BOOTSTRAP YOUR OWN NOISE: DENOISING AND LATENT PREDICTION COOPERATE ONLY WHEN NOISE IS INFORMATIVE AND ALIGNMENT IS GATED

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction and latent prediction teach a vision encoder different things. Reconstruction from masked or noised inputs yields features that attend locally. Predicting the embedding of a clean view yields features that attend globally, so combining the two is a natural goal. We show that the combination is not additive. We pre-train a ViT-B for 400 epochs on ImageNet-1K. Adding Gaussian latent noise to masked image modeling leaves fine-tuning accuracy unchanged (82.86% against 82.89%), and adding a bootstrapped latent-prediction loss on top lowers it to 82.01% and costs 4.6 mIoU on ADE20K. We trace both failures to a single cause and remove it with two small changes. *Image-adaptive noise* sets the noise level of each image from its token saliency, and a *reliability gate* scales the latent-prediction loss by the current reconstruction error. The gate alone returns accuracy to 82.84%, adaptive noise alone reaches 83.16%, and the two together reach 83.56%. In other words, the latent-prediction loss helps only after the noise carries information about the image and the target it aligns to can be trusted. Under the same budget, the resulting model, *Bootstrap Your Own Noise* (BYON), improves over masked and denoising pre-training by 0.7 points on ImageNet-1K after fine-tuning and by 2.2 under a linear probe, by 1.4 mIoU on ADE20K, and by 1.7 box AP on COCO. It gains 0.8 to 3.5 points on six fine-grained benchmarks, and its margin over MAE does not shrink at 800 epochs or with ViT-L. Our evidence is empirical and covers models up to ViT-L trained on ImageNet-1K; code is provided in the supplementary material.

1 INTRODUCTION

Reconstruction and latent prediction teach a vision encoder different things. Masked image modeling (Bao et al., 2022; He et al., 2022; Xie et al., 2022) and its denoising relatives (Vincent et al., 2008; Wei et al., 2023; Chen et al., 2025) train an encoder to restore a corrupted input, and the resulting attention is local (Xie et al., 2023; Park et al., 2023). Joint-embedding methods train an encoder to predict the embedding that a slowly updated copy of itself assigns to another view (Grill et al., 2020; Caron et al., 2021; Assran et al., 2023), and the resulting attention is long-range and object-centric but coarse (Park et al., 2023). Dense prediction and fine-grained recognition need both properties, which is why a long line of work adds a joint-embedding loss to a reconstruction loss (Zhou et al., 2022; Dong et al., 2022; Assran et al., 2022; Tao et al., 2023; Huang et al., 2024).

Recent theory explains why the two objectives differ and suggests that joining them is delicate. Balestriero & LeCun (2024) prove that reconstruction spends capacity on the directions of largest input variance, which carry little information for perception, and that masking shifts this allocation whereas additive Gaussian noise does not. Littwin et al. (2024) show, for deep linear networks, that predicting in latent space instead favors *influential* features, those with large regression coefficients, and that this bias appears only when the encoder is deep. The two results share one reading. A reconstruction loss and a latent-prediction loss pull the same encoder toward different subspaces, and what the corruption removes from the input decides which of them prevails. Whether the two losses cooperate in practice is therefore an empirical question, and Jiang et al. (2023) already report conflicting gradients when contrastive learning and masked image modeling share all layers.

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.

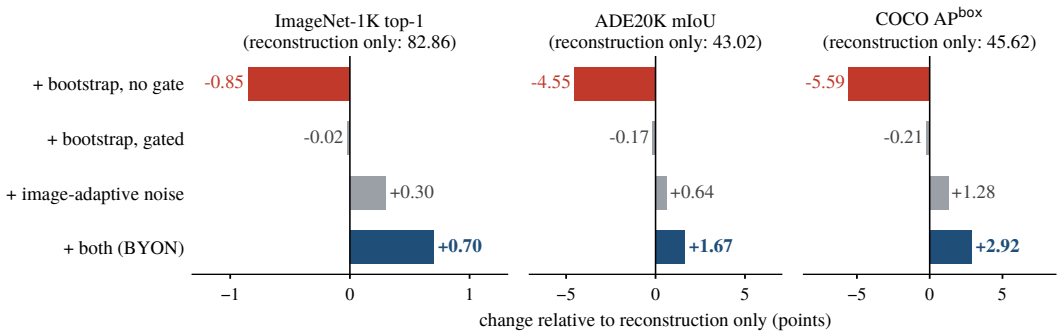

Figure 1: **Latent prediction helps reconstruction only when noise is informative and alignment is gated.** Each bar is the change relative to a ViT-B that reconstructs masked patches from uniformly noised token embeddings (400 epochs, ImageNet-1K pre-training, fine-tuned on each benchmark). Adding a bootstrapped latent-prediction loss lowers ImageNet-1K accuracy by 0.85 points and ADE20K mIoU by 4.55; gating that loss by reconstruction error makes it neutral; image-adaptive noise alone adds 0.30 points and 0.64 mIoU. Gate and adaptive noise together add 0.70 points, 1.67 mIoU, and 2.92 box AP, more than the sum of their separate effects on every benchmark. Takeaway: the two objectives conflict by default, and two small changes turn the conflict into cooperation.

We study this question in the setting where it is sharpest: masked image modeling with noise injected into the token embeddings, the latent counterpart of diffusion-style pre-training (Wei et al., 2023; Zheng et al., 2024; Chen et al., 2025). Under a fixed budget (ViT-B, 400 epochs on ImageNet-1K, identical code and hardware for every method), we find that neither addition helps by itself (Figure 1). Uniform Gaussian noise on top of masking changes ImageNet-1K fine-tuning accuracy from 82.89% to 82.86% and lowers COCO box AP from 46.04 to 45.62, as the analysis of Balestriero & LeCun (2024) predicts. A BYOL-style loss between the class token of the corrupted view and that of a clean view, added to this model, lowers accuracy further to 82.01%, ADE20K mIoU from 43.02 to 38.47, and COCO box AP from 45.62 to 40.03. The drop of 0.85 points is four times the run-to-run standard deviation we measure over five seeds.

We argue that the two failures have one cause. Uniform noise is independent of image content, so the restoration task it creates rewards the same high-variance directions as before. An encoder trained on this task offers the latent-prediction loss little that is worth aligning. Early in training the reconstruction is also poor. The latent-prediction loss then aligns a corrupted view whose content the encoder has not yet recovered, and its gradient competes with the reconstruction gradient instead of refining it. If this account is right, two changes should suffice: make the noise depend on the image, and let alignment begin only once reconstruction is reliable.

We make exactly these two changes. *Image-specific adaptive noise* (IAN) sets the noise level of each image from the saliency of its tokens, measured by the attention each token receives. Images with more salient content are corrupted more strongly. IAN extends saliency-guided masking (Kakogeorgiou et al., 2022; Li et al., 2022; Choi et al., 2024b), which decides *which* tokens to hide, with a decision about *how strongly* to perturb the tokens that remain. *Semantic uncertainty estimation* (SUE) converts the current reconstruction error into a weight between zero and one on the latent-prediction loss, which acts as a curriculum (Bengio et al., 2009) that needs no additional parameters. We call the resulting model *Bootstrap Your Own Noise* (BYON).

