SELECT WHAT RESISTS COMPRESSION: RECONSTRUCTION CODE LENGTH FOR LABEL-FREE DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting training data typically requires labels, captions, or an externally pretrained encoder to assess data quality. A pool of unlabeled images offers none of these, and what makes an image useful for self-supervised pretraining is itself unsettled. We show that the per-image loss of a corruption–reconstruction network trained on the pool is a conditional code length: the number of nats needed to encode the image given a masked and noised view of it. We rank the pool by this quantity, which we call the CORE score. In other words, we keep the images that a model of the pool compresses worst. On ImageNet-1k, MAE pretrained on the top-scored 50% at half the gradient steps reaches 62.1% under linear probing against 62.3% for the full set, whereas a random 50% reaches 56.7%; at equal steps the top-scored half exceeds the full set by 5.9 linear-probe points and 1.1 ADE20K mIoU, and a random half by 3.3 and 0.8. The ordering top $>$ random $>$ bottom holds under fine-tuning for five pixel-prediction backbones and for DINO, which never reconstructs pixels, and a scorer trained on half of the pool ranks the other half almost identically ($\rho_s = 0.89$). A 1.28M-image subset selected from ImageNet-21k exceeds the human-curated ImageNet-1k by 0.32 points on average. The score correlates with lossless compressed size ($\rho = 0.61$), yet the lowest-scored images include high-frequency repetitive textures: heavy masking discounts the bits that a generic compressor counts. We report the one-time cost of the scorer together with the number of pretraining runs after which it pays off. Code is provided in the supplementary material.

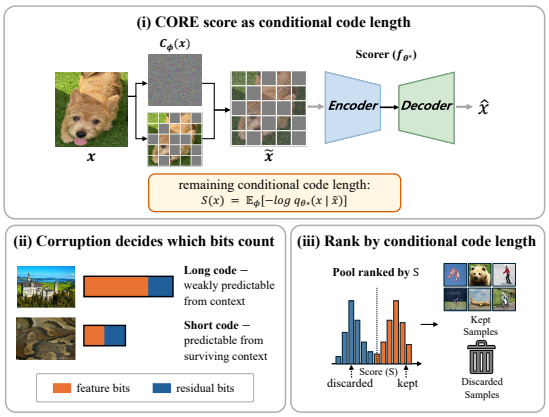
1 INTRODUCTION

Not all samples are worth the same pretraining compute. Large pools repeat themselves: pruning redundant examples improves on power-law scaling of error with dataset size (Sorscher et al., 2022), and de-duplicating and diversifying a web corpus improves language-model pretraining at a fixed token budget (Tirumala et al., 2023). Compute-optimal training ties the number of examples to model size (Hoffmann et al., 2022), so an example that teaches little still has a price.

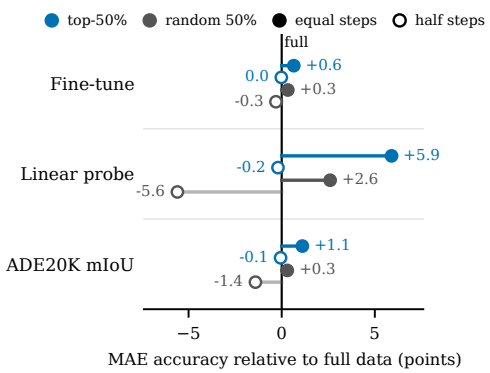
The tools for choosing examples assume side information. Given labels, one can count forgetting events (Toneva et al., 2019), measure error norms early in training (Paul et al., 2021), or sample in proportion to gradient norm (Katharopoulos & Fleuret, 2018). Labels also allow coresets that match the full gradient or cover the feature space (Mirzasoleiman et al., 2020; Killamsetty et al., 2021; Zheng et al., 2023). Given captions, one can threshold image–text similarity under CLIP (Radford et al., 2021; Gadre et al., 2023) or train a filtering network on trusted pairs (Fang et al., 2024). Given a strong pretrained encoder, one can cluster its embeddings and rebalance the pool (Vo et al., 2024; Oquab et al., 2024). Language-model pipelines likewise resample toward a target distribution (Xie et al., 2023a) or learn an influence model against a reference task (Yu et al., 2024). A pool of unlabeled images from a new domain has no labels, no captions, and no reference task, and an external encoder scores such a pool by the distribution it was trained on.

A second gap is conceptual: it is not settled what makes an example useful when there are no labels. For contrastive learning, the examples that contribute most are the ones that contribute least to supervised learning (Joshi & Mirzasoleiman, 2023), and whether to keep hard or easy examples

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.



(a)



(b)

Figure 1: **Keep the samples that the pool's own model compresses worst.** (a) The per-image loss of a network trained to invert masking and noise is a conditional code length: the bits a receiver still needs after seeing the corrupted view (i). The corruption decides which bits are charged, so a texture that continues from the visible patches is cheap while an object the context does not determine is not (ii), and the pool is ranked by this length with the top half kept (iii). (b) MAE accuracy relative to pretraining on all of ImageNet-1k. At half the gradient steps a random half loses 5.6 linear-probe points and the top-scored half 0.2; at equal steps both exceed the full pool, the top-scored half by 5.9 linear-probe points and 1.1 mIoU, 3.3 and 0.8 above the random half. The same order holds for five pixel-prediction learners and for DINO, which never reconstructs pixels (Tables 1 and 3).

depends on how much data is available (Sorscher et al., 2022). High loss also marks noise and outliers, which is why loss-based selection in supervised learning subtracts an irreducible-loss term (Mindermann et al., 2022). A label-free score therefore has to answer two questions: what quantity it measures, and for which learner that quantity predicts utility.

We ask what can be measured from the pool alone, and our answer is a code length. We train one corruption–reconstruction network on the pool. The network sees an image with most patches masked and the remaining tokens noised, and it predicts the pixels, combining masked image modeling (Devlin et al., 2019; Bao et al., 2022; He et al., 2022; Xie et al., 2022) with denoising (Vincent, 2011; Ho et al., 2020) in a single encoder (Choi et al., 2026). After training we freeze the network and define the CORE *score* of an image as its expected reconstruction loss over random corruptions. A reconstruction loss is a negative log-likelihood up to constants—squared error under a Gaussian observation model, absolute error under a Laplace one—and a negative log-likelihood is a code length (Rissanen, 1978; Grünwald, 2007). That is, the score counts the nats that a receiver who holds the corrupted view and the scorer still needs to recover the image. Kolmogorov complexity is uncomputable (Kolmogorov, 1965); a network trained on the pool gives a computable code length relative to it (Hinton & van Camp, 1993; Blier & Ollivier, 2018; Delétang et al., 2024), exact in two idealized cases: for Gaussian noise in input space with squared error (Guo et al., 2005; Kong et al., 2023; Kingma et al., 2021; Kingma & Gao, 2023), and for masking alone (Sahoo et al., 2024). Our scorer applies both corruptions at once, so we use these identities as a reference point and claim only the conditional code length. Selecting the top-scored images is then a concrete rule: keep what the pool's own model compresses worst.

A pixel-level code length has a known weakness, and we test for it. Likelihoods of generative models track the size of a losslessly compressed file (Serrà et al., 2020). Reconstruction objectives fit the high-variance directions of pixel space first, although those directions are the least useful for perception (Balestriero & LeCun, 2024), whereas predicting in latent space favors features with high regression coefficients (Littwin et al., 2024; Assran et al., 2023; Bardes et al., 2024). We make the concern exact: the code length splits into *feature bits*, which any representation must also pay, and *residual bits*, which a learner that predicts features never sees (Equation (7)). If the CORE score were dominated by residual bits, it would behave like a texture detector. We find that it does not. The score correlates with lossless compressed size at $\rho = 0.61$, leaving more than 60% of its variance unexplained by compressibility, and the lowest-scored images include snake

skin, sawdust, and stone inscriptions: high-frequency textures that are predictable from the visible patches. Balestriero & LeCun (2024) prove that masking can align reconstruction with perception whereas additive Gaussian noise cannot, and our ablation has the matching order (masking only 82.55%, noise only 82.41%, both 82.90%), although the first gap is within the seed-level spread of a random subset.

