.

3.5 Level-Set Walks Reveal Greater Downstream Underdetermination Than Pretraining

We next use Algorithm 1 to measure how much classification underdetermination lies beyond that observed during pretraining. We first evaluate walks from matched-loss starts along the pretraining trajectories of the three decoder sizes at masking ratio 0.75. At each of three evenly spaced nominal loss levels, $\bar{r} \in \{0.463, 0.476, 0.489\}$, we select five starts for each decoder size, yielding 45 matched-loss walks. To evaluate converged models, we also initialize walks from six final checkpoints for

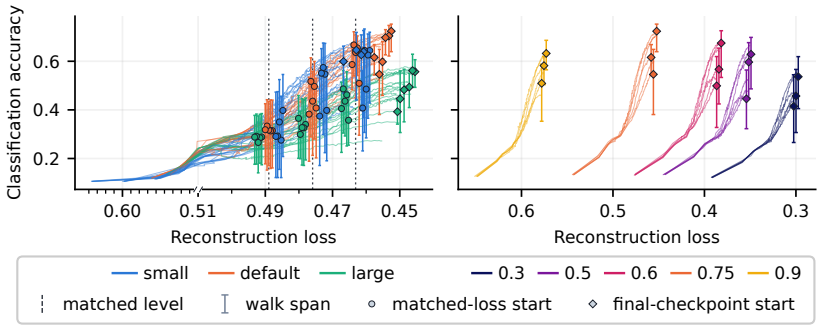

Figure 3: **Linear-probe accuracy along pretraining trajectories and level-set walks vs reconstruction loss**. Curves show pretraining trajectories; each bar spans a walk's accuracy range within $S^{\varepsilon}_{r_0}$, with its marker indicating the start. **(Left)** Decoder sizes at masking ratio 0.75. **(Right)** Default decoder at five masking ratios. Across all settings, level-set walks expose substantially wider downstream underdetermination than pretraining, both during training and at convergence.

each decoder size at masking ratio 0.75, giving 18 walks. Finally, we initialize walks from three final checkpoints at each of the five masking ratios with the default decoder size, giving 15 walks. The full hyperparameter specification is in Table 5, and an example walk is shown in Figure 7.

Across all walks, for every decoder size we observe a larger accuracy spread than the pretraining checkpoints at the same loss (Figure 3). Importantly, across nine walks the respective spread is 1.23–$2.05\times$ the pretraining spread and the walks reach 1.3–10.3 pp above the best pretraining accuracy and 3.7–15.0 pp below the worst (Table 2). The same effect holds at convergence. Across the three decoder sizes, the six final-checkpoint walks per decoder span 1.91–$2.09\times$ the final-checkpoint spread, reaching 2.9–8.0 pp above and 8.9–17.7 pp below the corresponding final-checkpoint range (Table 3). Across the five masking ratios, the three final-checkpoint walks per ratio span 1.71–$2.70\times$ the final-checkpoint spread, extending 2.9–7.0 pp above and 10.7–17.0 pp below it (Table 4). Thus, across decoder sizes and masking ratios, level-set walks expose substantially more downstream underdetermination than pretraining, both during training and at convergence.

Next, we turn our attention to tracking the cosine similarity between $\mathbf{g}_{\text{rec}}$ and $\mathbf{g}_{\text{cls}}$ and report our results in (Figure 8). Importantly, we observe that the two gradients are typically nearly orthogonal with only weak positive alignment. To show that this pattern is not specific to the selected starts, we examine seven representative pretraining trajectories: all three decoder capacities at masking ratio 0.75 and the default decoder at five masking ratios. Strikingly, the cosine remains near zero with a weak positive bias throughout training (Figure 4). Thus, across decoder sizes and masking ratios, the classification gradient retains a substantial component tangent to the reconstruction level set throughout pretraining, indicating widespread downstream underdetermination and further illustrating the insufficiency of reconstruction learning alone for selecting strong classification features.

Finally, to test whether wide underdetermination is an artifact of our ImageNet-100 pretraining sweep, we apply the same procedure to the public ImageNet-1k MAE checkpoints`facebook/vit-mae-base` and `facebook/vit-mae-large`, using a 100,000-image gradient subset and the same ε. For ViT-B/16, the walks improve linear-probe accuracy from 65.6% to 69.7% and degrade it to 4.2%, yielding a 65.5 pp spread. For ViT-L/16, they improve accuracy from 72.1% to 75.7% and degrade it to 2.3%, yielding a 73.3 pp spread. Held-out reconstruction loss remains within 1.5×10^{-3} of its starting value throughout.

3.6 A Spectral Signature of Downstream Underdetermination

Although the results above establish widespread downstream underdetermination, they leave open why it arises so consistently. We hypothesize a mismatch between the spectral structure relevant to reconstruction and classification: directions with little variance in the reconstruction targets contribute little to MSE, yet may carry substantial class-discriminative variation. We therefore ask whether level-set walks act preferentially in these directions. To investigate this question, we assign each image a fixed mask and collect all corresponding target patches. We compute their PCA basis

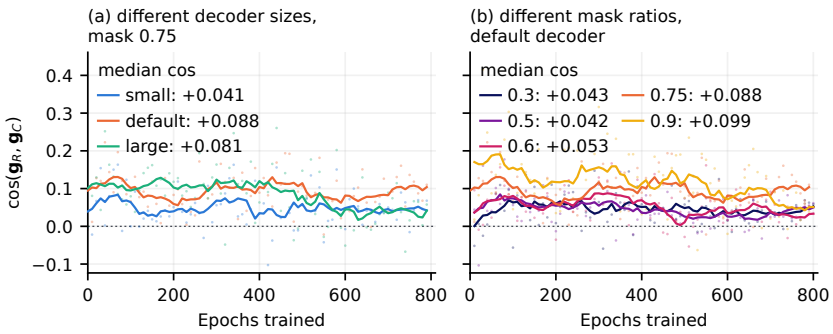

Figure 4: **Cosine between the reconstruction and classification gradients throughout pretraining**. Dots show measurements every 10 epochs; lines show 90-epoch running means. **(a)** Three decoder sizes at masking ratio 0.75. **(b)** Five masking ratios with the default decoder. All runs use learning rate 3×10^{-4} and weight decay 0.05. Across decoder sizes and masking ratios, the gradients remain nearly orthogonal and weakly positively aligned throughout pretraining.

and let $\mathbf{u}_k$ denote its eigenvectors, ordered by increasing target variance λ_k. At every pretraining checkpoint and walk step, we project the predicted masked patches onto this same basis and compute $\widehat{\lambda}_k := \mathrm{Var}\left(\mathbf{u}_k^\top \widehat{\mathbf{x}}\right)$ over the masked validation patches. The target spectrum $\{\lambda_k\}$ gives the variance available along each direction, while the prediction spectrum $\{\widehat{\lambda}_k\}$ shows how strongly each direction is expressed in the model's predictions and how this changes during pretraining and level-set walks.

During pretraining, the prediction spectrum approaches the target spectrum most closely in high-variance directions, while the low-variance tail remains strongly suppressed (Figure 9, left). This suppression is expected under MAE's squared-error objective: its Bayes-optimal predictor is the conditional mean of the target patch given the visible context. The conditional mean averages out target variation that is unpredictable from the visible patches and therefore has no greater variance than the target along any direction (Section D.1). The low-variance tail is also the last part of the spectrum to settle. Across the seven runs, its prediction variance comes within 5% of its final value only at epochs 780–790 of 800, whereas the remainder of the spectrum does so by epochs 540–600. The improving and degrading walks leave the dominant spectrum nearly unchanged and concentrate their changes in this tail (Figure 9, center and right). Across decoder capacities and masking ratios, the affected directions contain only 0.06–0.11% of the total target variance (Figure 10).

Together, these observations explain why level-set walks can substantially change downstream performance while preserving reconstruction loss. High-variance directions dominate the reconstruction objective and are learned consistently, so the walks leave this dominant reconstructive structure nearly unchanged. Instead, they act primarily on a late-learned, low-variance tail that reconstruction weakly constrains but whose alteration substantially changes classification accuracy. This separation creates the conditions for downstream underdetermination: reconstruction-based learning can constrain the dominant output structure while remaining insufficient to determine good downstream features.

4 Conclusion

This work presents an extensive empirical investigation of underdetermination in reconstruction based pretraining through the lens of the reconstruction loss level sets. Through a wide range of design choices (decoder size, masking ratio, training hyperparameter), we establish that MAE pretraining, while improving downstream accuracy, consistently underexplores its local representation space. Importantly, our level-set walk algorithm finds nearly orthogonal directions to the reconstruction gradient that can be exploited to locally improve downstream classification performance.

Limitations. Our work is limited to in-distribution downstream classification at ImageNet-100 scale, leaving to future work extending our evaluation to a broader battery of downstream tasks and larger scales. While our theoretical analysis is non-linear, it is limited to first-order and local. We hope our work will serve as guidance and inspiration for a deeper non-linear theory, further characterizing the existence of classification-aligned directions w.r.t. the masking distribution.

AI Use Statement

We used AI to generate code for plotting our results, which we carefully inspected, as well as for polishing some sections of the paper. AI was also used to check the correctness of the proof of Lemma 4 and particularly the required smoothness assumptions. The authors take full responsibility for the scientific contributions and all written content.

Ethics Statement

This work focuses on understanding underdetermination of reconstruction-based pretraining for downstream classification. The work identifies shortcomings of an established self-supervised learning paradigm in informing feature selection for downstream performance. While our work has no direct societal or ethical impact, we hope that it will inform future work aimed at improving representation learning algorithms and particularly model selection and cross-validation from pretraining.

References

Sydney Naphtali Afriat. Theory of maxima and the method of lagrange. *SIAM Journal on Applied Mathematics*, 20(3):343–357, 1971.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Randall Balestriero and Yann Lecun. How learning by reconstruction produces uninformative features for perception. In *International Conference on Machine Learning*, pp. 2566–2585. PMLR, 2024.

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

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

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 12 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 22188–22214, 2023.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *International Conference on Learning Representations*, 2026.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Marcin Przewięźlikowski, Randall Balestriero, Wojciech Jasiński, Marek Śmieja, and Bartosz Zieliński. Beyond [cls]: Exploring the true potential of masked image modeling representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23442–23452, 2025.