The account survives the test. SUE alone returns accuracy to 82.84%, IAN alone raises it to 83.16%, and the two together reach 83.56%. The latent loss that was neutral under uniform noise (-0.02) is worth 0.40 points once the noise is adaptive, and the joint gain of 0.70 points exceeds the sum of the separate effects (-0.02 and $+0.30$). On dense prediction the pattern is stronger: the gated loss is neutral under uniform noise (-0.17 mIoU, -0.21 box AP) and worth 1.03 mIoU and 1.64 box AP under adaptive noise, and the gate still matters once the noise is adaptive, since without it the same model loses 0.34 points, 0.89 mIoU, and 1.63 box AP. The depth dependence that Littwin et al. (2024) derive is also visible in our models. In the last blocks the heads of masked and denoising encoders converge to nearly one attention distance (Raghu et al., 2021), whereas BYON keeps both

short- and long-range heads at every depth (Section 4.4). Its class-token attention overlaps ground-truth object masks with a mean IoU of 0.71, against 0.58 for DINO, 0.48 for BYOL, and 0.35 for MAE. The finding reaches beyond this model: whenever one encoder is trained with a generative and a predictive objective, what the corruption removes decides whether the two help each other.

In summary, we make three claims, each tested in the section indicated. (i) Under a fixed budget, uniform latent noise adds nothing to masked image modeling, and an ungated latent-prediction loss lowers ImageNet-1K accuracy by 0.85 points (Section 4.3). (ii) Image-adaptive noise and a reliability gate remove the conflict: their joint effect exceeds the sum of their separate effects by 1.2 mIoU on ADE20K and 1.9 box AP on COCO, and removing the gate from the full model costs 0.34 points, 0.89 mIoU, and 1.63 box AP (Section 4.3). (iii) Against MAE, SimMIM, DiffMAE, and MaskDiT, all re-trained with the same code and budget, BYON gains 0.7 points on ImageNet-1K (83.56 ± 0.16 against 82.86 ± 0.05 for the strongest of them, five seeds). It gains 2.2 points under a linear probe, 1.4 mIoU on ADE20K, 1.7 box AP on COCO, and 0.8 to 3.5 points on six fine-grained benchmarks, and its margin over MAE is 1.0 points at 800 epochs and 0.8 points with ViT-L (Section 4).

2 Related Work

Reconstruction from corrupted inputs. Denoising autoencoders learn features by restoring an input from a corrupted copy (Vincent et al., 2008; 2010), and context encoders replaced pixel noise with missing regions (Pathak et al., 2016). Masked language modeling (Devlin et al., 2019) and the Vision Transformer (Dosovitskiy et al., 2021) turned this recipe into masked image modeling, first over pixel sequences (Chen et al., 2020a) and then over patches, with discrete tokens (Bao et al., 2022), raw pixels (He et al., 2022; Xie et al., 2022), hand-crafted features (Wei et al., 2022), or latent codes (Chen et al., 2024) as targets. Theory followed practice. Zhang et al. (2022) relate masked autoencoding to an implicit alignment of mask-induced views, and Kong et al. (2023) show that the masking ratio and patch size select which level of a latent hierarchy is recovered. Balestriero & LeCun (2024) prove that reconstruction favors high-variance directions and that masking, unlike additive Gaussian noise, changes this preference. These analyses study one objective in isolation. We take the last result as a hypothesis about what happens when noise is added to masking, confirm it at ViT-B scale, and ask what kind of noise escapes it.

Noise and diffusion as a pretext task. Diffusion models restore data from Gaussian noise at many levels (Ho et al., 2020; Nichol & Dhariwal, 2021), operate in a learned latent space (Rombach et al., 2022), and scale with Transformers (Peebles & Xie, 2023). Masking has been added to them to improve or accelerate generation (Gao et al., 2023; Hatamizadeh et al., 2024; Zheng et al., 2024). Their encoders carry recognizable semantics (Xiang et al., 2023), which motivated denoising as a pre-training signal: through a bottleneck (Hudson et al., 2024), inside an autoregressive model (Li et al., 2024), or on the masked patches of an autoencoder (Wei et al., 2023). Chen et al. (2025) deconstruct a diffusion model into a classical denoising autoencoder and find that noising a latent space, not the pixels, is the component that matters. In all of these methods the noise level is drawn independently of the image. We show that such noise, even in latent space, adds nothing to masked reconstruction under a matched budget, and that tying the level to image content is what makes it useful. REPA (Yu et al., 2025) is the mirror image of our setting: it aligns the features of a noised input to those of a clean input from a frozen, externally pre-trained encoder to help generation. We align to a target that the model bootstraps from itself, for recognition, and find that this alignment has to be gated.

Latent prediction and self-distillation. Contrastive methods learn embeddings that are invariant across views by separating positives from negatives (Chen et al., 2020b; He et al., 2020). BYOL removed the negatives by predicting the output of a momentum copy of the encoder (Grill et al., 2020), and later work simplified the recipe (Chen & He, 2021), moved it to Transformers (Caron et al., 2021; Chen et al., 2021), and replaced the asymmetry with explicit regularizers (Bardes et al., 2022). Joint-embedding predictive architectures predict the embeddings of hidden regions from visible ones in images (Assran et al., 2023) and video (Bardes et al., 2024). On the theory side, Tian et al. (2021) explain why a predictor and a stop-gradient prevent collapse, and Garrido et al. (2023b) show that contrastive and non-contrastive criteria are close to equivalent. Littwin et al. (2024) show that latent prediction in deep linear networks favors influential features over high-variance ones. We

do not propose a new architecture of this family. We use its simplest member, a BYOL-style loss on the class token, as a probe, and ask under which conditions its bias toward influential features survives next to a reconstruction loss.

Combining reconstruction with a joint-embedding loss. MST (Li et al., 2021) and iBOT (Zhou et al., 2022) add masked prediction to self-distillation, and BootMAE (Dong et al., 2022) and data2vec (Baevski et al., 2022) regress the features of a momentum encoder. MSN (Assran et al., 2022) and SiameseIM (Tao et al., 2023) match a masked view to an intact one, CMAE (Huang et al., 2024) adds a contrastive branch to MAE, dBOT (Liu et al., 2024) distills in stages, and DINOv2 (Oquab et al., 2024) combines image- and patch-level distillation at scale. These methods sum the losses with fixed weights from the first step and study masking only; Park et al. (2023) report that even a linear combination of a contrastive and a masked objective helps, whereas Jiang et al. (2023) observe conflicting gradients between the two and assign them to different layers. We find that once the corruption includes noise the sum turns harmful, by 0.85 points, and resolve the conflict without splitting the network: the weight of the latent loss follows the reliability of the reconstruction, and the noise is made informative so that there is something to align.

Content-adaptive corruption. Random masking ignores image content. AttMask hides the tokens that a teacher attends to most (Kakogeorgiou et al., 2022), ADIOS learns masks adversarially (Shi et al., 2022), and SemMAE masks semantic parts from easy to hard (Li et al., 2022). SBAM scores tokens by the attention they receive and adapts the masking ratio to the image (Choi et al., 2024b), and Choi et al. (2024a) study the masked token itself. All of these decide *which* tokens to remove. IAN reuses a saliency score of this kind for a different decision, *how strongly* to perturb the image, and its gain of 0.72 to 0.79 points is constant across masking ratios of 50% to 70%, which indicates that the two decisions are complementary.

3 METHOD

We first fix notation and state the two objectives (Section 3.1). We then show, in a linear model, why noise whose level ignores the image cannot change what reconstruction learns, and why noise whose level depends on the image can (Section 3.2). The two components of BYON follow: image-specific adaptive noise (Section 3.3) and the reliability gate (Section 3.4).

