DEGRADE, THEN ATTACK: COMPOSITION ORDER IS PART OF THE THREAT MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Inputs reach a deployed classifier after lossy transport or degradation, and may also have been attacked. Evaluating a model against both means composing two operators, which means choosing which comes first, and that choice is usually made by the evaluation instrument rather than reported. Attacking a released corruption benchmark computes the attack on an already degraded image; evaluating a lossy transformation as a defence applies it to an image attacked first. The two lines carry opposite defaults, and a composed robustness number does not say which it used. Across 300 pairs of a CIFAR-10 model and a corruption type, degrading an image and then attacking it leaves accuracy a median of 12.40 points below attacking it and then degrading it, with the sign consistent on all 20 models and all 15 corruption types, and +15.50 points on native 1000-way ImageNet. Splitting the difference isolates the one term whose sign is predictable in advance, a recovery of accuracy on inputs already attacked. At the severity we evaluate it is -1.73 points with a confidence interval excluding zero, and it falls further as the corruption strengthens, so its sign is a property of the strength rather than of degradation. The effect is carried by the other term, that a corrupted image is easier to attack, and it tracks boundary geometry rather than the frequency-band overlap commonly invoked for this interaction. Most of the gap is attacker information: on the 5 differentiable corruption types, an attacker told the realised corruption but required to commit before it recovers 89.4% of the gap, or 70.3% dropping cells with an unstable small denominator. An attacker optimising over the corruption distribution lands between the two orders, making degrading before attacking the conservative endpoint. The two orders are two threat models differing in when the attack happens and in what the attacker sees, and they do not rank the same models in the same way, so a robustness number for a composed threat is under-specified by 12 points unless it states which order it measured.

1 INTRODUCTION

An image that reaches a deployed classifier has usually been through something on the way. It has been compressed for transport, captured through weather or a dirty lens, or resampled by a pipeline the model never sees. It may also have been attacked. Evaluating a model against both possibilities means composing two operators, and composing them means putting one first. Which one comes first follows from where in the pipeline the adversary sits. Upstream of the degradation, the perturbation is committed first and the degradation acts on it: an image perturbed and then uploaded, re-encoded and resized, or a printed pattern photographed through rain, dirt and motion blur. Downstream of it, the adversary receives an image that has already been degraded and computes the perturbation on that image: a compromised link between a camera's preprocessing and the model it feeds, or stored imagery modified after acquisition and compression. An adversary that probes the deployed system without seeing the realised degradation sits between the two, and we measure that case as well (Section 5).

The two robustness literatures meet at exactly this point, and the composition is fixed by the instrument rather than chosen. Corruption robustness is measured on benchmarks released as fixed sets of already degraded images (Hendrycks & Dietterich, 2019), so an attack evaluated on one of them is necessarily computed after the degradation, which is the arrangement used where the two axes are measured together at scale (Li et al., 2024). Attack robustness is measured by ensembles applied

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.

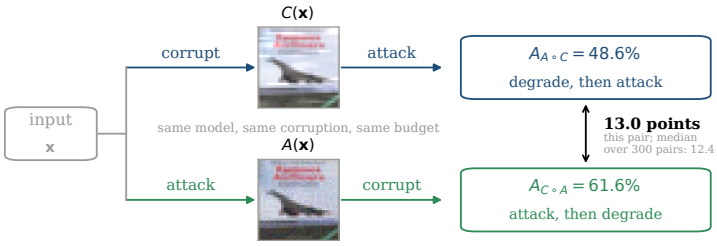

Figure 1: The two orders on one CIFAR-10 image, for the model of Carmon et al. (2019) under snow at severity 3: the corruption is plainly visible and the perturbation at $\varepsilon = 8/255$ is not. The image is illustrative; the 13.0 points are an accuracy difference over the 500-image subset for this model and corruption type, close to the median of 12.40 over all 300 pairs.

to clean inputs (Croce & Hein, 2020; Croce et al., 2021), and the opposite arrangement is what a lossy transformation applied as a defence evaluates (Guo et al., 2018; Nie et al., 2022), studied separately and for a different purpose. Each default follows from how its evaluation is built rather than from a threat being modelled, neither is stated as part of what is measured, and the two defaults are opposite. A reader of a composed robustness number cannot recover which of them produced it.

How far apart the two answers are has not been quantified across models and corruption types, and neither has the question of what makes them differ. This paper measures both. We evaluate 20 CIFAR-10 models on 15 corruption types at a fixed severity, attacking with PGD-50 at $\varepsilon = 8/255$, and run both orders on the same images.

The two answers are far apart, and the gap has the same sign on almost every pair of a model and a corruption type. Inserting the adversarial accuracy on clean inputs splits that difference into two terms, and the sign of one of them can be predicted in advance: if degrading an attacked image returns accuracy to the defender, an attacked-then-degraded image is classified better than one that was only attacked. At the severity we evaluate that term is negative, because the corruption breaks more of the inputs the attack had failed on than it restores of those it had won. The effect comes almost entirely from the other term, which says a corrupted image is easier to attack, and it tracks the geometry of the decision boundary: corruption moves inputs closer to it and into regions with more competing classes, and the same budget then carries them across.

We make the following contributions.

- **The size of the order effect, and that it changes conclusions.** A median of 12.40 points over 300 model-corruption pairs, with the same sign on every model and every corruption type, surviving a stronger attack and reappearing on native 1000-way ImageNet. It is not a common shift: the two orders rank the same models differently (Section 4).
- **A directional result about the usual explanation.** The decomposition isolates the one term whose sign can be predicted in advance, a recovery of accuracy on inputs that were already attacked. At the severity we evaluate it is negative rather than positive, at -1.73 points with a confidence interval excluding zero, and it approaches zero as the corruption weakens (Section 6).
- **A candidate mechanism, with the competing one controlled.** The effect tracks the collapse of a margin distance measured on the corrupted input before the attack, reproduced from an independent distance estimate and as a dose-response within a corruption type; curvature stops acting once that distance is held fixed. The frequency-band account, for which we specified a negative prediction in advance and measured a positive correlation, ranks last of the three (Section 7).
- **What the gap is made of, and what to report.** An attacker told the realised corruption but required to commit before it recovers most of the gap, so the effect is mainly one of attacker information rather than of timing, and adaptive attackers land between the two orders. The two orders are different threat models rather than two implementations of one (Sections 5 and 8).

2 Related work

Robustness to common corruptions was made measurable by fixing the perturbations in advance: CIFAR-10-C and ImageNet-C release 15 corruption types at five severities as static sets of already degraded images (Hendrycks & Dietterich, 2019). The methods built on that instrument are augmentation schemes, from mixing chains of simple operations (Hendrycks et al., 2020) to compositions of random primitives (Modas et al., 2022), and the benchmark has since been carried to new architectures, tasks and model classes (Vaish et al., 2024; Ranabhat et al., 2025; Siedel et al., 2026; Zhang et al., 2025b;a; 2024c; Yi et al., 2024; Oei et al., 2025). Because the degraded images are the dataset, anything applied to them comes after. Robustness to adversarial perturbations was made measurable the other way round, by fixing a threat model and searching inside it: a norm ball around the clean input, training against the worst case in it (Madry et al., 2018; Zhang et al., 2019), and a standardised ensemble of attacks with a public leaderboard (Croce & Hein, 2020; Croce et al., 2021), an axis since pushed with generated data, scale and loss design (Wang et al., 2023; 2024b; Cui et al., 2024; Zhang et al., 2024a; Bartoldson et al., 2024) and served by tooling that spans both (Mao et al., 2025; Pan & Yao, 2026). That search starts at the clean image, so the perturbation is committed before anything else acts on it.

The two lines have been connected at the level of error rates, adversarial examples and errors under random noise being two views of the same test error (Gilmer et al., 2019), a connection drawn within one arrangement of the operators. Adversarial training with a small budget is known to improve corrupted accuracy (Kireev et al., 2022), the relation between the two kinds of perturbation has been surveyed and modelled probabilistically (Karunanayake et al., 2024; Rossolini, 2026), and new model families are reported on both axes as a matter of course (Malik et al., 2025; Medi et al., 2024). They are measured together at scale by OODRobustBench, which runs one adversarial evaluation per shifted dataset over 706 robust models and reports a strong positive linear relation between in-distribution and out-of-distribution adversarial robustness (Li et al., 2024). Every one of those evaluations attacks data that has already been shifted, which fixes the arrangement without making it a reported quantity. We reproduce that relation on an independent model set and return in Section 8 to what it does and does not predict.

Composition itself has been treated as a design question, with the composed operators belonging to one of the two parties. On the attack side the strongest composite is found by searching over sequences of attack operators (Mao et al., 2021), adversarial training has been generalised to composite semantic perturbations with the ordering as one of the attacker's degrees of freedom (Hsiung et al., 2023), and corruption is used as a component of the attack itself, where augmenting the inputs an attack is built on makes the resulting examples transfer better to defended models (Huang et al., 2023). An attacker that optimises against a distribution of transformations rather than one draw is the expectation-over-transformation construction of Athalye et al. (2018b), which falls between the two orders here. On the defence side, input transformations apply a lossy operator to an input that may already carry a perturbation (Guo et al., 2018; Nie et al., 2022) and how to evaluate them is itself a subject of study (Athalye et al., 2018a; Lee & Kim, 2023); there the operator is the defender's and acts last. What has not been systematically quantified across models and corruption types is how far apart the two orders are when the corruption is neither the attacker's to choose nor the defender's to apply, but drawn from a deployment distribution and specified independently of the model, and what makes them differ.

3 Composing a corruption with an attack

Let f be a classifier, let C be a corruption operator drawn from a corruption distribution, and let A_f be an attack on f. The two operators differ in a way that matters for what follows: C is stochastic and does not depend on f, while A_f is computed from f at the point it is applied to. Composing them therefore requires choosing an order, and the choice is not a matter of notation. Writing $\mathbf{1}\{\cdot\}$ for the indicator of a correct prediction,

$$
\begin{aligned}
A_{\text{clean}} &= \mathbb{E}_x[\mathbf{1}\{f(x) = y\}]\,, & A_{\text{corr}} &= \mathbb{E}_{x,C}[\mathbf{1}\{f(C(x)) = y\}]\,,\\
A_{\text{adv}} &= \mathbb{E}_x[\mathbf{1}\{f(A_f(x)) = y\}]\,, & A_{C\circ A} &= \mathbb{E}_{x,C}[\mathbf{1}\{f(C(A_f(x))) = y\}]\,,\\
A_{A\circ C} &= \mathbb{E}_{x,C}[\mathbf{1}\{f(A_f(C(x))) = y\}]\,. & &
\end{aligned}
\tag{1}
$$

$A_{C\circ A}$ is the accuracy when the image is attacked and then degraded, and $A_{A\circ C}$ is the accuracy when it is degraded and then attacked. The quantity this paper measures is their difference,

$$\mathrm{NC}(f) \;=\; A_{C\circ A}(f) \;-\; A_{A\circ C}(f). \tag{2}$$

3.1 A DECOMPOSITION WITH A SIGN PREDICTION IN IT

NC admits one decomposition that is more informative than the others, obtained by inserting the adversarial accuracy on clean inputs:

$$\mathrm{NC} \;=\; \underbrace{(A_{C\circ A} - A_{\mathrm{adv}})}_{\text{recovery}} \;+\; \underbrace{(A_{\mathrm{adv}} - A_{A\circ C})}_{\text{exposure}}. \tag{3}$$

The two terms answer different questions. The recovery term asks whether applying the corruption to an already-perturbed input recovers accuracy. The exposure term asks whether the same attack budget is more effective once the input has been corrupted. The decomposition is useful because the first term has a predicted sign, so measuring it is a directional test rather than a comparison of magnitudes.[1]

We also compared the composed error to the error independent failure events would produce: that gap is $+1.39$ pp for the $A\circ C$ order, below the 3 pp specified in advance and load-bearing for nothing here (Appendix C.4).

