ORDERED TOP-K UAPS: BENCHMARKING, DIAGNOSING, AND A SET-THEN-ORDER ATTACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordered top-K universal adversarial perturbations (UAPs) are a demanding and under-explored form of attack: *can one perturbation impose a prescribed ordered top-K list on held-out images, for an ensemble of models?* We benchmark four ordered top-K objectives, developed for per-image single-model attacks, in this setting over 3,180 UAPs spanning single models to an 18-model ensemble. The capability is real, but its reach shrinks as the ensemble grows: one perturbation controls ordered lists of up to eight classes on a single model, but only short lists across an ensemble. Fitted to many models, it also transfers: on seven unseen models, foundation-scale ones included, it imposes the attacker's class on up to two-thirds of images on average, and exact ordered pairs on fewer. No objective wins everywhere, and a single attack success rate (ASR) cannot say why. We therefore decompose ordered success into *set membership, ordering given membership, and cross-model agreement*. Ordering is set by how explicitly the loss encodes order — a control that encodes none sits on the chance line — while membership and agreement track attack strength; and each failure, whether under transfer, at large K, under adversarial training or against input defences, is one factor collapsing. Acting on the diagnosis, we propose *Set-then-Order*, which establishes membership with a distribution-matching loss and then orders it with hinges, at matched compute. With one schedule frozen across settings, it is never meaningfully worse than the hinge loss alone, and gains substantially where the membership bought in the first phase survives the second.

1 INTRODUCTION

Universal adversarial perturbations (UAPs) are a sharper threat than per-image attacks because one fixed δ, added to any image, fools a model on much of the data distribution (Moosavi-Dezfooli et al., 2017; Hendrik Metzen et al., 2017; Shafahi et al., 2020). Most UAP results are *untargeted*, and targeted universality is reported as hard to obtain at all (Liu et al., 2016; Zhang et al., 2020b; Weng et al., 2023). Ordered top-K attacks ask for far more than a label: that the K highest-scoring classes be exactly a prescribed list, in a prescribed order (Zhang & Wu, 2020; Paniagua et al., 2023; 2025). Systems that consume a ranked list rather than an argmax—retrieval, triage, moderation—are controlled by the adversary who controls that list.

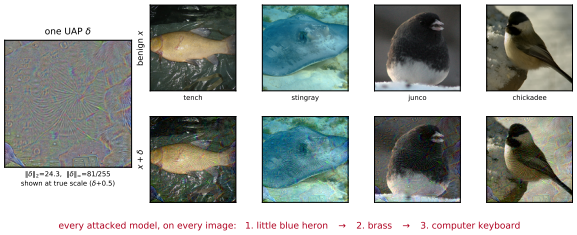

Figure 1: **Ordered top-K universality: one perturbation, many images, an exact prescribed ranking.** A single UAP δ (left), learned on a four-model ensemble at $K{=}3$ and added to four held-out images, makes *every* attacked model (ResNet-50, DenseNet-121, ViT-B/16 and DeiT-S) rank the same three classes in the prescribed order (bottom). δ and $x + \delta$ are shown at true scale (δ offset by 0.5 so that zero is mid-grey, not amplified). The $\|\delta\|_2 \leq 24.34$ budget used throughout bounds perturbation energy, not perceptibility, so the perturbation is visible (Appendix B). More learned UAPs: the HTML viewer in the supplementary.

A demanding problem that has been significantly under-explored. We study the two together: *ordered top-K UAPs* (Fig. 1), where one perturbation must impose the attacker's

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.

ordered list on held-out images, for every model in an ensemble at once. Existing ordered top-K attacks are per-image and single-model (Zhang & Wu, 2020; Paniagua et al., 2023; 2025). Whether the capability exists, and which objective obtains it, are open.