3.1 SETUP: TWO OBJECTIVES ON ONE ENCODER

An image $\boldsymbol{x}$ is split into N patches and embedded into tokens $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_N]^\top \in \mathbb{R}^{N \times D}$. A corruption operator $\mathcal{C}$ acts on tokens, an encoder f_θ maps the corrupted tokens to a class token and N patch tokens, and a light decoder g_ϕ maps patch tokens back to pixels. A reference encoder $f_{\bar{\theta}}$ sees the intact image, and its weights are an exponential moving average of θ (Grill et al., 2020).

Corruption. Let $\boldsymbol{m} \in \{0,1\}^N$ be a binary mask with $m_j = 1$ for hidden tokens, $\boldsymbol{\vartheta} \in \mathbb{R}^D$ a learned mask token, and $\bar{\alpha}_t \in (0,1)$ the signal rate of a variance-preserving diffusion process at level t (Ho et al., 2020). We define the corruption as follows:

$$\mathcal{C}_{\boldsymbol{m},t}(\boldsymbol{U}) \;=\; \sqrt{\bar{\alpha}_t}\,\big(\boldsymbol{m} \odot \boldsymbol{\vartheta} + (\boldsymbol{1} - \boldsymbol{m}) \odot \boldsymbol{U}\big) \;+\; \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{1}$$

Masking replaces hidden tokens by $\boldsymbol{\vartheta}$, and noise then perturbs the whole sequence, hidden positions included, in the embedding space rather than in pixels (Chen et al., 2025). The mask follows saliency-based masking (Choi et al., 2024b); the level t is the subject of Section 3.3.

Reconstruction. Let $(\boldsymbol{c}^s, \boldsymbol{Z}^s) = f_\theta(\mathcal{C}_{\boldsymbol{m},t}(\boldsymbol{U}), t)$ be the class and patch tokens of the corrupted view, and let $\hat{\boldsymbol{x}} = g_\phi(\boldsymbol{Z}^s)$. The reconstruction loss compares $\hat{\boldsymbol{x}}$ on the hidden patches $\mathcal{M} = \{j : m_j = 1\}$ with a pixel target $\boldsymbol{y}_t$ diffused to the same level, as follows:

$$\mathcal{L}_{\mathrm{rec}}(\theta, \phi) \;=\; \mathbb{E}_{\boldsymbol{x},\boldsymbol{m},t}\,\frac{1}{|\mathcal{M}|}\sum_{j \in \mathcal{M}} \big\|\hat{\boldsymbol{x}}_j - \boldsymbol{y}_{t,j}\big\|_1, \qquad \boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{x} + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}'. \tag{2}$$

Because $\boldsymbol{\epsilon}'$ is independent of the input and has zero median, the noise in the target cannot be predicted, and the decoder is trained toward the clean patch attenuated by the signal rate $\sqrt{\bar{\alpha}_t}$; heavily corrupted images therefore contribute a proportionally smaller target.

Latent prediction. Let $\boldsymbol{c}^r$ be the class token of $f_{\bar{\theta}}$ on the intact image. With a projector p, a predictor h, a momentum projector $\bar{p}$, and a stop-gradient sg, the latent-prediction loss is defined as follows:

$$\mathcal{L}_{\text{lat}}(\theta) \;=\; \mathbb{E}_{\boldsymbol{x},\boldsymbol{m},t}\left\|h\big(p(\boldsymbol{c}^s)\big) - \text{sg}\big[\bar{p}(\boldsymbol{c}^r)\big]\right\|_2^2. \tag{3}$$

This is the BYOL objective with one change: the two views are not two augmentations but a corrupted and an intact copy of the same image, so the loss asks the encoder to recover, in latent space, what the corruption removed. The naive combination studied in Section 1 is $\mathcal{L}_{\text{rec}} + \beta\,\mathcal{L}_{\text{lat}}$ with t drawn uniformly.

3.2 Why uniform noise is neutral and adaptive noise is not

We analyze the restoration task in the linear setting of Balestriero & LeCun (2024). Let $\boldsymbol{x} \in \mathbb{R}^d$ be zero-mean with covariance $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, and let a linear map $\boldsymbol{W}$ restore $\boldsymbol{x}$ from $\tilde{\boldsymbol{x}} = a\,\boldsymbol{x} + b\,\boldsymbol{\epsilon}$ by minimizing $\frac{1}{2}\mathbb{E}\|\boldsymbol{W}\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$ with gradient flow from $\boldsymbol{W}(0) = \boldsymbol{0}$.

Proposition 1 (Uniform noise preserves the order of learning)**.** *If $a > 0$ and $b > 0$ are constants, then $\boldsymbol{W}(\tau) = \boldsymbol{V}\,\text{diag}\big(g_i\,(1-e^{-s_i\tau})\big)\boldsymbol{V}^\top$ with gains $g_i = a\lambda_i/(a^2\lambda_i+b^2)$ and rates $s_i = a^2\lambda_i+b^2$. Both g_i and s_i increase with λ_i for every such (a, b), and the same holds when masking is applied before the noise (Section A).*

That is, the restorer shares the eigenvectors of the data, and it learns them in order of variance whatever the noise level. Noise of this kind rescales the problem and leaves its preferences intact, which is the behaviour we measure at ViT-B scale (Section 4.3).

Proposition 2 (Image-dependent noise reweights the second moments)**.** *If $a = a(\boldsymbol{x})$ and $b = b(\boldsymbol{x})$ depend on the image and $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{x}$, the minimizer is*

$$\boldsymbol{W}^\star = \boldsymbol{\Sigma}_a\big(\boldsymbol{\Sigma}_{a^2} + \bar{b}^2\boldsymbol{I}\big)^{-1}, \qquad \boldsymbol{\Sigma}_\omega = \mathbb{E}\big[\omega(\boldsymbol{x})\,\boldsymbol{x}\boldsymbol{x}^\top\big], \quad \bar{b}^2 = \mathbb{E}\big[b(\boldsymbol{x})^2\big], \tag{4}$$

and the rates of gradient flow are the eigenvalues of $\boldsymbol{\Sigma}_{a^2} + \bar{b}^2\boldsymbol{I}$.

In other words, the task is no longer governed by $\boldsymbol{\Sigma}$ but by moments in which each image is weighted by its own signal rate. When $a(\boldsymbol{x})$ correlates with image content, these moments need not share the eigenvectors of $\boldsymbol{\Sigma}$ or the order of its eigenvalues. Section A gives the proofs and a two-component example in which the direction of smaller variance becomes the one learned first. The propositions do not say that every image-dependent level helps perception. They say that an image-dependent level is the minimal change that leaves the regime of Proposition 1, and Section 3.3 specifies one such level.

3.3 Image-specific adaptive noise (IAN)

IAN chooses t from the image itself. Let $\boldsymbol{A} = \text{softmax}(\boldsymbol{U}\boldsymbol{U}^\top) \in \mathbb{R}^{N\times N}$ be the row-normalized token affinity. The saliency of token j is the total attention it receives, $s_j = \sum_i A_{ij}$, rescaled to $\bar{s}_j \in [0, 1]$ within the image (Choi et al., 2024b). We summarize an image by the fraction of its tokens that are salient, and we map this fraction to a noise level as follows:

$$\rho(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\big[\bar{s}_j > \delta\big], \qquad t(\boldsymbol{x}) = \min\big(\kappa\,\rho(\boldsymbol{x}),\, T_{\max}\big) + \xi, \quad \xi \sim \mathcal{U}[0, T_\xi]. \tag{5}$$

An image whose saliency is spread over many tokens receives a high level, and an image dominated by a few tokens receives a low one; the jitter ξ keeps the level stochastic. Substituting $t(\boldsymbol{x})$ into Equation (1) gives $a(\boldsymbol{x}) = \sqrt{\bar{\alpha}_{t(\boldsymbol{x})}}$, which is the setting of Proposition 2. IAN needs no additional network: the saliency comes from the token embeddings that the encoder already computes. Uniform noise is the special case in which t is drawn independently of $\boldsymbol{x}$. Section B reports the effect of IAN under different masking ratios.