3.2 SETUP

The main measurements use 20 CIFAR-10 models: 18 adversarially trained ones from RobustBench (Croce et al., 2021), spanning unlabelled data, early stopping, weight perturbation, proxy distributions and generated-data augmentation (Carmon et al., 2019; Rice et al., 2020; Wu et al., 2020; Sehwag et al., 2022; Rebuffi et al., 2021), and two standard-trained ones. The corruptions are all 15 CIFAR-10-C types at severity 3 (Hendrycks & Dietterich, 2019), and the attack is PGD-50 in ℓ_∞ at $\varepsilon = 8/255$ (Madry et al., 2018), run on the same 500 held-out images in both orders so that every comparison is paired at the image level. The subset is fixed at 500 because the design spends its budget on breadth, 300 configurations each run in both orders, at a cost of about 2.2 pp of standard error on a single configuration. The statistics throughout are medians and sign tests over configurations. The ImageNet replications, their model sets and their attack budgets are specified in Appendices B, G.1 and G.2; they test the mechanism chain rather than absolute levels.

3.3 THE POPULATION THESE CLAIMS ARE ABOUT

The recovery term is bounded by $1 - A_{\mathrm{adv}}$ above and $-A_{\mathrm{adv}}$ below, so a model with almost no adversarial accuracy admits a large positive value for arithmetic reasons and a corruption that by itself destroys accuracy admits only a negative one, and neither describes a model whose robustness is in question (Appendix C.1). The population the paper's claims are about is therefore models with adversarial headroom, $A_{\mathrm{adv}} \geq 0.20$, on configurations with corruption headroom, $A_{\mathrm{corr}} \geq 0.50$. Both bounds were fixed after inspecting the accuracy ranges, and every affected quantity below is reported with and without them. On CIFAR-10 they leave 18 of 20 models and remove 16 of their 270 configurations, all contrast.

Every other criterion in this paper was written into the analysis script before the run it judges. Appendix A lists each with the threshold specified in advance and its outcome, including the four that came out against the prediction we had registered.

4 THE ORDER EFFECT IS LARGE AND CLUSTER-ROBUST

Across the 300 pairs of a model and a corruption type the median value of NC is $+12.40$ pp and 99% of the pairs are positive: degrading an image and then attacking it leaves the classifier substantially less accurate than the reverse (Figure 2a). The effect is not an artifact of the attack being weak:

[1] Means are additive under Eq. (3) and medians are not, so decomposition tables report means. The per-configuration share $\mathrm{median}_i(\mathrm{exposure}_i/\mathrm{NC}_i)$ is not the ratio of tabulated medians.

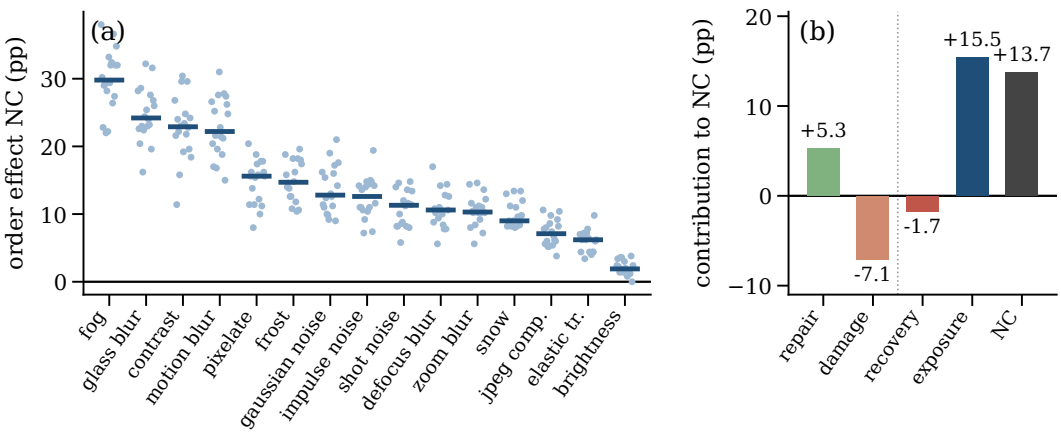

Figure 2: **(a)** The order effect per corruption type, one point per model over the 18 models with headroom, bars at the type median. **(b)** The decomposition of Eq. (3) as means over the population of Section 3: the corruption repairs less of the evaluation set than it damages, and the effect is carried by exposure.

replacing PGD-50 with AutoAttack on 5 models leaves the median at $+25.8$ pp against $+17.3$ pp on the same cells, a factor of 1.49, with 0 models reversing.[2] Appendix I.1 plots the two accuracies against each other.

4.1 The effect under four clustering conventions

The 300 observations are not independent: they are nested in 20 models and cross-classified by 15 corruption types, and an intraclass correlation of 0.078 leaves them an effective sample size of 143.7, which is why the tests we rely on are at the cluster level and none is reported on the first row of Table 1. The effect survives all four conventions for that dependence, and per-model medians run from $+4.80$ to $+23.60$ pp.

Table 1: The order effect under four ways of treating the dependence among the 300 observations. Intervals are percentile bootstraps, two-level where the unit is a cluster; taking the median within each model first gives +11.30 pp on every row.

Unit of analysis	k	Median NC (configs)	Positive	Test	95% interval
Configuration	300	+12.40	99%	not reported	[+11.20, +13.90]
Model	20	+12.40	20 of 20	sign $p = 1.9e-06$	[+11.00, +14.40]
Model, with headroom	18	+12.10	18 of 18	sign $p = 7.6e-06$	[+10.90, +14.20]
Corruption type	15	+12.40	15 of 15	sign $p = 6.1e-05$	[+9.70, +17.70]

4.2 Neither ceiling produces the effect

Two arithmetic ceilings could manufacture a large NC with no mechanism behind it, and one of them is visible in our own data on the standard-trained models (Appendix C.1). This is the reason for the population defined in Section 3, and why the headline is reported both over all models and over the models with headroom (Table 1). A single budget also places models of different robustness at different points of their own accuracy curves; calibrating ε per model to a common clean adversarial accuracy leaves the effect positive on all 104 configurations, at a median of $+24.2$ pp (Appendix C.3). The corruption types were not selected for the result either: an earlier round on 4 representative types gives the same answer as the full set of 15, to within 1.70 pp (Appendix C.2).

[2]That comparison uses 20 cells, 4 types at $n = 256$ against the 15 at $n = 500$ used elsewhere, and includes the two types with the largest effect, which is why the paired PGD figure is quoted beside it.

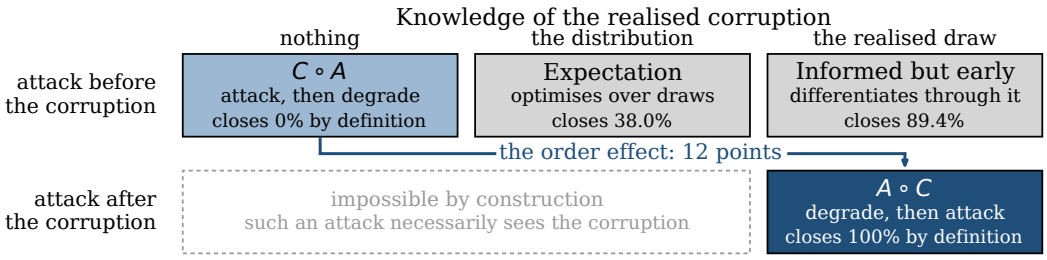

Figure 3: The four arrangements: columns vary what the attacker knows, rows vary when the attack is computed, and the two normally reported differ in both at once. The closed fractions are medians over the 25 shared cells, where the order effect is $+12.6$ pp (Appendix D).

4.3 The two orders rank models differently

A difference in level alone would leave comparisons between models untouched, and it does not. Ranking the 18 models with headroom by mean accuracy over the 15 corruption types gives two different orderings: Kendall's τ between them is 0.61, with an interval of $[0.48, 0.74]$ resampling corruption types, and 30 of 153 model pairs reverse, 14 of them in at least 95% of the resamples. The largest single move is 8 places out of 18: one model leads another by 5.3 pp when the attack comes first and trails it by 2.8 pp when the corruption does.[3] Which model looks more robust is therefore a function of an arrangement that neither ordering states; Figure 5 in Appendix C.1 draws the two orderings side by side.

5 Separating attacker knowledge from attack timing

That arrangement is two things at once. In $A \circ C$ the attack is computed on the corrupted input it will be judged on, while in $C \circ A$ it is computed beforehand, so the comparison varies when the attack happens and what the attacker sees together (Figure 3). This section separates them.

An attacker optimising the expected loss over the corruption distribution (Athalye et al., 2018b) closes a median of 31% of the gap on 5 models and 15 types, where the closed fraction q divides an arrangement's distance from $A_{C \circ A}$ by the order effect, so $q = 0$ is $C \circ A$ and $q = 1$ is $A \circ C$ (Eq. (6)). It does not settle the attribution: it knows the distribution but not the draw.

The fourth cell closes that gap in the design: its attacker is told the realised corruption and differentiates through it but must still commit before it is applied, so only the timing differs. It runs on the 5 types with a differentiable form, where the expectation attacker closes 38.0% of the same 25 cells. Moving to the realised instance takes it to 89.4%, the top band of the criterion set in advance, or 70.3% with the brightness cells dropped for an unstable ratio; both readings assign the majority of the gap to what the attacker knows, against the claim we registered. What the timing is worth separates by how the operator acts on a perturbation rather than by the accuracy it costs, in a grouping that was not specified in advance (Appendix D).

The expectation attacker also falls between the two orders, $A_{A \circ C} \leq A_{\mathrm{EOT}} \leq A_{C \circ A}$, on 70 of 75 cells. There the cheaper $A \circ C$ is already a conservative reading of an adaptive attacker, while $C \circ A$ is optimistic against it by a median of $+5.0$ pp. Composed robustness is therefore not well defined until the causal order and the attacker's knowledge are stated: where the deployment fixes the order, report the matching cell; where both timings are plausible, $A \circ C$ is the conservative endpoint. What is not defensible is reporting one without saying which.

[3] This comparison was not specified in advance, and the resampling is over corruption types, so the interval carries between-type variation and not image-level noise. We report the two orderings, not a verdict on any pair of models.