Benchmarking. We evaluate four ordered top-K objectives that span how explicitly the loss encodes order — CW-K's consecutive-pair hinges (Carlini & Wagner, 2016; Zhang & Wu, 2020), QuadAttacK's and RisingAttacK's inequality constraints (Paniagua et al., 2023; 2025), and AdvDistill's KL distribution matching (Zhang & Wu, 2020) — plus a no-order control, PGD-CE: targeted cross-entropy (Madry et al., 2017) against uniform $1/K$ mass on the same K classes, which rewards membership but not order. RisingAttacK, the strongest per-image method, has no per-example loss to average, so we *lift* it to the universal setting (§4). Over $3{,}180$ UAPs—19 configurations from single models to an 18-model ensemble, K up to 10 (5 for the 18-model ensemble), typically five seeds—the longest exact ordered list one perturbation imposes on at least 20% of held-out images is 3 to 8 classes on a single model but 3 on every ensemble; the 18-model perturbation also transfers, taking the top-1 prediction of seven unseen models, foundation-scale ones included, on up to two-thirds of images; and no objective wins everywhere: distribution matching leads jointly at $K{=}1$ and is last at $K{=}2$, and the per-image leader is third of four on the large ensemble at $K{=}1$ (§4).

Diagnosing. A single attack success rate (ASR) cannot say which of these outcomes is a failure to *find* the targets, to *arrange* them, or to do so on the *same images* across models. We factor it: on one model, ordered success is set success times the probability of the right order given the set; over a fleet it is an intersection, governed by how far per-model successes co-occur (§3). The factors explain why objectives succeed (§5.1–§5.2): ordering is supplied only by an explicit ordering term (the no-order control sits on the $1/K!$ chance line), while KL leads on membership on 11 of 11 backbones and on agreement in 19 of 20 cells. They also explain when objectives fail (§5.3): transfer and input defences remove membership while ordering survives, adversarial training removes membership almost entirely, and agreement is what large ensembles and a halved budget destroy.

Set-then-Order. The diagnosis places KL's and CW-K's strengths in different factors of one product, which suggests combining them. Set-then-Order (StO) establishes membership with KL and then orders it with CW-K, at matched compute (§6). With the switch point chosen once on disjoint images and frozen across ten settings, it is never more than 1.2 points worse than the hinge loss alone, and much better where the membership bought in the first phase survives the second: $+38$ points on ViT-B/16 and $+6.7$ on a four-model ensemble at $K{=}3$. How much of that membership survives, not how large it was, is what decides the gain.

Contributions. (i) **The problem**: ordered top-K attacks in the universal regime, across images *and* models. (ii) **A benchmark** of four objectives over $3{,}180$ UAPs, including a universal lifting of RisingAttacK. (iii) **A diagnosis** into membership, ordering and cross-model agreement, with the instrumentation that measures each. (iv) **A method**, Set-then-Order, derived from the diagnosis and tested with a frozen schedule in ten settings.

2 Related Work

UAPs. One perturbation that fools a model on most images was established for classification by Moosavi-Dezfooli et al. (2017) and Shafahi et al. (2020), and for segmentation by Hendrik Metzen et al. (2017); transfer across architectures and its ensemble construction go back to Liu et al. (2016). This line is almost entirely untargeted or top-1 targeted, and reports targeted universality as hard to obtain (Zhang et al., 2020b; Weng et al., 2023). The one universal treatment of top-K is untargeted, displacing the true class rather than installing a list (Tursynbek et al., 2022).

Ordered top-K attacks. Dictating the content and order of the top-K was posed by Zhang & Wu (2020), who also extend the Carlini–Wagner hinge (Carlini & Wagner, 2016) to consecutive-pair constraints, and developed by QuadAttacK (Paniagua et al., 2023) and RisingAttacK (Paniagua et al., 2025). All are per-image and single-model; we make them universal (§4) and combine them (§6).

Measuring and resisting attacks. Both literatures report one success rate, which for our setting is the product and intersection that §3 separates; transfer-oriented work likewise optimises the transferred rate rather than measuring whether successes co-occur (Dong et al., 2018; Xie et al., 2019). On the defence side we evaluate adversarial training (Madry et al., 2017; Salman et al., 2020), input transformations (Guo et al., 2018), and, as the adaptive threat we do not model, expectation over transformations (Athalye & Sutskever, 2017). Appendix A gives the extended discussion.

3 PROBLEM: ORDERED TOP-K UAPS

Notation. Let $\mathcal{Y}$ be the label set ($|\mathcal{Y}| = 1000$ for ImageNet-1k) and let $F_\Theta(x) \in \mathbb{R}^{|\mathcal{Y}|}$ denote the logits of a network with parameters Θ on an image $x \in [0,1]^{3\times H\times W}$. Write $\hat{Y}_\Theta(x)_{1:K}$ for the K classes with the largest logits, listed in descending logit order. A UAP is a single perturbation δ added to every image. We are given a set of *attacked* models $\mathcal{M} = \{\Theta_1, \ldots, \Theta_N\}$, a set of images $\mathcal{D}$, and an ordered target list $T = [t_1, \ldots, t_K]$, $t_k \in \mathcal{Y}$, sampled uniformly at random.