The main evidence is in Figure 1(b). On ImageNet-1k (Deng et al., 2009) we pretrain MAE (He et al., 2022) on the top-scored half, a random half, and the bottom-scored half of the images. At half the gradient steps of the full-data run, the three reach 62.1%, 56.7%, and 51.5% under linear probing against 62.3% for the full set, and 47.22, 45.87, and 42.64 mIoU against 47.28 when the encoders are transferred to semantic segmentation on ADE20K (Zhou et al., 2017). At equal steps a random half exceeds the full set (64.9%, 47.6 mIoU) and the top-scored half exceeds it by more (68.2%, 48.4 mIoU). The same order holds under full fine-tuning for five pixel-prediction backbones (He et al., 2022; Xie et al., 2022; Wei et al., 2023; Zheng et al., 2024; Choi et al., 2026), where the margins shrink to 0.23 points because fine-tuning on 1.28M labels narrows the differences between encoders. When both the class set and the images are chosen by score from ImageNet-21k (Ridnik et al., 2021), the resulting 1.28M-image subset exceeds ImageNet-1k by 0.32 points on average. The scorer also ranks COCO (Lin et al., 2014), CUB (Wah et al., 2011), and NABirds (Van Horn et al., 2015) without retraining (Section E).

All of these learners predict pixels, as the scorer does, so their agreement could reflect a shared objective rather than data utility; contrastive and masked encoders learn different features (Park et al., 2023). We therefore repeat the comparison with DINO (Caron et al., 2021), a self-distillation learner that predicts in latent space (Grill et al., 2020; Chen et al., 2020; Zhou et al., 2022). The top-scored half matches the full set and stays 1.8 to 2.0 points above a random half, which is in turn 2.3 to 2.6 points above the bottom-scored half. Training the default scorer costs 0.92 of one full MAE pretraining run in FLOPs, so selection pays off from the second pretraining run on the pool; a ViT-S scorer pays off from the first, at about half of the gain (Table 5).

Contributions. (i) We show that the per-image loss of a corruption–reconstruction network is a conditional code length. The description-length motivation for data selection thereby becomes a quantity computed from pixels alone, with no labels, captions, or external encoder. (ii) We show which part of that code length matters. It splits exactly into feature bits and residual bits; heavy masking discounts texture, and weak masking removes the gain, as the analysis predicts. (iii) On ImageNet-1k the top-scored half matches the full set at half the compute and exceeds it at equal compute by 5.9 linear-probe points and 1.1 mIoU, 3.3 and 0.8 above a random half at the same compute; the ordering top $>$ random $>$ bottom holds for five pixel-prediction backbones and for DINO, and a score-selected 1.28M subset of ImageNet-21k exceeds ImageNet-1k by 0.32 points. A scorer trained on half of a pool ranks the other half with $\rho_s = 0.89$ agreement, so the score also applies to pools that are still growing.

2 Reconstruction Code Length

Problem. Let $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ be a pool of unlabeled images, $\mathcal{A}$ a self-supervised learner, and U a downstream utility. Given a budget $K < N$, we seek

$$\mathcal{S}^\star \;=\; \arg\max_{\mathcal{S}\subset\mathcal{D},\ |\mathcal{S}|=K} U\big(\mathcal{A}(\mathcal{S})\big). \tag{1}$$

U is unknown at selection time, so we restrict attention to rules of the form $\mathcal{S} = \text{top-}K(\mathcal{D}; S)$ for a per-image score $S : \mathcal{X} \to \mathbb{R}$ that is computed from $\mathcal{D}$ alone: no labels, no captions, and no model trained outside $\mathcal{D}$.

2.1 Corruption as a Channel

Let $\boldsymbol{X} \in \mathbb{R}^{L\times d}$ be the token sequence of $\boldsymbol{x}$. A corruption $\phi = (\boldsymbol{m}, t, \boldsymbol{\epsilon})$ consists of a binary mask $\boldsymbol{m} \in \{0,1\}^L$ that keeps a fraction $1 - r$ of the tokens, a noise level t with signal rate $\bar{\alpha}_t$, and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The corrupted view is as follows:

$$\tilde{\boldsymbol{X}} \;=\; C_\phi(\boldsymbol{X}) \;=\; \underbrace{\boldsymbol{m} \odot \big(\sqrt{\bar{\alpha}_t}\,\boldsymbol{X} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}\big)}_{\text{visible, noised}} \;+\; \underbrace{(\mathbf{1}-\boldsymbol{m}) \odot \boldsymbol{\theta}_{\mathrm{m}}}_{\text{masked}}, \qquad \phi \sim \pi, \tag{2}$$

where $\boldsymbol{\theta}_{\mathrm{m}}$ is a learned mask token and π is the corruption distribution. We treat C_ϕ as a channel through which a receiver observes the image.

2.2 The CORE score is a conditional code length

A single network f_θ is trained on the pool to invert the channel:

$$\theta^\star \;=\; \arg\min_\theta \; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\, \mathbb{E}_{\phi\sim\pi}\; d\big(\boldsymbol{x},\; f_\theta(C_\phi(\boldsymbol{x}), t)\big), \tag{3}$$

where d is a reconstruction loss of the form $d(\boldsymbol{x}, \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_p^p$. Any such loss defines an observation model

$$q_\theta(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) \;=\; Z_\beta^{-1} \exp\big(-d(\boldsymbol{x}, f_\theta(\tilde{\boldsymbol{x}}, t))/\beta\big), \tag{4}$$

which is Gaussian for $p = 2$ and Laplace for $p = 1$, with a normalizer Z_β that does not depend on $\boldsymbol{x}$ or $\tilde{\boldsymbol{x}}$. We freeze $\theta^\star$ and define the CORE *score* as follows:

$$S(\boldsymbol{x}) \;=\; \mathbb{E}_{\phi\sim\pi}\; d\big(\boldsymbol{x},\; f_{\theta^\star}(C_\phi(\boldsymbol{x}), t)\big) \;=\; \beta\, \mathbb{E}_{\phi\sim\pi}\big[-\log q_{\theta^\star}(\boldsymbol{x} \mid C_\phi(\boldsymbol{x}))\big] \;-\; \beta \log Z_\beta. \tag{5}$$

Proposition 1 (Code length)**.** *For any sufficiently fine discretization of $\mathcal{X}$ there exists a prefix code that transmits $\boldsymbol{x}$ to a receiver who holds $C_\phi(\boldsymbol{x})$ and $\theta^\star$, and whose expected length is within* $\log 2$ *nats of $\beta^{-1}S(\boldsymbol{x}) + c$, where c does not depend on $\boldsymbol{x}$. Ranking the pool by S is therefore ranking it by conditional code length.*

This is the standard correspondence between likelihood and code length; we state it to fix what the score measures. The score counts how much of an image remains to be described after the pool's own model has seen a corrupted view of it, and the cost of describing $\theta^\star$, the first part of a two-part code (Rissanen, 1978; Grünwald, 2007), is shared by all images and leaves the ranking unchanged.