3.4 Semantic uncertainty estimation (SUE): a reliability gate

The latent loss in Equation (3) is informative only if the class token of the corrupted view already encodes the image. We measure this with the quantity that reconstruction provides for free. Let $\hat{\mathcal{L}}_{\text{rec}}^{(k)}$

Table 1: **Comparison under a matched budget** (ViT-B/16, 400 epochs, ImageNet-1K pre-training). ImageNet-1K and ADE20K are mean ± std over five seeds; COCO is a single run per method. BYON improves over the strongest baseline in each column by 0.70 points, 1.44 mIoU, and 1.7 box AP.

Corruption	Method	ImageNet-1K top-1	ADE20K mIoU	COCO AP^{box}
Masking	SimMIM	82.54 ± 0.15	42.62 ± 0.16	46.8
	MAE	82.78 ± 0.20	43.02 ± 0.12	46.0
Masking + uniform noise	DiffMAE	82.60 ± 0.33	42.68 ± 0.48	44.9
	MaskDiT	82.86 ± 0.05	42.90 ± 0.11	43.3
Masking + adaptive noise	BYON	$\mathbf{83.56 \pm 0.16}$	$\mathbf{44.46 \pm 0.21}$	**48.5**

be the reconstruction loss of the minibatch at step k. SUE maps it to a weight in $(0, 1]$, and BYON minimizes the gated objective, as follows:

$$w_k = 1 - \tanh\big(\gamma\, \hat{\mathcal{L}}_{\text{rec}}^{(k)}\big), \qquad \mathcal{L}_{\text{BYON}}^{(k)} = \hat{\mathcal{L}}_{\text{rec}}^{(k)} + \beta\, w_k\, \hat{\mathcal{L}}_{\text{lat}}^{(k)}. \tag{6}$$

The gate is a decreasing function of the reconstruction error, so it is close to zero while the encoder cannot yet restore the input and approaches one as restoration improves. Two limits clarify its role: $w_k \equiv 1$ recovers the naive sum, and $w_k \equiv 0$ recovers reconstruction alone. Between them, Equation (6) is a curriculum (Bengio et al., 2009) whose pace is set by the model and not by a schedule. We test this reading in Section 4.3 by replacing w_k with a fixed ramp and with an off-the-shelf uncertainty estimator (van Amersfoort et al., 2020).

IAN and SUE interact through $\mathcal{L}_{\text{rec}}$. IAN raises the difficulty of restoration for content-rich images, which delays the gate for the batches that contain them, and the gate in turn admits the latent loss only when the corrupted view is recoverable. The reference branch adds one forward pass without gradients. In our setting, pre-training takes 97 GPU hours against 84 for DiffMAE (Section 4.5).

4 Experiments

4.1 Setup

Protocol. Every method is pre-trained with the same code base, data pipeline, and budget: ViT-B/16 (Dosovitskiy et al., 2021), 400 epochs on ImageNet-1K (Deng et al., 2009) without labels, and four A100 GPUs. We compare with masked image modeling (SimMIM (Xie et al., 2022), MAE (He et al., 2022)), denoising pre-training (DiffMAE (Wei et al., 2023), MaskDiT (Zheng et al., 2024)), and joint-embedding methods with and without masking (BYOL (Grill et al., 2020), DINO (Caron et al., 2021), iBOT (Zhou et al., 2022), CMAE (Huang et al., 2024)). Absolute numbers under this budget are below those of the original papers, which use longer schedules and larger batches; the comparison is between methods at equal cost, since numbers obtained with 8 to 64 GPUs would confound the method with the budget.

Evaluation. We fine-tune for 100 epochs on ImageNet-1K with AdamW (Loshchilov & Hutter, 2019) and report top-1 accuracy. We also evaluate the frozen encoder with a linear probe, which measures the representation without adapting it. We transfer to ADE20K (Zhou et al., 2017) for semantic segmentation (mIoU), to COCO (Lin et al., 2014) with Mask R-CNN (He et al., 2017) for detection (box AP), and to six fine-grained benchmarks (Wah et al., 2011; Van Horn et al., 2015; 2018; Krause et al., 2013; Maji et al., 2013). ImageNet-1K and ADE20K results in Tables 1 and 2 are the mean and standard deviation of five seeds unless stated otherwise. Section C lists the protocols.

4.2 Main results

Table 1 tests claim (iii) against the reconstruction methods, and Table 2 places all nine methods side by side. The two denoising baselines do not improve on masking: DiffMAE and MaskDiT stay within 0.2 points of MAE after fine-tuning and fall below it on COCO, which is the neutrality that Proposition 1 describes. Adding an ungated joint-embedding loss to masking does not help either (iBOT 82.03%, CMAE 82.89%); BYON is the only method above 83%, with a margin of 0.67 points

Table 2: **All methods under the same budget** (ImageNet-1K top-1, %). FT: fine-tuning; iBOT, CMAE, and the methods of Table 1 are means over five seeds, BYOL and DINO single runs. Linear: a linear classifier on the frozen encoder, one recipe for every method (Section C), single run; absolute values are below published ones for every method for that reason.

	BYOL	DINO	SimMIM	MAE	DiffMAE	MaskDiT	iBOT	CMAE	BYON
Masked reconstruction			✓	✓	✓	✓	✓	✓	✓
Latent noise					✓	✓			✓
Joint-embedding loss	✓	✓					✓	✓	✓
FT	80.27	82.81	82.54	82.78	82.60	82.86	82.03	82.89	**83.56**
Linear	57.1	62.7	61.4	62.3	62.0	61.9	60.3	62.4	**64.5**

Table 3: **Fine-grained classification** (top-1, %). BYON improves over the strongest baseline on all six benchmarks, by 0.8 to 3.5 points.

Method	CUB	NABirds	iNat-2017	iNat-2018	Cars	Aircraft
SimMIM	75.4	73.6	65.2	69.1	88.4	73.8
MAE	79.1	77.8	68.9	70.8	92.1	81.6
DiffMAE	78.5	76.5	68.8	64.3	90.3	79.7
MaskDiT	78.8	77.3	68.3	68.8	92.0	80.9
BYON	**82.5**	**81.3**	**69.7**	**72.8**	**92.9**	**84.9**

over CMAE and 0.70 over MaskDiT, three or more times the larger standard deviation in each pair, and 1.44 mIoU over MAE on ADE20K. The frozen encoder shows a larger gap: under the linear probe BYON leads MAE by 2.2 points and DINO by 1.8, while every other method stays within 0.4 points of MAE or below it.

Table 3 shows the largest margins where classes differ by small parts (CUB $+3.4$, NABirds $+3.5$, Aircraft $+3.3$); Section B extends the comparison to a vision-language model.

4.3 Ablation: a factorial test of the two changes

Uniform noise adds nothing. Table 4 tests claims (i) and (ii). Row U against row N isolates the noise: accuracy moves from 82.89% to 82.86% and COCO box AP falls by 0.42. This agrees with Proposition 1 and with the denoising baselines of Table 1.

