ATTACKING A LINEAR PROBE MEASURES THE HEAD AS MUCH AS THE ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial evaluations of pretrained vision encoders often freeze the encoder, train a linear classifier, and attack the resulting model. How much of the reported robustness belongs to the classifier head? Across seven ImageNet-1k probes, we retrain six heads without their BatchNorm layer while holding the encoder and seed fixed. Under the AutoAttack ensemble, this intervention raises MoCo v3's robust accuracy from 3.88% to 6.26%: a change equal to 30% of the original between-encoder spread. In all six pairs, the sign of the change matches whether the original BatchNorm layer's mean feature-wise scale is below one (attenuation) or above one (amplification), with three differences significant under corrected paired tests. Cross-entropy PGD exaggerates the effect enough to reverse one clear pairwise encoder ordering; the ensemble does not resolve a rank reversal. A second intervention isolates attack sensitivity: dividing a trained head's logits by a positive constant preserves its decision function, yet changes PGD accuracy by up to 28.7 percentage points on MoCo v3. The extreme increase reflects gradient masking, not improved robustness. The two trained heads also differ under an attack with a scale-invariant objective. By contrast, the BatchNorm intervention changes ImageNet-C accuracy by at most 2.0 percentage points, leaving the large gaps between encoders essentially unchanged. These results support controlling the head and checking the attack when using frozen probes to compare representations. They establish dependence on the trained probe, while leaving the roles of head scaling and optimisation only partly separated.

1 INTRODUCTION

A frozen linear probe combines a pretrained encoder with a trained classifier head. The encoder maps an image to features; the head maps those features to class logits. To evaluate adversarial robustness, an attack searches for a small input perturbation that makes the resulting classifier wrong. When the resulting accuracy is used to compare encoders, a basic measurement question arises: how much can it change if the encoder stays fixed and only the head changes?

This question matters for comparisons of self-supervised Vision Transformers (Dosovitskiy et al., 2021; Caron et al., 2021; Oquab et al., 2024; He et al., 2022; Assran et al., 2023; Chen et al., 2021). Frozen-probe work reports objective- and architecture-dependent robustness rankings (Çağatan et al., 2025). A broader evaluation of released pre-trained classifiers instead reports higher adversarial susceptibility for contrastive self-supervision (Gupta et al., 2023). The published probing recipes for the models also differ: MAE specifies non-affine BatchNorm before the linear map (He et al., 2022), whereas DINOv2 and MoCo v3 specify a bare linear map (Oquab et al., 2024; Chen et al., 2021). A comparison assembled across such recipes varies the head alongside the encoder. Our experiments measure the size of this confound in our setup; whether it explains disagreements across studies remains open.

We distinguish two ways the measurement can change. First, retraining a different head changes the classifier $f = \mathrm{head} \circ \mathrm{encoder}$, even though the representation is fixed. We test this by removing BatchNorm and retraining the head. Second, an attack's estimate can change even when the classifier's decisions do not. We test this by positively rescaling a trained head's logits. This transformation preserves the predicted class for every input and leaves true worst-case robustness unchanged.

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.

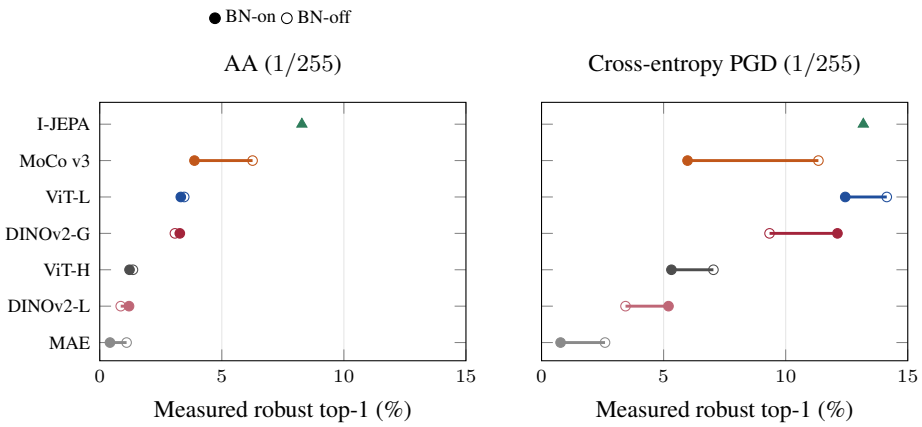


Figure 1: **Changing the head changes measured robustness; PGD magnifies the effect.** Each segment joins a BatchNorm-equipped head to a retrained BatchNorm-free head on the same frozen encoder (5,000 shared images). I-JEPA (triangle) is unchanged because its head already has no BatchNorm. AA shifts magnitudes without a clear rank reversal; PGD moves MoCo v3 past DINOv2-G. Rows use the same order in both panels. Exact values and paired-test outcomes appear in Table 2; the full sweep is in Appendix A.5.

Figure 1 shows the first effect. Removing BatchNorm changes accuracy under our reference evaluation, the `torchattacks` AutoAttack ensemble (AA) at $1/255$, and the differences are larger under cross-entropy PGD. We interpret the PGD ranking reversal as protocol sensitivity and use AA for the main comparison.

We make three contributions. **(1)** A paired head ablation measures how much adversarial comparisons can move without changing the encoder, including effects that remain under AA (Section 4). **(2)** A prediction-preserving logit rescaling tests whether an attack responds to logit scale even when the classifier's decisions stay fixed. A scale-invariant control shows that this sensitivity alone does not explain the head-ablation effect (Section 5). **(3)** Sensitivity, attack-transfer and corruption comparisons establish the scope of the finding: the same BatchNorm intervention changes ImageNet-C much less relative to the between-encoder spread, and corruption and adversarial evaluation favour different probes (Sections 6–7).

2 Related Work

Evaluating frozen representations. Readout choices and evaluation protocols affect measured accuracy and rankings (Li et al., 2023; Lee et al., 2023; Marks et al., 2025), including in speech (Zaiem et al., 2023) and under distribution shift (Al-Zyoud & El Saddik, 2026). In particular, Lee et al. (2023) recommend BatchNorm for evaluation stability. Petrov & Kwiatkowska (2022) note that prior encoder-robustness work evaluates through a downstream classification task and propose model- and task-agnostic, label-free measures. Here, paired head interventions and stronger attacks test how much changing the head alters measured robustness.

Self-supervision and robustness. Prior work reports improved adversarial robustness from auxiliary self-supervision (Hendrycks et al., 2019), differing outcomes for contrastive representations (Çağatan et al., 2025; Gupta et al., 2023), and adversarial weakness associated with pixel-reconstruction targets (Huang et al., 2023). Robustness also depends on the shift and downstream task (Chhipa et al., 2023; Kowalczuk et al., 2024). Adversarial pre-training changes the representation itself (Kim et al., 2020; Gowal et al., 2021; Addepalli et al., 2024); our paired comparison changes the downstream classifier head on a frozen encoder.

BatchNorm and attack reliability. BatchNorm affects adversarial robustness, transferability and corruption robustness in end-to-end trained networks (Galloway et al., 2019; Benz et al., 2021b; Wang et al., 2022; Benz et al., 2021a). There, changing normalisation also changes the learned

representation; here the encoder remains frozen. Gradient masking, stochasticity and attacks that do not account for the evaluated model can inflate robust accuracy (Athalye et al., 2018; Carlini et al., 2019; Tramèr et al., 2020). RobustBench (Croce et al., 2021) uses AutoAttack (Croce & Hein, 2020) to reduce such failures. Local loss-geometry diagnostics offer another view of attack behaviour (Moosavi-Dezfooli et al., 2019; Andriushchenko & Flammarion, 2020); our small FGSM-direction diagnostic does not explain the iterative-attack results (Appendix C.3). The rescaling experiment connects attack reliability to ordinary, undefended linear probes.

3 Experimental Setup

Models and heads. We study seven public backbones based on the Vision Transformer architecture (Dosovitskiy et al., 2021). I-JEPA, MAE and supervised ViT-H use Huge models; DINOv2-G uses Giant; DINOv2-L, MoCo v3 and supervised ViT-L use Large. Their pre-training objectives are latent prediction for I-JEPA, pixel reconstruction for MAE, self-distillation for DINOv2, contrastive learning for MoCo v3 and supervision for ViT-H and ViT-L. Appendix A records the full setup, including the checkpoint inventory in Table 3. Each encoder is frozen; its head is trained on ImageNet-1k for 8 epochs with cross-entropy and AdamW (Loshchilov & Hutter, 2019), using a cosine schedule with warm-up. Six baseline heads have affine BatchNorm (Ioffe & Szegedy, 2015) before the linear map; I-JEPA has none. This is a controlled baseline, not a reconstruction of each model's published head: MAE specifies non-affine BatchNorm, while DINOv2 and MoCo v3 specify no BatchNorm.

I-JEPA uses concatenated mean-pooled features from its last four blocks, before the encoder's final LayerNorm; the other six use the final-layer `[CLS]` token, including timm's (Wightman, 2019) default readout for MoCo v3. Training uses ImageNet-1k/ILSVRC-2012 (Russakovsky et al., 2015) through `benjamin-paine/imagenet-1k-256x256`; its dataset card states that images were center-cropped to a square and rescaled to 256^2 with Lanczos resampling.[1] Evaluation applies `Resize(256)` then `CenterCrop(224)`. Input normalisation is inside the model wrapper, so attacks operate on pixels in $[0, 1]$. Optimiser and augmentation details are in Appendix A.1.