Bill Psomas, Dionysios Christopoulos, Eirini Baltzi, Ioannis Kakogeorgiou, Tilemachos Aravanis, Nikos Komodakis, Konstantinos Karantzalos, Yannis Avrithis, and Giorgos Tolias. Attention, please! revisiting attentive probing through the lens of efficiency. In *International Conference on Learning Representations*, volume 2026, pp. 77155–77186, 2026.

Arthur Sard. The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*, 48, 12 1942.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 19250–19286, 2022.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

APPENDIX

A FORMAL STATEMENTS

This section provides the formal version of the theoretical statements appearing in Sections 2 and 3.3, complete with their mathematical proof. We begin by stating our results in the linear regime (Section A.1) and then we conclude with our local non-linear analysis (Section A.2).

Notation. All vectors are column vectors and are denoted with $\mathbf{v} = (v_1, \dots, v_d)^\top$. Matrices are represented by $\mathbf{V} = (v_{ij})$, while sets S are denoted by upper-case Latin. For a symmetric positive semi-definite matrix, $\mathbf{M} = \mathbf{P_M} \mathbf{D_M} \mathbf{P_M}^\top$ denotes its eigendecompositon. For a vector-valued multivariate function $\mathbf{f}(\mathbf{x})$, we denote its Jacobian as $(D_\mathbf{x}\mathbf{f})_{ij} = \left(\frac{\partial f_i}{\partial x_j}\right)$. For a scalar-valued function f, we write $\nabla_\mathbf{x} f := (D_\mathbf{x} f)^\top$ for its gradient. For batched variables $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$, matrix-valued differentiation is carried out by implicit vectorization $D_\mathbf{Z} f := \frac{\partial f}{\partial \operatorname{vec}(\mathbf{Z})}$ and reshaping. Finally, unless otherwise stated, we will use the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\operatorname{tr}(\mathbf{A}^\top \mathbf{A})}$ throughout our derivations.

A.1 LINEAR AUTOENCODERS

This section studies linear encoders $\mathbf{g}_1(\mathbf{X}, \boldsymbol{\phi}) := \mathbf{V}\mathbf{X}$ and decoders $\mathbf{g}_2(\mathbf{Z}, \boldsymbol{\psi}) := \mathbf{U}\mathbf{Z}$, for $\mathbf{V} \in \mathbb{R}^{k \times d}$, $\mathbf{U} \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\phi} = \operatorname{vec}(\mathbf{V})$, $\boldsymbol{\psi} = \operatorname{vec}(\mathbf{U})$. In this setting, the classification-guided reconstruction objective of Equation (1) reads

$$\mathcal{L}(\mathbf{V}, \mathbf{U}, \mathbf{W}) = \frac{1}{2n} \|\mathbf{U}\mathbf{V}\mathbf{X} - \mathbf{X}\|_F^2 + \frac{\lambda}{2n} \|\mathbf{W}\mathbf{V}\mathbf{X} - \mathbf{Y}\|_F^2 \tag{12}$$

We begin by stating Proposition 3, providing equations for the critical points of linear autoencoders.

Proposition 3 (Linear encoder/decoder (Balestriero & Lecun (2024); Baldi & Hornik (1989)))**.** *Let*

$$\mathbf{A} := \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{Y}^\top \mathbf{Y}\right) \mathbf{X}^\top \tag{13}$$

$$\mathbf{H} := \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} \mathbf{P}_{\mathbf{X}\mathbf{X}^\top}^\top \mathbf{A} \mathbf{P}_{\mathbf{X}\mathbf{X}^\top} \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} \tag{14}$$

Then, Equation (12) *admits the critical points*

$$\mathbf{V} \textit{spans } \mathbf{P}_{\mathbf{X}\mathbf{X}^\top} \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} (\mathbf{P}_\mathbf{H})_{\cdot, I} \quad \textit{for } I = \{i_1, \dots, i_k\} \subset [d] \tag{15}$$

and is minimized by

$$\mathbf{W}^\star = \mathbf{Y}\mathbf{X}^\top \mathbf{V}^{\star\top} \left(\mathbf{V}^\star \mathbf{X}\mathbf{X}^\top \mathbf{V}^{\star\top}\right)^{-1} \tag{16}$$

$$\mathbf{U}^\star = \mathbf{X}\mathbf{X}^\top \mathbf{V}^{\star\top} \left(\mathbf{V}^\star \mathbf{X}\mathbf{X}^\top \mathbf{V}^{\star\top}\right)^{-1} \tag{17}$$

$$\mathbf{V}^\star \textit{spans } \mathbf{P}_{\mathbf{X}\mathbf{X}^\top} \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}} (\mathbf{P}_\mathbf{H})_{\cdot, 1:k} \tag{18}$$

Next, we provide close-form equations for the reconstruction and classification error, assuming we are given a critical point of Equation (12).

Proposition 4 (Close-form Error at Critical Points)**.** *In the setting of Proposition 3, let* $I = \{i_1, \dots, i_k\} \subset [d]$ *denote a set of* k *distinct eigendirections*

$$\mathbf{Q} = (\mathbf{P}_\mathbf{H})_{\cdot, I} \tag{19}$$

$$\mathbf{V} = \left(\mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{Q}\right)^\top \tag{20}$$

with the shorthand $\mathbf{B} := \mathbf{X}\mathbf{X}^\top = \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B} \mathbf{P}_\mathbf{B}^\top$. *Let* $\tilde{\mathbf{X}} = \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top \mathbf{X}$ *represented the whitened data. Then, for* $\boldsymbol{\Pi}_I = \mathbf{Q}\mathbf{Q}^\top$ *denoting the projector onto the eigendirections indexed by* I*, we have the following close-form expressions*

$$\mathcal{L}_{\text{cls}}(I) = \|\mathbf{Y}\|_F^2 - \operatorname{tr}(\mathbf{Y}\tilde{\mathbf{X}}^\top \boldsymbol{\Pi}_I \tilde{\mathbf{X}} \mathbf{Y}^\top) \tag{21}$$

$$\mathcal{L}_{\text{rec}}(I) = \operatorname{tr} \mathbf{D}_\mathbf{B} - \operatorname{tr}(\boldsymbol{\Pi}_I \mathbf{D}_\mathbf{B}) \tag{22}$$

where we made the dependence on I *explicit.*

Proof. We begin by computing the classification MSE in close form. Starting from the optimal linear probe (Equation (16)):

Step 1.

$$\mathbf{V}\mathbf{B}\mathbf{V}^\top = \mathbf{Q}^\top \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B} \mathbf{P}_\mathbf{B}^\top \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{Q} \tag{23}$$

$$= \mathbf{Q}^\top \mathbf{Q} \tag{24}$$

$$= \mathbf{I}_k \tag{25}$$

We then move on to the term $\mathbf{Y}\mathbf{X}^\top \mathbf{V}^\top \mathbf{V}\mathbf{X} - \mathbf{Y}$.

Step 2.

$$\mathbf{V}^\top \mathbf{V} = \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{\Pi}_I \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top \tag{26}$$

Step 3. Finally,

$$\mathbf{Y}\mathbf{X}^\top \mathbf{V}^\top \mathbf{V}\mathbf{X} - \mathbf{Y} = \mathbf{Y}\tilde{\mathbf{X}}^\top \mathbf{\Pi}_I \tilde{\mathbf{X}} - \mathbf{Y} \tag{27}$$

$$= \mathbf{Y}\left(\tilde{\mathbf{X}}^\top \mathbf{\Pi}_I \tilde{\mathbf{X}} - \mathbf{I}_n\right) \tag{28}$$

$$= \|\mathbf{Y}\|_F^2 - \operatorname{tr}\left(\mathbf{Y}\tilde{\mathbf{X}}^\top \mathbf{\Pi}_I \tilde{\mathbf{X}} \mathbf{Y}^\top\right) \tag{29}$$

We now turn our attention to the reconstruction MSE in Equation (12).

Step 4. Recall that $\mathbf{V}\mathbf{B}\mathbf{V}^\top = \mathbf{I}_k$. Then

$$\mathbf{U} = \mathbf{B}\mathbf{V}^\top \tag{30}$$

$$= \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B} \mathbf{P}_\mathbf{B}^\top \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{Q} \tag{31}$$

$$= \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{Q} \tag{32}$$

Step 5. We can now solve $\mathbf{U}\mathbf{V}\mathbf{X}$:

$$\mathbf{U}\mathbf{V}\mathbf{X} = \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{Q}\mathbf{Q}^\top \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B} \mathbf{P}_\mathbf{B}^\top \tag{33}$$

$$= \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{Q}\mathbf{Q}^\top \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top \tag{34}$$

Step 6.

$$\mathcal{L}_{\text{rec}} = \|\mathbf{U}\mathbf{V}\mathbf{X} - \mathbf{X}\|_F^2 \tag{35}$$

$$= \left\| \left(\mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{\Pi}_I \mathbf{D}_\mathbf{B}^{-\frac{1}{2}} \mathbf{P}_\mathbf{B}^\top - \mathbf{I}_d\right) \mathbf{X} \right\|_F^2 \tag{36}$$

$$= \|\mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \mathbf{\Pi}_I \tilde{\mathbf{X}} - \mathbf{P}_\mathbf{B} \mathbf{D}_\mathbf{B}^{\frac{1}{2}} \tilde{\mathbf{X}}\|_F^2 \tag{37}$$

$$= \|\mathbf{D}_\mathbf{B}^{\frac{1}{2}} \left(\mathbf{\Pi}_I - \mathbf{I}_d\right) \tilde{\mathbf{X}}\|_F^2 \tag{38}$$

Step 7. Finally, by developing the trace:

$$\|\mathbf{U}\mathbf{V}\mathbf{X} - \mathbf{X}\|_F^2 = \operatorname{tr}\left(\tilde{\mathbf{X}}^\top \left(\mathbf{\Pi}_I - \mathbf{I}_d\right) \mathbf{D}_\mathbf{B} \left(\mathbf{\Pi}_I - \mathbf{I}_d\right) \tilde{\mathbf{X}}^\top\right) \tag{39}$$

$$= \operatorname{tr}\left(\left(\mathbf{\Pi}_I - \mathbf{I}_d\right) \mathbf{D}_\mathbf{B} \left(\mathbf{\Pi}_I - \mathbf{I}_d\right)\right) \tag{40}$$

$$= \operatorname{tr}\left(\mathbf{\Pi}_I \mathbf{D}_\mathbf{B} \mathbf{\Pi}_I - \mathbf{\Pi}_I \mathbf{D}_\mathbf{B} - \mathbf{D}_\mathbf{B} \mathbf{\Pi}_I + \mathbf{D}_\mathbf{B}\right) \tag{41}$$

$$= \operatorname{tr}\left(\mathbf{\Pi}_I - \mathbf{D}_\mathbf{B}\right) - 2\operatorname{tr}\left(\mathbf{\Pi}_I \mathbf{D}_\mathbf{B}\right) + \operatorname{tr} \mathbf{D}_\mathbf{B} \tag{42}$$

$$= \operatorname{tr} \mathbf{D}_\mathbf{B} - \operatorname{tr} \mathbf{\Pi}_I \mathbf{D}_\mathbf{B} \tag{43}$$

concluding the proof. □

We now consider two different projectors $\mathbf{\Pi}_I, \mathbf{\Pi}_{I'}$, corresponding to two distinct index sets $I \neq I'$, both of size k. The next result provides conditions for the two different solutions, corresponding to stationary points of the linear autoencoder (Proposition 3), to admit the same reconstruction and downstream classification error.