Learning. Learning an ordered top-K UAP is the maximum-satisfiability problem

$$\begin{aligned} \underset{\delta\in\mathbb{R}^{3\times H\times W}}{\text{minimize}} \quad & \|\delta\|_p^2, \qquad (1)\\ \text{s.t.} \quad & F_\Theta(x^+)_{t_k} > F_\Theta(x^+)_{t_{k+1}}, \quad F_\Theta(x^+)_{t_K} > F_\Theta(x^+)_c,\ \forall k \in [1, K-1],\ \forall c \in \mathcal{Y}\setminus T,\\ & x^+ = \mathrm{Clamp}(x+\delta), \quad \forall x \in \mathcal{D},\ \forall \Theta \in \mathcal{M}, \end{aligned}$$

whose satisfaction is exactly $T = \hat{Y}_\Theta(x^+)_{1:K}$ for every image and every model. The constraint set is enormous—$O(|\mathcal{D}| \cdot N \cdot |\mathcal{Y}|)$ inequalities—and, crucially, it is *heterogeneous*: the first family fixes an order among the targets, the second fixes membership of the top-K, and the quantifier over Θ demands that one δ satisfy both families simultaneously on every model. Those are the three difficulties the paper separates.

Two decompositions. Fix a model Θ and an image x, and let $x^+ = \mathrm{Clamp}(x+\delta)$. Define the two success events

$$S_\Theta(x) \;=\; \big\{\, \{\hat{Y}_\Theta(x^+)_{1:K}\} = \{T\} \,\big\}, \qquad O_\Theta(x) \;=\; \big\{\, \hat{Y}_\Theta(x^+)_{1:K} = T \,\big\}, \qquad (2)$$

i.e. the targets occupy the top-K *as a set*, and the targets occupy the top-K *in the prescribed order*. Since both lists have length K, set success is equality of sets, and $O_\Theta \subseteq S_\Theta$ always. Hence, over the image distribution,

$$\underbrace{\Pr[O_\Theta]}_{\text{ordered ASR}} \;=\; \underbrace{\Pr[S_\Theta]}_{\text{set ASR}} \cdot \underbrace{\Pr[O_\Theta \mid S_\Theta]}_{\text{order-conditional term}}. \qquad (3)$$

Equation 3 is an identity, not a modelling assumption; what makes it useful is that, empirically, the two factors are controlled by different things (§5.1). The second family of constraints in Eq. 1 governs $\Pr[S_\Theta]$, the first governs $\Pr[O_\Theta \mid S_\Theta]$, and the pairwise ordering inequalities number $O(K^2)$ per image–model pair against $O(|\mathcal{Y}|)$ membership inequalities—so the two factors are not merely conceptually distinct, they scale differently in K.

Second, universality over a model set is an *intersection*. The joint success event is

$$J_\mathcal{M}(x) \;=\; \bigcap_{\Theta\in\mathcal{M}} O_\Theta(x), \qquad \Pr[J_\mathcal{M}] \;\le\; \min_{\Theta\in\mathcal{M}} \Pr[O_\Theta]. \qquad (4)$$

Two attacks with identical per-model rates $\Pr[O_\Theta]$ can therefore have very different joint rates, depending on whether the models succeed on the *same* images. Per-model strength and cross-model agreement are separate quantities, and Eq. 4 says only the second can explain a gap between them.