An ungated latent loss hurts, and the gate removes the harm. Adding Equation (3) with a constant weight lowers accuracy by 0.85 points (row U+L), about three standard deviations of a difference between two runs, and the damage is far larger on transfer: 4.55 mIoU on ADE20K and 5.59 box AP on COCO. With the gate the loss is neutral (row U+gL, -0.02 points, -0.17 mIoU, -0.21 box AP): the gate prevents damage, but with uniform noise it has nothing to add.

The gain appears once the noise is adaptive. IAN alone adds 0.30 points, 0.64 mIoU, and 1.28 box AP (row A), and IAN with the gated loss adds 0.70 points, 1.67 mIoU, and 2.92 box AP (row A+gL). The gated loss that is neutral under uniform noise (-0.02, -0.17, -0.21) is therefore worth 0.40 points, 1.03 mIoU, and 1.64 box AP under IAN. The joint effect of the two changes exceeds the sum of their separate effects by 0.42 points on ImageNet-1K, 1.20 mIoU on ADE20K, and 1.85 box AP on COCO; the first is within the noise of four single runs, the last two are several times the seed deviation of Table 1. Without noise the gated loss gives a smaller gain (row N+gL, $+0.13$ points, $+0.42$ mIoU, $+1.15$ box AP over row N).

The gate is still needed under adaptive noise. Row A+L removes the gate from the full model. Adaptive noise removes the harm of the ungated loss (row A+L is within 0.14 of row A on every benchmark) but not the need for the gate: the gated model is 0.34 points, 0.89 mIoU, and 1.63 box AP higher, so informative noise and a reliable target are separate conditions.

Table 4: **Factorial ablation.** Rows U and A use a uniform or an image-adaptive (IAN) noise level; rows N mask without noise. The latent loss is absent, ungated, or gated. Δ is the change on ImageNet-1K relative to row U. Single runs; the seed deviation in Table 1 is 0.05 to 0.33.

Row	Noise level	Latent loss	ImageNet-1K	Δ	ADE20K	COCO AP^{box}
U	uniform	none	82.86	–	43.02	45.62
U+L	uniform	ungated	82.01	−0.85	38.47	40.03
U+gL	uniform	gated (SUE)	82.84	−0.02	42.85	45.41
A	IAN	none	83.16	+0.30	43.66	46.90
A+L	IAN	ungated	83.22	+0.36	43.80	46.91
A+gL	IAN	gated (SUE)	**83.56**	+0.70	**44.69**	**48.54**
A+ramp	IAN	fixed ramp	82.93	+0.07	–	–
A+DUQ	IAN	gated (DUQ)	83.02	+0.16	–	–
N	none	none	82.89	+0.03	42.91	46.04
N+gL	none	gated (SUE)	83.02	+0.16	43.33	47.19

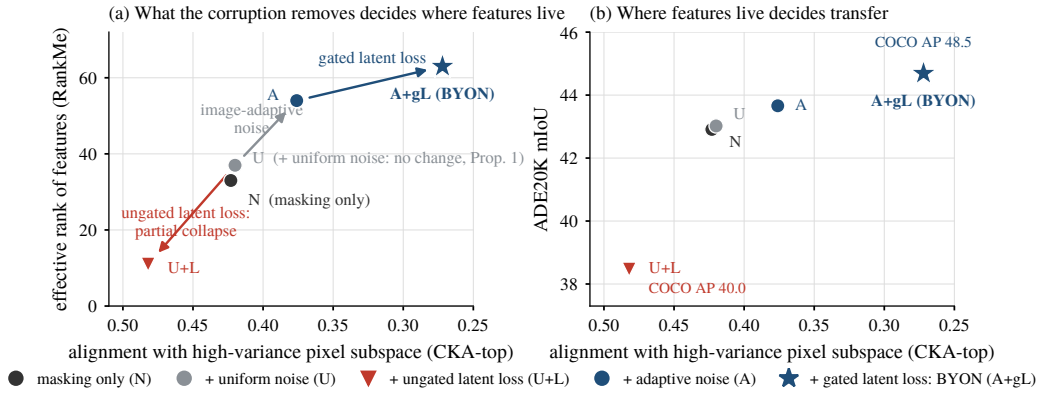

Figure 2: **What the corruption removes decides where the features live, and where they live decides transfer.** Five configurations of Table 4, measured on the frozen encoder: alignment of the class-token features with the pixel subspace that explains 90% of image variance (CKA-top; lower is further from it) against their effective rank (a) and against ADE20K mIoU (b). Uniform noise leaves masking where it was, as Proposition 1 predicts; an ungated latent loss collapses the features toward the high-variance subspace (rank 11) and costs 4.55 mIoU; adaptive noise moves them away, and the gated latent loss moves them furthest (rank 63), where transfer is best. The five points fall on one monotone curve in (b).

The form of the gate. A linear ramp from 0 to 1 over the first 120 epochs, where the measured gate passes 0.5 (Figure 3), reaches 82.93% (row A+ramp), below SUE by 0.63 points and below no gate at all (row A+L); an uncertainty estimator that does not see the reconstruction task (row A+DUQ) reaches 83.02%. The gate is neither a warm-up schedule nor a generic uncertainty weight. Section B gives the full component study, the effect of IAN across masking ratios, and a comparison of one and two decoder heads.

4.4 Analysis: Testing the Account

Depth. If the latent loss acts through a bias that needs depth (Littwin et al., 2024), encoders with and without it should differ in the last blocks and not in the first. Figure 4 (Section D) plots the mean attention distance (Raghu et al., 2021) of every head: in blocks 1 to 5 the three encoders look alike, from block 7 on the heads of the masked and, more strongly, of the denoising encoder converge to nearly one distance, whereas BYON keeps heads of short and long range up to block 12. Its class-token attention also overlaps ground-truth object masks more closely: mean IoU 0.71 against 0.58 for DINO, 0.48 for BYOL, and 0.35 for MAE.

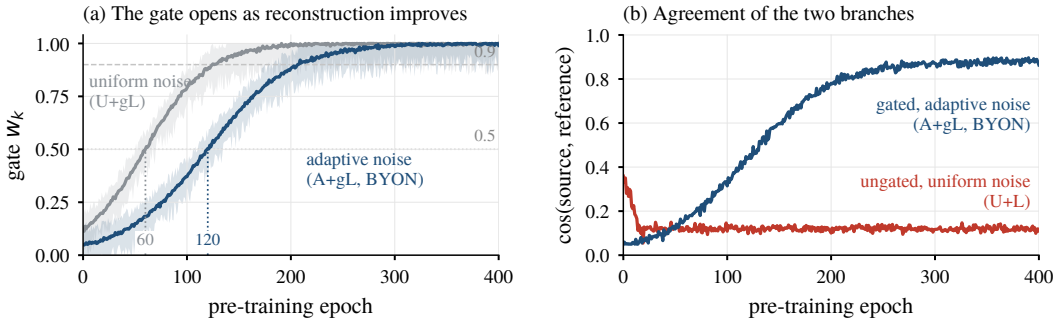

Figure 3: **The gate over training.** (a) Mean and range over batches of w_k per epoch: the gate opens as reconstruction improves, and later under adaptive noise (0.5 at epoch 120, 0.9 at 207) than under uniform noise (60, 125). (b) Cosine similarity between the source and reference projections: gated, it rises from 0.05 to 0.86 as the gate opens; ungated, it falls within 20 epochs and stays near 0.11.