Proposition 5 (No Underdetermination)**.** *Let* $A = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{Y}^\top\mathbf{Y}\right)\mathbf{X}^\top$ *and* $\mathbf{H} = \mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}}\mathbf{P}_{\mathbf{X}\mathbf{X}^\top}^\top\mathbf{A}\mathbf{P}_{\mathbf{X}\mathbf{X}^\top}\mathbf{D}_{\mathbf{X}\mathbf{X}^\top}^{-\frac{1}{2}}$ *(Equation* (13)*). Let* $d_i^{\mathbf{B}} := (\mathbf{D}_{\mathbf{B}})_{ii}$*. Similarly, denote by* $d_i^{\mathbf{H}}$ *the* i*-th eigenvalue of* H*.*

For two index sets $I \neq I'$*, assume* $\mathcal{L}_{\text{rec}}(I) = \mathcal{L}_{\text{rec}}(I')$*. Then,*

$$\mathcal{L}_{\text{rec}}(I) = \mathcal{L}_{\text{rec}}(I') \iff \sum_{i\in I} d_i^{\mathbf{B}} = \sum_{i\in I'} d_i^{\mathbf{B}} \tag{44}$$

Furthermore, the respective classification errors coincide if

$$\begin{cases} \mathcal{L}_{\text{cls}}(I) = \mathcal{L}_{\text{cls}}(I') \iff \sum\limits_{i\in I} d_i^{\mathbf{H}} = \sum\limits_{i\in I'} d_i^{\mathbf{H}} & (\text{if } \lambda > 0) \\ \mathcal{L}_{\text{cls}}(I) = \mathcal{L}_{\text{cls}}(I') \iff \sum\limits_{i\in I} \frac{\|\mathbf{Y}\mathbf{X}^\top\tilde{\mathbf{q}}_i\|_F^2}{d_i^{\mathbf{B}}} = \sum\limits_{i\in I'} \frac{\|\mathbf{Y}\mathbf{X}^\top\tilde{\mathbf{q}}_i\|_F^2}{d_i^{\mathbf{B}}} & (\text{if } \lambda = 0) \end{cases} \tag{45}$$

where $\tilde{\mathbf{q}}_i := \mathbf{Q}_i^\top\mathbf{D}_{\mathbf{B}}^{-\frac{1}{2}}\mathbf{P}_{\mathbf{B}}^\top$.

Proof. We begin by expanding

$$\mathbf{A} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{Y}^\top\mathbf{Y})\mathbf{X}^\top \tag{46}$$

$$= \mathbf{P}_{\mathbf{B}}\mathbf{D}_{\mathbf{B}}^2\mathbf{P}_{\mathbf{B}}^\top + \lambda\mathbf{X}\mathbf{Y}^\top\mathbf{Y}\mathbf{X}^\top \tag{47}$$

Next, we define

$$\mathbf{S} := \tilde{\mathbf{X}}\mathbf{Y}^\top\mathbf{Y}\tilde{\mathbf{X}}^\top \tag{48}$$

so that we can rewrite

$$\mathbf{H} = \mathbf{D}_{\mathbf{B}} + \lambda\mathbf{S} \tag{49}$$

Case $\lambda \neq 0$**.** We have $\mathbf{S} = \frac{1}{\lambda}(\mathbf{H} - \mathbf{D}_{\mathbf{B}})$.

Next, we notice that the classification error can be expressed as

$$\|\mathbf{Y}\|_F^2 - \operatorname{tr}(\mathbf{Y}\tilde{\mathbf{X}}^\top\mathbf{\Pi}_I\tilde{\mathbf{X}}\mathbf{Y}^\top) = \|\mathbf{Y}\|_F^2 - \operatorname{tr}(\mathbf{\Pi}_I\mathbf{S}) \tag{50}$$

$$= \|\mathbf{Y}\|_F^2 - \frac{1}{\lambda}\operatorname{tr}(\mathbf{\Pi}_I\mathbf{D}_{\mathbf{B}}) - \frac{1}{\lambda}\operatorname{tr}(\mathbf{\Pi}_I\mathbf{H}) \tag{51}$$

$$= \|\mathbf{Y}\|_F^2 + \frac{1}{\lambda}[\operatorname{tr}\mathbf{D}_{\mathbf{B}} - \mathcal{L}_{\text{rec}}(I)] - \frac{1}{\lambda}\operatorname{tr}(\mathbf{\Pi}_I\mathbf{H}) \tag{52}$$

where the last derivation used Equation (43). Finally, by using $\mathcal{L}_{\text{rec}}(I) = \mathcal{L}_{\text{rec}}(I')$, we can deduce $\mathcal{L}_{\text{cls}}(I) = \mathcal{L}_{\text{cls}}(I') \iff \operatorname{tr}(\mathbf{\Pi}_I\mathbf{H}) = \operatorname{tr}(\mathbf{\Pi}_{I'}\mathbf{H})$, as desired.

Case $\lambda = 0$**.** We begin by noting that from $\mathbf{H} = \mathbf{D}_{\mathbf{B}} + \lambda\mathbf{S}$, then $\mathbf{H} = \mathbf{D}_{\mathbf{B}}$ and hence $\mathbf{Q} = \mathbf{P}_{\mathbf{H}} = \mathbf{I}_d$. But then

$$\mathbf{\Pi}_I = \mathbf{Q}\mathbf{Q}^\top = \operatorname{diag}(\mathbf{1}_{[i\in I]}) \tag{53}$$

Lastly, by expanding the component of the classification error that depends on I

$$\operatorname{tr}(\mathbf{Y}\tilde{\mathbf{X}}^\top\mathbf{\Pi}_I\tilde{\mathbf{X}}\mathbf{Y}^\top) = \operatorname{tr}(\mathbf{Q}^\top\tilde{\mathbf{X}}\mathbf{Y}^\top\mathbf{Y}\tilde{\mathbf{X}}^\top\mathbf{Q}) \tag{54}$$

$$= \operatorname{tr}(\mathbf{Q}^\top\mathbf{D}_{\mathbf{B}}^{-\frac{1}{2}}\mathbf{P}_{\mathbf{B}}^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y}\mathbf{X}^\top\mathbf{P}_{\mathbf{B}}\mathbf{D}_{\mathbf{B}}^{-\frac{1}{2}}\mathbf{Q}) \tag{55}$$

$$= \sum_{i\in\mathbf{I}} \frac{\tilde{\mathbf{q}}_i^\top\mathbf{X}\mathbf{Y}^\top\mathbf{Y}\mathbf{X}\tilde{\mathbf{q}}_i}{d_i^{\mathbf{B}}} \tag{56}$$

yielding the desired result. □

Lastly, for two distinct critical points $I \neq I'$, we use the close form error equations to express the difference in classification error at equal reconstruction error. The following proposition states the formal version of Proposition 1.

Proposition 6 (Underdetermination in linear autoencoders)**.** *Consider two distinct index sets $I \neq I'$ and assume that the corresponding encoder features admit the same reconstruction error $\mathcal{L}_{\mathrm{rec}}(I) = \mathcal{L}_{\mathrm{rec}}(I')$, as given by Equation* (44)*.*

Define $\mathbf{S}$ *as in Equation* (48)*. Since* $\mathbf{S} \succeq 0$*, let* $s_1 \geq \ldots \geq s_d \geq 0$ *denote its eigenvalues. Then*

$$|\mathcal{L}_{\mathrm{cls}}(I) - \mathcal{L}_{\mathrm{cls}}(I')| = \left| \sum_{i \in I} d_i^{\mathbf{H}} - \sum_{i \in I'} d_i^{\mathbf{H}} \right| \qquad (\lambda \neq 0) \tag{57}$$

$$|\mathcal{L}_{\mathrm{cls}}(I) - \mathcal{L}_{\mathrm{cls}}(I')| \leq \left| \sum_{j=1}^{k} s_j - \sum_{j=d-k+1}^{d} s_j \right| \qquad (\lambda = 0) \tag{58}$$

Proof. At equal reconstruction error, the difference in classification errors of solutions I and I' can be obtained directly from Proposition 5.

Case $\lambda \neq 0$**.** Equation (45) gives

$$|\mathcal{L}_{\mathrm{cls}}(I) - \mathcal{L}_{\mathrm{cls}}(I')| = |\operatorname{tr}\left((\mathbf{\Pi}_I - \mathbf{\Pi}_{I'})\,\mathbf{H}\right)| = \left| \sum_{i \in I} d_i^{\mathbf{H}} - \sum_{i \in I'} d_i^{\mathbf{H}} \right| \tag{59}$$

which is the difference in selected eigenvalue mass in $\mathbf{H}$.

Case $\lambda = 0$**.** In this case, Equation (45) yields

$$|\mathcal{L}_{\mathrm{cls}}(I) - \mathcal{L}_{\mathrm{cls}}(I')| = |\operatorname{tr}\left((\mathbf{\Pi}_I - \mathbf{\Pi}_{I'})\,\mathbf{S}\right)| \tag{60}$$

By the Ky Fan principle, if I and I' range uniformly over choices of k indices

$$\sum_{j=d-k+1}^{d} s_j \leq \operatorname{tr}(\mathbf{\Pi}_I \mathbf{S}) \leq \sum_{j=1}^{k} s_j \tag{61}$$

Combining the bounds for I and I' and taking the absolute value gives the result. □

A.2 Locally Characterizing Underdetermination

We begin by stating our theoretical assumptions, which are collected in Section A.2.1. We then derive the reduced loss $\bar{\mathcal{L}}_{\mathrm{cls}}$ (Equation (7)) and present its level-set projected gradient flow (Proposition 2). Finally, we describe the construction behind the reduction $\bar{\mathcal{L}}_{\mathrm{rec}}$ and conclude by proving Theorem 1.