Attacks and accuracy. The ℓ_∞ budget ε limits the change to each pixel channel: $1/255$ allows a normalized channel to change by at most one integer level in its 8-bit representation. FGSM (Goodfellow et al., 2015) takes one gradient-sign step; PGD (Madry et al., 2018) repeatedly recomputes the gradient and projects the result back into the allowed perturbation set. We use `torchattacks` (Kim, 2020): FGSM and 20-step, single-restart PGD at $\{1, 2, 4, 8\}/255$, and `torchattacks.AutoAttack(version="standard")` (AA), its APGD-CE, APGD-DLR-T, FAB-T and Square-5000 ensemble, at $\{1, 4\}/255$. This implementation differs slightly in component settings and execution from the official AutoAttack protocol; we use its AA results as our reference evaluation. The rescaling control uses APGD-DLR with 20 steps and one restart; the MAE masking check below uses standalone Square.

Reported robust top-1 is the fraction of evaluation images correctly classified under the specified attack, not a certified worst-case guarantee. The main attack comparisons share the same ordered $5{,}000$-image RobustBench-IN1k validation subset (Croce et al., 2021). The logit-rescaling experiments use its first $1{,}000$ images; the transfer and CKA analyses use a fixed, seeded, class-balanced sample of $1{,}000$ images from it. The sensitivity analysis in Section 6 uses 256 stratified images, and the loss-geometry diagnostic in Appendix C.3 uses 12 class-stratified images. Every comparison uses the same images across configurations.

MAE masking check. With the default configuration of `facebook/vit-mae-huge`, HuggingFace's `transformers` (Wolf et al., 2020) MAE implementation samples a fresh 75% patch mask even at evaluation. We set `mask_ratio=0` for all main comparisons. As a gradient-free check, we ran Square Attack (Andriushchenko et al., 2020) with $1{,}000$ queries at $4/255$ on the same $1{,}000$ images under each setting. Measured robust top-1 was 67.2% with default masking (clean: 69.6%) and 30.3% with masking disabled (clean: 76.8%). The discrepancy extends beyond gradient-based attacks, though disabling the mask also changes the evaluated classifier. Appendix A.3 discusses the noisy score comparisons.

[1] `https://huggingface.co/datasets/benjamin-paine/imagenet-1k-256x256`

Probe	Clean	AA $1/255$	AA $4/255$	PGD $1/255$	PGD $2/255$
DINOv2-G	85.88	3.28	0.00	12.12	1.04
DINOv2-L	84.90	1.20	0.00	5.20	0.14
ViT-L	80.46	3.32	0.00	12.44	1.98
MAE	76.76	0.42	0.00	0.78	0.02
MoCo v3	75.32	3.88	0.00	5.98	0.32
ViT-H	75.26	1.22	0.00	5.32	0.34
I-JEPA	73.46	8.28	0.00	13.18	2.20

Table 1: **The comparison is informative only at small budgets.** Top-1 accuracy (%) on $5{,}000$ shared images, with the baseline heads. AA denotes the `torchattacks` AutoAttack ensemble defined in Section 3; PGD uses 20 steps. AA and PGD column labels give the ℓ_∞ perturbation budget ε. The full attack sweep is Table 4.

Interventions. For each BN-equipped probe, we remove BatchNorm and retrain the head on the same frozen encoder with seed 0. At evaluation BatchNorm is affine, so both head architectures can express the same class of affine maps; retraining can nevertheless produce different maps. We summarise the original BatchNorm gain by

$$\bar{s} = \frac{1}{d} \sum_{j=1}^{d} \frac{\gamma_j}{\sqrt{\sigma_j^2 + \epsilon_{\mathrm{BN}}}}. \tag{1}$$

Here d is the number of features, γ_j is the learned scale, σ_j^2 is the stored variance, and ϵ_{BN} is the numerical stabiliser, distinct from the attack budget. All saved γ_j are positive. Thus $\bar{s}$ is the average scale applied across feature coordinates; it does not measure the change in norm for a particular feature vector or perturbation. The I-JEPA-last1 control retrains a separate head using the mean-pooled final output after LayerNorm, reducing feature dimension from 5120 to 1280. The baseline last-four features precede that LayerNorm. This comparison changes block selection, feature dimension and final normalisation while retaining mean pooling (Appendix B).

Uncertainty. Differences are in percentage points (pp). When describing gaps and rankings, we use the unpaired binomial standard error of a difference and do not treat gaps below about two standard errors as a clear separation. For the six paired AA ablations, we use two-sided exact conditional McNemar tests (equivalently, conditional binomial tests on the discordant counts under $X \sim \mathrm{Binomial}(b + c, 1/2)$) with Bonferroni correction at family-wise $\alpha = 0.05$ (McNemar, 1947; Fagerland et al., 2013). Rank correlations are descriptive and uncorrected. These tests quantify image-level uncertainty for the trained heads; each head has only one training seed. Appendix C.1 reports the full tests.

3.1 Choosing an Informative Attack Budget

At $4/255$, AA reduces all seven probes to zero measured accuracy; PGD-20 leaves at most 0.18%. This budget cannot distinguish their robustness. At $1/255$, AA leaves a measurable spread, from MAE's 0.42% to I-JEPA's 8.28% (Table 1). We therefore use AA-$1/255$ as the reference and restrict subsequent adversarial accuracy comparisons to $\varepsilon \leq 2/255$. All measured robust accuracies remain low even in this more informative regime. The full FGSM/PGD sweep and the I-JEPA-last1 row are retained in Appendix A.5; larger FGSM budgets are used only for diagnostics.

4 Changing the Head Changes Adversarial Accuracy

The effect remains under AA. Removing BatchNorm changes MoCo v3's AA-$1/255$ accuracy from 3.88% to 6.26%, while the encoder is identical. The 2.38 pp increase is 30% of the original 7.86 pp range from the lowest to the highest baseline encoder. MAE also improves, whereas DINOv2-L declines. These three paired differences are significant after Bonferroni correction; the other three do not clear even the uncorrected threshold (Table 2). Thus the head affects the reference measurement, although AA does not show a clear ranking reversal: the only observed swap is between the two least robust probes, within noise.

Probe	BN scale	Clean	AA $1/255$	ΔAA	PGD $1/255$	ΔPGD
MoCo v3	18.81	$75.32 \to 73.56$	$3.88 \to 6.26$	$+2.38^{\dagger}$	$5.98 \to 11.34$	$+5.36$
MAE	11.17	$76.76 \to 73.76$	$0.42 \to 1.10$	$+0.68^{\dagger}$	$0.78 \to 2.60$	$+1.82$
ViT-H	7.35	$75.26 \to 75.32$	$1.22 \to 1.36$	$+0.14$	$5.32 \to 7.04$	$+1.72$
ViT-L	1.66	$80.46 \to 80.78$	$3.32 \to 3.46$	$+0.14$	$12.44 \to 14.14$	$+1.70$
DINOv2-G	0.54	$85.88 \to 85.86$	$3.28 \to 3.08$	-0.20	$12.12 \to 9.34$	-2.78
DINOv2-L	0.51	$84.90 \to 84.26$	$1.20 \to 0.86$	$-0.34^{\dagger}$	$5.20 \to 3.44$	-1.76

Table 2: **Paired head retraining on six fixed encoders.** Arrows give BN-on $\to$ BN-off top-1 accuracy (%), on the same 5,000 images; Δ columns give the change in percentage points. BN scale is the original BatchNorm layer's mean feature-wise scale $\bar{s}$ (Eq. 1). † marks a significant exact McNemar test after correction over six pairs. Full attack columns and test counts are in Tables 5 and 7.

Mean BatchNorm scale tracks the sign of the change. All four probes with $\bar{s} > 1$ improve under both AA and PGD; both probes with $\bar{s} < 1$ decline. PGD changes are larger: $+1.70$ to $+5.36$ pp when $\bar{s} > 1$, compared with -2.78 and -1.76 pp when $\bar{s} < 1$. Nearly identical scales on the two DINOv2 heads (0.54 and 0.51) accompany different changes, so scale does not determine their magnitude. The two $\bar{s} < 1$ cases are both DINOv2, and no observed scale lies between 0.54 and 1.66. The sign agreement is descriptive; six observations cannot establish a threshold or predictive rule (Figure 4). Clean accuracy changes by at most 3.0 pp, and by less than 0.7 pp on four of the six pairs.

PGD can change the apparent encoder ordering. Under PGD-$1/255$, MoCo v3 moves from 5.98% to 11.34%, passing DINOv2-G, which moves from 12.12% to 9.34%. MoCo v3 becomes third among the seven probes, 3.3 standard errors above DINOv2-G. ViT-L also passes I-JEPA numerically, but by only 1.4 standard errors, which we do not treat as a clear reordering. MoCo v3's 5.36 pp change is 43% of the original 12.4 pp PGD spread. AA places MoCo v3 second even before the intervention, so this PGD reordering does not appear in our reference evaluation.

Head convention matters when comparing published results. Çağatan et al. (2025) report MoCo v3 as more robust than their DINO probe and caution that this result does not establish a general contrastive-versus-distillation rule. Our PGD comparison with DINOv2-G can take either direction depending on the head. Their models and protocols differ from ours, leaving the head's role in their specific ranking unresolved.

What BatchNorm removal changes. The encoder is fixed, but removing BatchNorm changes how the head is trained as well as the affine map eventually learned. The mean coordinate scale does not describe changes in individual feature directions. It is also coupled to encoder feature scale, and the BN-on and BN-off training runs reach different losses. The experiment establishes dependence on the trained head without separating the contributions of scaling, conditioning and optimisation. We next examine uniform output rescaling in isolation, without retraining.

5 Logit Rescaling Changes Measured Accuracy Under Attack

An unchanged classifier can have a different measured accuracy under attack. Dividing all logits by a positive constant preserves the largest logit, and therefore the predicted class, for every input. We implement this as $W \leftarrow W/c$ and $b \leftarrow b/c$ in a trained head. No weights are retrained. The decision boundary and true worst-case robustness stay fixed, but the softmax probabilities used by a cross-entropy attack change. This provides a direct check of whether the measured robust accuracy depends on logit scale.