Proposition 2 (What the score charges for)**.** *Let p be the distribution of the pool and $p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})$ the conditional law of $\boldsymbol{x}$ given $\tilde{\boldsymbol{x}} = C_\phi(\boldsymbol{x})$ under $\boldsymbol{x} \sim p$. Then*

$$\beta^{-1}S(\boldsymbol{x}) + \log Z_\beta \;=\; \underbrace{\mathbb{E}_\phi\big[-\log p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})\big]}_{\textit{conditional surprisal}} \;+\; \underbrace{\mathbb{E}_\phi\Big[\log \frac{p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})}{q_{\theta^\star}(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})}\Big]}_{\textit{excess code length}}, \tag{6}$$

and the pool average of the second term is $\mathbb{E}\,\mathrm{KL}\big(p(\cdot \mid \tilde{\boldsymbol{x}}) \,\|\, q_{\theta^\star}(\cdot \mid \tilde{\boldsymbol{x}})\big) \geq 0$, the quantity that Equation (3) minimizes. Moreover, if $C_{\phi'}$ is a degraded version of C_ϕ, then $H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}') \geq H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}})$, with $H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}) \to 0$ as the corruption vanishes and $H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}) \to H(\boldsymbol{X})$ as it becomes total.

Proposition 2 states which bits are charged. Content that the surviving context predicts is free; content that it does not predict is paid for. The corruption therefore selects the notion of information. With no corruption every image costs nothing and the score is uninformative. With total corruption the score is the marginal code length $-\log q(\boldsymbol{x})$, which for pixel models tracks low-level complexity (Serrà et al., 2020). Between the two, masking removes context in space, so a texture that continues from its neighbors stays cheap while an object that the context does not determine is charged. This yields two testable predictions: weak masking should select worse than strong masking, and a noise-only score should select worse than a mask-only score (Section 3).

Pixel bits and feature bits. For any deterministic representation g, the conditional surprisal of Equation (6) splits exactly:

$$-\log p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) \;=\; \underbrace{-\log p\big(g(\boldsymbol{x}) \mid \tilde{\boldsymbol{x}}\big)}_{\text{feature bits}} \;\; \underbrace{-\log p\big(\boldsymbol{x} \mid g(\boldsymbol{x}), \tilde{\boldsymbol{x}}\big)}_{\text{residual bits}}. \tag{7}$$

A learner that predicts representations pays only the first term; a learner that predicts pixels pays both. The residual bits describe what g discards, such as the exact realization of a texture or sensor noise. Equation (7) is the formal content of the argument that prediction should take place in feature space (Assran et al., 2023; Bardes et al., 2024; Littwin et al., 2024; Balestriero & LeCun, 2024), and it has two consequences for selection. First, a pixel-space score ranks a pool correctly for a latent learner only to the extent that residual bits do not reorder it. Masking keeps them small



Figure 2: **The two ends of the ranking on ImageNet-1k.** Top row: highest CORE scores (0.52–0.54); bottom row: lowest (0.20–0.23). The lowest-scored images include high-frequency repetitive textures, which a frequency statistic would rank high. More examples and the score–compressibility scatter are in Section C.

wherever detail continues from the visible context, whereas additive noise charges them in every image. Second, the feature bits can be estimated without an external model. With g the encoder of the frozen scorer,

$$S_{\mathrm{lat}}(\boldsymbol{x}) \;=\; \mathbb{E}_{\phi\sim\pi}\,\big\|g(\boldsymbol{x}) - g\big(f_{\theta^\star}(C_\phi(\boldsymbol{x}),t)\big)\big\|_2^2 \tag{8}$$

is a proxy for the first term of Equation (7) under a Gaussian observation model in feature space, at the cost of one additional forward pass. We use S throughout; S_{lat} makes precise what a latent learner would charge for, and Section 4.3 tests whether the two rankings agree where it matters. Proofs, and two special cases in which the code length is exact, are in Section A.

2.3 SELECTION AND ITS COST

Given a keep ratio ρ, the selected subset is

$$\mathcal{D}_{\mathrm{CORE}} \;=\; \text{top-}\,K\big(\mathcal{D};S\big), \qquad K=\lfloor \rho N \rfloor. \tag{9}$$

To curate from a labeled super-pool under a class budget, we first rank classes by $\bar{S}_c = |\mathcal{D}_c|^{-1}\sum_{\boldsymbol{x}\in\mathcal{D}_c} S(\boldsymbol{x})$ and then apply Equation (9) within the retained classes.

Selection is not free. Let C_{s} be the cost of training the scorer and scoring the pool, and C_{p} the cost of one pretraining run on the full pool. At fixed epochs a run on $\mathcal{D}_{\mathrm{CORE}}$ costs $\rho\,C_{\mathrm{p}}$, so selection pays off after

$$n^\star \;=\; \left\lceil \frac{C_{\mathrm{s}}}{(1-\rho)\,C_{\mathrm{p}}} \right\rceil \tag{10}$$

pretraining runs on the pool. We report C_{s}, C_{p}, and $n^\star$ for every scorer in Section 4.5. Architectures and schedules are in Section B.

3 WHAT THE SCORE MEASURES

Proposition 2 says that the corruption decides which bits are charged. This section tests that statement on ImageNet-1k.

The score is related to compressibility but is not a frequency statistic. We compress every image losslessly and correlate file size with the score (Figure 3). The Pearson correlation is $\rho = 0.612$, so compressed size explains 37% of the variance of the score. The remaining variance is not noise around a frequency statistic, as the two ends of the ranking show.

Masking discounts texture. Figure 2 shows the two ends of the ranking. The highest-scored images contain several objects and structured scenes: instrument clusters, machinery, cribs, crowds. The lowest-scored images include snake skin, sawdust, and stone inscriptions. These have high pixel variance and high-frequency content, yet they score low, because the visible patches predict the masked ones. The figure shows images from the two ends of the ranking; the pattern holds throughout the top and bottom percentiles.

Table 1: **ImageNet-1k fine-tuning top-1 (%).** Top block: pretraining on 50% of ImageNet-1k. Paired t-test over the five backbones, top against random: $t = 5.82$, $p = 0.004$. Bottom block: two 1.28M-image, 1,000-class corpora drawn from ImageNet-21k.

Pretraining set	MAE	SimMIM	DiffMAE	MaskDiT	ECC	Mean
ImageNet-1k, full (100%)	82.92	83.10	82.18	82.89	83.31	82.88
CORE top 50%	82.90	83.11	82.15	82.71	83.28	82.83
Random 50%	82.61	82.84	82.02	82.55	82.93	82.59
CORE bottom 50%	82.14	82.77	81.91	81.68	82.49	82.20
Δ top $-$ random	$+0.29$	$+0.27$	$+0.13$	$+0.16$	$+0.35$	$+0.24$
ImageNet-1k (human-curated)	82.92	83.10	82.18	82.89	83.31	82.88
CORE-1k (score-selected)	83.37	83.34	82.60	83.02	83.67	83.20
Δ	$+0.45$	$+0.24$	$+0.42$	$+0.13$	$+0.36$	$+0.32$

Corruption strength and type behave as predicted. Proposition 2 predicts that weak corruption leaves little to charge for. With MAE pretrained on the top 50%, a scorer with a mask ratio of 25% yields 81.31%, below a random half (82.61%); 50% yields 82.65% and 75% yields 82.90%. It also predicts that masking, which removes context in space, is the more useful corruption. A mask-only scorer yields 82.55%, a noise-only scorer 82.41%, and the combination 82.90%, in agreement with the result of Balestriero & LeCun (2024) that masking can align reconstruction with perception whereas additive Gaussian noise cannot. The gap between the two single corruptions is within the seed-level spread of a random subset (± 0.18), so we read their order as consistent with the prediction and not as evidence for it; the effect of the mask ratio (1.5 points) is outside that spread.