6 CORRUPTION APPLIED AFTER THE ATTACK DOES NOT RECOVER ACCURACY AT THE SEVERITY WE EVALUATE

The recovery term of Eq. (3) is the one part of the decomposition whose sign can be predicted in advance: a transformation applied after an attack can in principle return accuracy to the defender, and that is what a positive value would mean. It is what the purification defences of Section 2 are built to do, from bit-depth reduction and compression to diffusion purification and the procedures built on it (Zhang et al., 2026; Tang & Zhang, 2024; Wang et al., 2024a; Zhang et al., 2024b; Lin et al., 2024; Dai et al., 2025), and perturbations printed and re-photographed lose part of their effect (Kurakin et al., 2017). The corruptions measured here are deployment degradations from a standard benchmark rather than purpose-built purification, and the experiments below concern the former.

On the population defined in Section 3, the recovery term is -1.73 pp with a 95% confidence interval of $[-2.27, -1.21]$, and the exposure term is $+15.45$ pp with an interval of $[+14.12, +16.81]$ (Figure 2b). Applying the corruption after the attack does not return accuracy to the defender; it lowers it further. Table 2 gives the same quantities without the corruption-headroom bound: the bound reduces the magnitude and leaves the sign and the significance unchanged. Over all 300 pairs the exposure term is what tracks the total: rank correlation $+0.747$ with NC against -0.303 for recovery, per-configuration share 105% against -5%.

6.1 THE NET FIGURE IS A DIFFERENCE BETWEEN TWO FLOWS

A negative recovery term does not on its own say that the perturbation survived the corruption, because the corruption misclassifies inputs of its own. The paired outcomes separate the two flows. Among the inputs the attack had defeated, applying the corruption restores a correct prediction on a median of 12.1% of them; among the inputs that survived the attack it breaks a median of 9.3%. The two rates have different denominators, and the flow that is larger as a rate is the smaller of the two once both are put over the whole evaluation set: there the two average $+5.34$ and -7.08, and the recovery term is their net.

Repair therefore happens and is the smaller of the two flows. What this rules out is the operational claim, that a lossy transformation applied after an attack returns accuracy to the defender; how much of the perturbation survives is not resolved by outcome counts. Where the term is positive the ceiling of Section 4 explains it (Appendix C.1).

The sign is a statement about a corruption strength, not about corruption in general. Read off the same three-severity runs used for the dose-response of Section 7, on the cells that clear the corruption-headroom bound, the recovery term falls from $+0.17$ pp at severity 1 to -2.76 and -4.65 at severities 3 and 5, decreasing on all 5 of 5 models, and at severity 1 it is not distinguishable from zero (3 of 5 positive, $p = 1$). Dropping the bound moves the levels and not the direction (Appendix E). The claim is therefore about degradations strong enough to cost accuracy, and the ImageNet replication sits where that predicts: -2.80 pp at severity 1 but $+0.97$ pp once the strength is matched down (Appendices E and G.1). What the recovery term leaves unexplained is the rest of the order effect, and the question becomes what makes a corrupted input easier to attack.

7 CORRUPTION MOVES INPUTS TOWARD THE DECISION BOUNDARY

The exposure term says that the same attack budget goes further on a corrupted input. This section asks why that point is easier to attack, a different question from what it costs to reach it from before the corruption, which is what Section 5 asked. The chain we test runs from the displacement the corruption causes in feature space, to the boundary geometry around the displaced point, to the effectiveness of the attack, every link measured on the pairs used above so they can be compared with one another.

7.1 BOUNDARY GEOMETRY PREDICTS EXPOSURE

The first link is the easiest to check: the distance a corruption moves an input in feature space, $\|\varphi(C(x)) - \varphi(x)\|$, correlates with the exposure term on all 18 models with headroom, at $+0.582$ to



(a) the reading

(b) the ranking

Figure 4: **(a)** The reading: the corruption leaves the input on the same side of the boundary but closer to it, so the same budget reaches across (a schematic; Appendix F.5). **(b)** Squared Spearman correlation with the exposure term for the three candidates, one dot per model, bars at the median, all within-model.

$+0.932$, and at or above $+0.70$ on 17 of them.[4] Displacement alone does not say where the input has moved to. We measure how far an input sits from the nearest decision boundary with the first-order estimate

$$d(x) \;=\; \frac{z_y(x) - \max_{j \neq y} z_j(x)}{\left\| \nabla_x \left(z_y(x) - \max_{j \neq y} z_j(x) \right) \right\|_2}, \tag{4}$$

the linearised distance used by DeepFool (Moosavi-Dezfooli et al., 2016), and report the ratio $d(C(x))/d(x)$ per corruption type, computed on the corrupted input before any attack so that it predicts what the attack will achieve instead of summarising it.[5] To check that the chain is not simply the proxy predicting itself, we recompute the ratio from a search-based estimate that iterates rather than extrapolating once, on 5 models over all 15 types (0.000 unresolved); it correlates with the exposure term at -0.950 to -0.864, negative on every model.

The ratio correlates with the exposure term at -0.989 to -0.907 across the 18 models, negative on every one of them: the smaller the remaining margin, the larger the exposure. One corruption type, contrast, drives the median margin negative, meaning the corruption alone has already moved most inputs across the boundary, and removing it leaves the correlations at -0.987 to -0.886 (Appendix F.3). The number of classes competing near the input moves with the distance on 17 of the 18 models, at $+0.922$ down to $+0.618$; the remaining model gives -0.047, which is indistinguishable from no relation rather than a reversal. Figure 4(a) draws the reading these numbers support.

A corruption that misclassifies more inputs on its own will both shrink the margin and raise the exposure term, so part of this correlation could be the severity of the corruption acting on both. Holding the corrupted accuracy fixed, the partial correlation is -0.741 at the median and negative on 18 of 18 models (-0.950 to -0.243); restricting both to the inputs the corruption still classifies correctly gives -0.710 on 18 of 18. Either control weakens the relation and neither changes its sign on any model. One level down these aggregates are a monotone sequence over corruption types (Appendix F.2).

7.2 Curvature usually stops acting once boundary distance is controlled

Local curvature is the other geometric quantity that could explain the exposure term, and on its own it does: the curvature ratio between the corrupted and clean neighbourhood correlates with the extra drop on all 18 models, at $+0.656$ to $+0.888$. Controlling for the margin distance removes it on most of the population: the partial correlations between curvature and the exposure term have median $+0.084$ and 16 of the 18 models fall within ± 0.40, with two exceptions at $+0.508$ and $+0.918$; all 18 values are in Appendix F.1. Curvature therefore usually stops acting as an independent variable once distance is controlled, and not always, which leaves the margin distance as the only one of the two whose sign is stable across all 18 models and is why the chain is stated in terms of it.

[4] The endpoint specified in advance was $|\mathrm{NC}|$ rather than the exposure term, which gives $+0.625$ to $+0.901$; both are in Appendix B.1.

[5] The estimate uses the model's own gradients and the ℓ_2 denominator DeepFool defines while the attack is ℓ_∞; both caveats are collected in Section 8 and Appendix B.

The correlations above are computed across corruption types, so they could reflect a stable ordering of types rather than a response to how much corruption is applied. Varying severity separates the two: holding the type fixed and raising severity, the order effect rises on 5 of 6 types and the margin distance ratio falls on all 6, a dose-response inside a single type.

7.3 Frequency is ranked last among the candidates

Frequency-domain complementarity is a common account of how corruptions and adversarial perturbations interact (Yin et al., 2019), and it remains the working picture in recent analyses of both phenomena (Zhang et al., 2024d; Chen et al., 2025; Vaish et al., 2024; Zhang et al., 2025b). We specified a prediction from it before running: the overlap between the band a corruption damages and the band a perturbation occupies should correlate negatively with the composition gap, because a corruption that shares bands with the perturbation should be better at destroying it. The measured correlations run from $+0.064$ to $+0.596$, positive on all 18 models and against the sign specified in advance; the per-model values are in Appendix A. One reading is that band overlap measures whether the corruption acts on the bands the model is sensitive to, which makes it a facet of displacement rather than a capacity to destroy. By squared Spearman correlation with the exposure term the three candidates order as margin distance 82 to 98 per cent, displacement 34 to 87 per cent, band overlap 0.2 to 27.3 per cent, all within model across corruption types rather than fitted with held-out prediction; dynamic range compression is a partial mediator at the strong end (Appendix F.4).

7.4 The order effect holds on native ImageNet-1k

The order effect is not a property of CIFAR-10. On the native 1000-way ImageNet task, with no class masking, 10 robust models spanning 5 architectures from 7 independent groups give a median order effect of $+15.50$ pp, positive on all 10 of them, with a model-level bootstrap interval of $[+14.22, +17.25]$ pp and a sign test at $p = 0.00195$. Its size is not comparable to CIFAR-10, where the budget, the strength, the class count and the model set all differ (Appendix B); what carries across is the sign, the consistency and the chain. At the official severity all three links reproduce on every model, and at matched strength on most but not all, so the unanimity, like the recovery term, is a property of the strength rather than of the dataset. Under a second convention for choosing it the effect is $+9.83$ pp on all 10 models. Table 5 in Appendix G.1 sets every link of the chain on the three datasets side by side.

8 Discussion and limitations

A robustness number for a composed threat is under-specified by 12 points unless it names its order: a model robust to corruption and robust to attack has not thereby been shown robust to their composition.

The population on which either order can be measured is narrow: 5 defended models appear on both standard leaderboards (Croce et al., 2021) and none on ImageNet. On it we reproduce the OODRobustBench relation of Section 2, at $r = +0.861$ over the 18 models with headroom (accuracies spanning only 0.184) and $r = +0.988$ over all 20; it is estimated across models while NC is a difference within one, so it still says nothing about which order was measured (Appendix H).

What to report. A composed robustness number needs three fields that are not currently printed: which operator is applied first, what the attacker knows about the corruption when it commits, and the strength the corruption is measured at. The first two name a cell of Figure 3; the third matters because the recovery term changes sign with it. Where both timings are plausible, $A \circ C$ bounds the adaptive attacker on 93% of cells and is the endpoint to quote.

Limitations. Every claim is made within a pair. The mechanism evidence is associational: no intervention moves boundary geometry at fixed displacement, and both distance estimates share the attack's gradients and are ℓ_2 while the attack is ℓ_∞. Each configuration is 500 images at one CIFAR-10 severity; AutoAttack covers only 5 models and 4 types of it; the fourth cell covers 5 differentiable types, one by a frozen residual, grouped after seeing the result. On ImageNet the mechanism chain and the recovery term's sign depend on the corruption strength (Appendices B, G.1, E).

REPRODUCIBILITY STATEMENT

The criteria that decide every claim in this paper were specified before the runs they judge, and Appendix A lists each one with its threshold and its outcome, including the four that came out against the prediction we had specified. One further run passed its criterion but moved two quantities the criterion did not cover, and both are reported in Appendix G.1 rather than repaired. We call these criteria pre-specified rather than pre-registered: they were written into the analysis scripts ahead of the runs, but those scripts were not deposited in a timestamped external archive, so a reader cannot verify from outside that a threshold was never revised afterwards. Two thresholds were introduced after inspecting accuracy ranges rather than before, the adversarial-headroom bound $A_{\mathrm{adv}} \geq 0.20$ and the corruption-headroom bound $A_{\mathrm{corr}} \geq 0.50$; Appendix C.1 states what motivated them, what each removes, and reports every affected quantity with and without them. The models are public checkpoints from RobustBench (Croce et al., 2021) and the corruption data is CIFAR-10-C (Hendrycks & Dietterich, 2019), both listed per model in Appendix I. The attack configurations, the 500-image evaluation subset and the seeds are fixed in the supplementary code, and every measured quantity in the text, every table and every figure is produced by a single script from the stored JSON outputs of the runs. The literal numbers left in the text are the design constants: the size of the evaluation subset, the two headroom bounds above, and the thresholds fixed in advance in Appendix A. Correlations computed on six or fifteen points are reported with $\sum d^2$ and the number of attainable coefficient values, following Appendix B.1.