Estimators. All quantities are estimated on held-out images (ImageNet-1k `val`). For a model set $\mathcal{M}$ we report the *per-model* rate $\frac{1}{N}\sum_\Theta \widehat{\Pr}[O_\Theta]$ and the *joint* rate $\widehat{\Pr}[J_\mathcal{M}]$, and we summarize agreement three ways. The first is the fraction of per-model success that survives the intersection,

$$r \;=\; \frac{\widehat{\Pr}[J_\mathcal{M}]}{\frac{1}{N}\sum_\Theta \widehat{\Pr}[O_\Theta]}, \qquad \bar{\phi} \;=\; \binom{N}{2}^{-1} \sum_{\Theta\neq\Theta'} \mathrm{corr}\big(\mathbf{1}[O_\Theta],\, \mathbf{1}[O_{\Theta'}]\big), \qquad (5)$$

with $\bar{\phi}$ the mean pairwise ϕ coefficient: for each pair of models, the correlation across held-out images of their $0/1$ per-image success indicators (1: fooled on exactly the same images; 0: unrelated).

r alone does not measure agreement: under independence with equal marginals $p_\Theta \equiv \Pr[O_\Theta] = p$, $r = p^{N-1}$ falls with N and with attack weakness. We therefore also report the strength-controlled

$$D \;=\; \Big(\widehat{\Pr}[J_\mathcal{M}] \;-\; \prod_\Theta \widehat{\Pr}[O_\Theta]\Big) \Big/ \Big(\min_\Theta \widehat{\Pr}[O_\Theta] \;-\; \prod_\Theta \widehat{\Pr}[O_\Theta]\Big), \qquad (6)$$

which places the observed joint rate between independence (D=0) and the most co-occurrence its marginals permit (D=1), so objectives with different per-model rates are comparable (Appendix I).

We estimate the order-conditional term per model as $\widehat{\Pr}[O_\Theta]/\widehat{\Pr}[S_\Theta]$ and average over models with enough set successes (Appendix I). The joint rate for arbitrary $\mathcal{M}$, and $\bar{\phi}$, require per-image indicators $\mathbf{1}[O_\Theta(x)]$ rather than their means; Appendix B describes the instrumentation that retains them.

Table 1: **Ordering is a separable cost, and the loss sets it.** Perturbations learned on the 18-model ensemble, scored on the same 5,000-image held-out `val` re-scoring set (mean±std over 5 seeds) **PGD-CE** is the chance row: targeted at the same K classes with uniform $1/K$ mass and *no* ordering term, it matches the best set rate and lands at the analytic $1/K!$ line (K=3 on four seeds). At K=1 set and ordered success coincide; PGD-CE is run from K=2.

		$K=1$	$K=2$			$K=3$			$K=4$			$K=5$		
	objective	ASR	set	ord	$P(\text{o}\|\text{s})$	set	ord	$P(\text{o}\|\text{s})$	set	ord	$P(\text{o}\|\text{s})$	set	ord	$P(\text{o}\|\text{s})$
seen 18 models	CW-K	91.3 ± 3.4	68.1 ± 7.4	61.6 ± 9.7	89.6 ± 4.9	38.5 ± 10.8	29.3 ± 10.0	70.7 ± 7.3	13.6 ± 6.1	7.0 ± 3.6	40.6 ± 10.0	2.5 ± 1.8	0.9 ± 0.8	24.9 ± 8.9
	Rising	88.5 ± 5.6	64.5 ± 10.0	54.9 ± 12.4	83.9 ± 7.8	36.4 ± 9.8	21.9 ± 8.7	57.1 ± 11.1	15.1 ± 4.7	4.1 ± 2.0	24.0 ± 9.1	7.1 ± 3.5	0.9 ± 0.6	9.2 ± 4.1
	QP	81.8 ± 7.1	52.1 ± 11.8	42.2 ± 13.6	78.2 ± 10.0	22.6 ± 8.9	14.3 ± 6.6	54.2 ± 8.5	7.7 ± 4.1	3.0 ± 2.2	27.2 ± 10.4	2.8 ± 0.9	0.6 ± 0.4	15.8 ± 7.0
	KL	91.7 ± 3.1	73.2 ± 5.3	43.4 ± 5.2	58.8 ± 3.8	47.3 ± 8.2	15.7 ± 3.8	32.8 ± 4.7	27.5 ± 6.8	4.6 ± 1.5	16.3 ± 3.9	11.4 ± 3.6	1.0 ± 0.7	9.1 ± 4.3
	PGD-CE	–	73.0 ± 6.2	35.0 ± 4.8	47.8 ± 4.2	49.9 ± 7.7	8.4 ± 1.7	16.7 ± 1.2	29.3 ± 6.2	1.5 ± 0.6	5.1 ± 2.0	16.9 ± 4.2	0.2 ± 0.1	1.0 ± 0.3
unseen 7 models	CW-K	62.8 ± 10.9	25.5 ± 10.5	17.9 ± 8.9	68.2 ± 18.0	4.8 ± 5.2	2.9 ± 2.7	61.2 ± 17.3	0.6 ± 0.9	0.1 ± 0.1	28.1 ± 26.1	0.0 ± 0.0	0.0 ± 0.0	–
	Rising	56.5 ± 12.5	24.2 ± 9.6	16.5 ± 7.6	65.6 ± 16.6	5.4 ± 5.9	1.9 ± 1.6	38.0 ± 10.0	0.6 ± 0.7	0.1 ± 0.1	12.1 ± 10.8	0.1 ± 0.1	0.0 ± 0.0	3.5
	QP	46.7 ± 11.6	15.3 ± 11.2	9.9 ± 6.8	61.3 ± 20.4	2.2 ± 2.4	1.1 ± 1.1	48.8 ± 15.1	0.3 ± 0.5	0.1 ± 0.1	21.4	0.0 ± 0.0	0.0 ± 0.0	–
	KL	68.1 ± 7.9	30.6 ± 10.4	13.5 ± 7.9	44.3 ± 30.6	8.3 ± 7.4	2.2 ± 1.2	25.4 ± 14.6	2.4 ± 2.4	0.3 ± 0.2	12.8 ± 15.4	0.2 ± 0.3	0.0 ± 0.0	4.1 ± 3.9
	PGD-CE	–	27.8 ± 6.3	12.6 ± 8.1	43.5 ± 31.6	10.2 ± 6.0	1.8 ± 1.5	11.8 ± 8.4	1.6 ± 0.9	0.2 ± 0.3	9.7 ± 11.2	0.6 ± 0.6	0.0 ± 0.0	0.2 ± 0.3