We sweep the MoCo v3 pair on 1,000 images under PGD-$1/255$, with $c \in [0.05, 20]$ for the BN-on head and $c \in [0.05, 5]$ for the BN-off head. At $c = 1$, accuracies differ from the corresponding 5,000-image results by at most 0.22 pp. Figure 2 shows that PGD accuracy spans 6.2% to 34.9% for the BN-on head and 11.3% to 30.0% for the BN-off head: swings of 28.7 and 18.7 pp with no change in either decision function. Each head's lowest measured PGD accuracy occurs at its trained scale, $c = 1$. The grid is coarse near $c = 1$, so the location of the true minimum remains unresolved.

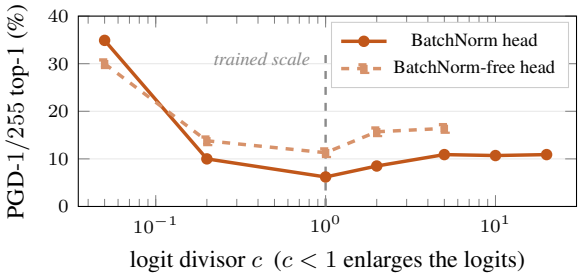


Figure 2: **The same decision function receives different PGD accuracies.** Dividing MoCo v3's logits by c preserves every prediction. Gradient masking causes the high apparent accuracies at $c = 0.05$. Both heads have their lowest measured accuracy at the trained scale ($c = 1$), over the tested grid. Each curve uses 1,000 images; Table 10 gives the values and APGD-DLR control.

Why cross-entropy PGD changes. PGD uses the sign of the gradient, so multiplying that gradient by a positive scalar does not change the step. Logit rescaling instead changes the relative contributions of the classes through the softmax probabilities. Smaller logits distribute gradient weight across more classes; sufficiently large logits saturate the softmax on confident images and can make the numerical gradient vanish. The latter is *gradient masking*: the attack becomes less effective and overstates robustness. It accounts for most of the large swings here and is related to the failure of defensive distillation (Papernot et al., 2016; Carlini & Wagner, 2017). Appendix D gives the gradient expression and complete sweep.

The DLR loss addresses this dependence by dividing one logit difference by another, so uniform positive scale cancels (Croce & Hein, 2020). On the BN-on head, APGD-DLR returns 4.60%, 4.30% and 4.30% at $c = 1, 5, 20$, where cross-entropy PGD changes by 4.7 pp. The objective is exactly scale-invariant; the measured implementation differs slightly (Appendix D). We use this 20-step, single-restart run solely to diagnose scale sensitivity.

The trained heads still differ under the scale controls. At their trained scales ($c = 1$), PGD accuracy is 6.20% for the BN-on head and 11.30% for the BN-off head, a gap of 5.10 pp. These are also the lowest values in their respective scale sweeps, so minimising over the tested logit scales leaves the same gap. At $c = 1$, APGD-DLR gives a 5.70 pp gap, within about a third of a standard error of the PGD gap. AA on $5{,}000$ images agrees in direction ($+2.38$ pp), although its APGD-CE component is not scale-invariant.

A third control rescales the BN-on head until the two heads have the same mean clean logit margin. Their PGD accuracies then remain 4.30 pp apart, compared with 5.10 pp at their trained scales. The retained fraction, about five sixths, depends on the matching statistic and is not an additive decomposition of the BatchNorm effect; Appendix D contrasts operator-norm matching. The scale sweep covers MoCo v3 alone; both DINOv2 sign reversals remain untested. Uniform logit scale therefore strongly affects cross-entropy PGD. The two trained MoCo v3 heads still differ under APGD-DLR, whose objective removes that dependence.

6 Sensitivity and Attack Transfer

To examine the accuracy shifts, we measure where input perturbations are amplified and how the head intervention affects attacks transferred between encoders. Appendix C gives the paired head-ablation tests and sensitivity values behind this analysis. These diagnostics describe the tested probes without yielding a general robustness ordering.

Separating feature and logit changes. A perturbation first changes the encoder's features, then the head maps that feature change into a logit change. For head weights W and the evaluation-time BatchNorm diagonal $S = \mathrm{diag}(\gamma_j / \sqrt{\sigma_j^2 + \epsilon_{\mathrm{BN}}})$, with $S = I$ without BatchNorm,

$$\|\Delta\mathrm{logit}\|_2 \le \underbrace{\sigma_{\max}(WS)}_{\|\mathrm{head}\|_{\mathrm{op}}} \|\Delta\mathrm{feat}\|_2. \tag{2}$$

Biases cancel in the difference. We compute the head norm from its checkpoint and measure the feature and logit shifts per unit input shift, denoted s_{enc} and s_{total}, along each configuration's own

FGSM-$4/255$ direction on 256 images. These are realised directional sensitivities, giving lower bounds on Lipschitz constants rather than estimates of those constants.

Across the seven baseline probes, the head norm spans $320\times$ and s_{enc} spans $32\times$, but s_{total} spans only $1.8\times$. Large changes in one stage can therefore be offset by the other. Encoder sensitivity alone has no detectable association with AA-$1/255$ accuracy ($\rho = -0.07$, $p = 0.88$), and its correlation changes sign under PGD. After BN removal, head norms change by factors of 1.2–20, while s_{enc} changes by at most $+7.7\%$ or -3.6%. Four of six feature-sensitivity shifts are significant in corrected paired t-tests. Their magnitudes remain small relative to the range across probes. Fixed encoders do not force identical values here, because each head defines its own FGSM direction. Appendix C.2 gives the values, averaging conventions and exceptions, including why these quantities do not consistently order the I-JEPA readout ablation.

Attacks transfer differently from self-attacks. A transfer attack is crafted on one probe and evaluated on another. Repeating the FGSM-$4/255$ transfer matrix on the six BN-off heads (Liu et al., 2017) changes accuracy against self-attacks by 4.5 pp on average, versus 1.2 pp for attacks from different encoders, with no consistent direction in the latter (Appendix E). The absolute movement is smaller off the diagonal, but those attacks also transfer poorly: relative to the accuracy already lost from clean, the changes are 7.3% for self-attacks and 11.5% for transfer. The 1.2 pp shift is below the ≈ 1.6 pp absolute difference expected under an unpaired sampling approximation; the paired noise level is unknown because per-image transfer outcomes were not retained. We therefore report limited absolute movement, not evidence that transferable structure is unchanged.

FGSM also ranks probes differently from iterative attacks. Appendix C.3 examines this discrepancy through a 12-image loss-geometry diagnostic: curvature along the initial gradient-sign direction correlates with FGSM accuracy across the seven probes, but not detectably with PGD or AA. We therefore treat the loss-geometry result as specific to FGSM; the principal head-ablation results use iterative attacks.

7 The Same Head Change Has Less Effect on Corruption Accuracy

Common corruptions test a different kind of robustness: accuracy under prescribed changes such as noise, blur and weather, rather than perturbations optimised against a classifier. We evaluate all thirteen head configurations on ImageNet-C (Hendrycks & Dietterich, 2019), averaging 15 corruptions at 5 severities over the same $5{,}000$-image RobustBench subset used for the main attacks. Figure 3 compares this accuracy with AA.

Removing BatchNorm changes overall ImageNet-C accuracy by only -1.6 to $+2.0$ pp. These shifts leave the much larger gaps between encoders intact: DINOv2 scores 73.5–78.7%, ViT-L scores 62.6–63.3%, and the other four backbones score 47.7–52.4%. On MoCo v3, the intervention is 6% of the ImageNet-C spread, versus 30% of the AA spread and 43% of the PGD spread: the adversarial measures are roughly five to seven times more sensitive by this relative comparison. Five of the six changes agree in sign across corruption and adversarial evaluation; MAE is the exception. We do not report significance tests for ImageNet-C because those runs retained aggregate accuracies rather than per-image outcomes (Appendix F).

The rankings also disagree. I-JEPA leads the baseline AA comparison at 8.28% but has the lowest ImageNet-C accuracy, 47.7%; DINOv2-G leads ImageNet-C but is fourth under AA. Across the seven probes, the rank correlation is $\rho = -0.54$ (asymptotic $p = 0.21$). This coefficient is inconclusive, so the observed rank reversals carry the comparison. The small BatchNorm effect on ImageNet-C does not extend to every head change: using only I-JEPA's last block lowers ImageNet-C accuracy by a further 6.0 pp (Appendix B). Corruption accuracy provides a complementary comparison whose values can still depend on the head.

8 Discussion and Conclusion

Adversarial accuracy from a frozen linear probe depends on both the trained head and the attack used to evaluate it. The BatchNorm intervention changes the classifier while preserving the encoder; the logit-rescaling intervention preserves the decision function while exposing an attack failure. AA

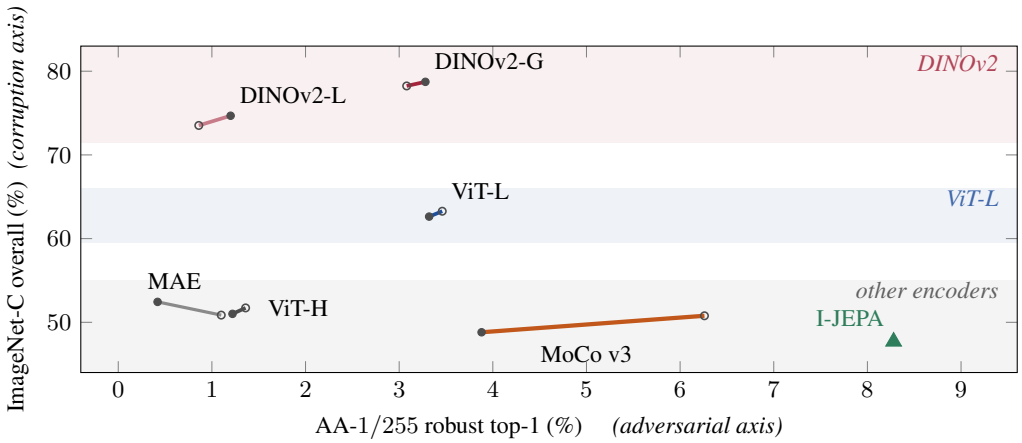


Figure 3: **Removing BatchNorm barely changes ImageNet-C accuracy.** Segments join BN-on (filled) and BN-off (open) heads on the same encoder; I-JEPA (triangle) has no BatchNorm. All configurations use 5,000 images. The shaded ranges highlight the large gaps that remain between DINOv2, ViT-L and the other encoders. The two axes also disagree: I-JEPA leads under AA, whereas DINOv2-G leads on ImageNet-C. Per-severity and per-category values appear in Appendix F.

reduces the dramatic PGD effects, but measurable head dependence remains. These findings support controlling the probe when interpreting differences between representations.