A.2.1 Assumptions

To make our analysis well-posed, for a level set S_r we assume that $r > 0$ is a regular value almost everywhere according to Sard's Lemma, which we state for scalar functions below.

Lemma 2 (Sard's Lemma (Sard, 1942))**.** *Let $f : \mathbb{R}^p \to \mathbb{R} \in \mathcal{C}^t(\mathbb{R}^p)$, for $t \geq p$. Let $C = \{\theta \in \mathbb{R}^p : \nabla_\theta f = \mathbf{0}\}$. Then $f(C)$ has zero Lebesgue measure in the codomain $\mathbb{R}$.*

Furthermore, in Section A.2.2 and Section A.2.4 we will make use of the following statement, which allows for writing the gradients of the reduced losses $\bar{\mathcal{L}}_{\mathrm{cls}}$ and $\bar{\mathcal{L}}_{\mathrm{rec}}$ as a function of $\boldsymbol{\phi}$ only.

Lemma 3 (Envelope Theorem (Afriat, 1971))**.** *Let $f(\mathbf{u}(\mathbf{z}), \mathbf{z}) : \mathbb{R}^m \times \mathbb{R}^q \to \mathbb{R}$ be $\mathcal{C}^1$ jointly in $\mathbf{u}$ and $\mathbf{z}$. For fixed $\mathbf{z}$, let $\mathbf{u}^\star(\mathbf{z}) \in \operatorname{argmin}_{\mathbf{u}} \mathbf{f}(\mathbf{u}, \mathbf{z})$ s.t.*

$$D_{\mathbf{u}} f(\mathbf{u}^\star(\mathbf{z}), \mathbf{z}) = 0 \tag{62}$$

i.e. $\mathbf{u}^\star$ is a stationary point for f. Finally, introduce the value function $v(\mathbf{z}) := f(\mathbf{u}^\star(\mathbf{z}), \mathbf{z})$. Then, whenever the mapping $\mathbf{z} \mapsto \mathbf{u}^\star(\mathbf{z})$ is locally $\mathcal{C}^1$, the value function $v(\mathbf{z})$ is differentiable with

$$D_{\mathbf{z}} v(\mathbf{z}) = D_{\mathbf{z}} f(\mathbf{u}, \mathbf{z})\big|_{\mathbf{u}=\mathbf{u}^\star(\mathbf{z})} \tag{63}$$

Equivalently, when computing the total derivative in Equation (63)*, the term $D_{\mathbf{u}} f\, D_{\mathbf{z}} \mathbf{u}^\star$ vanishes due to Equation* (62)*, which also proves the lemma.*

Finally, we will make the following assumptions:

1. The encoder $\boldsymbol{\phi} \mapsto \mathbf{g}_1(\cdot, \boldsymbol{\phi})$ and decoder $\boldsymbol{\psi} \mapsto \mathbf{g}_2(\cdot, \boldsymbol{\psi})$ are sufficiently smooth, so that $\mathcal{L}_{\text{rec}} \in C^t$ with $t \geq p$ according to Lemma 2.
2. The batched embeddings $\mathbf{Z}(\boldsymbol{\phi})$ have locally constant rank near $\boldsymbol{\phi}$.
3. The differential operators $D_{\boldsymbol{\phi}}\mathbf{Z}$ and $D_{\mathbf{Z}}\mathbf{g}_2$ have locally constant rank near $\boldsymbol{\phi}$. Together with the above assumptions, this ensures that $\bar{\mathcal{L}}_{\text{cls}}$ and $\bar{\mathcal{L}}_{\text{rec}}$ smoothly depend on $\boldsymbol{\phi}$.

A.2.2 Optimal Linear Probe

Next, we restate Equation (7) and detail its derivation:

$$\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\phi}) := \mathcal{L}_{\text{cls}}(\boldsymbol{\phi}, \mathbf{W}^\star) \quad \text{for } \mathbf{W}^\star \in \operatorname*{argmin}_{\mathbf{W}} \mathcal{L}_{\text{cls}}(\boldsymbol{\phi}, \mathbf{W}) \tag{64}$$

In the above equation, the Ordinary Least-Squares (OLS) solution can be computed in closed form as $\mathbf{W}^\star = \mathbf{Y}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top)^\dagger$, where $\mathbf{M}^\dagger$ denotes the Moore–Penrose pseudoinverse. By considering the singular value decomposition $\mathbf{Z} = \mathbf{U}_\mathbf{Z}\mathbf{\Sigma}_\mathbf{Z}\mathbf{V}_\mathbf{Z}^\top$, with $\mathbf{\Sigma}_\mathbf{Z} = \text{diag}(\sigma_1(\mathbf{Z}), \ldots, \sigma_n(\mathbf{Z}))$, the classification error can be written in close form as (Hotelling, 1936):

$$\begin{aligned}
\mathcal{L}_{\text{cls}}(\mathbf{W}^\star, \boldsymbol{\phi}) &= \frac{1}{2n}\|\mathbf{Y}(\mathbf{I}_n - \mathbf{Z}^\dagger\mathbf{Z})\|_F^2 && (65)\\
&= \frac{1}{2n}\|\mathbf{Y}\|_F^2 - \frac{1}{2n}\|\mathbf{Y}\mathbf{V}_\mathbf{Z}\mathbf{\Sigma}_\mathbf{Z}^\dagger\mathbf{\Sigma}_\mathbf{Z}\|_F^2 && (66)\\
&= \frac{1}{2n}\|\mathbf{Y}\|_F^2 - \frac{1}{2n}\sum_{i,j}\mathbf{1}\{\sigma_i(\mathbf{Z}) > 0\}\langle(\mathbf{V}_\mathbf{Z})_i, (\mathbf{V}_\mathbf{Y})_j\rangle^2\,\sigma_j(\mathbf{Y})^2 && (67)
\end{aligned}$$

thereby expressing the task variance that is not linearly readable from $\mathbf{Z}$.

Remark. If $\mathcal{L}_{\text{cls}}$ is $\mathcal{C}^1$ in $\mathbf{Z}$, since $\mathcal{L}_{\text{cls}}(\mathbf{W}, \mathbf{Z})$ is convex and quadratic in $\mathbf{W}$, then $\mathbf{W}^\star = \mathbf{Y}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top)^\dagger$ satisfies $D_\mathbf{W}\mathcal{L}_{\text{cls}} = 0$, and the fitted values $\mathbf{W}^\star\mathbf{Z} = \mathbf{Y}\mathbf{Z}^\dagger\mathbf{Z}$ are independent of the choice of minimizer. Wherever $\text{rank}(\mathbf{Z})$ is locally constant, Lemma 3 gives

$$\mathbf{G}_{\text{cls}} := \nabla_\mathbf{Z}\mathcal{L}_{\text{cls}}\big|_{\mathbf{W}^\star} = \frac{1}{n}(\mathbf{W}^\star)^\top(\mathbf{W}^\star\mathbf{Z} - \mathbf{Y}) \tag{68}$$

A.2.3 Projected Gradient Flow

We are now ready to prove Proposition 2.

Proposition 2 (First-order stationarity conditions)**.** *Let* $\mathcal{T}_{\boldsymbol{\theta}}S_r = \{\mathbf{u} \in \mathbb{R}^p : \langle\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}, \mathbf{u}\rangle = 0\}$. *Solving Equation* (8) *is equivalent at first order to the projected gradient flow*

$$\dot{\boldsymbol{\theta}} = -\mathbf{P}_\mathcal{T}\nabla_{\boldsymbol{\theta}}\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) \qquad \textit{for} \qquad \mathbf{P}_\mathcal{T} := \mathbf{I}_p - \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}\,\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}^\top}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}\|_2^2} \tag{9}$$

Particularly, the flow is stationary for $\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) = 0$ *or* $\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) \in \text{span}(\mathcal{L}_{\text{rec}}(\boldsymbol{\theta}))$ *and* $\nabla_\psi\mathcal{L}_{\text{rec}} = 0$.

Proof. We begin by writing the constrained optimization problem:

$$\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta}) + \mu(\mathcal{L}_{\text{rec}} - r) = 0 \tag{69}$$

Solving the Lagrangian multiplier yields

$$\mu = -\frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}^\top\nabla_{\boldsymbol{\theta}}\bar{\mathcal{L}}_{\text{cls}}(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}\|^2} \tag{70}$$

substituting in the constrained formulation and regrouping gives

$$\mathbf{P}_\mathcal{T} = \mathbf{I}_d - \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}^\top}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{rec}}\|^2} \tag{71}$$

as claimed. Lastly, by imposing stationarity, $\nabla_\psi\mathcal{L}_{\text{rec}} = 0$, concluding the proof. □

A.2.4 LOCALLY OPTIMAL DECODER

To represent the reconstruction loss for the optimal local decoder, $\bar{\mathcal{L}}_{\text{rec}}$, in the setting of Lemma 3, for each $\boldsymbol{\phi} \in \mathbb{R}^{p_1}$, we consider $\boldsymbol{\psi}^\star(\boldsymbol{\phi}) \in \operatorname{argmin}_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi})$, as detailed in the following lemma.

Lemma 4 (Decoder reduction)**.** *Let $(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and*

$$H_{\boldsymbol{\psi}\boldsymbol{\psi}} := D_{\boldsymbol{\psi}} \nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}, \quad H_{\boldsymbol{\psi}\boldsymbol{\phi}} := D_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}, \quad H_{\boldsymbol{\phi}\boldsymbol{\phi}} := D_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{rec}}, \tag{72}$$

all evaluated at $(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0)$, with $H_{\boldsymbol{\phi}\boldsymbol{\psi}} = H_{\boldsymbol{\psi}\boldsymbol{\phi}}^\top$. If

$$\nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) = 0 \tag{73}$$

and

$$H_{\boldsymbol{\psi}\boldsymbol{\psi}} \succ 0 \tag{74}$$

then $\exists\, U, V$ respectively neighbourhoods of $\boldsymbol{\phi}_0$ and $\boldsymbol{\psi}_0$ on which

(i) $\boldsymbol{\psi}^\star(\boldsymbol{\phi}_0) = \boldsymbol{\psi}_0$ is the unique map on U that minimizes $\mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \cdot)$ on V.