4 Benchmarking Ordered Top-K UAPs

Protocol in brief. We attack 28 ImageNet-1k networks in 19 training configurations, from single models to an 18-model ensemble, and measure transfer on 7 held-out models never seen during learning. Each UAP is optimised for 50 epochs on 5 training images per class *without any data augmentation*, under one energy budget $\|\delta\|_2 \leq 24.34$ (not an imperceptibility constraint; the perturbations are visible, Fig. 1), and is scored at its *final* epoch on held-out `val` images; the corruptions of §5.3 are applied only at evaluation. The benchmark comprises **3,180** completed UAPs (four objectives, K up to 10, or 5 on the 18-model ensemble, typically five seeds per cell; Table 17) plus a half-budget replication; **Appendix B gives the full protocol.** *Our PyTorch code will be released upon publication.*

Four objectives. We benchmark four existing ordered top-K objectives, chosen because they span how explicitly the loss encodes order (Appendix B gives each loss). **CW-K** carries $K-1$ explicit consecutive-pair hinges plus a boundary hinge. **QP** (QuadAttacK, Paniagua et al., 2023) and **RisingAttacK** (Paniagua et al., 2025) impose order as inequality constraints, solved as a quadratic program in feature space and as a sequential QP over a Jacobian subspace respectively. **KL** (after AdvDistill, Zhang & Wu, 2020) matches a target distribution in which the targets carry descending logits a fixed 0.3 apart, so order is a small preference inside a matched distribution rather than a constraint. CW-K, QP and KL are differentiable in the logits and become universal by averaging over images and models. As a control we add **PGD-CE**, optimized with the same projected AdamW schedule as the four objectives: it encodes membership and *no* order, so it calibrates what the ordering terms are worth (§5.1).

Lifting RisingAttacK to a universal perturbation. RisingAttacK, the strongest per-image ordered top-K attack, is not a loss but a sequential-QP solver for one image and one model, so there is no per-example objective to average. We lift it by *stacking* the attacked models' Jacobians into one constrained system, so the search subspace is one the models share, and by turning the solver's converged iterate $x^\star$ into an outer objective $\|x^\star - \text{Clamp}(x+\delta)\|_2^2$ on the shared perturbation (Appendix C.1). We claim the lifting preserves the method, not that it is the best way to universalize it. The lifting adds no ordering terms of its own, so the ordering behaviour we report for RisingAttacK is a property of the objective being lifted.