Implications for evaluation. Our results suggest that encoder comparisons are easier to interpret when they use a common head architecture. A head ablation can help assess whether an apparent encoder difference depends on that choice. A common architecture controls one design choice, although the learned heads can still differ through scale and optimisation; head-free measures offer a complementary approach (Petrov & Kwiatkowska, 2022). Reporting corruption and adversarial accuracy together may also help because they evaluate different vulnerabilities. Rankings at budgets where every probe reaches the attack's accuracy floor contain little information. As one descriptive check, we compare the between-encoder spread with the largest BatchNorm-ablation effect: here their ratio is 2.3 under PGD and 3.3 under AA at $1/255$. At $4/255$ the baseline AA spread is zero, leaving no encoder ranking to interpret. We use these ratios only as a diagnostic, not as a universal threshold.

Scope and limitations. Each head has one training seed. The paired tests quantify image-level differences between those fitted classifiers, without estimating head-retraining variance. The BN-off runs end at higher training loss than the BN-on runs for all four probes with $\bar{s} > 1$, and at lower loss for both DINOv2 probes with $\bar{s} < 1$. A training schedule that matched optimisation outcomes across heads would help separate this effect from their fitted scales. The logit-scale control covers only MoCo v3. Mean BatchNorm scale is also closely tied to encoder feature sensitivity in these six probes, so its association with the ablation's sign does not locate the effect solely in the head. The checkpoints differ in size, pre-training data and pooling; our comparisons do not isolate pre-training objective.

I-JEPA retains an advantage after the tested controls: its last-block probe, which replaces the baseline four-block feature concatenation with the mean-pooled final block, reaches 7.80% under AA. This is 1.54 pp above the best BN-free comparison probe, versus the original 4.4 pp lead. I-JEPA still mean-pools patch tokens while the other probes use CLS tokens, so this residual difference does not isolate an encoder effect. Our PGD runs use one restart, AA covers two budgets, and the ImageNet-C comparison uses the 5,000-image subset rather than the full benchmark or its mCE metric. We have not tested whether the BatchNorm effect persists under full fine-tuning. Frozen-probe adversarial scores are most informative when the head configuration is controlled and the attack is checked for sensitivity.

AI USE STATEMENT

We used generative AI tools (large language models, as a coding and writing assistant) at several stages of this work, and disclose that use below against the categories set out in the ICLR AI policy. Every experimental result reported here was produced by running the described pipeline on the described checkpoints and data (Section 3) on our own hardware. No measurement, table entry, or dataset in this paper was generated, imputed, or synthesised by an AI tool.

Under the policy's *required* categories we disclose the following uses. First, *discovery*: AI tools helped us with discovering some related work we had missed initially. Second, *implementing methods*: parts of the evaluation and analysis codebase, specifically the attack and probe drivers, the sensitivity decomposition, and the plotting utilities, were written or edited with AI assistance, under our direction and review. Finally, *refining hypotheses and giving feedback on methodology*: AI tools helped inform which experiments we judged strong enough to report. All such decisions were made by the authors. We did not use AI tools to propose, design, or execute any experiment. This work contains no theoretical results or proofs.

Under the policy's *recommended* categories, an AI assistant drafted manuscript text that the authors then revised, generated the figures as plotting code from numbers in our result tables, and identified some references which we had not previously cited.

We have reviewed all AI-assisted work. Every number in this paper was checked by the authors against the raw per-configuration result files. We take responsibility for the final content of this work, including all text, claims, and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates the adversarial and corruption robustness of publicly released, pre-trained image classifiers on a public benchmark. It involves no human subjects, no personal data, and no new data collection. The attacks we use are standard, long-published, and available in open-source libraries; we introduce no new attack. Our findings are defensive in orientation. The primary result is that the evaluation protocol used by the studies in Section 2 can produce non-comparable robustness rankings, and the secondary result is that a default in a reference implementation can make an undefended model appear defended. We consider surfacing both to reduce rather than increase risk, since inflated robustness estimates are themselves a safety hazard. We note that ImageNet-1k carries known distributional and annotation biases which we do not address and which our accuracy numbers inherit.

REPRODUCIBILITY STATEMENT

Section 3 specifies the seven checkpoints, the probe-training setup, evaluation transform, attack configurations, and the fixed image subsets shared across configurations. Section 4 states the BatchNorm ablation protocol, including what is held fixed. Section 6 defines the feature and logit sensitivities and how $\|\mathrm{head}\|_{\mathrm{op}}$ is computed from a saved checkpoint; Appendix C.2 defines the alignment ratio and averaging conventions, and Appendix C.3 gives the loss-landscape probe's construction and per-probe values. Appendix A.4 describes the compute environments and the resume mechanism used for the long AA runs, and Appendix E gives the full transfer and centred kernel alignment (CKA, Kornblith et al., 2019) matrices summarised in Section 6. The single non-obvious step needed to reproduce our MAE numbers is disabling eval-time patch masking, which Appendix A.3 describes. The paired tests in Table 7 use per-image correctness retained by the AA resume checkpoints over the shared image list. The BatchNorm-free rows of Table 8 were computed in a separate run from the BatchNorm-equipped rows, using the same script and image subset. We will release the anonymized code needed to run the reported evaluations; checkpoints and raw result artifacts will not be part of the submission package.

REFERENCES

Sravanti Addepalli, Priyam Dey, and R. Venkatesh Babu. ProFeAT: Projected feature adversarial training for self-supervised learning of robust representations. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2406.05796.

Izaldein Al-Zyoud and Abdulmotaleb El Saddik. Rethinking feature conditioning for robust forged media detection in edge AI sensing systems, 2026. URL https://arxiv.org/abs/2603.26387.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://arxiv.org/abs/2007.02617.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, pp. 484–501. Springer, 2020. doi: 10.1007/978-3-030-58592-1_29.

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. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://arxiv.org/abs/2301.08243.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 2018.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 494–503, 2021a. doi: 10.1109/WACV48630.2021.00054. URL https://arxiv.org/abs/2010.03630.

Philipp Benz, Chaoning Zhang, and In So Kweon. Batch normalization increases adversarial vulnerability and decreases adversarial transferability: A non-robust feature perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7798–7807, 2021b. doi: 10.1109/ICCV48922.2021.00772. URL https://arxiv.org/abs/2010.03316.

Ömer Veysel Çağatan, Ömer Faruk Tal, and M. Emre Gürsoy. Adversarial robustness of discriminative self-supervised learning in vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2313–2324. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00223.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. URL https://arxiv.org/abs/1902.06705.

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. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00950.

Prakash Chandra Chhipa, Johan Rodahl Holmgren, Kanjar De, Rajkumar Saini, and Marcus Liwicki. Can self-supervised representation learning methods withstand distribution shifts and corruptions? In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4469–4478. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00481.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html.

Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 4421–4435, 2022. URL https://arxiv.org/abs/2202.13711.

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. URL https://arxiv.org/abs/2010.11929.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91, 2013. doi: 10.1186/1471-2288-13-91. URL https://doi.org/10.1186/1471-2288-13-91.

Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W. Taylor. Batch normalization is a cause of adversarial vulnerability. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. URL https://arxiv.org/abs/1905.02161.

Carlos Gomes. Measuring the false sense of security, 2022. URL https://arxiv.org/abs/2204.04778.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6572.

Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=bgQek2O63w.

Rohit Gupta, Naveed Akhtar, Ajmal Mian, and Mubarak Shah. Contrastive self-supervised learning leads to higher adversarial susceptibility. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 14838–14846, 2023. doi: 10.1609/aaai.v37i12.26733.

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. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://arxiv.org/abs/1906.12340.

Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, and Nenghai Yu. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1600–1610. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00154. URL https://arxiv.org/abs/2308.10315.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

Hoki Kim. Torchattacks: A PyTorch repository for adversarial attacks, 2020. URL https://arxiv.org/abs/2010.01950.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 2983–2994, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Antoni Kowalczuk, Jan Dubiński, Atiyeh Ashari Ghomi, Yi Sui, George Stein, Jiapeng Wu, Jesse C. Cresswell, Franziska Boenisch, and Adam Dziedzic. Benchmarking robust self-supervised learning across diverse downstream tasks. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL https://arxiv.org/abs/2407.12588.

Jae-Hun Lee, Doyoung Yoon, ByeongMoon Ji, Kyungyul Kim, and Sangheum Hwang. Rethinking evaluation protocols of visual representations learned via self-supervised learning, 2023. URL https://arxiv.org/abs/2304.03456.

Yazhe Li, Jorg Bornschein, and Marcus Hutter. Evaluating representations with readout model switching. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09579.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.02770.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Keane Lucas, Matthew Jagielski, Florian Tramèr, Lujo Bauer, and Nicholas Carlini. Randomness in ML defenses helps persistent attackers and hinders evaluators, 2023. URL https://arxiv.org/abs/2302.13464.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.06083.

Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, 133(8):5013–5025, 2025. doi: 10.1007/s11263-025-02402-w. URL https://arxiv.org/abs/2407.12210.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9078–9086. IEEE, 2019. doi: 10.1109/CVPR.2019.00929.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, 2016. doi: 10.1109/SP.2016.41. URL https://arxiv.org/abs/1511.04508.

Aleksandar Petrov and Marta Kwiatkowska. Robustness of unsupervised representation learning without labels, 2022. URL https://arxiv.org/abs/2210.04076.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1633–1645, 2020. URL https://arxiv.org/abs/2002.08347.

Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. Removing batch normalization boosts adversarial training. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 23433–23445, 2022. URL https://proceedings.mlr.press/v162/wang22ap.html.

Ross Wightman. PyTorch image models, 2019. URL https://github.com/rwightman/pytorch-image-models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. Speech self-supervised representation benchmarking: Are we doing it right? In *Interspeech*, pp. 2873–2877, 2023. doi: 10.21437/Interspeech.2023-1087.

A Experimental Setup Detail

A.1 Probe-Training and Attack Hyperparameters

Every head is trained with AdamW (Loshchilov & Hutter, 2019) (β_1=0.9, β_2=0.999), weight decay 0.05, a learning rate scaled linearly with batch size, a cosine schedule with warm-up over $\max(500, 5\%)$ of training steps, and mixed precision. Training-time augmentation is `RandomResizedCrop` (scale $[0.8, 1.0]$) followed by a horizontal flip with probability 0.5; the evaluation transform (Section 3) uses no random augmentation. Every PGD result in the paper uses step size $\alpha = 2\varepsilon/\text{steps}$ and a random start inside the ε-ball.

FGSM uses $x_{\text{adv}} = \text{Clip}_{[0,1]}\big(x + \varepsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y))\big)$; PGD repeats a gradient-sign update with projection onto the allowed ε-ball and clipping to $[0, 1]$. The descriptive standard error of an accuracy difference is $\sqrt{\hat{p}_1(1-\hat{p}_1)/n + \hat{p}_2(1-\hat{p}_2)/n}$, the unpaired binomial approximation. Paired AA claims instead use the exact tests in Appendix C.1.

A.2 Backbone Inventory

Table 3 lists the seven backbones, their checkpoints, parameter counts and pooling conventions, together with the I-JEPA last-block ablation below the rule.

The backbone families also differ in pre-training data scale. DINOv2 uses 142M curated images, I-JEPA and the supervised ViTs use roughly 14M, and MAE and MoCo v3 use 1.28M. Consequently, comparisons between objective families, including I-JEPA's lead in Section 8, also reflect this uncontrolled difference. The paired head ablation does not remove it.

short name	model	checkpoint	params	feature	objective family	pre-train data
I-JEPA	I-JEPA	`facebook/ijepa_vith14_22k`	632 M	last-4 GAP	MIM (latent prediction)	ImageNet-22k
MAE	MAE	`facebook/vit-mae-huge`	632 M	CLS	MIM (pixel reconstruction)	ImageNet-1k
ViT-H	ViT-Huge	`google/vit-huge-patch14-224-in21k`	632 M	CLS	supervised (ImageNet-21k)	ImageNet-21k
ViT-L	ViT-Large	`google/vit-large-patch16-224-in21k`	304 M	CLS	supervised (ImageNet-21k)	ImageNet-21k
DINOv2-G	DINOv2-Giant	`facebook/dinov2-giant`	1.13 B	CLS	discriminative (self-distillation)	LVD-142M
DINOv2-L	DINOv2-Large	`facebook/dinov2-large`	304 M	CLS	discriminative (self-distillation)	LVD-142M
MoCo v3	MoCo v3	`nyu-visionx/moco-v3-vit-l`	304 M	CLS (timm)	discriminative (contrastive)	ImageNet-1k
I-JEPA-last1	I-JEPA (last-1)	`facebook/ijepa_vith14_22k`	632 M	last-1 GAP	last-block ablation of I-JEPA	ImageNet-22k

Table 3: Frozen-backbone inventory. All seven probes train linear heads on the same ImageNet-1k split; six heads use BatchNorm and I-JEPA's does not. The ruled-off I-JEPA-last1 row shares I-JEPA's encoder weights and reads only its final block. MIM means masked image modelling; GAP means mean pooling over patch tokens; "CLS (timm)" is the MoCo v3 checkpoint's class-token readout. Parameter counts describe the backbones.

A.3 MAE's Eval-Time Patch Masking

With the default `facebook/vit-mae-huge` configuration, the HuggingFace `ViTMAEModel` forward pass draws a fresh random binary mask of ratio 0.75 on *every* call, including at evaluation. This makes a standard deterministic gradient attack unreliable: the gradient in the inner loop belongs to one sampled function while the scoring pass uses an independent one. Pixels in dropped patches receive zero gradient, so $\operatorname{sign}(0) = 0$ leaves those pixels unchanged. Query-based attacks can also be distorted when each query sees a different mask. This is an instance of stochastic evaluation obscuring the attack signal (Athalye et al., 2018).

Under the Square attack at $4/255$ on 1,000 validation images, masked MAE scores 67.2% robust top-1 against a clean 69.6%, a gap of only 2.4 pp. Disabling the mask drops robust accuracy to 30.3% against a clean 76.8%, a gap of 46.5 pp and a swing of 36.9 pp in the reported number. Square uses no input gradient, yet its greedy acceptance rule compares losses from independently sampled masks. The masked result therefore does not establish stronger robustness: the attack is acting on noisy scores. Masking also reduces clean accuracy because the head was trained on unmasked features. We use `mask_ratio`$=0$ throughout.

Evaluation-time stochasticity can also arise from dropout, stochastic depth or randomised token dropping in other pre-trained backbones. Each can inflate gradient-based and query-based attack accuracy. Evaluators can attack in expectation over the randomness or remove it (Croce et al., 2022; Lucas et al., 2023); finding it in this case required inspecting the checkpoint's forward pass.

A.4 Compute and Reproducibility Detail

Hardware and runtime. Each probe-training or evaluation run used one GPU (RTX A5000, RTX A6000, RTX 6000 ADA or V100 depending on availability). Runs were distributed across several environments, primarily Linux servers; all used Python 3.12 and dependencies from the common requirements file, but exact package builds varied. Training a head is inexpensive because the encoder is frozen. The fourteen reported `torchattacks` AA runs at $\varepsilon{=}1/255$ and $5{,}000$ images total roughly five GPU-weeks; the I-JEPA last-block run alone took 4.8 days.

Saved results. Long attack runs write resume checkpoints. A completed run records robust and clean top-1, its attack label and nominal settings, checkpoint path, seed, image count and a per-image correctness boolean. The metadata does not preserve the instantiated AA subattack settings, which may differ from the nominal step count. The correctness booleans support the paired McNemar tests in Table 7. The records also omit an ordered image-ID hash, so equal array lengths alone cannot prove that two runs evaluated the identical sequence; the paired analysis relies on the shared evaluation commands and subset selection.

Image subsets. The main attack comparisons select the same ordered $5{,}000$-image RobustBench-IN1k list, fixed by published image IDs rather than a random seed. Transfer and CKA use a fixed, seeded sample of $1{,}000$ images selected across classes from that list. Logit rescaling uses its first $1{,}000$ images. Comparisons within each analysis are intended to be paired at the image level.

A.5 Complete Attack Results

Tables 4 and 5 retain the complete attack columns behind the compact main-text Tables 1 and 2. All attack accuracies here use the shared $5{,}000$-image subset.

model	Clean	FGSM 1	FGSM 2	FGSM 4	FGSM 8	PGD 1	PGD 2	PGD 4	PGD 8	AA 1	AA 4
DINOv2-G	**85.88**	**48.14**	**38.02**	**30.40**	**26.04**	12.12	1.04	0.00	0.00	3.28	0.00
DINOv2-L	84.90	34.60	25.00	20.56	20.48	5.20	0.14	0.00	0.00	1.20	0.00
ViT-L	80.46	42.12	34.66	29.42	25.96	12.44	1.98	0.08	0.00	3.32	0.00
MAE	76.76	18.86	12.00	7.52	5.88	0.78	0.02	0.00	0.00	0.42	0.00
MoCo v3	75.32	19.24	7.80	3.40	1.88	5.98	0.32	0.00	0.00	3.88	0.00
ViT-H	75.26	22.10	14.04	9.48	7.48	5.32	0.34	0.00	0.00	1.22	0.00
I-JEPA	73.46	32.74	21.08	14.14	11.68	**13.18**	**2.20**	0.18	0.06	**8.28**	0.00
I-JEPA-last1	67.92	33.58	23.58	17.82	15.42	13.34	2.72	0.08	0.00	7.80	0.00

Table 4: Multi-budget attack sweep on the $5{,}000$-image RobustBench-IN1k subset. Cells are robust top-1 accuracy (%). Budgets in units of $1/255$; PGD uses 20 steps; AA is the `torchattacks` AutoAttack ensemble defined in Section 3. MAE runs with `mask_ratio`$=0$ (Appendix A.3). Bold marks the column maximum over the seven probes, excluding PGD 4, PGD 8 and AA 4, where every probe is at or below the noise floor. The last row is the I-JEPA last-block ablation (Appendix B), excluded from the bolding.