The selection does not hinge on one scorer. Accuracy on the selected half is stable once the scorer has converged: 82.70%, 82.90%, and 82.85% after 200, 400, and 800 epochs, and 82.73%, 82.90%, and 82.86% for ViT-S, ViT-B, and ViT-L. Scorers trained for 100 epochs or fewer fall below random selection (82.35% and 81.98%). The score therefore reflects difficulty that persists after training, not the transient difficulty of an under-trained model. All ablations are in Section G; gradient-norm dynamics, attention maps, and class balance are in Section C.

4 EXPERIMENTS

Protocol. The scorer is a ViT-B/16 (Dosovitskiy et al., 2021) trained on the pool that it later scores. At a keep ratio of 50% we form three subsets of ImageNet-1k (Deng et al., 2009): the top-scored half, a uniformly random half, and the bottom-scored half. Each learner is pretrained on a subset with its published recipe and a fixed number of epochs, so a half-size subset receives half of the gradient steps. Evaluation uses the full labeled training set and the original validation set (Section B).

4.1 WHICH HALF IS KEPT MATTERS MORE THAN HALVING THE DATA

Table 1 (top) reports five learners that predict pixels: MAE (He et al., 2022), SimMIM (Xie et al., 2022), DiffMAE (Wei et al., 2023), MaskDiT (Zheng et al., 2024), and ECC (Choi et al., 2026). The column to read is the last one. A random half loses 0.29 points relative to full data; the top-scored half loses 0.05; the bottom-scored half loses 0.68. The order top $>$ random $>$ bottom holds for every backbone, and the top–bottom gap (0.63) is twice the cost of discarding half of the data at random. In a separate five-seed run of MAE that varies the initialization and the data order (Section C), the means are 82.92 ± 0.01 (full), 82.85 ± 0.02 (top), 82.53 ± 0.18 (random), and 82.14 ± 0.02 (bottom); Welch's t-test for top against random gives $p = 0.016$.

Curating from a larger pool. ImageNet-1k was drawn from ImageNet-21k (Ridnik et al., 2021) by human curation. We train the scorer on ImageNet-21k and select 1,000 classes and 1.28M images by score alone (Section 2.3), matching ImageNet-1k in both counts. The comparison in Table 1 (bottom) is between two pipelines, because the class set is an output of each. The score-selected corpus is better for all five backbones, by 0.32 points on average.

Table 2: **MAE under a linear probe and transfer to ADE20K.** Top block: 800 epochs, so a half receives half the steps of the full run; bottom block: 1600 epochs, equal steps.

Pretraining set	FT	Linear	ADE20K
Full (100%)	82.92	62.3	47.28
CORE top 50%	82.90	62.1	47.22
Random 50%	82.61	56.7	45.87
CORE bottom 50%	82.14	51.5	42.64
CORE top, 2× ep.	**83.57**	**68.2**	**48.39**
Random, 2× ep.	83.25	64.9	47.58

Table 3: **DINO on the same subsets.** ImageNet-1k top-1 (%) with frozen features; ViT-S/16, 400 epochs.

Pretraining set	k-NN	Linear
Full (100%)	74.15	75.3
CORE top 50%	74.11	75.3
Random 50%	72.29	73.3
CORE bottom 50%	69.95	70.7
Δ top $-$ random	$+1.82$	$+2.0$

4.2 THE MARGIN IS SET BY THE EVALUATION AND THE STEP BUDGET, NOT BY THE SCORE

Full fine-tuning on 1.28M labeled images narrows the differences between pretrained encoders, and masked image modeling gains little from additional data unless the model overfits (Xie et al., 2023b). The margins of Table 1 are therefore small for every row: a random half loses only 0.29 points. We evaluate the MAE encoders of Table 1 in two ways that do not retrain them on ImageNet labels: a linear probe on frozen features, and transfer to semantic segmentation on ADE20K (Zhou et al., 2017) with a segmentation head, both following the MAE protocol.

Table 2 shows the same four encoders separating as the evaluation becomes less able to repair them. At half the steps, the top-scored half is within 0.2 points of the full set under the linear probe, the random half loses 5.6 points, and the bottom-scored half loses 10.8; on ADE20K the order is the same. Both halves received the same steps, so the difference is the images; the bottom-scored deficit on segmentation matches what those images are (Figure 2), textures and uniform regions. The last two rows give each half the steps of the full-data run. A random half then exceeds the full pool by 2.6 linear-probe points, as expected from the insensitivity of masked image modeling to data size (Xie et al., 2023b): at this budget, more passes over fewer images is the better use of compute. The top-scored half exceeds the full pool by 5.9 linear-probe points, 1.1 mIoU, and 0.65 fine-tuning points, and the random half at the same steps by 3.3, 0.8, and 0.3. The number of distinct images is not what limits this learner; which images it sees is.

4.3 A LEARNER THAT DOES NOT RECONSTRUCT PIXELS

Every learner so far shares the pixel-prediction objective of the scorer, so their agreement could reflect objective alignment. Contrastive and masked encoders learn different features (Park et al., 2023), and for contrastive learning the most valuable examples are not the hard ones (Joshi & Mirzasoleiman, 2023), so we pretrain DINO (Caron et al., 2021), a self-distillation learner that never reconstructs pixels, on the same three subsets and evaluate with the k-NN and linear protocols of that paper. Table 3 shows the same picture as Table 2. The top-scored half matches the full set on both metrics, the random half is 1.8 and 2.0 points below it, and the bottom-scored half is a further 2.3 and 2.6 points below random. The ranking that a pixel-space code length induces is therefore not specific to learners that pay for pixels. In the terms of Equation (7), the residual bits that DINO never sees do not reorder the pool enough to change which half is worth keeping.

4.4 COMPARISON WITH LABEL-FREE SELECTORS

Every row of Table 4 shares the protocol of Table 1, and CORE is the only label-free selector above random selection. Distance to a self-supervised prototype (Sorscher et al., 2022) falls 0.7 points below random; that metric was designed for supervised training on the subset, and the images far from a cluster center are not the ones a reconstruction learner needs. CORE also exceeds the reconstruction loss of a publicly pretrained MAE (82.55%), which uses masking alone and a model trained outside the pool. Supervised coreset methods, which use labels that CORE does not, are compared in Section F.

Table 4: **Label-free selectors at a 50% budget** (MAE, protocol of Table 1). Ext.: needs a model trained outside the pool.

Selector	Ext.	Top-1
Random	no	82.61
Laplacian var. (Pech-Pacheco et al., 2000)	no	80.51
SSL prototypes (Sorscher et al., 2022)	no	81.90
DINOv2 clustering (Vo et al., 2024)	yes	80.90
Pretrained MAE loss	yes	82.55
CORE (ours)	no	82.90

Table 5: **Cost and break-even** (Equation (10), $\rho = 0.5$; FLOPs, Section D).

Scorer	Top-1	C_s/C_p	$n^\star$
ViT-B, 400 ep.	82.90	0.92	2
ViT-B, 200 ep.	82.70	0.46	1
ViT-S, 400 ep.	82.73	0.24	1
ViT-B, 100 ep.	82.35	0.23	–
Random 50%	82.61	0	–

A scorer trained on ImageNet-1k also ranks other datasets without retraining. Keeping the top-scored half of the training set of ADE20K, COCO, CUB, or NABirds is above keeping a random half on all five metrics; it matches the full set to within 0.16 points on four of them and loses 1.25 mask AP on COCO (Section E).

4.5 WHAT SELECTION COSTS

A selected half is cheaper to train on only after the scorer has been paid for. The default scorer costs 0.92 of one full pretraining run, so selection pays off from the second run on the pool (Table 5). Cheaper scorers pay off from the first run but keep only 0.09 to 0.12 points of the 0.29-point fine-tuning gain; a scorer trained for 100 epochs is below random selection and never pays off. The scorer is trained once and reused, here for five learners and four datasets; for a single run, random selection is cheaper. At equal compute the scorer buys accuracy rather than a saving (Table 2).