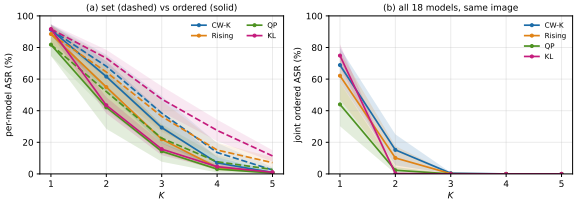

Figure 2: **The benchmark as K grows.** 18-model ensemble at the fixed budget; bands are ±1 std over seeds. **(a)** per-model ASR under the set criterion (dashed) and the ordered criterion (solid); the vertical gap between a pair of curves is the ordering cost. It is widest for KL at every K, and as a fraction of set-successes CW-K retains the most order at every K (Table 1). **(b)** joint ordered ASR, requiring all 18 models to emit the exact list on the same image. KL leads at K=1 and is last by K=2—the metric-dependence of the ranking, visible as a crossing.

What the benchmark shows. Figure 2 summarises the 18-model ensemble, the hardest setting. *(i) The capability exists, and its reach shrinks as the ensemble grows.* At K=1 every objective fools the attacked models on 82–92% of held-out images on average, and all eighteen at once on 44–75%; at K=2 the best objective (CW-K) still imposes the exact ordered pair on 61.6% per model and on all

eighteen simultaneously on 15.2%. By $K{=}5$ no objective exceeds 1% per model. Smaller model sets support longer lists. On the full held-out split, the largest K at which the best objective still reaches 20% per model is 3 for every ensemble we train — two, four or eighteen models — but 3 to 8 for single backbones: 6 for ConvNeXtV2-L and ViT-B/CLIP and 8 for DINOv3-7B, which still takes an exact ordered list of eight classes on 20% of images (Table 3; ensembles: Appendix C.2). ***(ii) It transfers to models it never saw.*** Scored on seven held-out models — including the foundation-scale ViT-B/CLIP, EVA-Giant/CLIP and DINOv3-7B — the 18-model perturbation at $K{=}1$ imposes the attacker's class on 47–68% of images per model, depending on the objective (Table 1), and on a majority of the seven for up to 76%; even the most resistant, DINOv3-7B, is fooled on 18–31% (Table 21). At $K{=}2$ it still imposes the exact ordered pair on 10–18% of images per unseen model. This is a property ensembling buys: a perturbation fitted to one backbone transfers to its three sibling architectures on only 4% of images (CW-K; Table 4). *(iii) No objective wins everywhere.* KL leads joint success at $K{=}1$ (74.9%) and is last at $K{=}2$ (0.2%, against CW-K's 15.2%), and the per-image literature's leader, RisingAttacK, is third of four jointly at $N{=}18$, $K{=}1$. A single ordered rate does not say why these outcomes differ; §5 applies the decomposition of §3 to find out.

5 Diagnosis: Why Objectives Succeed, and When They Fail

5.1 Why Objectives Succeed: Membership and Ordering Are Bought Separately

Equation 3 splits ordered success into placing the targets in the top-K and arranging them there. The benchmark objectives are strong on different factors, and for identifiable reasons: the set term tracks how strong the attack is—it responds to budget and is what transfer removes—while $P(\text{order} \mid \text{set})$ tracks how explicitly the *loss* encodes order, and barely tracks the architecture at all.

Ordering is a large, separable cost. As shown in Figure 2(a) and Table 1, enforcing the prescribed order is strictly harder than hitting the target set, and the gap is substantial: averaged over objectives and $K \in \{2, 3\}$, about a third (34%) of set-successes carry the wrong order. This validates the ordered formulation directly. An ordered top-K attack is not merely a top-K set attack with more targets—about a third of the difficulty at moderate K lies in arrangement, and a metric that scores only set membership would report a materially easier problem.

The order-conditional term is set by the loss. Reading down the $P(\text{o}|\text{s})$ columns of Table 1 at $K{=}2$ and $K{=}3$, the four order-encoding objectives fall in the same sequence, and it is exactly the sequence in which they encode order (Appendix B); CW-K stays first and KL last among them at every K shown, with PGD-CE below all four, while Rising and QP swap at $K \geq 4$, where the estimate rests on few set-successes (see the boundary below). At $K{=}2$: CW-K 89.6, with its $K{-}1$ explicit consecutive-pair hinges; RisingAttacK 83.9 and QP 78.2, which carry ordering as explicit inequality constraints; and KL 58.8, which expresses order only as a 0.3 logit margin inside a matched target distribution. A thirty-point spread in the ordering term, produced by objectives whose *set* rates are closer together, is a loss-design→behaviour link rather than a property of the attacked models.