configuration	BN scale	clean	FGSM 1	FGSM 2	FGSM 4	PGD 1	PGD 2	AA 1
MoCo v3, BN-on	18.81	**75.32**	19.24	7.80	3.40	5.98	0.32	3.88
MoCo v3, BN-off		73.56	**30.92**	**16.74**	**7.98**	**11.34**	**1.08**	**6.26**
Δ		-1.76	$+11.68$	$+8.94$	$+4.58$	$+5.36$	$+0.76$	$+2.38$
MAE, BN-on	11.17	**76.76**	18.86	12.00	7.52	0.78	0.02	0.42
MAE, BN-off		73.76	**28.62**	**20.72**	**14.40**	**2.60**	**0.14**	**1.10**
Δ		-3.00	$+9.76$	$+8.72$	$+6.88$	$+1.82$	$+0.12$	$+0.68$
ViT-H, BN-on	7.35	75.26	22.10	14.04	9.48	5.32	0.34	1.22
ViT-H, BN-off		**75.32**	**25.42**	**17.48**	**13.34**	**7.04**	**0.74**	**1.36**
Δ		$+0.06$	$+3.32$	$+3.44$	$+3.86$	$+1.72$	$+0.40$	$+0.14$
ViT-L, BN-on	1.66	80.46	42.12	34.66	29.42	12.44	1.98	3.32
ViT-L, BN-off		**80.78**	**44.14**	**36.48**	**31.94**	**14.14**	**2.94**	**3.46**
Δ		$+0.32$	$+2.02$	$+1.82$	$+2.52$	$+1.70$	$+0.96$	$+0.14$
DINOv2-G, BN-on	0.54	**85.88**	**48.14**	**38.02**	**30.40**	**12.12**	**1.04**	**3.28**
DINOv2-G, BN-off		85.86	44.50	33.54	26.06	9.34	0.54	3.08
Δ		-0.02	-3.64	-4.48	-4.34	-2.78	-0.50	-0.20
DINOv2-L, BN-on	0.51	**84.90**	**34.60**	**25.00**	**20.56**	**5.20**	**0.14**	**1.20**
DINOv2-L, BN-off		84.26	29.26	20.46	17.10	3.44	0.02	0.86
Δ		-0.64	-5.34	-4.54	-3.46	-1.76	-0.12	-0.34

Table 5: BatchNorm ablation on the six BN-equipped probes, $5{,}000$-image subset. Attack columns are labelled by budget in units of $1/255$. Each pair gives the BN-on baseline, the BN-off retrain, and a Δ row in percentage points. "BN scale" is the mean of $\gamma/\sqrt{\sigma^2 + \epsilon_{\mathrm{BN}}}$ over the original head's BN units. Bold marks the larger of each pair. The encoder is frozen and identical within each pair; only the head differs.

B THE I-JEPA LAST-BLOCK ABLATION

I-JEPA leads at small ε and is also the one probe that reads its encoder unconventionally, pooling the last four blocks into a 5120-dimensional feature where the others take a 1024- to 1536-dimensional `[CLS]` token. This unusual readout could contribute to I-JEPA's lead. The I-JEPA-last1 configuration uses the identical frozen encoder and a separately trained BatchNorm-free linear head with seed 0, reading the mean-pooled final output (1280 dimensions). The last-four block outputs precede the encoder's final LayerNorm; the final output follows it. Table 6 collects this compound readout comparison.

axis	quantity	I-JEPA	I-JEPA-last1	Δ
pooling	feature dimension	5120	1280	$\div 4$
clean	top-1 (%)	73.46	67.92	-5.54
adversarial	FGSM $4/255$ (%)	14.14	17.82	$+3.68$
	PGD $1/255$ (%)	13.18	13.34	$+0.16$
	PGD $2/255$ (%)	2.20	2.72	$+0.52$
	AA $1/255$ (%)	8.28	7.80	-0.48
corruption	ImageNet-C overall (%)	47.7	41.7	-6.0
	ImageNet-C severity 5 (%)	25.4	21.2	-4.2
mechanism	$\|\mathrm{head}\|_{\mathrm{op}}$	14.7	24.5	$\times 1.7$
	s_{enc}	2.26	2.32	$+0.06$
	realised s_{total}	6.91	8.28	$+1.37$
	align	0.209	0.145	-0.064
	CKA to I-JEPA	1.00	0.83	

Table 6: The I-JEPA readout comparison. Mechanism scalars are defined in Section 6 and Table 8. Both configurations use the identical frozen `facebook/ijepa_vith14_22k` encoder and a separately trained BatchNorm-free linear head with seed 0. The last-four readout uses block outputs before final LayerNorm; the last-one readout uses the final output after it. Attack and corruption numbers are on the $5{,}000$-image subset, mechanism scalars on 256 images.

The last-block probe retains its adversarial lead. PGD-$1/255$ is unchanged within noise and AA-$1/255$ falls 0.48 pp, leaving I-JEPA-last1 at 7.80%: still 3.92 pp clear of the best BN-on probe and 1.54 pp clear of the best BatchNorm-free one. Because the readout change also removes 5.54 pp of clean accuracy, the robust fraction rises, from 17.9% of clean under PGD-$1/255$ to 19.6%. The same intervention costs 6.0 pp of ImageNet-C, enough to put I-JEPA-last1 below every configuration in Table 14 at every severity. The four-block configuration has higher clean and corruption accuracy, with little difference in measured adversarial accuracy. Block selection, final normalisation, feature dimension and the fitted head all change together.

Transfer. In Table 11, attacks crafted on I-JEPA take I-JEPA-last1 to 19.9% and the reverse takes I-JEPA to 23.7%, within 0.4 and 8.8 pp of their respective targets' self-attack diagonals and better than any pair of distinct backbones. Their encoder weights are identical, while separately trained heads read features of different dimensions; their pooled features have CKA 0.83. This pair shows stronger transfer than pairs of distinct encoders. Section 6 reports the smaller absolute changes in cross-encoder transfer after the BatchNorm intervention, together with the limits of that comparison.

Limits of the sensitivity decomposition. s_{enc} moves by 0.06 between the two readouts, despite their different feature dimensions and final normalisation. The head term moves more: $\|\mathrm{head}\|_{\mathrm{op}}$ rises from 14.7 to 24.5 as the same 1,000-way map is fitted on a $4\times$ narrower input, realised s_{total} rises to 8.28, and align falls to 0.145. Reading s_{total} as a robustness predictor would suggest worse small-ε accuracy, yet here the direction agrees only at AA-$1/255$, where the 0.48 pp fall is within noise; PGD-$1/255$ and FGSM-$4/255$ move the other way. Realised s_{total} tracks head rescaling across six paired retrains but does not provide a general small-ε robustness ordering.

Caveats. Both I-JEPA configurations mean-pool patch tokens, leaving the pooling difference with the six [CLS] probes unresolved. We did not build a [CLS]-token I-JEPA probe. The two readouts differ in final LayerNorm and feature dimension. Their separately fitted heads also have different function classes: a 1280-dimensional head cannot express what a 5120-dimensional one can. The 5.54 pp clean-accuracy gap and the corruption delta reflect this combined intervention.

C Paired Tests and the Sensitivity Decomposition

C.1 Paired Tests for the BatchNorm Ablation

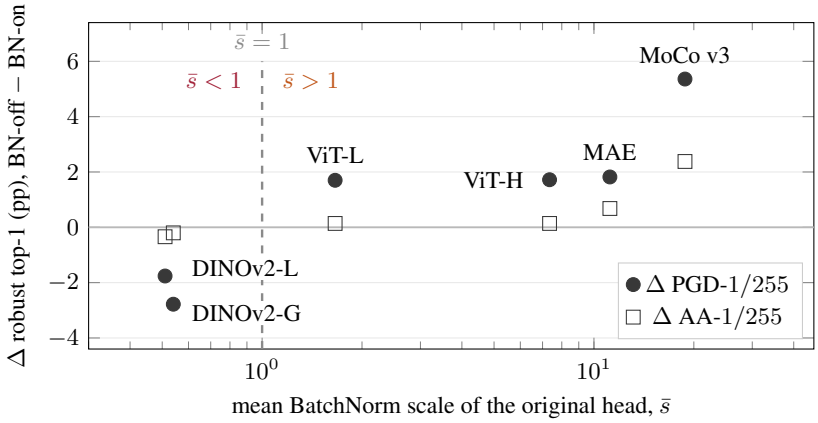


Figure 4: **The sign of the head effect splits at mean BatchNorm scale 1.** Change in robust top-1 from removing the head's BatchNorm layer, against the original head's mean coordinate scale $\bar{s}$ (log axis). All four probes with $\bar{s} > 1$ gain accuracy when the layer is removed; the two with $\bar{s} < 1$ lose accuracy. ΔAA-$1/255$ (squares) is monotone non-decreasing in $\bar{s}$ across all six probes (the two DINOv2 squares, -0.20 and -0.34, nearly coincide), and an exact McNemar test finds at least one probe significant on each side of the split (Table 7). Values are from Table 5; $\bar{s}$ does not measure the realised norm change of a particular feature shift.

Figure 4 plots the sign split discussed in Section 4, and Table 7 gives the exact McNemar tests summarised there.

probe	BN scale	ΔAA-$1/255$	BN-on only	BN-off only	exact p	verdict
MoCo v3	18.81	$+2.38$	15	134	4.6×10^{-25}	significant
MAE	11.17	$+0.68$	4	38	5.7×10^{-8}	significant
ViT-H	7.35	$+0.14$	4	11	0.12	not significant
ViT-L	1.66	$+0.14$	14	21	0.31	not significant
DINOv2-G	0.54	-0.20	30	20	0.20	not significant
DINOv2-L	0.51	-0.34	21	4	9.1×10^{-4}	significant

Table 7: Exact McNemar tests on the AA-$1/255$ ablation pairs, 5,000 shared images. "BN-on only" counts images robustly classified correctly only by the BN-on head; "BN-off only" counts those classified correctly only by the BN-off head. The two-sided test conditions on these discordant correctness outcomes. "Significant" uses the Bonferroni threshold $0.05/6 = 0.00833$. Pairing roughly halves the descriptive standard error of the difference relative to treating the two runs as independent samples.