5 RELATED WORK

Selection from labels and training dynamics. Supervised selectors range from curricula and hard-example mining (Bengio et al., 2009; Shrivastava et al., 2016) through coresets and proxy models (Sener & Savarese, 2018; Mirzasoleiman et al., 2020; Killamsetty et al., 2021; Coleman et al., 2020) to statistics of a training run such as influence, gradient norm, forgetting, and early error norms (Koh & Liang, 2017; Katharopoulos & Fleuret, 2018; Toneva et al., 2019; Paul et al., 2021). Two later results qualify the rule that hard examples are the useful ones: high loss also marks noise (Mindermann et al., 2022), and at high pruning rates hard-only subsets fall below random selection (Zheng et al., 2023). All of these scores are functions of a label. CORE keeps the proxy-model structure of Coleman et al. (2020), but the proxy is trained without labels and its score is a conditional code length (Proposition 1), defined without reference to a supervised run. A longer account is in Section I.

Selection without labels. Sorscher et al. (2022) show that pruning can improve on power-law scaling and introduce a label-free metric: the distance to a k-means prototype in the embedding space of a pretrained encoder. Their learners are supervised, and whether to keep hard or easy examples depends on how much data is available. For contrastive pretraining, Joshi & Mirzasoleiman (2023) prove that the most valuable examples are those whose augmentations resemble other examples, which are the ones that matter least for supervised learning. Curation pipelines cluster or retrieve in the embedding space of a strong encoder (Oquab et al., 2024; Vo et al., 2024; Tirumala et al., 2023), image–text pipelines threshold CLIP similarity or train a filtering network on trusted pairs (Radford et al., 2021; Gadre et al., 2023; Fang et al., 2024), and language-model pipelines resample toward a target distribution or learn influence on a reference task (Xie et al., 2023a; Yu et al., 2024). Each signal comes from something that must already exist: an encoder that is good on the pool, a set of trusted pairs, or a reference task. Our score needs only a reconstruction network trained on the pool itself, and it is defined in the space of the data, not in the space of a particular encoder. We evaluate it on self-supervised learners of two kinds (Sections 4.1 and 4.3) and compare it with prototype distance and embedding clustering under one protocol (Table 4).

Description length, likelihood, and denoising. The minimum description length principle equates learning with compression (Rissanen, 1978; Grünwald, 2007), with Kolmogorov complexity as its uncomputable ideal (Kolmogorov, 1965). Deep learning has applied it to weights (Hinton & van Camp, 1993), to labels under a network (Blier & Ollivier, 2018), and to language models as compressors (Delétang et al., 2024). Training a denoising autoencoder (Vincent et al., 2008) is a form of score matching (Vincent, 2011), and diffusion models train one denoiser across noise levels (Ho et al., 2020). The resulting loss is a variational bound on likelihood (Kingma et al., 2021; Kingma & Gao, 2023), exact under the pointwise I-MMSE relation (Guo et al., 2005; Kong et al., 2023), with an analogous bound for masking (Sahoo et al., 2024). In pixel space, per-image likelihood is a poor semantic signal. Generative models assign high likelihood to out-of-distribution images (Nalisnick et al., 2019); Serrà et al. (2020) trace this to input complexity and correct it by subtracting the length of a generic compressor. These works measure code length per dataset, or per image for outlier detection. We use a per-image code length to rank a pool for pretraining, and we handle the complexity confound inside the channel instead of by subtraction: the corruption decides which bits are charged (Proposition 2), and masking leaves predictable texture cheap (Section 3).

Reconstruction and latent prediction. Masked prediction moved from language (Devlin et al., 2019) to images (Bao et al., 2022; He et al., 2022; Xie et al., 2022) and was then combined with diffusion noise (Wei et al., 2023; Zheng et al., 2024; Choi et al., 2026). A second family predicts representations instead of pixels, through contrast (Chen et al., 2020), self-distillation (Grill et al., 2020; Caron et al., 2021; Zhou et al., 2022), or prediction of masked regions in feature space (Assran et al., 2023; Bardes et al., 2024). Recent analyses explain why the two families differ. Reconstruction fits high-variance pixel directions first although they are the least useful for perception, and masking, unlike additive noise, reduces this misalignment (Balestriero & LeCun, 2024). Latent prediction is biased toward features with high regression coefficients (Littwin et al., 2024). Contrastive and masked encoders use different frequencies and rely on different layers (Park et al., 2023), and masked image modeling gains little from more data unless the model overfits (Xie et al., 2023b). This literature asks what an objective does to a representation; we ask what its loss says about the data, and the analyses above become predictions that we test: masking, not noise, should carry the useful part of the score (Section 3), and the split into feature and residual bits (Equation (7)) says when a pixel-space ranking should serve a learner that predicts features (Table 3).

6 Limitations

The linear-probe, ADE20K, and equal-compute results are for MAE, and DINO is the one latent-prediction learner we test; contrastive learners, whose most useful examples are known to differ (Joshi & Mirzasoleiman, 2023), and a scorer that predicts features directly (Equation (8)) remain open. The score is defined relative to the pool the scorer saw, and on web-scale image–text pools it is one stage of a pipeline rather than a filter of its own (Section H). We also evaluate a single keep ratio; whether the gap over random selection widens as more of the pool is discarded, as the coverage results of Zheng et al. (2023) would suggest for the bottom of the ranking, is left to future work.

7 Conclusion

The per-image loss of a corruption–reconstruction network is a conditional code length, and the corruption decides which bits it charges for. On ImageNet-1k, the half of the pool with the longest code matches the full pool at half the compute and exceeds it at equal compute by 5.9 linear-probe points and 1.1 ADE20K mIoU, 3.3 and 0.8 above a random half at the same compute; the ordering holds for five pixel-prediction learners and for DINO. A corpus selected by code length from ImageNet-21k exceeds ImageNet-1k by 0.32 points, the gain disappears under weak masking as the analysis predicts, and a scorer trained on half of a pool ranks the rest of it. The description-length principle thereby becomes a selection rule that needs nothing but the pool, and the same construction in feature space (Equation (8)) is the natural next step for learners that predict representations rather than pixels. For a practitioner, the recipe is one network trained on the pool, one scoring pass, and a threshold, with no labels, captions, or external encoder in the loop.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing of the manuscript, including improving the grammar, clarity, readability, and phrasing of author-written text. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure that they accurately reflect the intended scientific meaning. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses public datasets (ImageNet-1k, ImageNet-21k, ADE20K, COCO, CUB-200-2011, NABirds, DataComp) and involves no human subjects and no new data collection. A selector decides what a model learns from. If the code length correlates with demographic or domain attributes in a deployed pool, selection can shift their representation. On ImageNet-1k the selected half retains all classes with a small departure from the class balance of a random half (Section C); we recommend auditing the selected subset for the attributes that matter in a given deployment.

REPRODUCIBILITY STATEMENT

The score, the selection rule, and the cost model are defined in Section 2, with proofs of both propositions in Section A. The configuration of the scorer is in Section B, and the FLOPs accounting behind Table 5 is in Section D. All learners use the published recipes of their original papers, and all datasets are public. The supplementary material contains our code for training the scorer.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 2566–2585, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48, 2009.

Léonard Blier and Yann Ollivier. The description length of deep learning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 1597–1607, 2020.