(ii) $\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi})\big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^\star(\boldsymbol{\phi})}$

(iii) The intersection $\mathcal{T}_{\boldsymbol{\phi}_0}(S_r \cap U) = \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi})^\perp$ is well defined.

Proof. We begin by applying the implicit function theorem to $F : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}^{p_2}$, s.t. $F(\boldsymbol{\phi}, \boldsymbol{\psi}) := \nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi})$.

Step 1. By assumption (Section A.2.1), F is C^1 near $(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0)$. Furthermore, by stationarity, $F(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) = 0$, and its partial Jacobian in $\boldsymbol{\psi}$ is $D_{\boldsymbol{\psi}} F(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) = H_{\boldsymbol{\psi}\boldsymbol{\psi}}$, which is assumed invertible.

Then, the implicit function theorem yields neighbourhoods $U_1 \ni \boldsymbol{\phi}_0$, $V_1 \ni \boldsymbol{\psi}_0$ and a unique C^1 map $\boldsymbol{\psi}^\star : U_1 \to V_1$ with $\boldsymbol{\psi}^\star(\boldsymbol{\phi}_0) = \boldsymbol{\psi}_0$ and

$$F(\boldsymbol{\phi}, \boldsymbol{\psi}^\star(\boldsymbol{\phi})) = 0 \qquad \text{for all } \boldsymbol{\phi} \in U_1, \tag{75}$$

such that the zero set of F in $U_1 \times V_1$ is exactly the graph of $\boldsymbol{\psi}^\star$.

Finally, by the chain rule $D_{\boldsymbol{\phi}} F + D_{\boldsymbol{\psi}} F \, D_{\boldsymbol{\phi}} \boldsymbol{\psi}^\star = 0$, yielding

$$D_{\boldsymbol{\phi}} \boldsymbol{\psi}^\star = -H_{\boldsymbol{\psi}\boldsymbol{\psi}}^{-1} H_{\boldsymbol{\psi}\boldsymbol{\phi}} \tag{76}$$

Step 2. We now prove that $\boldsymbol{\psi}^\star(\boldsymbol{\phi})$ is a minimizer of $\bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi})$. This is done by taking $U \subseteq U_1$ and $V \subseteq V_1$ convex so that $D_{\boldsymbol{\psi}} \nabla_{\boldsymbol{\psi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi}) \succ 0$ for all $(\boldsymbol{\phi}, \boldsymbol{\psi}) \in U \times V$. Then, by continuity of $\mathcal{L}_{\text{rec}}(\boldsymbol{\phi})$, its critical point $\boldsymbol{\psi}^\star(\boldsymbol{\phi})$ is its unique minimizer on V.

Step 3. Since $\bar{\mathcal{L}}_{\text{rec}} = \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\psi}^\star(\boldsymbol{\phi}))$ is C^1 on U, then Lemma 3 guarantees that

$$\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{rec}}(\boldsymbol{\phi}, \boldsymbol{\phi})\big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^\star(\boldsymbol{\phi})} \tag{77}$$

proving point (ii).

Finally, it remains to prove that the restriction of $\mathcal{T}_{(\boldsymbol{\phi},\boldsymbol{\psi})} S_r$ to $\mathcal{T}_{\boldsymbol{\phi}_0}(S_r \cap U)$ is well defined. To do so, it is sufficient to note that, by Lemma 2 and by $\bar{\mathcal{L}}_{\text{rec}}(\boldsymbol{\phi}_0) = r$, r is a regular value on U. Since $\|\bar{\mathcal{L}}_{\text{rec}}\| \neq 0$ on U, then the projection to $\mathcal{T}_{\boldsymbol{\phi}_0}(S_r \cap U)$ is well defined, concluding the proof. □

A.2.5 FREE ENCODER DIRECTIONS

Finally, we state the formal version of Theorem 1.

Theorem 5 (Free encoder directions)**.** *Let the encoder $\boldsymbol{\phi} \mapsto \mathbf{g}_1(\cdot, \boldsymbol{\phi})$ and decoder $\boldsymbol{\psi} \mapsto \mathbf{g}_2(\cdot, \boldsymbol{\psi})$ be sufficiently smooth (so that $\mathcal{L}_{\text{rec}} \in C^t$ with $t \geq p_1$, as required by Lemma 2), and fix $\boldsymbol{\phi} \in S_r$. Consider the level-set normal*

$$\nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\text{rec}} = \frac{1}{n} \left(D_{\boldsymbol{\phi}} \mathbf{Z}\right)^\top \left(D_{\mathbf{Z}} \mathbf{g}_2\right)^\top \operatorname{vec}\big(\hat{\mathbf{X}} - \mathbf{X}\big)\big|_{\boldsymbol{\psi}^\star} \tag{78}$$

Assume that:

1. $\bar{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\phi})$ *is given according to Lemma 4.*

2. $\mathbf{u} \in \mathbb{R}^{p_1}$ *with* $\mathbf{u} \neq \ker(D_{\boldsymbol{\phi}}\mathbf{Z})$.

Then, $\mathbf{u} \in \mathcal{T}_{\boldsymbol{\phi}} S_r^{\mathrm{enc}}$*, i.e.* $\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\mathrm{rec}}, \mathbf{u} \rangle = 0$*, if and only if*

$$D_{\mathbf{Z}}\mathbf{g}_2 \, D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u} \;\perp\; \mathrm{vec}(\hat{\mathbf{X}} - \mathbf{X}) \quad \text{(error redistribution);} \tag{79}$$

and each of the following implies Equation (79)*:*

(i) $D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u} \in \ker\left(D_{\mathbf{Z}}\mathbf{g}_2\right)$ (decoder invariant)*;*

(ii) $\exists$ *a sequence* $\delta_k \downarrow 0$ *with* $\bar{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\phi} + \delta_k \mathbf{u}) = r$ *for all* k (decoder expressible).

Proof. Taking the adjoint twice in (10),

$$\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\mathrm{rec}}, \mathbf{u} \rangle = \frac{1}{n} \big\langle \mathrm{vec}(\hat{\mathbf{X}} - \mathbf{X}),\, D_{\mathbf{Z}}\mathbf{g}_2 \, D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u} \big\rangle \tag{80}$$

giving Equation (79).

(i) If $D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u} \in \ker(D_{\mathbf{Z}}\mathbf{g}_2)$, then $D_{\mathbf{Z}}\mathbf{g}_2 D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u} = \mathbf{0}$, which is orthogonal to every vector, so Equation (79) holds.

(ii) Since $\bar{\mathcal{L}}_{\mathrm{rec}}$ is differentiable at $\boldsymbol{\phi}$ and $\bar{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\phi}) = r$,

$$\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\mathrm{rec}}, \mathbf{u} \rangle = \lim_{k \to \infty} \frac{\bar{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\phi} + \delta_k \mathbf{u}) - r}{\delta_k} = 0, \tag{81}$$

every quotient being zero by hypothesis; then Equation (79) follows by Equation (80). □

Lastly, we note that Equation (79), if written column-wise as

$$\frac{1}{n} \sum_{i=1}^{n} \langle \boldsymbol{\rho}_i, \xi_i(\mathbf{u}) \rangle = 0 \tag{82}$$

for $\boldsymbol{\rho}_i = \hat{\mathbf{x}}_i - \mathbf{x}_i$ and $\xi_i(\mathbf{u})$ being the induced perturbation on the output $\hat{\mathbf{x}}_i$, allows to distinguish two separate subcases, which we discuss in the following corollary.

Corollary 5.1 (Residual-reallocating directions)**.** *In the setting of Theorem 1, let* $\boldsymbol{\rho}_i := \hat{\mathbf{x}}_i - \mathbf{x}_i$ *and let* $\boldsymbol{\eta}_i(\mathbf{u}) \in \mathbb{R}^d$ *be defined by* $\mathrm{vec}\left(\boldsymbol{\eta}(\mathbf{u})\right) := D_{\mathbf{Z}}\mathbf{g}_2 \, D_{\boldsymbol{\phi}}\mathbf{Z}\,\mathbf{u}$*. Then*

$$\left\langle \nabla_{\boldsymbol{\phi}} \bar{\mathcal{L}}_{\mathrm{rec}}, \mathbf{u} \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \left\langle \boldsymbol{\rho}_i, \boldsymbol{\eta}_i(\mathbf{u}) \right\rangle \tag{83}$$

so $\mathbf{u} \in \mathcal{T}_{\boldsymbol{\phi}} S_r^{\mathrm{enc}}$ *if and only if the right-hand side vanishes. Particularly, this holds under the following subcases, in which* $\boldsymbol{\eta}_i(\mathbf{u}) \neq \mathbf{0}$ *for some* i*:*

(a) (per-sample rotation) $\langle \boldsymbol{\rho}_i, \boldsymbol{\eta}_i(\mathbf{u}) \rangle = 0$ *for every* i*, so each* $\|\boldsymbol{\rho}_i\|$ *is unchanged to first order;*

(b) (cross-sample cancellation) $\sum_i \langle \boldsymbol{\rho}_i, \boldsymbol{\eta}_i(\mathbf{u}) \rangle = 0$ *with individual terms nonzero, so reconstruction error is redistributed across samples at fixed mean.*

Proof. Direct computation, by noticing that cases (a) and (b) are instances of the RHS vanishing. □

B Experimental Setup and Pretraining Curves

B.1 Dataset, Architecture, and Pretraining

Dataset and preprocessing. ImageNet-100 consists of the first 100 WordNet-ordered classes in the ILSVRC-2012 training archive, with the official validation images for those classes. This yields 129,395 training images and 5,000 validation images. During pretraining, images are randomly resized and cropped to 224×224 with crop scale in $[0.2, 1.0]$, followed by a horizontal flip with probability 0.5. For validation and probe-feature extraction, we use a deterministic resize to 256 pixels followed by a 224×224 center crop. All inputs use ImageNet mean and standard-deviation normalization.

Architecture. The encoder is ViT-B/16, initialized from scratch. We vary decoder capacity using the three configurations in Table 1; width, depth, and attention-head count vary jointly.

Table 1: Decoder configurations used with the same ViT-B/16 encoder.