Spectrum. If uniform noise leaves the preference for high-variance directions intact and IAN does not, the features should align differently with the principal subspace of the pixels. Figure 2 confirms this: uniform noise leaves masking where it was, the ungated latent loss collapses the features toward the high-variance subspace (rank 11), adaptive noise moves them away, and the gated latent loss moves them furthest (CKA 0.420 to 0.272, rank 37 to 63); across the five configurations, distance from that subspace and ADE20K mIoU rise together (Table 9).

Pace. If the gate is a self-paced curriculum, it should open gradually, later when restoration is harder, and the branches should agree only once it does. In Figure 3 it passes 0.5 at epoch 120 under adaptive noise, twice as late as under uniform noise, so its pace is set by the difficulty of restoration; ungated, branch agreement collapses within 20 epochs and never recovers, the training-time counterpart of the rank collapse in Figure 2.

4.5 Scale and cost

The margin over MAE is 0.78 points at 400 epochs (83.56 against 82.78), 1.0 at 800 (84.4 against 83.4), and 0.8 with ViT-L (85.7 against 84.9), so BYON at 400 epochs matches MAE at 800 (Table 8, Section C); the reference branch costs 40% more FLOPs per step and 15% more GPU hours than DiffMAE (63.8 GFLOPs and 97 hours against 45.5 and 84).

5 Limitations

Our propositions concern a linear restorer and explain the neutrality of uniform noise; why the latent loss profits from adaptive noise is established empirically, and a theory of that half remains open.

6 Conclusion

Reconstruction and latent prediction do not cooperate by default. Under a matched budget, uniform latent noise added nothing to masked image modeling, and a bootstrapped latent loss on top of it cost 0.85 points. Two changes reversed this: a noise level that depends on the image, and a gate that admits the latent loss only as far as reconstruction is reliable. With both, the same latent loss improved ImageNet-1K, ADE20K, COCO, and six fine-grained benchmarks, moved the features away from the high-variance pixel subspace, and kept its margin at 800 epochs and with ViT-L. A linear analysis explains the first half of this result, since image-independent noise cannot change what a restorer learns first; the second half is the lesson we expect to carry: whenever a generative and a predictive objective share an encoder, what the corruption removes decides whether they help each other.

Reproducibility statement

The supplementary material contains the code and configuration of BYON; the baselines use their official implementations where available, all under the protocol of Section 4.1 and Section C. ImageNet-1K and ADE20K results in Tables 1 and 2 report mean and standard deviation over five seeds. The assumptions and complete proofs of Propositions 1 and 2 are given in Section A, together with a numerical example. The measurements of Section 4.4 are defined in Section D.

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.

References

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, 2022.

Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 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*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 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*, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

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

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020b.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations*, 2025.

Hyesong Choi, Hunsang Lee, Seyoung Joung, Hyejin Park, Jiyeong Kim, and Dongbo Min. Emerging property of masked token for effective pre-training. In *European Conference on Computer Vision*, 2024a.

Hyesong Choi, Hyejin Park, Kwang Moo Yi, Sungmin Cha, and Dongbo Min. Salience-based adaptive masking: Revisiting token dynamics for enhanced pre-training. In *European Conference on Computer Vision*, pp. 343–359, 2024b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision BERT pretraining. In *European Conference on Computer Vision*, pp. 247–264, 2022.

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.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *IEEE/CVF International Conference on Computer Vision*, pp. 23164–23173, 2023.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, 2023a.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *International Conference on Learning Representations*, 2023b.

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, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. DiffiT: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pp. 37–55, 2024.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2506–2517, 2024.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. SODA: Bottleneck diffusion models for representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23115–23127, 2024.

Ziyu Jiang, Yinpeng Chen, Mengchen Liu, Dongdong Chen, Xiyang Dai, Lu Yuan, Zicheng Liu, and Zhangyang Wang. Layer grafted pre-training: Bridging contrastive learning and masked image modeling for label-efficient representations. In *International Conference on Learning Representations*, 2023.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7918–7928, 2023.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. SemMAE: Semantic-guided masking for learning masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yazhe Li, Jörg Bornschein, and Ting Chen. Denoising autoregressive representation learning. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27417–27438, 2024.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. MST: Masked self-supervised transformer for visual representation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13165–13176, 2021.

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*, 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*, volume 37, 2024.

Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. In *International Conference on Learning Representations*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. Technical report, arXiv:1306.5151, 2013.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 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*, 2023.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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 *International Conference on Machine Learning*, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Yuge Shi, N. Siddharth, Philip H. S. Torr, and Adam R. Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, 2022.

Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, 2021.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, 2020.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11, 2010.

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, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 16284–16294, 2023.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *IEEE/CVF International Conference on Computer Vision*, pp. 15802–15812, 2023.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*, 2018.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14475–14485, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

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.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 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*, 2022.

APPENDIX

Section A proves the two propositions of Section 3.2 and gives a worked example. Section B contains the full component study and the ablations referred to in Sections 3.3 and 4.3. Section C lists the training and transfer protocols. Section D defines the measurements of Section 4.4 and shows further attention maps.

A PROOFS AND AN EXAMPLE FOR SECTION 3.2

Setting. $\boldsymbol{x} \in \mathbb{R}^d$ is zero-mean with covariance $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1 \geq \cdots \geq \lambda_d > 0)$. The corrupted input is $\tilde{\boldsymbol{x}} = a\,\boldsymbol{x} + b\,\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ independent of $\boldsymbol{x}$. A linear restorer $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ minimizes $\ell(\boldsymbol{W}) = \frac{1}{2}\mathbb{E}\|\boldsymbol{W}\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$ by gradient flow $\dot{\boldsymbol{W}} = -\nabla\ell(\boldsymbol{W})$ from $\boldsymbol{W}(0) = \boldsymbol{0}$. Writing $\boldsymbol{S} = \mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top]$ and $\boldsymbol{C} = \mathbb{E}[\boldsymbol{x}\tilde{\boldsymbol{x}}^\top]$, we have $\ell(\boldsymbol{W}) = \frac{1}{2}\operatorname{tr}(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}^\top) - \operatorname{tr}(\boldsymbol{W}\boldsymbol{C}^\top) + \frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}$ and $\dot{\boldsymbol{W}} = -(\boldsymbol{W}\boldsymbol{S} - \boldsymbol{C})$.

Proof of Proposition 1. With constant a, b, independence and $\mathbb{E}\boldsymbol{\epsilon} = \boldsymbol{0}$ give $\boldsymbol{S} = a^2\boldsymbol{\Sigma} + b^2\boldsymbol{I}$ and $\boldsymbol{C} = a\boldsymbol{\Sigma}$. Both are diagonal in the basis $\boldsymbol{V}$, so the substitution $\boldsymbol{W} = \boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{V}^\top$ decouples the flow entrywise: $\dot{\tilde{W}}_{ij} = -\big(s_j\tilde{W}_{ij} - a\lambda_i\delta_{ij}\big)$ with $s_j = a^2\lambda_j + b^2$. From $\tilde{\boldsymbol{W}}(0) = \boldsymbol{0}$ the off-diagonal entries remain zero, and the diagonal entries solve $\tilde{W}_{ii}(\tau) = g_i\,(1 - e^{-s_i\tau})$ with $g_i = a\lambda_i/s_i$. Hence $\boldsymbol{W}(\tau) = \boldsymbol{V}\operatorname{diag}\big(g_i(1 - e^{-s_i\tau})\big)\boldsymbol{V}^\top$. Finally, $\partial s_i/\partial\lambda_i = a^2 > 0$ and $\partial g_i/\partial\lambda_i = ab^2/s_i^2 > 0$ for every $a, b > 0$, so the gain and the rate of each direction increase with its variance. □