C.2 SENSITIVITY DECOMPOSITION

Table 8 gives the head/encoder sensitivity decomposition summarised in Section 6.

configuration	BN scale	s_{enc}	$\|\text{head}\|_{\text{op}}$	realised s_{total}	align
MoCo v3, BN-on	18.81	0.15	**1808.9**	7.59	0.027
MoCo v3, BN-off		0.17	89.5	**3.39**	**0.229**
MAE, BN-on	11.17	0.45	323.0	8.05	0.056
MAE, BN-off		0.46	49.2	**4.81**	**0.211**
ViT-H, BN-on	7.35	0.45	204.5	7.00	0.077
ViT-H, BN-off		0.46	34.4	**3.82**	**0.242**
ViT-L, BN-on	1.66	1.75	36.4	7.57	0.119
ViT-L, BN-off		1.75	14.4	**5.79**	**0.230**
I-JEPA	n/a	2.26	14.7	6.91	0.209
I-JEPA-last1	n/a	2.32	24.5	8.28	0.145
DINOv2-G, BN-on	0.54	4.88	**5.58**	4.37	0.162
DINOv2-G, BN-off		4.70	6.82	5.59	0.181
DINOv2-L, BN-on	0.51	4.95	13.1	6.73	0.104
DINOv2-L, BN-off		4.89	6.80	8.86	0.267

Table 8: Head/encoder sensitivity on 256 stratified images under FGSM at $\varepsilon{=}4/255$. "BN scale" is $\bar{s}$; $\|\text{head}\|_{\text{op}} = \sigma_{\max}(WS)$ comes from the saved head. s_{enc} and s_{total} measure realised feature- and logit-space changes per unit input perturbation along each configuration's FGSM direction. Alignment is the fraction of the head's maximum singular gain realised by that feature shift. These are directional measurements, not Lipschitz constants. Bold in the last two columns marks BN-off rows with lower s_{total} and higher alignment; bold head norms mark column extremes. I-JEPA has no BN scale.

For each image with a nonzero denominator, $\text{align} = s_{\text{total}}/(\|\text{head}\|_{\text{op}} s_{\text{enc}}) \in [0, 1]$. The table averages this ratio over images; it is not the ratio of the displayed row means. The two calculations agree within 1% on most rows, but need not be equal. Textual percentage changes use unrounded values, so they may differ slightly from calculations using the printed s_{enc} values.

Across images, the coefficient of variation of s_{enc} ranges from 16% to 52%. For each fixed configuration, the standard error of its 256-image mean is about 1%–3% of that mean (0.002 on MoCo v3 to 0.16 on DINOv2-G), below the $32\times$ spread between probe means. These standard errors describe image sampling for the fitted probes; they do not cover retraining or the small number of backbone families.

C.3 Loss-Landscape Ridge Detail

We test whether loss along FGSM's fixed update direction helps explain why PGD can do more damage at a smaller budget. FGSM takes one step along $u = \mathrm{sign}(\nabla_x \mathcal{L})$, the sign of the clean-point gradient. PGD recomputes the gradient after each step and can leave this direction. The geometry below is a diagnostic of that difference; it does not establish where PGD's path goes.

For each image we draw a standard-normal direction v, project it orthogonal to u, and rescale it to the same ℓ_∞ magnitude. The draw is shared across probes. On the same 12 images from distinct classes per probe, we evaluate cross-entropy on a 21×21 grid, $\mathcal{L}(\alpha, \beta) = \mathcal{L}(\mathrm{clip}_{[0,1]}(x + \alpha u + \beta v), y)$, for $(\alpha, \beta) \in [-8/255, 8/255]^2$. A separate 65-point sweep resolves the line $\mathcal{L}(\alpha, 0)$ more finely.

Table 9 gives loss at the clean point and at $\alpha = 4/255$ and $8/255$ along u. It also gives median per-image directional curvature at $\alpha = 0$, estimated by central finite differences with step $h = 1/510$ (Moosavi-Dezfooli et al., 2019). For scale, chance-level cross-entropy on 1,000 classes is $\ln 1000 = 6.91$ nats.

probe	$\mathcal{L}(0,0)$	$\mathcal{L}(4/255,0)$	$\mathcal{L}(8/255,0)$	median curvature
DINOv2-L	0.179	4.486	4.467	35,979
ViT-L	0.162	4.338	4.496	3,397
I-JEPA	0.587	4.218	4.803	112,940
DINOv2-G	0.129	4.251	4.925	14,511
MAE	0.180	5.293	5.494	175,568
ViT-H	0.310	6.313	6.922	252,538
MoCo v3	0.246	8.250	9.411	276,470

Table 9: Cross-entropy loss along $u = \mathrm{sign}(\nabla_x \mathcal{L})$ at the clean point and at $\alpha = 4/255$ and $8/255$, and the median per-image curvature at $\alpha = 0$, on the same 12 images from distinct classes per probe. Losses are means across images; curvature is the median. Rows are sorted by the $8/255$ column. The analysis below correlates the curvature column with FGSM/PGD/AA accuracy; the $8/255$ column is called the ridge height.

Holding $\alpha = 0$ and sweeping β across $[-8/255, 8/255]$ changes loss by at most 0.05 nats on six probes along this one random orthogonal direction (ViT-L reaches 0.06; these values come from the grid and are not tabulated). Along u, loss instead reaches 4.47–9.41 nats at $+8/255$. One random direction cannot rule out greater loss in another direction. DINOv2-L's mean loss at $8/255$ (4.467) is slightly below its value at $4/255$ (4.486); on 12 images we treat this as a plateau rather than evidence of a decrease.

Across the seven probes, median clean-point curvature is associated with FGSM-$2/255$ accuracy ($\rho = -0.93$, exact two-sided permutation $p = 0.0067$). The corresponding associations with PGD-$1/255$ ($\rho = -0.39$, $p = 0.396$) and AA-$1/255$ ($\rho = 0.00$, $p = 1.00$) are not resolved in this seven-probe sample. The FGSM coefficient is unchanged at $4/255$ because the seven probes have the same accuracy ranks at both budgets. These exploratory model-level associations leave the attack trajectories unexplained.

Ridge height has a weaker association with FGSM-$2/255$ accuracy ($\rho = -0.71$, exact $p = 0.088$). Both quantities come from FGSM's own direction, so this comparison is less independent than the curvature diagnostic. MoCo v3 has the highest ridge and is a case where FGSM is especially damaging; I-JEPA has one of the three lowest ridges and still favours FGSM. Ridge height leaves the different attack rankings unexplained. The curvature association is the stronger descriptive result, consistent with the local-geometry perspective of Andriushchenko & Flammarion (2020) and Gomes (2022).

D LOGIT-RESCALING DETAIL

Why rescaling acts through the softmax temperature. PGD-ℓ_∞ steps along $\mathrm{sign}(\nabla_x \mathcal{L})$, and the rescaling contributes only an overall positive factor to that gradient. Writing $z = z_0/c$ for the rescaled logits (z_0 the original logits) and p for the softmax of z, the chain rule gives $\nabla_x \mathcal{L} = c^{-1}(\partial z_0/\partial x)^\top (p - e_y)$ for one-hot target e_y. The positive scalar c^{-1} drops out of the sign, so c enters only through the temperature that sets p, which fixes how much weight the non-true classes carry. The 0.30 pp residue in the APGD-DLR check of Section 5 corresponds to three images. Although the objective is algebraically scale-invariant, the numerical implementation can change APGD's discrete step decisions. We report the observed discrepancy rather than treating the measured scores as exactly invariant.

Section 5 rescales a trained head's logits by a positive divisor c, $W \leftarrow W/c$ and $b \leftarrow b/c$, which divides every logit by c and leaves the argmax of every image untouched. The script asserts that per image and aborts if any prediction moves, so clean accuracy is constant along each row of Table 10. The check is per image within a run; rows assembled from separate runs agree to the reported precision. Nothing is retrained and the encoder is never loaded for training; Figure 2 in Section 5 plots the two sweeps.

head	attack	c=0.05	0.2	1	1.469	2	5	10	20
BatchNorm	PGD-$1/255$	34.90	10.00	**6.20**	7.00	8.50	10.90	10.70	10.90
BatchNorm-free	PGD-$1/255$	30.00	13.80	**11.30**	n/m	15.70	16.40	n/m	n/m
BatchNorm	APGD-DLR-$1/255$	n/m	n/m	4.60	n/m	n/m	4.30	n/m	4.30
BatchNorm-free	APGD-DLR-$1/255$	n/m	n/m	10.30	n/m	n/m	n/m	n/m	n/m

Table 10: Logit rescaling on both arms of the MoCo v3 pair, robust top-1 (%) on the shared $1{,}000$-image subset. Larger c shrinks the logits; $c < 1$ enlarges them. Clean accuracy is 76.00% on every BatchNorm row and 74.50% on every BatchNorm-free row, unchanged by the rescaling. Bold marks each arm's lowest measured value over c under cross-entropy PGD, which falls at the trained scale in both cases. $c = 1.469$ is the point at which the BatchNorm head's mean clean logit margin equals the BatchNorm-free head's. The DLR objective is invariant to positive logit rescaling; APGD-DLR is a diagnostic control. Cells marked *n/m* are grid points we did not run; Section 5 states which values of c each row covers.

The shape is consistent with softmax temperature affecting the cross-entropy gradient, as Section 5 describes: enlarging the logits saturates $p \to e_y$ on confident images and $\mathrm{sign}(\nabla_x \mathcal{L})$ vanishes on them, so at $c = 0.05$ roughly half the clean-correct images survive an attack that defeats them at the trained scale. This is gradient masking in the sense of Athalye et al. (2018). It arises in an undefended model after post-training rescaling that leaves every prediction fixed.