ETHICS STATEMENT

This work evaluates existing publicly released models with existing published attacks and corruption benchmarks, on public image datasets with no personal or sensitive content. It introduces no new attack. The measurement it reports makes robustness evaluations of composed threats harder to overstate, which we take to be the direction that favours defenders: a reported number that does not name its composition order can be optimistic by 12 points.

AI USE STATEMENT

Large language models were used to assist with language editing, with writing and refactoring analysis and plotting code, and with searching for related work. They were not used to generate data, to formulate the hypotheses or the pre-specified criteria, or to decide what the results mean. Every retrieved reference was checked by the authors against the publisher or arXiv record before it was cited, the analysis scripts assert their outputs against the pre-specified values and abort on a mismatch, and every number, table and figure is regenerated by those scripts from the stored outputs of the runs rather than transcribed. The authors read the final text in full and take responsibility for the contents of the paper.

REFERENCES

Sravanti Addepalli, Samyak Jain, and Venkatesh Babu Radhakrishnan. Efficient and effective augmentation strategy for adversarial training. In *Advances in Neural Information Processing Systems*, 2022a.

Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, Shivangi Khare, and Venkatesh Babu Radhakrishnan. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, 2022b.

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*, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 80, pp. 284–293, 2018b.

Brian R. Bartoldson, James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. Adversarial robustness limits via scaling-law and human-alignment studies. In *International Conference on Machine Learning*, 2024.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

Lu Chen, Han Yang, Hu Wang, Yuxin Cao, Shaofeng Li, and Yuan Luo. Intriguing frequency interpretation of adversarial robustness for CNNs and ViTs. *arXiv preprint arXiv:2506.12875*, 2025.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 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 *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled Kullback-Leibler divergence loss. In *Advances in Neural Information Processing Systems*, 2024.

Xuelong Dai, Dong Wang, Xiuzhen Cheng, and Bin Xiao. Gradient-free adversarial purification with diffusion models. *arXiv preprint arXiv:2501.13336*, 2025.

Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. MMA training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library). `https://github.com/MadryLab/robustness`, 2019.

Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin D. Cubuk. Adversarial examples are a natural consequence of test error in noise. In *Proc. Int. Conf. Mach. Learn. (ICML)*, volume 97, pp. 2280–2289, 2019.

Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. Improving robustness using generated data. In *Advances in Neural Information Processing Systems*, 2021.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Lei Hsiung, Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Lifeng Huang, Chengying Gao, and Ning Liu. Erosion attack: Harnessing corruption to improve adversarial examples. *IEEE Trans. Image Process.*, 32:4828–4841, 2023. doi: 10.1109/TIP.2023.3251719.

Naveen Karunanayake, Ravin Gunawardena, Suranga Seneviratne, and Sanjay Chawla. Out-of-distribution data: An acquaintance of adversarial examples—a survey. *arXiv preprint arXiv:2404.05219*, 2024.

Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1012–1021, 2022.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations, Workshop Track*, 2017.

Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Lin Li, Yifei Wang, Chawin Sitawarin, and Michael Spratling. OODRobustBench: A benchmark and large-scale analysis of adversarial robustness under distribution shift. In *International Conference on Machine Learning*, 2024.

Guang Lin, Chao Li, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Adversarial training on purification (AToP): Advancing both robustness and generalization. *arXiv preprint arXiv:2401.16352*, 2024.

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*, 2018.

Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Shahbaz Khan, and Salman Khan. Towards evaluating the robustness of visual state space models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. In *AAAI Conference on Artificial Intelligence*, 2021.

Xiaofeng Mao, Yuefeng Chen, Rong Zhang, Hui Xue, Zhao Li, and Hang Su. EasyRobust: A comprehensive and easy-to-use toolkit for robust and generalized vision. *arXiv preprint arXiv:2503.16975*, 2025.

Tejaswini Medi, Julia Grabinski, and Margret Keuper. Towards class-wise robustness analysis. *arXiv preprint arXiv:2411.19853*, 2024.

Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *arXiv preprint arXiv:2210.07540*, 2022.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. PRIME: A few primitives can boost robustness to common corruptions. In *European Conference on Computer Vision*, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2574–2582, 2016.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022.

Victor Oei, Jenny Schmalfuss, Lukas Mehl, Madlen Bartsch, Shashank Agnihotri, Margret Keuper, Andreas Bulling, and Andrés Bruhn. RobustSpring: Benchmarking robustness to image corruptions for optical flow, scene flow and stereo. *arXiv preprint arXiv:2505.09368*, 2025.

Chao Pan and Xin Yao. FastAT benchmark: A comprehensive framework for fair evaluation of fast adversarial training methods. *arXiv preprint arXiv:2604.22853*, 2026.

Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 2022.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022.

Robin Narsingh Ranabhat, Longwei Wang, Amit Kumar Patel, and K. C. Santosh. Promoting shape bias in CNNs: Frequency-based and contrastive regularization for corruption robustness. *arXiv preprint arXiv:2509.11355*, 2025.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. In *Advances in Neural Information Processing Systems*, 2021.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 2020.

Giulio Rossolini. How worst-case are adversarial attacks? linking adversarial and perturbation robustness. *arXiv preprint arXiv:2601.14519*, 2026.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? In *Advances in Neural Information Processing Systems*, 2020.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022.

Georg Siedel, Rojan Regmi, Abhirami Anand, Weijia Shao, Silvia Vock, and Andrey Morozov. Stylized synthetic augmentation further improves corruption robustness. In *International Conference on Computer Vision Theory and Applications*, 2026.

Naman D. Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for ImageNet: Architectures, training and generalization across threat models. In *Advances in Neural Information Processing Systems*, 2023.

Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020.

Linyu Tang and Lei Zhang. Robust overfitting does matter: Test-time adversarial purification with FGSM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Puru Vaish, Shunxin Wang, and Nicola Strisciuglio. Fourier-basis functions to bridge augmentation gap: Rethinking frequency augmentation in image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Hanrui Wang, Ruoxi Sun, Cunjian Chen, Minhui Xue, Lay-Ki Soon, Shuo Wang, and Zhe Jin. Iterative window mean filter: Thwarting diffusion-based adversarial purification. *IEEE Transactions on Dependable and Secure Computing*, 2024a.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023.

Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, 2020.

Zhonghua Yi, Hao Shi, Qi Jiang, Yao Gao, Ze Wang, Yufan Zhang, Kailun Yang, and Kaiwei Wang. Benchmarking the robustness of optical flow estimation to corruptions. *arXiv preprint arXiv:2411.14865*, 2024.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, 2019.

Fuyuan Zhang, Qichen Wang, and Jianjun Zhao. AR2: Attention-guided repair for the robustness of CNNs against common corruptions. *arXiv preprint arXiv:2507.06332*, 2025a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Jiacheng Zhang, Feng Liu, Dawei Zhou, Jingfeng Zhang, and Tongliang Liu. Improving accuracy-robustness trade-off via pixel reweighted adversarial training. *arXiv preprint arXiv:2406.00685*, 2024a.

Jiancheng Zhang, Peiran Dong, Yongyong Chen, Yin-Ping Zhao, and Song Guo. Random sampling for diffusion-based adversarial purification. *arXiv preprint arXiv:2411.18956*, 2024b.

Jiaqi Zhang, Juntuo Wang, Zhixin Sun, John Zou, and Randall Balestriero. FastDINOv2: Frequency based curriculum learning improves robustness and training speed. In *Advances in Neural Information Processing Systems*, 2025b.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*, 2024c.

Wei Zhang, Xinyu Chang, Xiao Li, Yiming Zhu, and Xiaolin Hu. Defending against patch-based and texture-based adversarial attacks with spectral decomposition. *IEEE Transactions on Information Forensics and Security*, 2026.

Zhun Zhang, Yi Zeng, Qihe Liu, and Shijie Zhou. Towards a novel perspective on adversarial examples driven by frequency. *arXiv preprint arXiv:2404.10202*, 2024d.

A CRITERIA SPECIFIED IN ADVANCE AND THEIR OUTCOMES

The criteria below were written into the analysis scripts before the runs they judge. Four of them came out against the prediction specified in advance, and all four are reported in the main text: the composition gap, the frequency prediction, the strict form of the ImageNet-100 replication criterion, and the fourth cell.

Criterion	Threshold specified in advance	Outcome	Where
Order effect	median NC > 0, cluster-aware	$+12.40$ pp, 20 of 20, 15 of 15	Sec. 4
Composition gap	median $G > 3$ pp	$+1.39$ pp, not met	Sec. 3
Recovery term, per type	flag any type significantly > 0	0 of 15 flagged	Sec. 6
Displacement (H3)	$\rho > 0$ against $\lvert\mathrm{NC}\rvert$	$+0.625$ to $+0.901$, 18 of 18	Sec. 7
Boundary distance (H4a)	$\rho < 0$ against exposure	-0.989 to -0.907, 18 of 18	Sec. 7
Competing classes (H4b)	$\rho > 0$ against exposure	$+0.922$ down to $+0.618$, 17 of 18	Sec. 7
Competing classes on ImageNet-100	masking exception, report only	not comparable	Sec. 7
Curvature (H2)	$\rho > 0$ against extra drop	$+0.656$ to $+0.888$, 18 of 18	Sec. 7
Frequency overlap (H1)	$\rho < 0$ against the gap	$+0.064$ to $+0.596$ on all 18, sign reversed[‡]	Sec. 7
Replication, strict reading	recovery term < 0	not met on ImageNet-100	Sec. 6
Replication, specified text	not significantly > 0	met, 0 of 4 positive	Sec. 6
Retention rule	keep fraction $< 50\%$ leaves the main ordering	contrast, median 37%, 17 of 18	App. F.3
Specified before the follow-up round			
Adaptive third cell	closed fraction < 0.30, 0.30–0.80, > 0.80	0.309, middle band	Sec. 5
Expectation convergence	$\lvert\Delta\rvert \leq 0.10$ under a stronger attack	0.037, met	Sec. 5
Stronger attack	effect survives AutoAttack	$+25.8$ pp, 0 of 5 reversed	Sec. 4
Dose-response	NC rises with severity within a type	5 of 6 types	Sec. 7
Independent boundary estimate	$\rho < 0$ and $\lvert\rho\rvert \geq 0.7$ on all	-0.950 to -0.864	Sec. 7
Native ImageNet-1k	≥ 4 of 6 types inside the accuracy band	FLOOR on the selection sweep, 6 of 6 met on the full run[††]	App. G.1
Fourth cell, informed but early	closed fraction < 0.30, 0.30–0.80, > 0.80	$+0.894$, INFO-DOMINATED[§]	Sec. 5
Matched strength	effect survives, or it does not	$+9.83$ pp, 10 of 10, survives[¶]	App. G.1

[§]INFO-DOMINATED is the label the registered rule attaches to its top band and is used nowhere else. This criterion came out against us and the outcome is stated in Section 5: we withdrew the claim it tested. Its denominator is also too permissive on small gaps, which we report rather than repair after the fact (Appendix D).