A calibrated chance line, and a controlled pair. Comparing four objectives that all encode order cannot show what the ordering signal is *worth*, so we add a fifth that encodes none: PGD-CE. It is symmetric across the targets by construction, so its ordering can only be whatever falls out of optimizing membership. On the 18-model ensemble it lands on the analytic chance rate $1/K!$ at every K we run: $P(\text{o}|\text{s}) = 47.8 \pm 4.2$ against 50% at $K{=}2$, 16.7 ± 1.2 against 16.7% at $K{=}3$, 5.1 ± 2.0 against 4.2% at $K{=}4$, and 1.0 ± 0.3 against 0.8% at $K{=}5$ (Table 1; $K{=}3$ rests on four seeds). The objectives above it are therefore measured against a line we locate empirically rather than only derive, and each clears it at every K — KL's small 0.3 margin included (32.8 against 16.7% at $K{=}3$). PGD-CE and KL are moreover a controlled pair: their set rates are indistinguishable (73.0 ± 6.2 against 73.2 ± 5.3) while $P(\text{o}|\text{s})$ differs by eleven points (47.8 against 58.8), so a 0.3 logit margin alone moves the ordering term. The link also survives varying the architecture instead of the ensemble: across eleven single-model backbones at $K{=}3$, KL attains the highest set rate and the lowest ordering rate on 11 of 11, and CW-K the highest ordering rate on 10 of 11 (Table 19; Appendix C.2).

Under transfer it is membership, not ordering, that gives way first. The lower block of Table 1 scores the same perturbations on the seven held-out models. Both factors fall, but not comparably: at $K{=}2$ the target *set* is found on 58–71% fewer images than on the attacked models, while the order-conditional term loses only 22–25% of its value, with the objective ranking preserved at $K{=}2$. So a perturbation carried to an architecture it was not optimised against mostly stops *placing* the target

classes; when it still places them, it arranges them nearly as well as before. Ordering is the harder constraint on the attacked models, but it is the more transferable one. The account has a boundary: at $K \geq 4$ on the 18-model ensemble the joint set rate is itself near zero, so there membership, not ordering, is the binding constraint (Appendix C.2).

5.2 Why Objectives Succeed Jointly: Agreement Across Models

Universal success over a model set is an intersection (Eq. 4), so two objectives with identical per-model rates can differ sharply in what an attacker actually controls. We measure how much per-model success survives the intersection with the statistics of Eq. 5—the surviving fraction r, the strength-controlled dependence D, and the mean pairwise success correlation $\bar{\phi}$. Agreement is what large ensembles destroy, and it is where the objectives differ most.

Table 2: **Average per-model success overstates simultaneous success, increasingly so with N.** $K = 1$, 5,000 held-out `val` images, $\|\delta\|_2 \leq 24.34$.

config	objective	per-model	joint	r	D	$\bar{\phi}$
R50+D121 (N=2)	CW-K	92.0	89.3	0.970	0.787	0.638
	Rising	85.5	81.5	0.953	0.741	0.664
	QP	81.9	72.8	0.889	0.861	0.465
	KL	93.0	90.5	0.972	0.738	0.609
ViT-B+DeiT-S (N=2)	CW-K	91.9	87.9	0.957	0.570	0.479
	Rising	87.0	80.9	0.931	0.557	0.464
	QP	87.4	80.0	0.915	0.642	0.390
	KL	95.7	94.0	0.982	0.688	0.599
4-model (N=4)	CW-K	90.1	80.7	0.896	0.686	0.533
	Rising	54.9	33.0	0.601	0.566	0.493
	QP	71.7	40.2	0.561	0.748	0.368
	KL	91.3	83.9	0.919	0.743	0.573
18-model (N=18)	CW-K	91.3	68.9	0.755	0.745	0.510
	Rising	88.5	62.2	0.702	0.712	0.495
	QP	81.8	44.0	0.538	0.609	0.460
	KL	91.7	74.9	0.817	0.817	0.549
7 unseen (transfer)	CW-K	62.8	18.9	0.301	0.631	0.376
	Rising	56.5	13.8	0.244	0.544	0.380
	QP	46.7	8.8	0.189	0.507	0.397
	KL	68.1	21.7	0.319	0.680	0.387

Average per-model success overstates simultaneous success. Table 2 reports K=1, where ordered and set success coincide. Retention r falls steadily with N: the best objective keeps 0.97–0.98 of its per-model success at N=2, 0.92 at N=4, 0.82 at N=18, and 0.32 on seven models it never saw.