Dependence on the matching statistic. Section 5 reports the match at equal mean clean logit margin, $c = 1.469$, which leaves 4.30 pp of the 5.10 pp gap. Other matching statistics can leave a smaller gap. Matching on the head's operator norm, which falls $1808.9 \to 89.5$ across the ablation (Table 8), puts the BN-on arm near $c = 20$, where Table 10 reads 10.90% against the BN-off arm's 11.30% and only 0.40 pp of the gap survives. At this match point, BN-on PGD accuracy is on the plateau observed from $c = 5$ onward. Accuracy alone does not tell us what happened to its gradients or attack trajectory. The margin match is therefore one specified comparison; its 4.30 pp residual does not bound all matching rules. The APGD-DLR comparison measures a 5.70 pp difference between the trained heads using a scale-invariant objective and requires no match point.

Coverage. The rescaling was swept on the MoCo v3 pair only. The two probes whose sign reverses are both DINOv2, and neither was scale-controlled. Their sensitivity to logit scale remains to be tested. The grid is also coarse near its minimum (nothing was measured between $c = 0.2$ and $c = 1$), so $c = 1$ is only the lowest measured point; the BN-off arm was not run above $c = 5$. Finally, the rescaling sweeps use the $1{,}000$-image subset while the between-encoder spread they are compared against is measured on the $5{,}000$-image superset. The descriptive spread-to-head-effect ratios in Section 8 use the BatchNorm ablation. The rescaling sweep serves as an attack audit because PGD accuracy changes while the classifier's predictions stay fixed.

E CROSS-MODEL MATRICES

Table 11 gives the FGSM (ε=4/255) transfer matrix on the BN-on probes and Table 12 the same matrix recomputed on the BN-off retrains, on the same 1,000 images and the same seed. Section 6 summarises the paired comparison. Table 13 gives the linear CKA (Kornblith et al., 2019) between the frozen pooled features. The BN-off run reproduces CKA to two decimal places on every entry, consistent with unchanged pre-head features; rounded agreement alone does not establish parameter identity.

target ↓ source →	I-JEPA	MAE	ViT-H	DINOv2-G	ViT-L	DINOv2-L	MoCo v3	I-JEPA-last1
I-JEPA	**14.9**	58.6	63.6	65.0	68.2	65.7	63.2	23.7
MAE	63.1	**8.4**	69.3	65.5	70.8	66.5	66.6	65.9
ViT-H	61.4	64.1	**9.5**	68.5	67.5	70.7	61.8	64.5
DINOv2-G	80.0	77.4	83.3	**33.1**	81.5	57.8	79.2	81.5
ViT-L	72.6	73.0	70.9	73.6	**31.2**	74.9	54.3	74.9
DINOv2-L	79.1	75.6	80.9	51.8	80.2	**22.1**	78.0	79.7
MoCo v3	64.9	64.8	67.3	66.5	58.9	67.3	**3.8**	67.4
I-JEPA-last1	19.9	54.3	60.5	60.1	63.5	59.7	58.4	**19.5**

Table 11: FGSM (ε=4/255) transfer matrix, BN-on probes, on a fixed, seeded sample of 1,000 images selected across classes, with the last-block readout as the last row and column. Cells are the target's robust top-1 (%) on the source's adversarial inputs; diagonals (bold) are self-attack. Between distinct backbones the off-diagonals sit at 51.8 to 83.3%, close to clean accuracy: these perturbations transfer poorly. Two pairs fall below 60 in both directions, the DINOv2 pair (51.8, 57.8) and the MoCo v3 / ViT-L pair (54.3, 58.9). Diagonals use the same fixed 1,000-image sample and so do not equal Table 4's 5,000-image column; the two differ by up to 2.7 pp from subset sampling. The (I-JEPA, I-JEPA-last1) cells are far stronger (19.9 and 23.7, within 0.4 and 8.8 pp of their respective targets' self-attack diagonals), and they are the only off-diagonals where source and target share encoder weights.

target ↓ source →	MAE	ViT-H	DINOv2-G	ViT-L	DINOv2-L	MoCo v3
MAE	**16.0**	68.2	66.9	69.9	67.3	65.4
ViT-H	64.5	**13.2**	70.4	67.7	71.2	60.2
DINOv2-G	77.5	82.5	**27.6**	80.3	58.5	78.2
ViT-L	73.3	70.2	74.2	**33.7**	75.9	51.0
DINOv2-L	73.7	80.3	51.7	79.2	**18.2**	76.9
MoCo v3	63.5	65.4	69.8	60.5	69.7	**7.7**

Table 12: The same matrix on the BatchNorm-off retrains; I-JEPA is omitted as it has no BatchNorm to remove. Rows and columns follow Table 11, with I-JEPA and its last-block ablation dropped, so the two grids are read by label rather than by position. The diagonal moves by 4.5 pp on average and in the direction Table 5 predicts; the off-diagonal moves by 1.2 pp. Its paired noise level is unknown because per-image transfer outcomes were not retained (Section 6). The DINOv2 pair survives the ablation nearly unchanged (58.5 against 57.8; 51.7 against 51.8).

CKA	I-JEPA	MAE	ViT-H	DINOv2-G	ViT-L	DINOv2-L	MoCo v3	I-JEPA-last1
I-JEPA	1.00	0.15	0.45	0.39	0.39	0.39	0.46	0.83
MAE	0.15	1.00	0.12	0.10	0.09	0.10	0.16	0.17
ViT-H	0.45	0.12	1.00	0.61	**0.66**	0.57	0.47	0.40
DINOv2-G	0.39	0.10	0.61	1.00	0.62	**0.71**	0.44	0.35
ViT-L	0.39	0.09	**0.66**	0.62	1.00	0.58	0.43	0.35
DINOv2-L	0.39	0.10	0.57	**0.71**	0.58	1.00	0.43	0.35
MoCo v3	0.46	0.16	0.47	0.44	0.43	0.43	1.00	0.43
I-JEPA-last1	0.83	0.17	0.40	0.35	0.35	0.35	0.43	1.00

Table 13: Linear centred kernel alignment (CKA) (Kornblith et al., 2019) between clean pooled features, 1,000 images, with the last-block ablation ruled off. MAE is a strong outlier (0.09 to 0.17 against everything else). Among distinct backbones, the DINOv2 variants are the most similar pair (0.71) and the two supervised ViTs the second (0.66). The supervised ViTs transfer weakly relative to their self-attacks (Table 11), so feature similarity alone does not determine transfer strength. The two I-JEPA readouts have CKA 0.83 and transfer more strongly than any pair of distinct backbones here.

F IMAGENET-C DETAIL

Table 14 gives the per-severity ImageNet-C numbers, sorted by overall accuracy. Table 15 then breaks the BatchNorm effect down by corruption category, alongside the same backbone's ΔPGD-$1/255$ for comparison.

configuration	sev 1	sev 2	sev 3	sev 4	sev 5	overall
DINOv2-G	84.4	81.9	80.5	77.1	69.8	**78.7**
DINOv2-G, BN-off	84.0	81.5	80.0	76.4	69.2	78.2
DINOv2-L	82.9	79.7	77.6	72.1	61.1	74.7
DINOv2-L, BN-off	82.0	78.7	76.4	70.8	59.7	73.5
ViT-L, BN-off	76.5	70.9	66.7	57.8	44.6	63.3
ViT-L	76.1	70.3	66.0	57.0	43.7	62.6
MAE	69.0	61.7	55.5	44.4	31.7	52.4
ViT-H, BN-off	68.7	61.5	55.0	43.1	30.3	51.7
ViT-H	68.5	61.0	54.2	42.1	29.2	51.0
MAE, BN-off	66.7	59.8	53.9	43.3	30.6	50.9
MoCo v3, BN-off	67.2	59.6	53.2	43.2	30.7	50.8
MoCo v3	67.2	58.2	50.7	40.1	27.8	48.8
I-JEPA	66.2	58.2	51.1	37.9	25.4	**47.7**
I-JEPA-last1	59.9	51.4	44.2	32.0	21.2	41.7

Table 14: ImageNet-C top-1 accuracy under corruption (%) on the $5{,}000$-image subset, sorted by overall accuracy. Each severity column averages 15 corruptions; "overall" averages all 15 at all five severities. BN removal changes overall accuracy by at most 2.0 pp. Five BN-on/BN-off pairs remain adjacent; MAE moves within a group of five configurations spanning 1.6 pp. Bold marks the extremes among the thirteen configurations above the rule. The ruled-off I-JEPA last-block ablation loses a further 6.0 pp.

backbone	BN-on overall	BN-off overall	Δ overall	Δ PGD-$1/255$	per-category Δ (noise / blur / weather / digital)
MoCo v3	48.81	50.78	$+1.97$	$+5.36$	$+3.0$ / $+2.6$ / $+2.0$ / $+0.6$
ViT-H	51.01	51.72	$+0.71$	$+1.72$	$+1.1$ / $+1.1$ / $+0.5$ / $+0.3$
ViT-L	62.63	63.27	$+0.64$	$+1.70$	$+0.7$ / $+0.7$ / $+0.7$ / $+0.5$
DINOv2-G	78.73	78.24	-0.49	-2.78	-0.4 / -0.6 / -0.5 / -0.4
DINOv2-L	74.68	73.52	-1.16	-1.76	-1.2 / -1.4 / -1.1 / -1.0
MAE	52.44	50.86	-1.57	$+1.82$	$+0.2$ / -1.6 / -2.6 / -1.9

Table 15: BatchNorm ablation effect on ImageNet-C in percentage points, always BN-off minus BN-on. The final column lists noise, blur, weather and digital categories in that order. Five of six overall changes have the same sign as the PGD-$1/255$ change; MAE is the exception, with its largest losses in weather and digital corruptions.