Proof of Proposition 2. Let $a = a(\boldsymbol{x})$ and $b = b(\boldsymbol{x})$ be measurable functions of the image with finite second moments. Since $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{x}$ and zero-mean, the cross terms vanish: $\boldsymbol{S} = \mathbb{E}[a(\boldsymbol{x})^2\boldsymbol{x}\boldsymbol{x}^\top] + \mathbb{E}[b(\boldsymbol{x})^2]\,\boldsymbol{I} = \boldsymbol{\Sigma}_{a^2} + \bar{b}^2\boldsymbol{I}$ and $\boldsymbol{C} = \mathbb{E}[a(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\Sigma}_a$. $\boldsymbol{S}$ is positive definite, so the flow $\dot{\boldsymbol{W}} = -(\boldsymbol{W}\boldsymbol{S} - \boldsymbol{C})$ converges to the unique minimizer $\boldsymbol{W}^\star = \boldsymbol{C}\boldsymbol{S}^{-1} = \boldsymbol{\Sigma}_a(\boldsymbol{\Sigma}_{a^2} + \bar{b}^2\boldsymbol{I})^{-1}$, and in the eigenbasis of $\boldsymbol{S}$ each component converges at the rate of the corresponding eigenvalue of $\boldsymbol{S}$. □

Remark (masking before noise). BYON masks before it noises, so consider $\tilde{\boldsymbol{x}} = a\,\boldsymbol{M}\boldsymbol{x} + b\,\boldsymbol{\epsilon}$ with a random diagonal mask $\boldsymbol{M} \in \{0, 1\}^{d\times d}$ independent of $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$. Then $\boldsymbol{S} = a^2\,\mathbb{E}[\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}] + b^2\boldsymbol{I}$ and $\boldsymbol{C} = a\,\boldsymbol{\Sigma}\,\mathbb{E}[\boldsymbol{M}]$. Masking changes the problem, through $\mathbb{E}[\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}]$ and $\mathbb{E}[\boldsymbol{M}]$, but uniform noise still enters only as the scalar a and the isotropic term $b^2\boldsymbol{I}$: it shifts every rate by the same b^2 and rescales every gain, and cannot change the eigenvectors of $\boldsymbol{S}$ or the order of its eigenvalues. Whatever preference masking induces, image-independent noise on top of it leaves that preference where it is, which is the content of Proposition 1 in the setting of Section 4.3.

Remark (reduction). When $a(\boldsymbol{x})$ is constant on the support of $\boldsymbol{x}$, $\boldsymbol{\Sigma}_{a^2} = a^2\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_a = a\boldsymbol{\Sigma}$, and Proposition 2 reduces to Proposition 1. When $a(\boldsymbol{x})$ varies with the image, $\boldsymbol{\Sigma}_{a^2}$ is a covariance in which each image is weighted by its own signal rate, and it need not share the eigenvectors of $\boldsymbol{\Sigma}$ or the order of its eigenvalues. The propositions describe the restoration task; they do not characterize which weighting is best for perception, which is the empirical question of Section 4.3.

Example. Let $\boldsymbol{x} \in \mathbb{R}^2$ be drawn with equal probability from $\mathcal{N}(\boldsymbol{0}, \operatorname{diag}(4, \frac{1}{2}))$ or $\mathcal{N}(\boldsymbol{0}, \operatorname{diag}(\frac{1}{2}, 2))$, so that $\boldsymbol{\Sigma} = \operatorname{diag}(2.25, 1.25)$ and the first direction dominates. With a uniform signal rate $\bar{\alpha} = \frac{1}{2}$, the rates are $s = (1.625, 1.125)$: the first direction is learned first. Now assign $\bar{\alpha} = 0.1$ to samples of the first component and $\bar{\alpha} = 0.9$ to samples of the second, which keeps the mean signal rate at $\frac{1}{2}$. Then $\boldsymbol{\Sigma}_{a^2} + \bar{b}^2\boldsymbol{I} = \operatorname{diag}(0.925, 1.425)$, and the second direction, the one of smaller variance, is now learned first. The noise budget is identical in both cases; only its dependence on the image differs. A simulation with 4×10^5 samples returns $\boldsymbol{W}^\star = \operatorname{diag}(0.942, 0.721)$ and the two rates above to three decimals.

B ADDITIONAL ABLATIONS

Component study. Table 5 lists every combination of the five components on ImageNet-1K, with ADE20K and COCO for the configurations that Section 4.3 builds on. Rows of Table 4 correspond as follows: U = A7, U+L = A4, U+gL = A9, A = A8, A+gL = A10, N = A5, N+gL = A6. Two rows without masking are included for completeness. Noise alone (A2) reaches 80.14%, 2.75 points below masking alone (A5), and IAN raises it to 82.38% (A3), the largest single effect of IAN in the table: when noise is the only corruption, making it image-dependent matters most. A latent loss alone (A1, the BYOL objective on the class token) reaches 80.27%.

Table 5: **Component study** (ViT-B/16, 400 epochs). BS: bootstrapped latent prediction.

	Noise	Mask	BS	IAN	SUE	ImageNet-1K	ADE20K	COCO AP^{box}
A1			✓			80.27	–	–
A2	✓					80.14	36.72	39.21
A3	✓			✓		82.38	39.95	40.83
A4	✓	✓	✓			82.01	38.47	40.03
A5		✓				82.89	42.91	46.04
A6		✓	✓		✓	83.02	43.33	47.19
A7	✓	✓				82.86	43.02	45.62
A8	✓	✓		✓		83.16	43.66	46.90
A9	✓	✓	✓		✓	82.84	42.85	45.41
A10	✓	✓	✓	✓	✓	**83.56**	**44.69**	**48.54**

IAN across masking ratios. Table 6 shows that the gain from IAN is 0.79, 0.72, and 0.77 points at masking ratios of 50%, 60%, and 70%. The decision of how strongly to perturb is complementary to the decision of how much to hide, which is the distinction to saliency-guided masking drawn in Section 2.

Table 6: **IAN across masking ratios** (ImageNet-1K fine-tuning, top-1 %).

	50%	60%	70%
Uniform noise	82.25	82.84	82.77
IAN	83.04	83.56	83.54
Gain	+0.79	+0.72	+0.77

Strength of the saliency signal. The coefficient κ in Equation (5) sets how strongly the salient fraction $\rho(\boldsymbol{x})$ moves the noise level. Table 7 sweeps it. At $\kappa = 0$ the level no longer depends on the image and accuracy returns to the uniform-noise regime (82.94%); a mild dependence ($\kappa = 0.1$) gives the best result, and strong dependence ($\kappa = 1$) over-corrupts salient images and loses 1.9 points. The curve is consistent with Proposition 2: what matters is that the level covaries with content, and a moderate covariation is enough.

Table 7: **Saliency coefficient** κ (ImageNet-1K fine-tuning, top-1 %).

κ	0	0.1	0.5	1
Top-1	82.94	**83.56**	83.28	81.70

Gate estimators. Predictive uncertainty is usually learned as a loss attenuation (Kendall & Gal, 2017), estimated from ensembles (Lakshminarayanan et al., 2017), or read from a single deterministic network (van Amersfoort et al., 2020); these estimators add parameters or forward passes and are agnostic to the pretext task. Replacing SUE with the deterministic network of van Amersfoort et al. (2020), trained alongside the encoder without access to the reconstruction error, gives 83.02% against 83.56% (row A+DUQ of Table 4). SUE has no parameters and is a function of the pretext loss; the comparison indicates that it is this coupling, and not the sophistication of the estimator, that the gate needs.