[††]The 3-model selection sweep at 128 images reached the threshold at no severity, so it fell back to the mildest setting; the criterion is met on the full run at that setting. Appendix G.1 gives both.

[¶]Both outcomes were written down before this run, and the criterion applies only to the order effect. Two quantities it did not cover moved: the recovery term stops being negative and the mechanism links stop being unanimous. Appendix G.1 reports both.

[‡]The 18 per-model values specified in advance for H1 are $+0.064$, $+0.071$, $+0.104$, $+0.111$, $+0.114$, $+0.136$, $+0.139$, $+0.164$, $+0.179$, $+0.186$, $+0.209$, $+0.232$, $+0.243$, $+0.275$, $+0.307$, $+0.382$, $+0.389$, $+0.596$. They are listed here rather than in the table because every one of them is on the wrong side of the sign specified in advance, and a reader checking that claim should be able to see all of them.

The replication row records a discrepancy between the text specified in advance and its first implementation. The text read "negative, or at least not significantly positive" and the code implemented the strict inequality. Both readings are computed and both are reported, and the main text makes the weaker claim on ImageNet-100.

B How each quantity is computed

The attack. On CIFAR-10, PGD in ℓ_∞ at $\varepsilon = 8/255$, 50 steps, step size $\varepsilon/4$, one random start inside the ball, one restart, cross-entropy loss. The same configuration is used in both orders and on both sides of every paired comparison, so the order effect is a difference between two runs of one attack rather than between two attacks.

On native ImageNet-1k the budget and the corruption strength are different: $\varepsilon = 4/255$ with 20 steps at ImageNet-C severity 1, against $8/255$ with 50 steps at CIFAR-10-C severity 3. The smaller budget is the usual one for ImageNet-scale robust models and the severity is what the matching rule of Appendix G.1 selected. Both are held fixed across the two orders, so the paired comparison within ImageNet is unaffected, but the two datasets are measured under different settings and their effect sizes are not directly comparable to one another.

The corruption draw. Each corruption type is instantiated once per evaluation image from a fixed seed, and the two orders are run against that same instance: $C \circ A$ and $A \circ C$ differ only in whether the attack is computed before or after it, never in which draw of C they face. The fourth cell of Appendix D reuses the same seed for the same reason.

Feature displacement. $\|\varphi(C(x)) - \varphi(x)\|_2$ with φ the penultimate layer, taken as the input to the final linear classifier, averaged over the evaluation subset.

Curvature. The trace of the loss Hessian, estimated by finite differences along Rademacher directions, $\mathrm{tr}(H) \approx \mathbb{E}_v[(L(x + hv) - 2L(x) + L(x - hv))/h^2]$, with $h = 10^{-2}$ and 8 probe directions at a fixed seed. The curvature ratio is this quantity on the corrupted neighbourhood over its value on the clean one. The trace is not sign-constrained and the ratio is undefined when the clean value is near zero, so a configuration enters the curvature analysis only when both traces are positive and the clean one exceeds 10^{-3}; the excluded configurations are reported with the rest and are not dropped from any other quantity.

Competing classes. The number of logits within 10% of the gap between the top logit and the smallest, counted per input and averaged. The threshold is a fraction of the logit range rather than an absolute value, so it does not depend on the scale of a model's logits.

Frequency-band overlap. DCT coefficients aggregated into three radial bands and normalised to proportions of total energy. With r the radial index normalised so that the highest attainable frequency is 1, the bands are $r < 1/3$, $1/3 \le r < 2/3$ and $r \ge 2/3$, applied to the 32×32 DCT of each channel and summed over channels. The overlap is between the band profile of the corruption's effect on the input and that of the perturbation.

Partial correlations. The partial Spearman correlation of u and v given w is the Pearson correlation of the residuals of the rank vectors: each of $\mathrm{rank}(u)$, $\mathrm{rank}(v)$ is regressed linearly on $\mathrm{rank}(w)$ and the two residual vectors are correlated. It is computed within a model over the 15 corruption types, so each value rests on 15 points and Appendix B.1 applies to it as well.

Search-based boundary distance. The iterative DeepFool search over the five highest-scoring classes with at most 20 iterations, reporting the ℓ_2 norm of the resulting perturbation. It differs from Eq. (4) in removing the single-step linearisation, not in the norm.

B.1 Reporting correlations on few points

With n corruption types a Spearman coefficient takes a finite number of values. At $n = 6$ there are 36 attainable values spaced 0.0571 apart, so -0.9429 is the value at $\sum d^2 = 68$ against a maximum of 70, one adjacent transposition from perfect reversal. All four ImageNet-100 models report exactly -0.9429, and this is a grid collision rather than four independent measurements agreeing: their underlying orderings take three distinct forms on the margin side and three on the exposure side. One ordering differs in the middle: on the ResNet-50 the pair of defocus blur and Gaussian noise is reversed relative to the other three models, which follows from that model being an outlier on

Table 2: The decomposition of Eq. (3), as means over configurations with percentile bootstrap intervals. The first row includes the two standard models, which are bound by the ceiling of Section 4.

Population	n	Recovery (pp)	Exposure (pp)	NC (pp)
All configurations	300	-1.05 $[-2.20, +0.18]$	$+15.74$ $[+14.28, +17.24]$	$+14.69$
Adversarial headroom ($A_{\text{adv}} \geq 0.20$)	270	-3.11 $[-3.96, -2.31]$	$+17.31$ $[+15.78, +18.86]$	$+14.20$
Adversarial and corruption headroom	254	-1.73 $[-2.27, -1.21]$	$+15.45$ $[+14.12, +16.81]$	$+13.72$

Gaussian noise, at a corrupted accuracy of 0.180 against 0.588 to 0.846 elsewhere. What replicates is the direction and the identity of the two extremes, contrast and fog, which rank first on all four. We therefore report $\sum d^2$ and the number of attainable values alongside any coefficient computed on six points, and we do not treat a repeated value as corroboration.

The first link of the chain is reported against two endpoints. The endpoint specified in advance was $|\text{NC}|$, giving $+0.625$ to $+0.901$ across the 18 models; measured against the exposure term, which is the quantity the rest of Section 7 explains, the same displacement gives $+0.582$ to $+0.932$. Both are positive on all 18 models and the paper's ranking of candidate mechanisms uses the exposure endpoint for all three candidates so that they are comparable, as shown in Figure 4.

Finally, one identity is worth recording because it explains why the fraction of negative margins and the corrupted accuracy cannot be treated as separate evidence: over the 75 model-corruption pairs with adversarial headroom, the fraction of negative margins equals $1 - A_{\text{corr}}$ with $r = 0.9968$.

C Checks on the measurement

Four checks ask whether the order effect could be produced by how the measurement was set up rather than by the composition: the two population bounds, the choice of corruption types, the single attack budget, and the independence baseline.

C.1 Thresholds introduced after seeing data

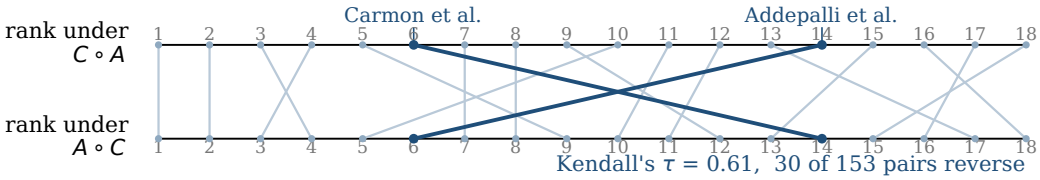

Figure 5: The 18 models ranked by mean accuracy over the 15 types under each order, rank 1 most accurate. Every crossing is a reversed pair; the highlighted lines move 8 places.

Two bounds define the population of Section 3 and both were set after inspecting accuracy ranges. The recovery term $A_{C \circ A} - A_{\text{adv}}$ is bounded above by $1 - A_{\text{adv}}$ and below by $-A_{\text{adv}}$, so a model with adversarial accuracy near zero admits a term of up to $+95$ pp for arithmetic reasons, and a configuration whose corrupted accuracy is near zero admits one that can only be negative. The first case is visible on our two standard models: grouped by how the models were trained, the mean recovery term is $+17.52$ pp on the standard-trained ones against -4.10 pp on the adversarially trained ones and -5.32 pp on those that add augmentation, which is the ceiling and not a property of the corruption. Per configuration the clearest case is JPEG compression on ResNet-18, whose adversarial accuracy is 4.2% to 6.2%: its recovery term reaches $+62.8$ pp, and that one model moves the mean for that corruption type from $+3.9$ to $+6.9$ pp. The second appeared on ImageNet-100, where contrast leaves a corrupted accuracy of 0.066 against a chance level of 0.01, driving the recovery term to approximately $-A_{\text{adv}}$. Neither case measures the interaction the paper is about.

Table 3: The four arrangements. The middle two hold the timing fixed and vary what the attacker knows. The last column is the fraction of the gap each closes, on the same 25 cells (Section 5).

Arrangement	Attack computed	Knowledge of the corruption	Closes
$C \circ A$	before the corruption	none	—
Expectation	before the corruption	the distribution	38.0%
Informed, early	before the corruption	the realised instance	89.4%
$A \circ C$	after the corruption	the realised instance	by definition

$A_{\text{adv}} \geq 0.20$ removes the 2 standard-trained models and leaves 270 configurations. $A_{\text{corr}} \geq 0.50$ removes a further 16 of those, leaving 254; all 16 belong to a single corruption type, contrast, and their corrupted accuracies run from 0.290 to 0.492 against a smallest retained value of 0.512. Every affected quantity appears with and without the second bound: the recovery term is -3.11 pp $[-3.96, -2.31]$ without it and -1.73 pp $[-2.27, -1.21]$ with it, and the exposure term is $+17.31$ pp and $+15.45$ pp respectively. The sign and the exclusion of zero are unchanged and the magnitude falls, which is what a floor component in the unbounded estimate would produce. A sensitivity sweep over $A_{\text{adv}} \in \{0.05, 0.20, 0.40\}$ and $A_{\text{corr}} \in \{0.0, 0.5, 0.6\}$ is in the supplementary JSON.

C.2 How the corruption types were chosen

An earlier round measured 4 representative corruption types on 7 models and found $+14.10$ pp over 28 configurations. Extending to all 15 types gives $+12.40$ pp, a difference of 1.70 pp. The decomposition shows the same stability: on the population of Section 3 the recovery term is -1.56 pp on the 4 original types and -1.80 pp on the 11 added ones, so the choice of representatives drove neither result. The comparison is available because the second round was run at $n = 500$ on all 15 types rather than extending the first.

C.3 The effect under a per-model calibrated budget

A single ε applied to models of different robustness places them at different points on their own accuracy curves, which could in principle manufacture the order effect. We therefore calibrated ε per model so that each model reaches the same adversarial accuracy on clean inputs, a target of 0.40, and re-ran both orders. The calibration succeeded on all 26 models measured here, with achieved adversarial accuracies between 0.371 and 0.414 and budgets spanning $0.83/255$ to $13.88/255$, a range of a factor of seventeen across the model set.