Hyesong Choi, Daeun Kim, Sungmin Cha, Kwang Moo Yi, and Dongbo Min. ECC: Encoder-centric corruption for fine-grained vision in VLMs. In *European Conference on Computer Vision (ECCV)*, 2026.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *International Conference on Learning Representations (ICLR)*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 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 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 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 (ICLR)*, 2021.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 27092–27112, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

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 (CVPR)*, pp. 16000–16009, 2022.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory (COLT)*, pp. 5–13, 1993.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Siddharth Joshi and Baharan Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 15356–15370, 2023.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 2525–2534, 2018.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1885–1894, 2017.

Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 91300–91336, 2024.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, 2022.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 6950–6960, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations (ICLR)*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 20596–20607, 2021.

José Luis Pech-Pacheco, Gabriel Cristóbal, Jesús Chamorro-Martínez, and Joaquín Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: A comparative study. In *Proceedings of the 15th International Conference on Pattern Recognition (ICPR)*, volume 3, pp. 314–317, 2000.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 8748–8763, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations (ICLR)*, 2020.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 761–769, 2016.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 19523–19536, 2022.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 53983–53995, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 10347–10357, 2021.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, 2015.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

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 (ICML)*, pp. 1096–1103, 2008.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *Transactions on Machine Learning Research (TMLR)*, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16284–16294, 2023.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34201–34227, 2023a.

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 (CVPR)*, pp. 9653–9663, 2022.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Zichun Yu, Spandan Das, and Chenyan Xiong. MATES: Model-aware data selection for efficient pretraining with data influence models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations (ICLR)*, 2023.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Transactions on Machine Learning Research (TMLR)*, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

A PROOFS

Throughout, $\mathcal{X} \subset \mathbb{R}^D$ is discretized into cells of side δ, as any stored image is, and $\log$ is the natural logarithm.

Proof of Proposition 1. For $d(\boldsymbol{x}, \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|_p^p$ the normalizer $Z_\beta = \int \exp(-\|\boldsymbol{x} - \boldsymbol{\mu}\|_p^p/\beta)\, \mathrm{d}\boldsymbol{x}$ is invariant to translation, so it depends on neither $\boldsymbol{\mu}$ nor $\tilde{\boldsymbol{x}}$, and Equation (4) is a density. For a sufficiently fine discretization the cell containing $\boldsymbol{x}$ has probability $P(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) = q_{\theta^\star}(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})\, \delta^D$ up to $o(\delta^D)$. The Shannon–Fano code for $P(\cdot \mid \tilde{\boldsymbol{x}})$ is a prefix code with length $\lceil -\log_2 P(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) \rceil$ bits, which is within one bit of

$$-\log q_{\theta^\star}(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) - D \log \delta \;=\; \beta^{-1} d\big(\boldsymbol{x}, f_{\theta^\star}(\tilde{\boldsymbol{x}}, t)\big) + \log Z_\beta - D \log \delta \tag{11}$$

nats. Both sender and receiver know $\tilde{\boldsymbol{x}} = C_\phi(\boldsymbol{x})$ and $\theta^\star$, so the code is decodable. Taking the expectation over $\phi \sim \pi$ gives $\beta^{-1} S(\boldsymbol{x}) + c$ with $c = \log Z_\beta - D \log \delta$, which does not depend on $\boldsymbol{x}$. □

Proof of Proposition 2. Since $\tilde{\boldsymbol{x}} = C_\phi(\boldsymbol{x})$ is a random function of $\boldsymbol{x}$, the pair $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has a joint law under $\boldsymbol{x} \sim p$ and $\phi \sim \pi$, and $p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})$ denotes its conditional; all expectations below are over this joint law. By Equation (5), $\beta^{-1} S(\boldsymbol{x}) + \log Z_\beta = \mathbb{E}_\phi[-\log q_{\theta^\star}(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})]$. Adding and subtracting $\log p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}})$ inside the expectation gives Equation (6). Averaging the second term over $\boldsymbol{x} \sim p$ and exchanging the order of expectation yields $\mathbb{E}_{\tilde{\boldsymbol{x}}} \mathrm{KL}\big(p(\cdot \mid \tilde{\boldsymbol{x}}) \,\|\, q_{\theta^\star}(\cdot \mid \tilde{\boldsymbol{x}})\big) \geq 0$. The same steps applied to Equation (3) show that the training objective equals $\beta\big(H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}) + \mathbb{E}\,\mathrm{KL}\big)$ up to a constant, so training minimizes the excess code length. If $C_{\phi'}$ is a degraded version of C_ϕ, then $\boldsymbol{X} \to \tilde{\boldsymbol{X}} \to \tilde{\boldsymbol{X}}'$ is a Markov chain, and the data-processing inequality gives $I(\boldsymbol{X}; \tilde{\boldsymbol{X}}') \leq I(\boldsymbol{X}; \tilde{\boldsymbol{X}})$, hence $H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}}') \geq H(\boldsymbol{X} \mid \tilde{\boldsymbol{X}})$. With no corruption $\tilde{\boldsymbol{X}}$ determines $\boldsymbol{X}$ and the conditional entropy is zero; with total corruption $\tilde{\boldsymbol{X}}$ is independent of $\boldsymbol{X}$ and it equals $H(\boldsymbol{X})$. □

Proof of Equation (7). Since g is deterministic, the event $\{\boldsymbol{X} = \boldsymbol{x}\}$ implies $\{g(\boldsymbol{X}) = g(\boldsymbol{x})\}$, so $p(\boldsymbol{x} \mid \tilde{\boldsymbol{x}}) = p(\boldsymbol{x}, g(\boldsymbol{x}) \mid \tilde{\boldsymbol{x}}) = p(g(\boldsymbol{x}) \mid \tilde{\boldsymbol{x}})\, p(\boldsymbol{x} \mid g(\boldsymbol{x}), \tilde{\boldsymbol{x}})$ by the chain rule, and taking $-\log$ gives the claim. □

Break-even point. At fixed epochs, n pretraining runs cost $n\, C_\mathrm{p}$ on the full pool and $C_\mathrm{s} + n\rho\, C_\mathrm{p}$ with selection. The second is smaller when $n > C_\mathrm{s}/((1-\rho)C_\mathrm{p})$, which gives Equation (10).

Remark 1 (Two exact special cases)**.** *For a Gaussian channel in input space and $p = 2$, the integral over noise levels of the pointwise error of the optimal denoiser equals $-\log p(\boldsymbol{x})$ up to a constant (Guo et al., 2005; Kong et al., 2023), and monotone reweightings of that integral remain variational bounds (Kingma et al., 2021; Kingma & Gao, 2023). For pure masking, the loss averaged over mask ratios is a variational bound of the same kind (Sahoo et al., 2024). Equation (5) contains both as limits. Our claims rest on Propositions 1 and 2, which hold for the combined channel and do not require the scorer to be optimal.*

B EXPERIMENTAL SETUP

The scorer is a ViT-B/16 trained for 400 epochs on the pool it scores, with a mask ratio of 75% and diffusion noise over $T = 1000$ levels; its ablations are in Section G. The five pixel-prediction learners in Table 1 follow the published recipes of MAE, SimMIM, DiffMAE, MaskDiT, and ECC, and the supervised experiment in Section H follows DeiT (Touvron et al., 2021). The linear probe and the ADE20K transfer in Table 2 follow the protocols of MAE, and the k-NN and linear evaluations in Table 3 follow those of DINO, with DINO trained for 400 epochs on each subset. The label-free selectors in Table 4 and the supervised coresets in Section F are all followed by the same MAE pretraining and fine-tuning as Table 1. Results in Tables 2 to 4, 7 and 8 are averages over three runs; the three random subsets are shared across all experiments.

C Additional Analysis

Seeds. Table 6 reports five seeds with MAE, from a run made before the three-run averages of Table 1; the two agree to within 0.1 points on every subset.