One or two decoder heads. Giving the noised and the hidden patches separate decoder heads changes accuracy from 83.56% to 83.59%, so we keep a single head.

Vision-language pre-training. To test whether the two changes depend on the vision-only setting, we apply them to CLIP (Radford et al., 2021). Starting from the released ViT-B/16 image encoder, we fine-tune on ImageNet-1K with IAN applied to the visual tokens and the loss gated by SUE computed from a visual reconstruction, against the same encoder fine-tuned without them. Top-1 accuracy rises from 83.35% to 85.28%.

C Protocols

Pre-training. All methods use ViT-B/16 at 224×224 resolution, 400 epochs on ImageNet-1K without labels, the same augmentation, and four A100 GPUs. BYON masks 60% of the tokens and uses the saliency-based mask of Choi et al. (2024b); the noise level follows Equation (5) with $\kappa = 0.1$; the reference encoder is an exponential moving average of the source encoder; the gate uses $\gamma = 1$ and the loss weight β in Equation (6) is fixed for all experiments. For methods with public code we use the official implementation; DiffMAE is re-implemented from its paper. The 800-epoch and ViT-L runs of Table 8 keep every other setting.

Fine-tuning on ImageNet-1K. Every method follows the recipe of Xie et al. (2022): 100 epochs with AdamW, base learning rate 1.25×10^{-3}, 20 warm-up epochs and cosine decay, weight decay 0.05, layer-wise learning-rate decay 0.65, and stochastic depth 0.1. The seeds of Table 1 change the pre-training and the fine-tuning seed together.

Linear probe. Features are the concatenation of the class token and the mean of the patch tokens of the last block, followed by batch normalization; a linear classifier is trained on them for 100 epochs with center-crop inputs, and the learning rate is chosen among three values on a held-out tenth of the training set, never on the validation set. The same recipe is applied to every method, which is why the absolute numbers in Table 2 are below those published with per-method recipes.

Transfer. Semantic segmentation uses UperNet (Xiao et al., 2018) on ADE20K at 512×512 for 160k iterations and reports single-scale mIoU. Detection uses Mask R-CNN (He et al., 2017) on COCO with the same schedule for every method and reports box AP. The six fine-grained benchmarks are fine-tuned from the same checkpoint with the ImageNet-1K recipe above at 224×224.

Table 8: **Longer schedules, a larger backbone, and cost.** Left: ImageNet-1K fine-tuning accuracy; the margin over MAE does not shrink. Right: pre-training cost (ViT-B/16, 400 epochs).

	ViT-B, 400 ep.	ViT-B, 800 ep.	ViT-L, 400 ep.
MAE	82.78	83.4	84.9
BYON	83.56	84.4	85.7
Margin	+0.78	+1.0	+0.8

	GFLOPs	GPU hours
DiffMAE	45.5	84
MaskDiT	43.7	83
BYON	63.8	97

Cost. GFLOPs are the theoretical forward-and-backward cost of one training step at batch size one; GPU hours are the wall-clock of the 400-epoch run on four A100 GPUs, multiplied by four.

D Details of the analysis

Attention distance. Mean attention distance follows Raghu et al. (2021): for one head, the pixel distance between a query patch and every key patch is averaged with the attention weights, then over queries, then over 1,000 ImageNet-1K validation images. Figure 4 plots this quantity for each of the 144 heads of the three encoders.

Attention overlap. For each image we take the class-token attention of the last block, average it over heads, threshold it at its mean, and compute the IoU with the union of the ground-truth

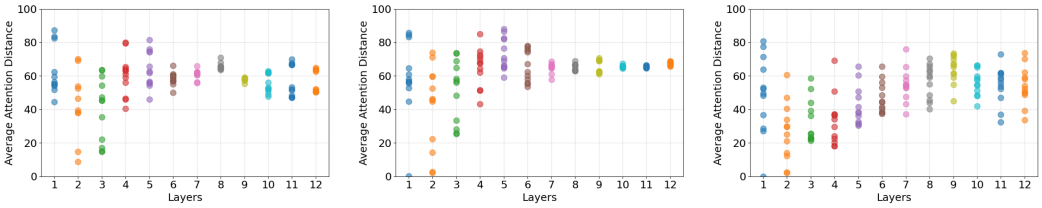

Figure 4: **Mean attention distance of every head, per block** (ViT-B/16). With masking (a) and with masking and uniform noise (b), the heads of the last blocks converge to one distance. BYON (c) keeps heads of short and long range at every depth. The encoders differ in the late blocks and not in the early ones, as a depth-dependent bias of latent prediction predicts.

object masks. The reported values (BYON 0.71, DINO 0.58, BYOL 0.48, MAE 0.35) average over images drawn at random from COCO, ADE20K, and ImageNet-1K with segmentation annotations; no image was selected by hand.

Spectral alignment. We resize ImageNet-1K training images to 64×64 and compute their principal components. The *top* subspace is spanned by the leading components that explain 90% of the pixel variance, and the *bottom* subspace is its orthogonal complement. For the validation images, let $\boldsymbol{P}_{\text{top}}$ and $\boldsymbol{P}_{\text{bot}}$ hold their coordinates in the two subspaces and $\boldsymbol{F}$ the frozen class-token features. We report linear CKA, $\|\boldsymbol{F}^\top \boldsymbol{P}\|_F^2 / (\|\boldsymbol{F}^\top \boldsymbol{F}\|_F \|\boldsymbol{P}^\top \boldsymbol{P}\|_F)$ on centered matrices, and RankMe (Garrido et al., 2023a), the exponential of the entropy of the normalized singular values of $\boldsymbol{F}$. Table 9 lists the values plotted in Figure 2.

Table 9: **Alignment of frozen features with pixel subspaces.** Linear CKA between the frozen class-token features and the coordinates of the image in the subspace that explains 90% of pixel variance (top) and in its complement (bottom), and the effective rank of the features (RankMe).

Row of Table 4	N	U	U+L	A	A+gL
CKA-top ↓	0.423	0.420	0.482	0.376	**0.272**
CKA-bottom ↑	0.021	0.022	0.012	0.034	**0.052**
RankMe ↑	33	37	11	54	**63**

Gate trajectory. The gate w_k of Equation (6) is logged at every step; Figure 3 plots its mean and range per epoch, together with the cosine similarity between the source and reference projections of Equation (3).

Additional attention maps. Figure 5 compares the class-token attention of four encoders on scenes with several objects, and Figure 6 shows BYON on objects of different scale. The images in both figures are drawn at random and not selected.

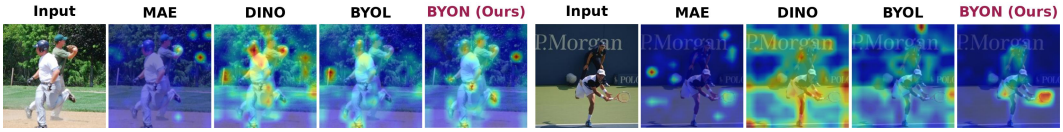

Figure 5: **Class-token attention on multi-object scenes** (COCO). MAE attends to scattered local spots, DINO and BYOL attend diffusely, and BYON follows the acting person and the held object.



Figure 6: **Class-token attention of BYON** on the last block, from large objects to thin structures.