Under that calibration NC is positive in 104 of 104 model-corruption configurations, with a median of $+24.2$ pp, which is larger than the $+12.40$ pp measured at the fixed budget. The four corruption types measured in this round give per-type medians of $+24.0$, $+32.4$, $+19.3$ and $+16.0$ pp for contrast, fog, Gaussian noise and defocus blur, all positive on all 26 models. This round uses 256 images and 20 attack steps rather than the 500 images and 50 steps of the main measurement, so it is a check on the sign and the order of magnitude rather than a replacement for it.

C.4 The additivity baseline

To ask whether the two vulnerabilities compound beyond independence, write $R_\bullet = 1 - A_\bullet$ for error rates and compare the composed error to the error that independent failure events would produce:

$$R_{\text{indep}} = 1 - (1 - R_{\text{corr}})(1 - R_{\text{adv}}), \qquad G = R_{\text{comp}} - R_{\text{indep}}. \tag{5}$$

We had specified in advance the expectation that G would be clearly positive, at a threshold of 3 pp. It is $+1.39$ pp for the $A \circ C$ order: the composed damage is roughly what independence predicts, the specified threshold is not met, and nothing in the paper is built on it.

D The fourth cell

What it is. Four arrangements differ in what the attacker knows and when it acts. In $C \circ A$ the attacker knows nothing about the corruption and commits before it. In the expectation cell it knows

the distribution but not the draw, and still commits before it. In $A \circ C$ it sees the realised corrupted image and has no uncertainty left. The fourth cell gives the attacker the realised corruption C_s and lets it differentiate through C_s, but requires the perturbation to be committed to the clean image: it maximises the loss of $f(C_s(x+\delta))$ over $\|\delta\|_\infty \le \varepsilon$. Its information equals that of $A \circ C$, so the accuracy it fails to reach is attributable to the timing rather than to knowledge.

How it is run. The cell is measured on 5 models and 5 corruption types: brightness, contrast, defocus blur, gaussian noise, shot noise. Four of them have an exact differentiable implementation and shot noise uses the frozen-residual approximation described below. They were chosen for that property and not for their effect sizes; contrast, the type with the largest effect, is among them. The grouping used in the main text is by how the operator acts on a perturbation: Gaussian noise and shot noise translate the input and are the additive group, contrast and defocus blur contract it, and brightness is reported with the rest but belongs to neither, being the type whose gaps are too small for the ratio. Each differentiable operator is checked against the corruption used everywhere else in the paper and agrees to 10^{-7}, both per image and in batch, where the random draw depends on position in the batch. Under that grouping the informed-but-early attacker recovers 95.2% of the gap on the additive types against 42.1% on contrast and defocus blur, with no overlap between the two ranges ($+0.874$ to $+1.064$ against $+0.390$ to $+0.580$) and the contracting group carrying the larger order effect, 14.5 against 11.9 pp. The grouping was read off the results rather than specified in advance, and the split rests on two types against two. The corruption instance in this cell uses the same seed as the one the $A \circ C$ arm attacks, so the two attackers face the same realised corruption. The attack is the main configuration, PGD at $\varepsilon = 8/255$ with 50 steps and step size $\varepsilon/4$, on the same 500 images.

The expectation cell. The third cell optimises the expected loss over the corruption distribution rather than a single draw. Raising it from 20 steps with 5 samples to 50 steps with 10 moves the closed fraction by 0.037, so the part it leaves open is not an artifact of an under-optimised expectation. That cell is measured on 5 models and all 15 corruption types, a wider set than the fourth cell. The comparison quoted in the main text restricts both to the 25 cells they share, on which the expectation attacker closes 38.0%; over all 15 types it closes 31%, and the difference between the two is the corruption types each covers, not the attacker.

The quantity. For a cell with accuracies $A_{C\circ A}$ and $A_{A\circ C}$, an intermediate arrangement with accuracy $A_\bullet$ closes the fraction

$$q \;=\; \frac{A_{C\circ A} - A_\bullet}{A_{C\circ A} - A_{A\circ C}}. \tag{6}$$

$q = 0$ means the arrangement gives the same accuracy as $C \circ A$ and closes nothing, $q = 1$ means it reaches $A \circ C$, and $q > 1$ means it attacks more effectively than $A \circ C$ does. The denominator is the order effect itself, so q is unstable where that effect is small, which is the defect described next.

A defect in our own criterion. The registered rule admits any cell whose gap exceeds 2 pp, and the closed fraction has that gap in its denominator. On 4 brightness cells the gap is about 3.0 pp and the ratio comes out above 3, which is an artifact of dividing by a small number rather than a fraction closed. We report the verdict under the rule as registered, $+0.894$ and INFO-DOMINATED, and note that dropping brightness gives $+0.703$ instead. The threshold was too permissive and we say so rather than change it after the fact; neither reading changes the conclusion that most of the gap is attacker information.

One approximation, and which way it cuts. Shot noise is a Poisson draw whose intensity depends on the input, so it cannot be differentiated through directly. We freeze the draw on the clean image and apply the resulting residual, which is exact at the clean point and a first-order approximation elsewhere. This makes the operator easier to attack than the true one, which pushes its closed fraction up, towards the conclusion that the gap is information rather than order. The approximation therefore works against the reading we would have preferred, not for it.

E THE RECOVERY TERM AGAINST CORRUPTION STRENGTH

Section 6 measures the recovery term at one severity. The dose-response runs of Section 7 cover severities 1, 3 and 5 on 5 models and 6 corruption types, and reading the recovery term off them instead of the order effect gives its dependence on strength directly. Averaging within a model first, the median across models is $+0.17$ pp at severity 1, -2.76 at severity 3 and -4.65 at severity 5, with median corrupted accuracies of 0.848, 0.803 and 0.746. The decrease holds on 5 of 5 models individually and on four of the 6 types; brightness and JPEG compression are flat until severity 5, and both are types whose corrupted accuracy barely moves. At severity 1 the term is not distinguishable from zero, at 3 of 5 models positive with a sign test at $p = 1$.

Those medians are over the cells meeting the corruption-headroom bound of Section 3, which is what the bound is for: raising severity drives some cells below it, and there the recovery term is bounded above by a number that shrinks with the corrupted accuracy, so a decrease would be produced by the ceiling of Appendix C.1 whether or not anything happened to the perturbation. The bound retains 30 of 30 configurations at severity 1, 25 at severity 3 and 20 at severity 5. Over all 30 regardless of the bound the medians are $+0.17$, -6.50 and -15.00 pp, decreasing on 5 of 5 models. The direction, the unanimity and the severity-1 null are the same either way; the levels at severities 3 and 5 are not, and the difference is concentrated in contrast, which at severity 3 gives a recovery term of -25.64 pp at a corrupted accuracy of 0.415. The claim in Section 6 is the direction, and it is quoted at the bounded levels.

No criterion was specified for this reading in advance; the rule, average within a model and then report the median and a sign test across models, is the one used throughout the paper, and the monotonicity is a count. Two things follow. The negative sign reported in Section 6 is a property of corruptions strong enough to cost accuracy and not of degradation in general, which is the form the claim is stated in. And the ImageNet result of Appendix G.1, where the term stops being negative once the strength is matched down, is what this dependence predicts rather than a conflict with it. The two datasets are not placed on a common strength axis: corrupted accuracy is not comparable across them (Appendix B), and the statement is made within each dataset separately.

F THE MECHANISM IN DETAIL

This section holds the per-model values behind Section 7, the one corruption type that behaves differently from the rest, the partial mediator at the strong end of the sequence, and the scope of the schematic.

F.1 SEPARATING THE TWO BOUNDARY QUANTITIES

The 18 partial correlations between curvature and the exposure term, holding the margin distance fixed, are -0.335, -0.258, -0.247, -0.233, -0.232, -0.113, -0.045, $+0.034$, $+0.081$, $+0.087$, $+0.152$, $+0.179$, $+0.180$, $+0.249$, $+0.351$, $+0.379$, $+0.508$, $+0.918$.

Controlling for the number of competing classes, the margin distance ratio keeps partial correlations with the exposure term of -0.607, -0.938, -0.972, -0.963, -0.730, -0.903, -0.860, -0.828, -0.882, -0.729, -0.847, -0.821, -0.922, -0.589, -0.940, -0.891, -0.905, -0.581. Controlling for distance, the number of competing classes keeps $+0.341$, -0.205, -0.409, -0.647, $+0.143$, $+0.390$, -0.300, -0.209, $+0.484$, $+0.023$, $+0.043$, $+0.042$, -0.369, $+0.161$, -0.358, $+0.266$, -0.407, -0.023. Distance is the only quantity whose sign is stable across all 18 models. The sign changes for competing classes and for curvature on the same two models, which are also the only augmented model and the only ResNet-18 backbone in the set, and model set cannot separate backbone from training recipe.

F.2 THE RELATION ON EVERY MODEL WITH HEADROOM

Figure 6 plots the exposure term against the margin distance ratio for every model with adversarial headroom, with no model set aside. The relation is negative on all 18 of them, at -0.989 to -0.907.

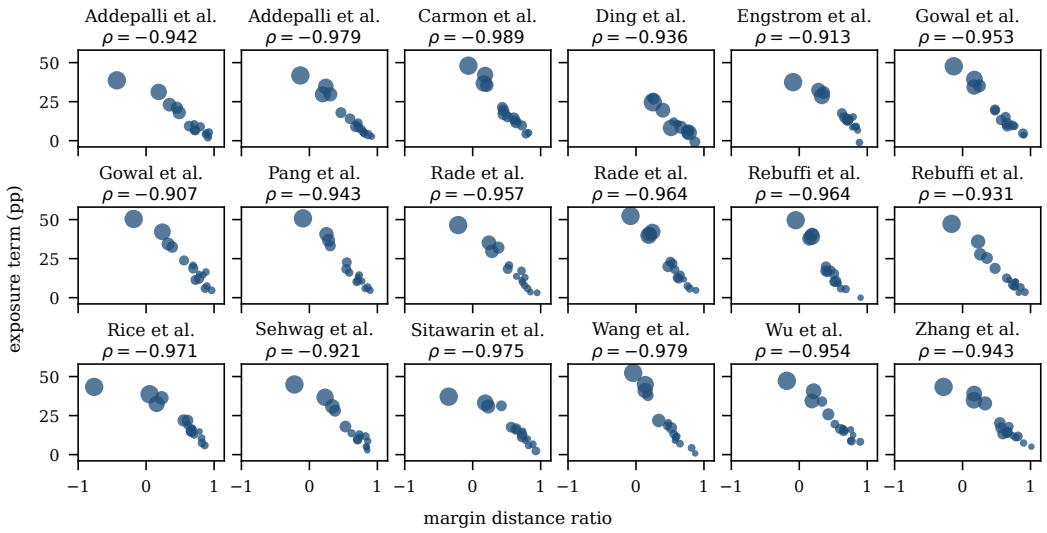

Figure 6: The exposure term against the normalised margin distance ratio for all 18 models with adversarial headroom, one point per corruption type. Marker size is the increase in the number of competing classes.