Decoder	Embedding dimension	Blocks	Attention heads
Small	256	4	8
Default	512	8	16
Large	768	12	16

Pretraining and hyperparameter sweep. Pretraining uses the MAE objective with per-patch normalized pixel targets (He et al., 2022). At the standard masking ratio of 0.75, we sweep learning rates in $\{10^{-4}, 3 \times 10^{-4}, 6 \times 10^{-4}\}$ and weight decays in $\{0, 0.01, 0.05, 0.15, 0.3, 0.6\}$ for each of the three decoder configurations, giving $3 \times 3 \times 6 = 54$ runs. To vary the pretext task, we additionally train the default decoder at masking ratios 0.3, 0.5, 0.6, and 0.9. At each masking ratio, we fix the learning rate at 3×10^{-4} and sweep the same six weight decays, adding $4 \times 6 = 24$ runs for 78 runs in total. Following the MAE optimization recipe, we train each configuration for 800 epochs using AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, 40 epochs of linear learning-rate warmup, and subsequent cosine decay. We reduce the batch size to 256 to fit our GPU memory budget and use bfloat16 autocasting. The random seed is fixed to 0 throughout.

B.2 Evaluation and Pretraining Curves

Evaluation. Validation reconstruction uses each run's training masking ratio, with masks fixed across checkpoints to provide a consistent objective for checkpoint comparisons and level-set walks. Every 10 pretraining epochs, we fit a linear probe to the frozen encoder and evaluate validation accuracy. The representation used for classification is the mean of the patch-token embeddings from the encoder's final layer, excluding the class token. Following the MAE linear-probing recipe (He et al., 2022), we train each probe by minimizing cross-entropy on the training set with LARS for 90 epochs, using a base learning rate of 0.1, 10 warmup epochs, momentum 0.9, and no weight decay. We reduce the batch size to 4096 to fit our GPU memory budget and replace random resized crops with deterministic center crops. The latter allows features to be reused during probe fitting, reducing the cost of frequent linear probing along level-set walks. Features are extracted in bfloat16 and converted to float32 for probe fitting, with the random seed fixed to 0. Pretraining curves for the 54-run decoder and optimizer sweep at masking ratio 0.75 are shown in Figure 5; the corresponding curves for the 24 additional masking-ratio runs are shown in Figure 6.

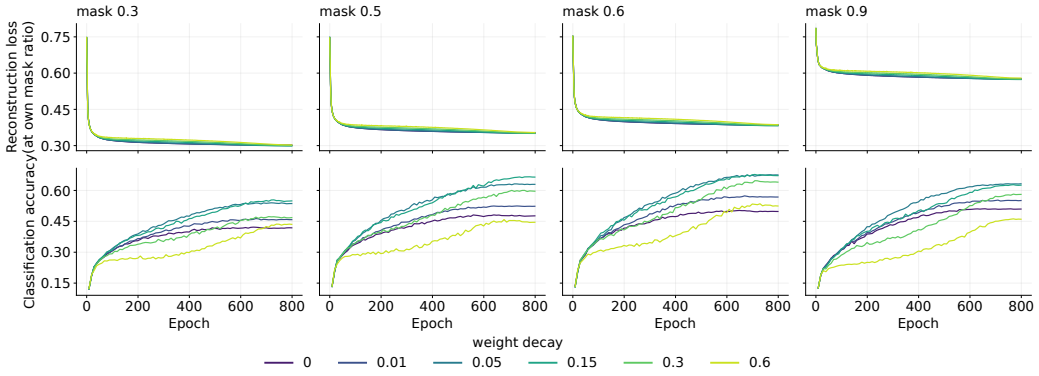

Figure 6: **Pretraining trajectories across masking ratios** for the default decoder at learning rate 3×10^{-4}. Each column shows one masking ratio, with held-out reconstruction loss evaluated at that ratio (top) and linear-probe accuracy (bottom) versus epoch; colors distinguish the six weight decays. **Across masking ratios, reconstruction loss decreases and linear-probe accuracy generally increases during training, while feature quality remains sensitive to weight decay.**

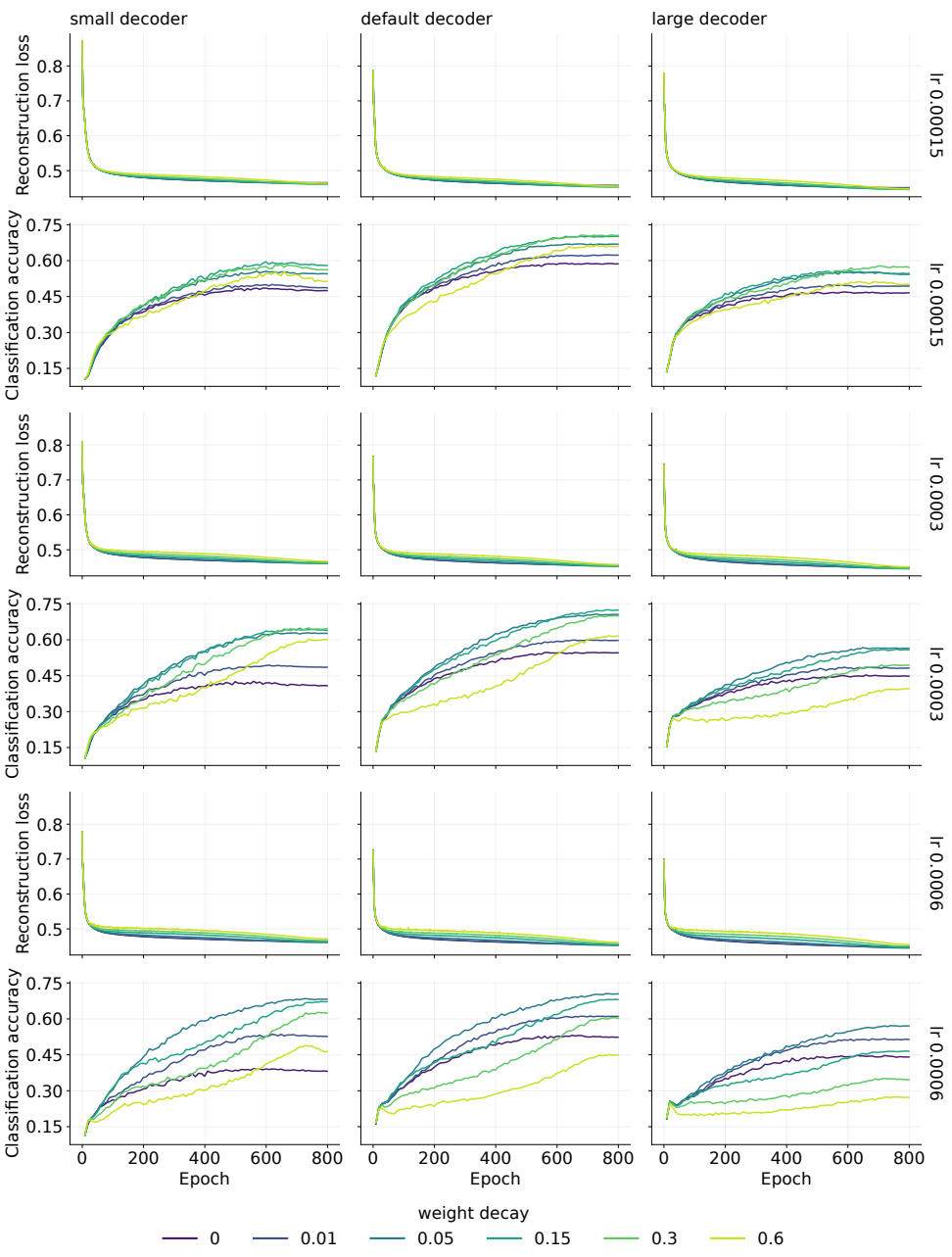

Figure 5: **Pretraining trajectories across decoder sizes and optimizer settings at masking ratio** 0.75. Each learning-rate block shows held-out reconstruction loss (top) and linear-probe accuracy (bottom) versus epoch for the small, default, and large decoders; colors distinguish the six weight decays. Level-set-walk starts are selected from the learning-rate-3 $\times 10^{-4}$ block. Over the final 100 epochs, 44 of 54 runs change in accuracy by less than 1 pp; the remaining ten all use weight decay 0.15 or higher. **Linear-probe accuracy has largely stabilized by 800 epochs except under the strongest weight decay.**

C Level-Set Walk Results and Details

C.1 Accuracy Spreads Along Level Sets

Table 2 reports the matched-loss comparisons along pretraining, while Tables 3 and 4 report the final-checkpoint comparisons across decoder sizes and masking ratios, respectively. Figure 7 illustrates the improving and degrading directions of a walk from a converged checkpoint.

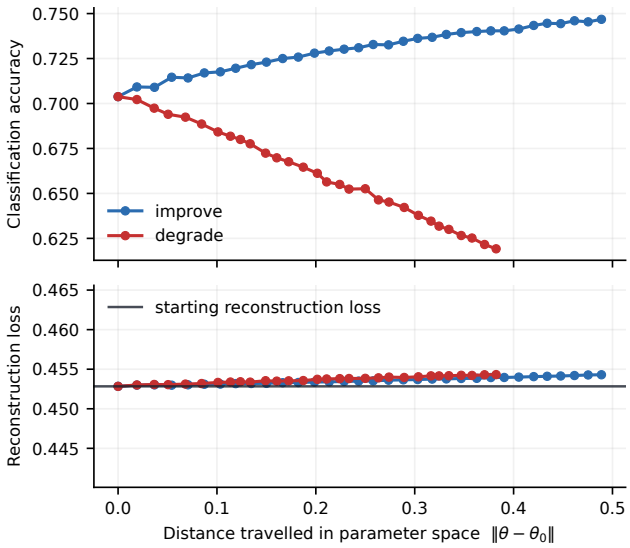

Figure 7: **Example improving and degrading level-set walks from a converged MAE checkpoint**. Both walks start from the 800-epoch checkpoint of the default-decoder run at masking ratio 0.75, learning rate 3×10^{-4}, and weight decay 0.05. The panels show linear-probe accuracy (top) and held-out reconstruction loss (bottom) versus distance $\|\theta - \theta_0\|$ from the start. The improving walk raises accuracy from 70.4% to 74.7% in 32 steps, while the degrading walk lowers it to 61.9% in 29 steps, yielding a 12.8 pp spread. Held-out reconstruction loss remains within 1.5×10^{-3} of its starting value of 0.4528. **Even at convergence, nearly unchanged reconstruction loss permits substantially different downstream accuracies.**