Table 6: ImageNet-1k fine-tuning top-1 (%) with MAE over five seeds (mean ± standard deviation).

Full	CORE top 50%	Random 50%	CORE bottom 50%
82.92 ± 0.01	82.85 ± 0.02	82.53 ± 0.18	82.14 ± 0.02

Class balance. The top-scored half retains all 1,000 classes. Its per-class distribution has a KL divergence of 0.015 from uniform, against 0.008 for a random half. Applying Equation (9) within each class changes accuracy by less than 0.1 points.

Compressibility. Figure 3 plots the score against losslessly compressed file size for ImageNet-1k images (Pearson $\rho = 0.612$); Figure 4 extends Figure 2 to fourteen images from each end of the ranking.

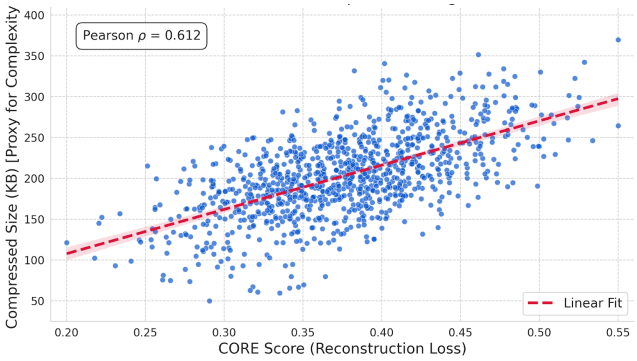


Figure 3: CORE score against lossless compressed size on ImageNet-1k (Pearson $\rho = 0.612$).

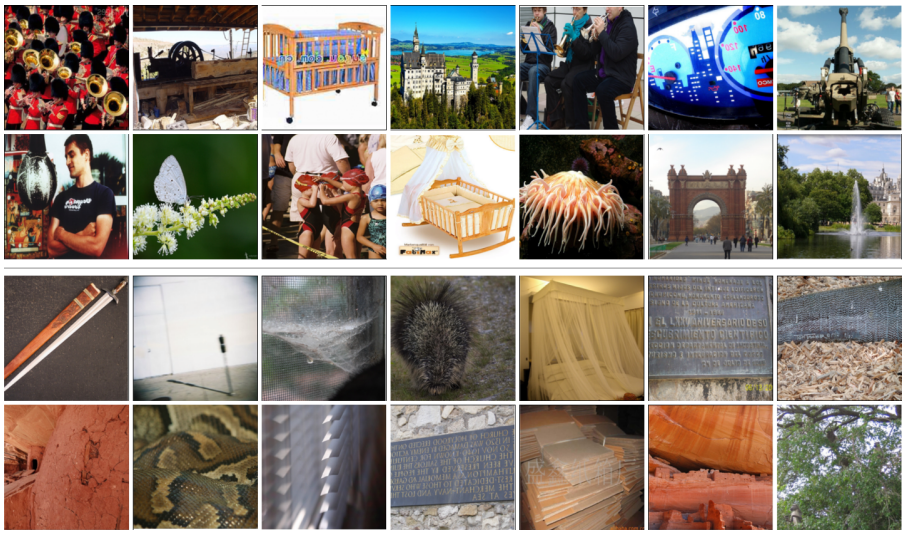


Figure 4: Highest-scored (top two rows) and lowest-scored (bottom two rows) ImageNet-1k images.

Scoring the training set. Every image is scored by a network that was trained on it, so part of a low score could be memorization rather than predictability from context. Two observations argue

against this reading. The corruption is drawn afresh at scoring time and removes 75% of the tokens, so the network never sees at scoring time an input it saw during training. Doubling the scorer's training from 400 to 800 epochs, which can only increase memorization, leaves the accuracy of the selected half unchanged (82.90% against 82.85%, Table 10), and the ranking transfers to four datasets the scorer never saw (Table 9). Table 7 adds the direct control: the pool is split into two disjoint halves, a scorer trained on one half scores the other, and the two score files are merged before selection, so that no image is scored by a network that saw it. The held-out ranking agrees with the in-sample ranking at $\rho_s = 0.89$, and the half it selects trains MAE to 82.80% fine-tuning and 61.7% linear-probe accuracy, within 0.4 points of the in-sample selection and 5 linear-probe points above a random half. Memorization therefore contributes little to the ranking, and a scorer trained on part of a pool can rank the rest of it.

Table 7: Held-out scoring on ImageNet-1k (MAE pretrained on the selected top 50%, protocol of Table 1). "In-sample" is the default scorer trained on the full pool; "held-out" trains one scorer per disjoint half and scores the other half. ρ_s is the Spearman correlation between the two rankings over all images.

Scorer	Fine-tune	Linear probe	ρ_s with in-sample
In-sample (default)	82.90	62.1	1
Held-out	82.80	61.7	0.89
Random 50%	82.61	56.7	–

Gradient norms. Figure 5 shows the mean gradient norm during the first 10 epochs of MAE pretraining on each subset. The top-scored half starts 28% above the random half and 109% above the bottom-scored half. Under a reconstruction loss a larger loss implies a larger gradient norm almost by construction, and MAE shares the objective of the scorer. We therefore report this as a consistency check and not as evidence of utility.

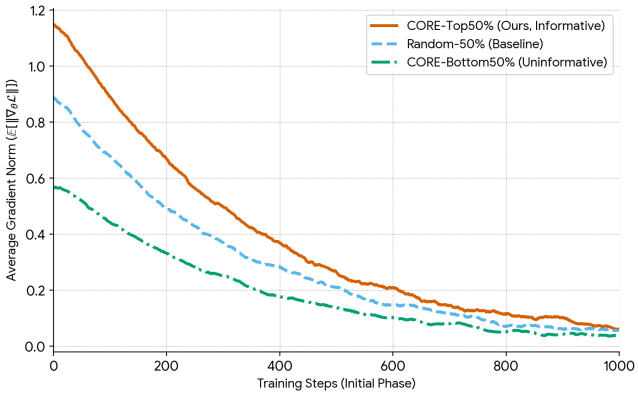


Figure 5: Mean gradient norm in early MAE pretraining on the three subsets.

Where the scorer attends. Figure 6 shows attention maps of the scorer for three high-scored COCO images. The three images are high-scored examples from the COCO validation set.

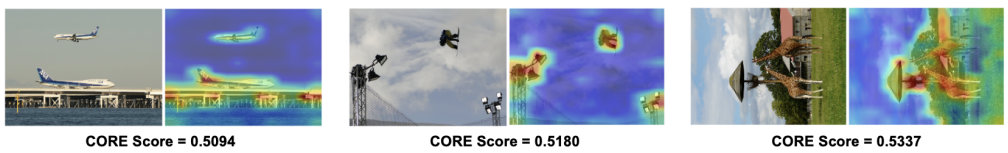


Figure 6: Attention maps of the scorer on high-scored COCO images (scores 0.509, 0.518, 0.534).

D COST ACCOUNTING

We count forward FLOPs per image and take a training step as three forward passes. The scorer processes all tokens of a ViT-B/16, 17.5 GFLOPs per image (4.6 for ViT-S/16). MAE processes the 25% visible tokens in its encoder and all tokens in its lightweight decoder, 9.5 GFLOPs per image in total. For a scorer trained for E_{s} epochs and an MAE run of 800 epochs on the same pool,

$$\frac{C_{\mathrm{s}}}{C_{\mathrm{p}}} = \frac{E_{\mathrm{s}} \cdot 17.5}{800 \cdot 9.5} + \varepsilon, \tag{12}$$

where $\varepsilon < 0.01$ is the cost of scoring the pool, a few forward passes per image. This gives 0.92 for the default scorer and the remaining entries of Table 5. The scorer used to build CORE-1k was trained on ImageNet-21k and costs about ten times as much; the corresponding break-even is against pretraining runs on that pool, not on ImageNet-1k. The ratio is smaller for learners that process all tokens, such as SimMIM, for which the default scorer costs about half of one pretraining run.