F.3 Contrast as a boundary case

Contrast is the only corruption type whose median margin changes sign, and it does so on 17 of the 18 models. A median of 63% of its samples sit on the wrong side of the boundary after the corruption, against per-type medians of 16% to 36% for every other type. A sign change means the semantics of the distance shifts from how far the attack must travel to how far the input has already crossed, so the two regimes are not on one axis. What removes contrast from the main ordering is the registered retention rule: the corruption alone leaves a median of 37% of the samples correctly classified (26% to 71%), below the 50% bar fixed in advance on 17 of 18 models, so the conditional quantities are computed on a small and self-selected remainder. Conditioning on that remainder does not systematically lower the exposure term, which is why we set the type aside rather than report a corrected value for it: the paired change has median $+1.5$ pp and is negative on only 6 of 18 models. Contrast does not fall out of the sequence: removing the sign-flipped type leaves the margin distance correlation at -0.987 to -0.886 across the 18 models.

One further observation is post hoc and labelled as such. The fraction of negative margins after the corruption is by construction $1 - A_{\text{corr}}$, so its agreement with how often the corruption alone misclassifies is an identity rather than evidence, and we do not count it. What is not an identity is that the samples the corruption leaves correct are also easier to attack: that correlation is positive on 18 of 18 models, but with a spread of $+0.47$ to $+0.97$ rather than the narrow range the earlier five-model round suggested. It is consistent with the chain, in which the corruption moves the whole distribution toward the boundary, but it was not specified in advance and it does not hold tightly.

F.4 Dynamic range as a partial mediator

Three operators separate the factors that make contrast and fog the strong end of the sequence. Contrast compresses dynamic range without introducing new frequency structure, brightness shifts the range without compressing it, and fog does both. Contrast reproduces the effect of fog, at $+47.4\%$ of the normalised gap averaged over the 18 models against $+49.8\%$ for fog, and brightness does not, at -8.3%. Renormalising the attack budget by the post-corruption contrast leaves $+26.4\%$ for fog and $+26.3\%$ for contrast, so 45% to 60% of the fog effect and 32% to 67% of the contrast effect survive the renormalisation. Dynamic range compression is therefore a partial mediator, and about half of the effect at the strong end is something else.

F.5 What the schematic does and does not assert

Figure 4(a) states the account of Section 7 in one drawing. It is a schematic and carries no data: the measurements behind it are the correlations of Section 7 and the controls of Appendix F.1, and the boundary distance in them is the first-order estimate of Eq. (4) rather than a searched distance. Nothing in the drawing is to scale, and the single boundary it shows stands for a decision region whose geometry we measure only through the two scalars of Appendix B.

G The ImageNet measurements

The native 1000-way measurement is the one the main text relies on. The masked 100-class replication came first and is retained as the earlier reading.

G.1 The native ImageNet-1k measurement

Why a matching step was needed. The corruption-headroom bound $A_{\mathrm{corr}} \geq 0.50$ of Section 3 is an absolute one. On CIFAR-10, where clean accuracy is near 0.87, it asks a corruption to leave about 57% of the clean accuracy standing. Applied unchanged to the ImageNet-1k robust models we swept, whose clean accuracies run from 0.59 to 0.67, the same number asks for 74% to 85%, which no meaningful corruption strength satisfies. The bound is not equally strong at the two accuracy levels, and the fix is to choose the corruption strength rather than to move the bound.

The matching rule. Before running anything on the native task we swept official ImageNet-C severities 1 to 5 on 128 images and 3 models, and asked for the smallest severity at which at least 4 of the 6 corruption types leave at least 80% of the models inside the band $[0.50, 0.90]$. No severity satisfied it on that sweep, the best being severity 2 with 2 types, so the sweep returned its fallback of the mildest setting and severity 1 is what the full run uses. On the full run of 10 models at 500 images that criterion is met, on 6 of 6 types and 53 of 60 cells. The rule was fixed before the sweep, it reads only A_{clean} and A_{corr}, and it never sees NC or either composed accuracy, so the choice of strength is independent of the quantity being claimed.

The model set. 10 models: the three ImageNet-Linf entries used in the earlier 100-class replication plus every entry in the largest architecture tier of the RobustBench ImageNet-Linf leaderboard, which is 5 architectures from 7 independent groups. The tier was named before downloading and every member of it was run; none was dropped after seeing a result. Table 4 lists all 10 of them.

Table 4: The 10 native ImageNet-1k models. Accuracies are over the same 500 evaluation images with no logit masking; the corruption strength is official ImageNet-C severity 1, fixed by the matching rule before any composed accuracy was computed. The last column is the median order effect within the model.

Model	Architecture	A_{clean}	A_{adv}	Median NC (pp)
`Amini2024MeanSparse_ConvNeXt-L`	ConvNeXt-L	0.778	0.606	$+12.20$
`Amini2024MeanSparse_Swin-L`	Swin-L	0.796	0.638	$+14.10$
`Liu2023Comprehensive_ConvNeXt-L`	ConvNeXt-L	0.790	0.588	$+15.10$
`Liu2023Comprehensive_Swin-L`	Swin-L	0.802	0.602	$+15.40$
`Mo2022When_ViT-B`	ViT-B	0.694	0.348	$+14.30$
`RodriguezMunoz2024Characterizing_Swin-L`	Swin-L	0.812	0.556	$+24.90$
`Salman2020Do_R50`	ResNet-50	0.654	0.378	$+16.00$
`Singh2023Revisiting_ConvNeXt-L-ConvStem`	ConvNeXt-L	0.780	0.590	$+14.40$
`Singh2023Revisiting_ConvNeXt-T-ConvStem`	ConvNeXt-T+ConvStem	0.738	0.518	$+13.10$
`Xu2024MIMIR_Swin-L`	Swin-L	0.800	0.608	$+17.00$

Result. The bounds of Section 3 are applied unchanged and the corruption bound is reported as a sensitivity condition rather than used to select cells, exactly as on CIFAR-10, so the medians quoted in the main text are over all 6 types on every model. All 10 models clear $A_{\mathrm{adv}} \geq 0.20$, with clean accuracies of 0.654 to 0.812 and adversarial accuracies of 0.348 to 0.638; 7 of them clear the corruption bound on all 6 types and the weakest clears it on 2. The order effect is $+15.50$ pp

at the median, positive on all 10 models, ranging from $+12.37$ to $+21.43$ pp, with a model-level bootstrap interval of $[+14.22, +17.25]$ pp and a sign test at $p = 0.00195$. The three mechanism links reproduce on every model: displacement $+0.829$ to $+1.000$, margin distance ratio -1.000 to -0.829, competing classes $+0.771$ to $+1.000$. The recovery term is -2.80 pp at the median, negative on 9 of 10 models and positive on one at $+2.23$ pp. That is the one quantity that does not survive a change in how the corruption strength is chosen: the next paragraph reruns the same measurement at matched strengths and it turns positive there, so what the ImageNet data support is a strength-dependent sign rather than a fixed one.

A second strength convention, and what changes under it. Severity 1 is a strength chosen so that at least four types land inside the accuracy band. A stricter convention is available: choose a per-type strength so that every type lands inside the band. We extended the corruption operators to a continuous strength t that interpolates between the official constant tables and reduces to the integer severities exactly, and ran the same selection rule on the same 128 images. It reads only A_{clean} and A_{corr}, as before. All 6 types are then matchable, none at a degenerate strength, and the median corrupted accuracy rises from 0.681 to 0.757. The strengths are not uniformly weaker: two types are matched above severity 1 and one exactly at it. We reran the full measurement on the same models, types and images at those strengths.

Three things follow, and only the first was specified in advance. *(i)* The order effect survives: $+9.83$ pp at the median, positive on all 10 models, $+7.60$ to $+13.13$ pp, bootstrap interval $[+8.75, +10.67]$ and sign test $p = 0.00195$, with the same verdict under clustering by group and by architecture. It is smaller than at severity 1, by -6.43 pp at the median, which is expected: the corruption is weaker on the types that were adjusted most. *(ii)* The recovery term does not keep its sign. At matched strength it is $+0.97$ pp and no longer negative on 9 of 10 models; paired within model it moves by $+3.60$ pp and in the same direction on all 10. Its change tracks the change in corrupted accuracy across the 60 cells at a rank correlation of $+0.82$, and the three types whose strength was reduced most are the three that account for the move. On ImageNet the sign of this term is therefore a function of corruption strength rather than a stable property. Its magnitude here, about one point, is the same size as the rerun jitter reported below, which is why we describe it as no longer negative and not as a net recovery. *(iii)* The unanimity of the mechanism links is also strength-dependent: displacement against the exposure term holds on 8 of 10 models rather than all, and the margin distance ratio on 9. The chain claim on ImageNet therefore holds at severity 1 and holds for most but not all models at matched strength.

One type, Gaussian noise, is matched at exactly $t = 1$, which is severity 1. It is therefore the same configuration run twice, and it gives a rerun jitter on the order effect of at most 1.00 pp per model. That number is why point (ii) is stated as a change of sign and not as a measured amount of recovery. The criteria, the two outcomes written for them and the checks on them are in the verdict record released with the code.

One link that does not carry across. On CIFAR-10 and on the masked 100-class replication the two corruption types with the lowest margin distance ratio are contrast and fog, which is what motivated reading dynamic range compression as the shared cause at that end. On the native task, over the same six corruption types, the two lowest are defocus blur and fog on eight of the 10 models and contrast and fog on the remaining two. Fog is common to all three datasets; contrast is not. The dynamic-range reading of Appendix F.4 is therefore a CIFAR-10 statement, and Table 5 reports the native ordering rather than the one that matches.

What this supersedes. The 100-class replication of Appendix G.2 masked the logits of the other 900 classes, which changes which class competes and therefore the margin distance itself. The native measurement does not, so it is the stronger form of the same claim and the masked one is retained only as the earlier reading.

G.2 The ImageNet-100 extension

Table 5 sets the three links of the chain side by side on the three datasets, so that what replicates and what does not can be read off one row at a time.

The four ImageNet-100 recovery terms are $+1.70$, -3.07, $+0.67$ and -2.90 pp, all within 3.1 pp of zero against an independent-samples bound of about 3.2 pp on the paired standard error at 500 images. None is significantly positive, which is what the criterion specified in advance asked, and none is significantly negative, so the stronger reading available on CIFAR-10 is not available here. Two differences between the datasets are relevant. The corruption bound removes more configurations on ImageNet-100, where contrast leaves a corrupted accuracy of 0.066 against a chance level of 0.01; and the four models are ImageNet-1k models evaluated through a logit mask, so their adversarial accuracies are higher than a model trained on the 100 classes would give.

Table 5: The chain measured on all three datasets. CIFAR-10 entries are the 18 models with adversarial headroom over 15 corruption types; ImageNet-100 entries are 4 models from 4 architecture families and 4 independent groups over 6 corruption types; native ImageNet-1k entries are 10 models over the same 6 types with no logit masking, so the competing-class link is measurable there and not on ImageNet-100. The ImageNet-1k column is at official severity 1; the recovery term there is strength-dependent and Appendix G.1 reports it under a second strength convention as well. Correlations are Spearman coefficients computed within a model across corruption types. The ImageNet-100 margin distance ratio is the same value on all four models, which a rank sum of $\sum d^2 = 68$ of 70 forces at this number of types.