Table 2: **Linear-probe accuracy ranges of pretraining and matched-loss walks** across three decoder sizes and three nominal loss levels $\bar{r}$. Each row compares pretraining checkpoints with the union of five walks in the corresponding decoder–level cell; all points lie within 2.6×10^{-3} of $\bar{r}$. Ranges are in %, and spreads and margins are in pp; *Higher* and *Lower* measure how far the walks extend above and below the pretraining range. **In every cell, level-set walks expose more downstream underdetermination than pretraining at the same loss.**

$\bar{r}$	Decoder	Pretraining range	Pretraining spread	Walk range	Walk spread	Ratio	Lower	Higher
0.463	small	38.1–68.5	30.4	23.1–73.3	50.2	1.65	15.0	4.8
	default	44.1–69.5	25.4	29.6–72.1	42.6	1.67	14.5	2.7
	large	26.2–53.5	27.3	22.5–56.2	33.7	1.23	3.7	2.7
0.476	small	30.0–61.0	31.0	17.1–67.5	50.4	1.63	12.9	6.5
	default	31.0–57.3	26.3	18.9–61.4	42.5	1.62	12.2	4.1
	large	23.9–44.6	20.7	17.1–45.9	28.8	1.39	6.7	1.3
0.489	small	23.6–44.3	20.7	12.1–54.6	42.4	2.05	11.4	10.3
	default	25.2–39.7	14.5	15.2–44.2	28.9	1.99	9.9	4.5
	large	20.5–34.9	14.4	14.1–38.5	24.4	1.70	6.4	3.6

Table 3: **Linear-probe accuracy ranges of final checkpoints and walks initialized from them across decoder sizes** at masking ratio 0.75 and learning rate 3×10^{-4}. Each row compares six final checkpoints with the union of their six walks. Ranges are in %, and spreads and margins are in pp; column definitions follow Table 2. **For every decoder, level-set walks expose substantially more downstream underdetermination than the final checkpoints from which they start.**

Val. loss	Decoder	Converged range	Converged spread	Walk range	Walk spread	Ratio	Lower	Higher
0.4607–0.4668	small	40.8–64.6	23.8	23.1–72.7	49.5	2.08	17.7	8.0
0.4526–0.4574	default	54.6–72.4	17.8	38.1–75.3	37.2	2.09	16.5	2.9
0.4455–0.4508	large	39.6–56.4	16.9	30.7–62.9	32.2	1.91	8.9	6.5

Table 4: **Linear-probe accuracy ranges of final checkpoints and walks initialized from them across masking ratios** with the default decoder and learning rate 3×10^{-4}. Each row compares six final checkpoints with the union of walks initialized from three of them. Ranges are in %, and spreads and margins are in pp; column definitions follow Table 2. **At every masking ratio, level-set walks expose substantially more downstream underdetermination than the final checkpoints from which they start.**

Val. loss	Mask ratio	Converged range	Converged spread	Walk range	Walk spread	Ratio	Lower	Higher
0.2990–0.3029	0.3	41.8–54.9	13.1	26.6–61.9	35.3	2.70	15.3	7.0
0.3504–0.3543	0.5	44.5–66.5	21.9	32.3–69.8	37.6	1.71	12.3	3.3
0.3824–0.3868	0.6	49.8–67.5	17.8	32.8–72.6	39.8	2.24	17.0	5.0
0.4526–0.4574	0.75	54.6–72.4	17.8	38.1–75.3	37.2	2.09	16.5	2.9
0.5736–0.5795	0.9	46.0–63.2	17.2	35.3–68.6	33.3	1.93	10.7	5.4

C.2 Gradient Geometry at Walk Starts

Figure 8 summarizes the alignment of the reconstruction and classification gradients at every unique checkpoint used to initialize a level-set walk.

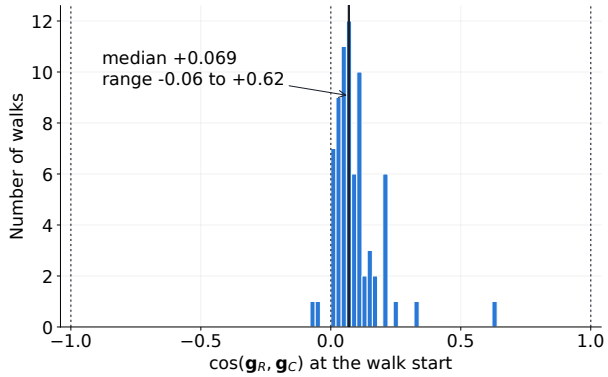

Figure 8: **Cosine between the reconstruction and classification gradients at 73 unique level-set-walk starts** in Figure 3: 45 matched-loss starts and 28 additional final-checkpoint starts. The solid line marks the median; dotted lines mark -1, 0, and 1. The cosine has median $+0.069$, is positive at 71 of 73 starts, and ranges from -0.06 to $+0.62$. **The gradients are generally nearly orthogonal and weakly positively aligned at the walk starts, so the level-set projection retains almost all of the classification gradient.**

C.3 Walk Implementation

Algorithms 2 and 3 expand the main-text procedure with the stopping rule for degenerate projected gradients, adaptive predictor step sizes, and the backtracked Newton corrector. Table 5 lists the values used in every walk reported in the paper.

Table 5: Level-set walk hyperparameters, as used in Algorithms 2 and 3.

Symbol	Value	Role
s_0	2×10^{-2}	initial predictor step length
$s_{\min}$	2.5×10^{-3}	step-length floor
τ	2×10^{-4}	corrector tolerance on the training-subset constraint
ε	1.5×10^{-3}	level-set tolerance: held-out drift budget, and the residual accepted at the step-length floor
ν	10^{-6}	degeneracy threshold on $\|\mathbf{v}\|_2/\|\mathbf{g}_C\|_2$
M, γ	30, 3×10^{-3}	early-stopping window (steps) and minimum validation-accuracy change in the intended direction over it
K, B	8, 6	corrector iterations per step, backtracking halvings per iteration
T	300	step cap per direction

The reconstruction constraint and gradient are evaluated on a fixed subset of 10,000 training images using fixed masks at the checkpoint's validation masking ratio, while the classification gradient is evaluated on unmasked inputs from the same subset. All three are computed in fp32. Held-out reconstruction loss and linear-probe accuracy are measured on the 5,000 validation images, and the head $\mathbf{W}^{\star}_{\mathrm{CE}}$ is re-fitted after every accepted step. As in the main text, validation reconstruction loss is used only to enforce the held-out level-set constraint, while validation accuracy is used only for compute-saving early stopping, not checkpoint selection.

We use validation reconstruction loss as a stopping criterion to ensure that the accuracy spread traced by the walks is measured within the same held-out reconstruction-loss band as the spread

obtained from the pretraining checkpoints, allowing the two to be compared fairly. The convergence of validation linear-probe accuracy is used only as an early-stopping criterion to avoid spending computation after progress in the predetermined direction becomes negligible. Because the reported spread is the difference between the maximum and minimum accuracy over all visited checkpoints, early stopping cannot increase it; it can only omit later checkpoints and thereby make the observed spread a more conservative lower bound. Thus, neither validation metric is used to select a checkpoint because it achieves favorable held-out performance, as in model selection for a learning algorithm.

Algorithm 2 Full level-set walk (one direction)

Require: start $\boldsymbol{\theta}_0$, direction $\sigma \in \{-1, +1\}$, initial step s_0, minimum step $s_{\min}$, tolerances τ, ε, degeneracy threshold ν, convergence window M and rate γ, corrector budget K, backtracking budget B, walk budget T

$r_{\mathrm{tr}} \leftarrow \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}_0), \quad r \leftarrow \widehat{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\theta}_0)$
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0, \quad s \leftarrow s_0, \quad \mathbf{W} \leftarrow \mathbf{W}^{\star}_{\mathrm{CE}}(\boldsymbol{\phi}_0)$
$\mathrm{Acc}^{\mathrm{val}}_0 \leftarrow \mathrm{Acc}(\mathbf{W}, \boldsymbol{\phi}_0)$
for $t = 1, \ldots, T$ **do**
  $\mathbf{g}_R \leftarrow \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}), \quad \mathbf{g}_C \leftarrow \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{cls}}(\mathbf{W}, \boldsymbol{\phi})$
  $\mathbf{v} \leftarrow \mathbf{g}_C - \dfrac{\mathbf{g}_C^{\top}\mathbf{g}_R}{\|\mathbf{g}_R\|_2^2}\mathbf{g}_R$
  if $\|\mathbf{v}\|_2 \le \nu \|\mathbf{g}_C\|_2$ **then**
    break ▷ $\mathbf{g}_C \in \mathrm{span}(\mathbf{g}_R)$
  end if
  $\mathit{accepted} \leftarrow \mathrm{false}, \quad \mathit{stalled} \leftarrow \mathrm{false}$
  repeat ▷ predictor, with adaptive step length
    $\widetilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta} + \sigma\, s\, \mathbf{v}/\|\mathbf{v}\|_2$
    $\widetilde{\boldsymbol{\theta}} \leftarrow \textsc{Correct}(\widetilde{\boldsymbol{\theta}}, r_{\mathrm{tr}}, \tau, K, B)$
    $\rho \leftarrow |\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\widetilde{\boldsymbol{\theta}}) - r_{\mathrm{tr}}|$
    if $\rho \le \tau$ **then**
      $\mathit{accepted} \leftarrow \mathrm{true}, \quad s \leftarrow \min(s_0,\, 1.5\, s)$ ▷ clean accept: regrow toward s_0
    else if $s > s_{\min}$ **then**
      $s \leftarrow \max(s_{\min},\, s/2)$ ▷ discard $\widetilde{\boldsymbol{\theta}}$; the step outran the local curvature
    else if $\rho \le \varepsilon$ **then**
      $\mathit{accepted} \leftarrow \mathrm{true}$ ▷ at the floor, still within the level-set tolerance; no regrowth
    else
      $\mathit{stalled} \leftarrow \mathrm{true}$
    end if
  until *accepted* **or** *stalled*
  if *stalled* **then**
    break ▷ constraint unmaintainable; never triggered in our runs
  end if
  $\boldsymbol{\theta} \leftarrow \widetilde{\boldsymbol{\theta}}, \quad \mathbf{W} \leftarrow \mathbf{W}^{\star}_{\mathrm{CE}}(\boldsymbol{\phi})$
  $\mathrm{Acc}^{\mathrm{val}}_t \leftarrow \mathrm{Acc}(\mathbf{W}, \boldsymbol{\phi})$
  if $|\widehat{\mathcal{L}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r| > \varepsilon$ **then**
    break ▷ left the validation reconstruction-loss band
  end if
  if $t \ge M$ **and** $-\sigma\,(\mathrm{Acc}^{\mathrm{val}}_t - \mathrm{Acc}^{\mathrm{val}}_{t-M}) < \gamma$ **then**
    break ▷ early stop: validation accuracy has converged
  end if
end for