Spending the saved compute. MAE keeps improving with longer schedules (He et al., 2022), so a half-size subset trained for twice the epochs must be compared with a random half under the same schedule. Table 8 gives every subset the same number of gradient steps as the full-data run. With equal steps a random half exceeds the full pool (83.25% against 82.92% fine-tuning, 64.9% against 62.3% linear probe), so its 5.6-point loss at half the steps was the steps and not the images, in line with the insensitivity of masked image modeling to data size (Xie et al., 2023b). The top-scored half gains more from the same steps: it exceeds the full pool by 5.9 linear-probe points and 1.1 mIoU, and the random half by 3.3 and 0.8. The number of distinct images is not what limits these learners; which images are seen is.

Table 8: Equal gradient steps (MAE, ImageNet-1k). Rows with 1600 epochs on a 50% subset see the same number of steps and FLOPs as the full-data run with 800 epochs.

Pretraining set	Epochs	Steps vs. full	Fine-tune	Linear probe	ADE20K
Full (100%)	800	1×	82.92	62.3	47.28
CORE top 50%	800	0.5×	82.90	62.1	47.22
Random 50%	800	0.5×	82.61	56.7	45.87
CORE top 50%	1600	1×	83.57	68.2	48.39
Random 50%	1600	1×	83.25	64.9	47.58

E SCORING OTHER DATASETS

Table 9 applies the scorer trained on ImageNet-1k to four other datasets without retraining and keeps 50% of each training set. The top-scored half is above the random half on all five metrics and above the bottom-scored half by up to 2.98 points. It matches the full set to within 0.16 points except for the mask AP on COCO, where it loses 1.25 points. This experiment tests the transfer of the scorer; the transfer of a pretrained encoder is in Table 2.

Table 9: Scoring other datasets with the ImageNet-trained scorer. Each task is trained on 50% of its own training set.

Dataset	Metric	Full	CORE top	Random	CORE bottom
ADE20K (Zhou et al., 2017)	mIoU	43.12	43.09	42.86	40.11
COCO (Lin et al., 2014)	AP^{box}	46.17	46.01	44.74	44.19
COCO (Lin et al., 2014)	AP^{mask}	40.21	38.96	38.28	38.01
CUB (Wah et al., 2011)	Acc.	79.10	79.12	78.40	76.27
NABirds (Van Horn et al., 2015)	Acc.	77.87	77.72	77.58	77.55

F SUPERVISED CORESET METHODS

CRAIG (Mirzasoleiman et al., 2020) and GLISTER (Killamsetty et al., 2021) select subsets with labels, which CORE does not use. At a 50% budget with MAE pretraining they reach 80.65% and 79.18%, against 82.90% for CORE and 82.61% for a random half. These methods were designed for supervised training on the subset, not for self-supervised pretraining followed by fine-tuning on all labels, so the comparison shows that label-based criteria do not carry over to this setting rather than that they are weak in their own. The same holds for the two label-free rows of Table 4 below random selection: Laplacian variance keeps textures, which Figure 2 shows to be the least useful images, and clustering in DINOv2 space was designed to balance a web-scale pool rather than to prune a curated one.

G SCORER ABLATIONS

Table 10 varies one design choice of the scorer at a time and reports MAE fine-tuning accuracy on the resulting top 50%.

Table 10: Scorer ablations (ImageNet-1k top-1, %, MAE pretrained on the top 50%). Defaults in bold. Random 50%: 82.61.

Factor	Setting	Top-1
Corruption	mask only / noise only / **both**	82.55 / 82.41 / **82.90**
Mask ratio	25% / 50% / **75%**	81.31 / 82.65 / **82.90**
Noise range T	100 / 250 / 500 / **1000**	82.67 / 82.68 / 82.77 / **82.90**
Architecture	ResNet-50 / ViT-S / **ViT-B** / ViT-L	82.45 / 82.73 / **82.90** / 82.86
Epochs	50 / 100 / 200 / **400** / 800	81.98 / 82.35 / 82.70 / **82.90** / 82.85

The mask-ratio row agrees with Proposition 2: weak corruption leaves little to charge for, and the resulting ranking is worse than random selection (81.31%). Scorers trained for 100 epochs or fewer are also below random selection, which bounds how far the cost in Table 5 can be reduced.

H SUPERVISED TRAINING AND IMAGE–TEXT DATA

Supervised training on the selected half. With a DeiT-style recipe (Touvron et al., 2021), a ViT-B/16 trained with labels on the top-scored half of ImageNet-1k reaches 81.11%, against 81.17% on the full set.

DataComp. Table 11 adds the score to existing filters on the small scale of DataComp (Gadre et al., 2023), a pool of 12.8M image–text pairs. The scorer was trained on ImageNet-1k and sees no text. Adding it to a filter changes ImageNet zero-shot accuracy by $+0.11$ to $+0.23$ points at an accuracy level of 3–4%. We regard this as preliminary: the differences are small in absolute terms, and the score is one stage of a pipeline here, not a filter of its own.

Table 11: ImageNet zero-shot accuracy (%) on DataComp small, with and without the CORE score as an additional filter.

Filter	Alone	+ CORE
No filtering	2.50	–
Basic	2.59	–
DFN (Fang et al., 2024)	3.11	3.34
Image-based ∩ CLIP score	3.50	3.71
Text-based	4.04	4.15

I Extended Related Work

Selection from labels and training dynamics. The earliest selectors order labeled examples by difficulty: curricula move from easy to hard (Bengio et al., 2009), and hard-example mining keeps the highest-loss candidates (Shrivastava et al., 2016). Coreset methods cast selection as approximation. They cover the feature space (Sener & Savarese, 2018) or match the full gradient with a weighted subset (Mirzasoleiman et al., 2020; Killamsetty et al., 2021), and a small proxy model can select for a larger one (Coleman et al., 2020). A parallel line reads the dynamics of a supervised run: influence on test loss (Koh & Liang, 2017), gradient norm (Katharopoulos & Fleuret, 2018), forgetting events (Toneva et al., 2019), and early error norms (Paul et al., 2021). Two later results qualify the rule that hard examples are the useful ones. High loss also marks noise, so Mindermann et al. (2022) subtract the loss of a model trained on held-out data. At high pruning rates, hard-only subsets lose coverage and fall below random selection (Zheng et al., 2023). All of these scores are functions of a label. CORE keeps the proxy-model structure of Coleman et al. (2020), but the proxy is trained without labels and its score is a conditional code length (Proposition 1), which is defined without reference to a supervised run.

Synthesizing instead of selecting. Dataset condensation learns a small synthetic set whose gradients match those of the full set (Zhao et al., 2021). It requires labels and an optimization over pixels for every target architecture, and the result is no longer a subset of the pool. CORE keeps real images, runs the scorer once, and reuses the ranking across learners (Table 1).

Ordering instead of selecting. Curricula (Bengio et al., 2009) and importance sampling (Katharopoulos & Fleuret, 2018) change the order or the weight of examples within one run and need the loss of the model being trained. Our ranking is computed before any learner is trained, so it cannot adapt to a learner, and it costs nothing during pretraining. Whether a code-length curriculum helps is outside the scope of this paper.

Text. Under the code-length reading, the counterpart of the CORE score for text is the token-level loss of a language model (Delétang et al., 2024). Existing selectors for language models read other signals, namely similarity to a target distribution (Xie et al., 2023a) or learned influence on a reference task (Yu et al., 2024). We do not evaluate on text.