Link	CIFAR-10 18 models, 15 types	ImageNet-100 4 models, 6 types	ImageNet-1k 10 models, 6 types
Feature displacement vs. exposure	$+0.582$ to $+0.932$	$+0.771$ to $+0.943$	$+0.829$ to $+1.000$
Margin distance ratio vs. exposure	-0.989 to -0.907	-0.943 (all four)	-1.000 to -0.829
Competing classes vs. exposure	-0.047 to $+0.922$	not comparable (masked)	$+0.771$ to $+1.000$
Recovery term (pp)	-1.73 $[-2.27, -1.21]$	$+1.70$, -3.07, $+0.67$, -2.90	-2.80 (9 of 10 negative)
Exposure share of NC	106%	186% to 368%	90% to 137%
Two lowest margin distance ratios	contrast, fog	contrast, fog (4 of 4)	defocus blur, fog (8 of 10)

H THE PUBLISHED MODEL POPULATION

The two standard leaderboards (Croce et al., 2021) carry 144 adversarial and 41 corruption entries. 7 models appear on both and 5 once undefended baselines are removed; on ImageNet the only model on both is undefended, and the five on CIFAR come from one research group. Two claims separate here and they are not the same. Absence from a board is evidence about who submitted what rather than about the model, so it is a statement about evaluation practice; that co-occurrence of the two kinds of robustness is rare is a claim about models, and it is supported here by our own measurements of 25 of them on both axes rather than by the count. One further thing follows, and it is not load-bearing for the claims of this paper. The budgets: the five most corruption-robust models we measured have adversarial accuracies between 0.000 and 0.160, and reaching 0.40 would take budgets of $1.13/255$ to $3.96/255$, against $9.1/255$ to $13.9/255$ for the adversarially trained ones. Whether the subpopulation is narrow because of what is achievable or because of what has been trained is a question about the training frontier, which counting boards does not settle and which this paper does not investigate.

I FULL PER-CONFIGURATION RESULTS

Table 6 lists the 20 CIFAR-10 models with their sources and their clean, corrupted and adversarial accuracies, and Tables 7 and 8 give NC for every one of the 300 model-corruption pairs behind the medians in the main text.

Table 6: The 20 CIFAR-10 models. Corrupted and adversarial accuracies are averaged over the 15 corruption types at severity 3 and over the same 500 evaluation images. The last column is the median order effect within the model.

Model	Source	A_{clean}	A_{corr}	A_{adv}	Median NC (pp)
`Addepalli2021Towards_RN18`	Addepalli et al. (2022b)	0.804	0.704	0.539	+8.00
`Addepalli2022Efficient_RN18`	Addepalli et al. (2022a)	0.850	0.761	0.537	+9.40
`Carmon2019Unlabeled`	Carmon et al. (2019)	0.888	0.791	0.597	+16.20
`Ding2020MMA`	Ding et al. (2020)	0.824	0.757	0.515	+11.00
`Engstrom2019Robustness`	Engstrom et al. (2019)	0.852	0.777	0.494	+14.80
`Gowal2020Uncovering_28_10_extra`	Gowal et al. (2020)	0.886	0.778	0.626	+11.40
`Gowal2021Improving_R18_ddpm_100m`	Gowal et al. (2021)	0.860	0.775	0.612	+12.60
`Pang2022Robustness_WRN28_10`	Pang et al. (2022)	0.878	0.793	0.636	+11.00
`Rade2021Helper_R18_ddpm`	Rade & Moosavi-Dezfooli (2022)	0.854	0.759	0.589	+10.40
`Rade2021Helper_extra`	Rade & Moosavi-Dezfooli (2022)	0.910	0.805	0.620	+14.80
`Rebuffi2021Fixing_70_16_cutmix_extra`	Rebuffi et al. (2021)	0.912	0.807	0.667	+14.60
`Rebuffi2021Fixing_R18_ddpm`	Rebuffi et al. (2021)	0.826	0.739	0.585	+8.40
`Rice2020Overfitting`	Rice et al. (2020)	0.858	0.728	0.550	+11.00
`Sehwag2021Proxy_R18`	Sehwag et al. (2022)	0.838	0.746	0.576	+11.20
`Sitawarin2020Improving`	Sitawarin et al. (2020)	0.862	0.759	0.500	+17.80
`Standard`	standard training (WRN-28-10)	0.948	0.705	0.000	+4.80
`Wang2023Better_WRN-28-10`	Wang et al. (2023)	0.932	0.823	0.678	+14.20
`Wu2020Adversarial`	Wu et al. (2020)	0.844	0.751	0.586	+11.00
`Zhang2019Theoretically`	Zhang et al. (2019)	0.844	0.748	0.527	+13.40
`resnet18`	standard training (ResNet-18)	0.942	0.705	0.051	+23.60
ImageNet-100 replication, six corruption types					
ResNet-50	Salman et al. (2020)	0.832	–	0.570	+6.50
XCiT-S12	Debenedetti et al. (2023)	0.878	–	0.610	+13.00
ConvNeXt-T+ConvStem	Singh et al. (2023)	0.882	–	0.684	+6.60
ViT-B	Mo et al. (2022)	0.870	–	0.552	+15.05

Table 7: The order effect NC in percentage points for every model and corruption type (models 1 of 2, in alphabetical order). Positive means that degrading before attacking costs accuracy.

Corruption	Addepal	Addepal	Carmon	Ding	Engstro	Gowal	Gowal	Pang	Rade	Rade
brightness	+3.6	+1.8	+2.4	+1.4	+3.0	+2.0	+1.6	+0.8	+1.4	+1.6
contrast	+15.8	+23.0	+26.8	+29.6	+30.4	+19.2	+23.4	+24.8	+21.8	+29.6
defocus blur	+5.6	+9.4	+17.0	+11.0	+14.2	+10.6	+10.0	+8.4	+7.8	+12.8
elastic transform	+3.4	+6.2	+7.0	+7.2	+7.8	+6.4	+4.4	+4.2	+4.0	+4.4
fog	+22.0	+28.2	+38.0	+22.2	+33.2	+29.4	+32.0	+32.4	+26.4	+36.6
frost	+12.6	+12.6	+18.8	+10.8	+18.2	+14.8	+16.2	+11.8	+10.4	+18.2
gaussian noise	+10.0	+9.8	+16.2	+9.2	+19.0	+13.0	+12.6	+11.2	+16.0	+17.2
glass blur	+16.2	+22.4	+28.2	+24.4	+32.2	+25.4	+23.0	+24.0	+23.2	+27.6
impulse noise	+7.2	+9.2	+14.2	+14.0	+14.8	+10.4	+14.4	+11.0	+15.0	+14.8
jpeg compression	+5.2	+5.2	+10.6	+8.6	+9.8	+7.4	+5.4	+6.8	+6.2	+6.0
motion blur	+16.8	+19.6	+26.6	+27.6	+31.0	+22.8	+21.4	+18.8	+21.2	+27.8
pixelate	+8.0	+11.4	+20.4	+15.4	+18.8	+17.4	+15.8	+12.2	+10.0	+17.8
shot noise	+5.8	+8.6	+14.0	+8.8	+12.0	+11.4	+11.2	+8.2	+13.6	+11.6
snow	+8.2	+8.6	+13.0	+9.0	+13.4	+8.0	+11.0	+8.2	+9.6	+8.4
zoom blur	+5.6	+8.4	+14.4	+11.0	+14.6	+10.2	+11.2	+9.2	+10.2	+11.0

Table 8: The order effect NC in percentage points for every model and corruption type (models 2 of 2, in alphabetical order). Positive means that degrading before attacking costs accuracy.

Corruption	Rebuffi	Rebuffi	Rice	Sehwag	Sitawar	Standar	Wang	Wu	Zhang	resnet18
brightness	+1.4	+1.2	+2.2	+3.6	+3.8	+0.2	+0.0	+3.4	+2.4	+2.2
contrast	+24.0	+22.8	+11.4	+22.2	+19.6	+0.0	+24.2	+21.6	+18.4	+21.0
defocus blur	+10.2	+7.8	+10.8	+10.6	+14.4	+4.8	+10.6	+8.8	+12.6	+43.2
elastic transform	+6.8	+4.4	+6.2	+7.2	+9.8	+0.0	+6.0	+4.4	+6.2	+12.4
fog	+29.0	+27.4	+22.8	+29.2	+32.0	+0.4	+34.8	+30.2	+32.0	+17.2
frost	+14.6	+10.6	+14.0	+13.8	+19.6	+0.0	+18.0	+15.8	+17.4	+8.0
gaussian noise	+15.4	+9.0	+11.0	+12.4	+21.0	+16.4	+14.2	+11.4	+17.6	+22.8
glass blur	+28.6	+19.6	+20.4	+23.0	+31.6	+19.8	+26.8	+22.6	+26.0	+23.6
impulse noise	+13.6	+7.4	+11.0	+10.6	+19.4	+8.0	+14.2	+11.0	+11.6	+28.8
jpeg compression	+8.0	+3.8	+5.6	+6.0	+9.2	+59.2	+8.2	+7.8	+10.4	+66.4
motion blur	+25.2	+15.0	+17.0	+21.6	+27.4	+16.8	+26.2	+20.4	+24.8	+27.6
pixelate	+16.2	+11.2	+15.6	+13.8	+17.8	+37.4	+16.2	+11.4	+15.6	+57.2
shot noise	+14.6	+8.0	+8.2	+10.0	+14.8	+10.2	+11.4	+8.2	+13.4	+23.8
snow	+9.0	+8.4	+8.8	+11.2	+13.4	+1.6	+9.8	+8.2	+12.0	+12.4
zoom blur	+10.4	+7.2	+11.6	+10.2	+13.6	+1.6	+10.2	+8.0	+12.2	+31.6

I.1 THE TWO ORDERS, CONFIGURATION BY CONFIGURATION

Figure 7 plots $A_{C \circ A}$ against $A_{A \circ C}$ for every one of the 300 configurations, so that the size of the order effect can be read off the distance from the identity line rather than from a summary statistic.

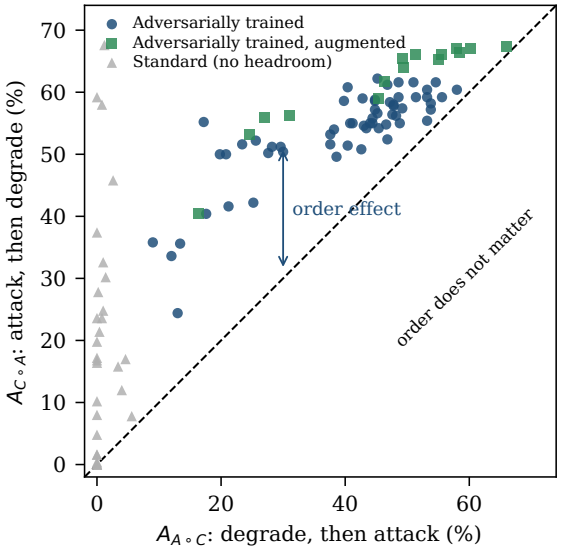

Figure 7: Accuracy under the two orders, one point per model and corruption type. Points above the identity line are configurations where degrading first costs accuracy.