Algorithm 3 Damped Newton corrector

function CORRECT($\boldsymbol{\theta}, r_{\mathrm{tr}}, \tau, K, B$)
  for $k = 1, \dots, K$ **do**
    $e \leftarrow \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}) - r_{\mathrm{tr}}$
    if $|e| \le \tau$ **then**
      return $\boldsymbol{\theta}$
    end if
    $\mathbf{g}_R \leftarrow \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta})$
    $\eta \leftarrow 1, \quad \mathit{improved} \leftarrow \mathrm{false}$
    for $b = 1, \dots, B$ **do** ▷ damping: halve until the residual shrinks
      $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} - \eta \, \dfrac{e}{\|\mathbf{g}_R\|_2^2} \mathbf{g}_R$
      if $|\widehat{\mathcal{L}}^{\mathrm{tr}}_{\mathrm{rec}}(\boldsymbol{\theta}') - r_{\mathrm{tr}}| < |e|$ **then**
        $\mathit{improved} \leftarrow \mathrm{true}$; **break**
      end if
      $\eta \leftarrow \eta / 2$
    end for
    if not $\mathit{improved}$ **then**
      return $\boldsymbol{\theta}$ ▷ every damping failed; stay put
    end if
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$
  end for
  return $\boldsymbol{\theta}$
end function

D Spectral Analysis

D.1 Why Lower Prediction Spectra Are Expected Under Squared Error

The prediction spectra in Figures 9 and 10 generally lie below the target spectrum. This phenomenon is expected because even the Bayes-optimal predictor under squared error has no more variance than the target in any direction. To see this, let $\mathbf{Y} \in \mathbb{R}^{d_p}$ denote a random target patch, and let $\mathbf{C}$ collect everything available to the model when predicting it, including the visible patches, their positions, and the target position. For any measurable predictor f, the squared-error risk decomposes as

$$\mathbb{E}\big[\|\mathbf{Y} - f(\mathbf{C})\|_2^2\big] = \mathbb{E}\big[\|\mathbf{Y} - \mathbb{E}[\mathbf{Y} \mid \mathbf{C}]\|_2^2\big] + \mathbb{E}\big[\|\mathbb{E}[\mathbf{Y} \mid \mathbf{C}] - f(\mathbf{C})\|_2^2\big] .$$

The second term is nonnegative and vanishes for the Bayes-optimal predictor

$$f^\star(\mathbf{C}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{C}].$$

The law of total covariance then gives

$$\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}\big(\mathbb{E}[\mathbf{Y} \mid \mathbf{C}]\big) + \mathbb{E}[\mathrm{Cov}(\mathbf{Y} \mid \mathbf{C})] .$$

Because the second term is positive semidefinite,

$$\mathrm{Cov}\big(f^\star(\mathbf{C})\big) \preceq \mathrm{Cov}(\mathbf{Y}).$$

Consequently, for every target-PCA direction $\mathbf{u}_k$ with eigenvalue λ_k,

$$\mathrm{Var}\big(\mathbf{u}_k^\top f^\star(\mathbf{C})\big) \le \mathrm{Var}\big(\mathbf{u}_k^\top \mathbf{Y}\big) = \lambda_k.$$

Thus, a trained MAE that approximates this Bayes-optimal predictor is likewise expected to have lower prediction variance than the ground-truth patches along each PCA direction.

D.2 Across Decoder Capacities and Masking Ratios

Figure 10 repeats the spectral analysis of Figure 9 across the three decoder capacities at masking ratio 0.75 and the default decoder at five masking ratios.

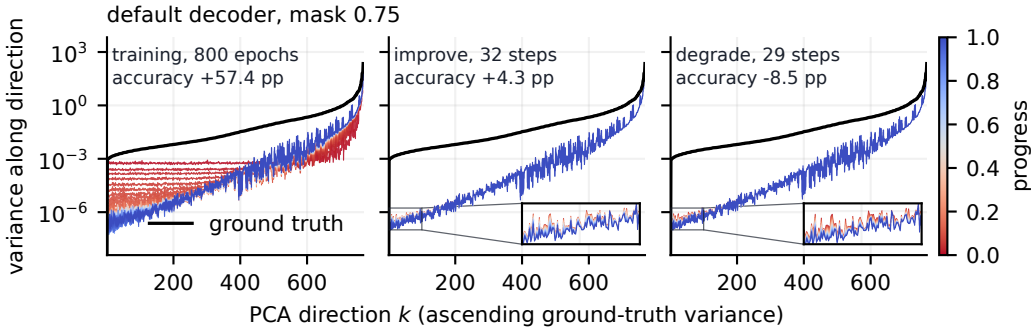

Figure 9: **Prediction spectra during pretraining and along level-set walks** for the default decoder at masking ratio 0.75, with both walks initialized from the final pretraining checkpoint. Directions are ordered by increasing target variance. The black curve shows the target spectrum; colored curves show prediction spectra, with color denoting progress from the start (red) to the end (blue). Insets magnify the lowest-variance directions. **Left:** pretraining. **Center:** improving walk. **Right:** degrading walk. The walks change accuracy by $+4.3$ and -8.5 pp at fixed reconstruction loss, while their spectral changes are concentrated in directions containing only 0.08% of the target variance. **Level-set walks alter only a low-variance, late-settling spectral tail yet substantially change downstream accuracy in either direction.**

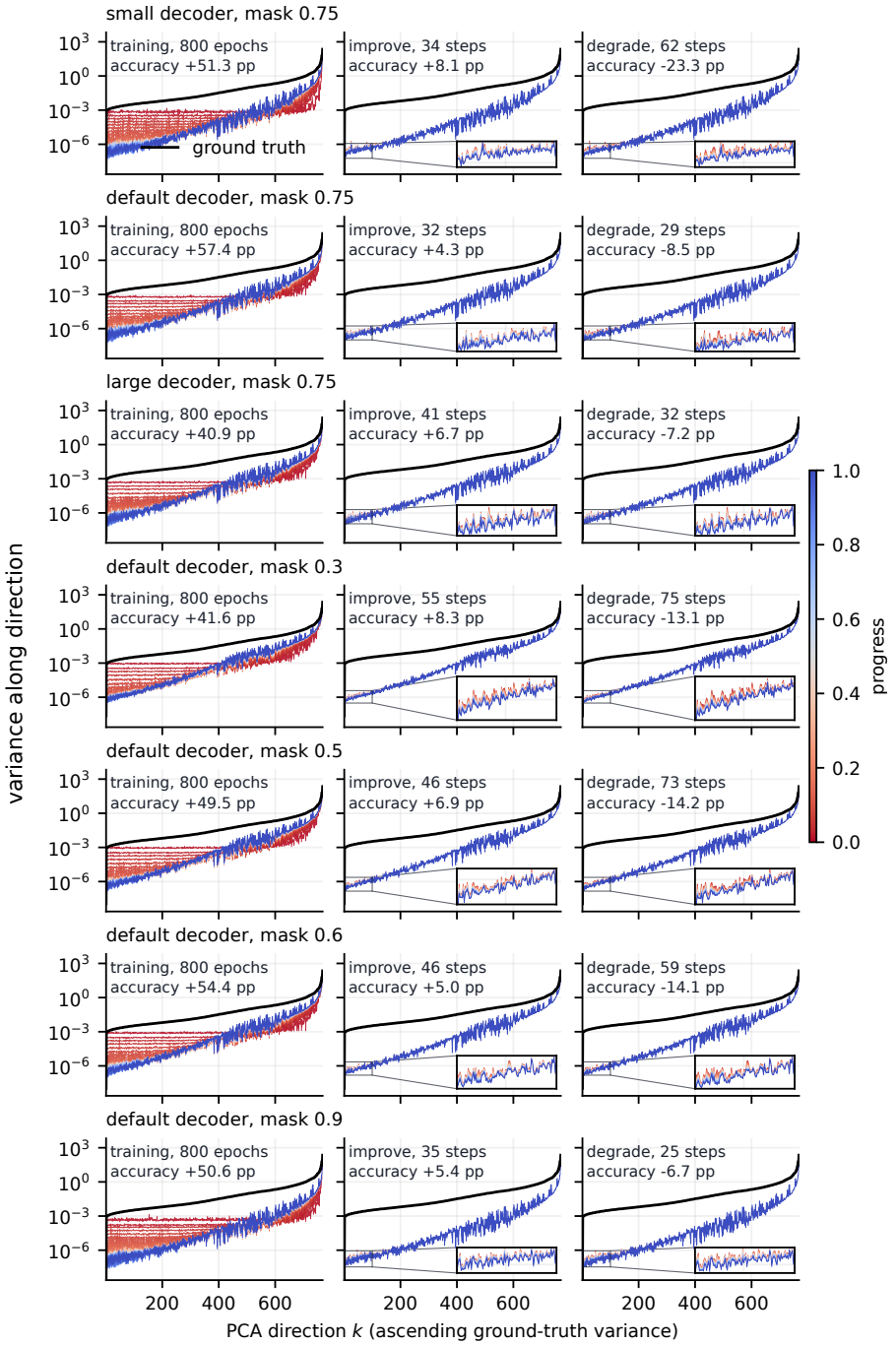

Figure 10: **Prediction spectra across decoder capacities and masking ratios during pretraining and level-set walks**. The seven runs use learning rate 3×10^{-4} and weight decay 0.05, with each pair of walks initialized from the corresponding final checkpoint. Rows show the three decoder capacities at masking ratio 0.75 and the default decoder at masking ratios 0.3, 0.5, 0.6, and 0.9; columns show pretraining, the improving walk, and the degrading walk. Curves, colors, and insets follow Figure 9, and all rows share one vertical scale. Across settings, the walks change accuracy by $+4.3$ to $+8.3$ pp and -6.7 to -23.3 pp at fixed reconstruction loss, while their spectral changes are concentrated in directions containing at most 0.11% of the target variance. **Across decoder capacities and masking ratios, level-set walks alter only a low-variance, late-settling spectral tail yet substantially change downstream accuracy in either direction.**