How much of that is dependence? Not all of it, and r cannot tell us: for independent successes with equal per-model rate p it equals p^{N-1}, so it falls with N *and* with attack strength at fixed dependence. The D column removes the strength component, placing the observed joint rate between independence and the most co-occurrence the same marginals allow. At N=18 the objectives differ in D in the same order as r (0.82 for KL against 0.61 for QP), so the agreement claim holds where we make it. At N=2 and N=4 the D ordering does not reproduce the r ordering, so the retention gap in the smaller ensembles is a strength effect, and we confine the agreement claim to the large-ensemble regime. The pairwise correlation $\bar{\phi}$ recovers the r ordering at N=4 and N=18 but not on R50+D121 or under transfer (Appendix C.3).

Per-model success is a poor proxy for what an attacker gets. At N=18, K=1 the four objectives reach per-model rates within ten points of each other (81.8–91.7%) but joint rates spanning thirty-one (44.0–74.9%): the intersection amplifies a 9.9-point difference into a 30.9-point one. Two mechanisms do that, and we separate them rather than attribute the whole effect to agreement. Compounding alone accounts for much of it—substituting the observed marginals into $\prod_{\Theta} \widehat{\Pr}[O_{\Theta}]$ gives an independence-predicted joint spread of 18.0 points (Appendix C.3), because a small per-model deficit is raised to the eighteenth power. The remaining amplification is dependence, and it is real: D separates the objectives over 0.61–0.82 in the same order. The consequence for method selection is the same under either mechanism. **The ranking the per-image literature establishes is not the ranking that holds here:** RisingAttacK, the strongest published per-image ordered top-K attack, is third of four in joint success, while the two objectives that lead—distribution matching and consecutive-pair hinges—are not the per-image leaders. Choosing an objective for a universal, multi-model attack on the strength of its per-image results is therefore unreliable, and at N=18 it is the joint rate, not the per-model rate, that an attacker actually obtains.

Isolating agreement from ordering. Recomputing the same statistics on *set* success, which divides the ordering factor out (Table 20), shows a second, independent decay: r_{set} falls monotonically with K in all sixteen (config, objective) cells, so **lengthening the target list costs cross-model agreement in its own right** (Appendix C.3).

Membership and ordering pull against each other. Reading Tables 19 and 20 together resolves what looked, in a single ordered ASR, like a puzzle. The KL objective attains the highest r_{set} in **19 of 20** (config, K) cells of Table 20, and the highest set rate on 11 of 11 backbones (Table 19). By both factors that make up universality— finding the targets, and finding them on the same images across models—KL is the strongest objective we measure. Its notorious collapse at $K \geq 2$, from a joint ordered rate of 75% at K=1 to 0.2% at K=2, is therefore *entirely* an ordering failure and

not a universality failure at all. The opposite holds for CW-K, which leads on ordering (Table 1) while trailing KL on set membership and agreement. No objective in our study leads on both factors, and a single ordered ASR reports their product without revealing that a design choice traded one for the other. Because the two strengths sit in different factors of a product, they are candidates to be *combined*; §6 does so.

5.3 When Objectives Fail: K, Transfer, Budget, and Robust Models

This section characterises where objectives stop succeeding, and which factor gives way when they do. Because our seven held-out models are attacked *both* as white-box targets and by the 18-model perturbation they never saw, we can also ask whether white-box and transfer resistance are the same property. They are not.

Table 3: **The white-box ordered top-K frontier is architecture-dependent, and adversarial training can block it.** Per-model ordered ASR%, best over the four objectives, final epoch, full held-out `val` split, $\|\delta\|_2 \leq 24.34$. $K^\star$: the largest K reaching 20%. Every K and configuration: Table 17.

model	K=1	K=2	K=3	K=5	K=7	K=10	$K^\star$
ResNet-50	92	78	55	13	1	0	4
DenseNet-121	97	74	52	13	2	0	4
ViT-B/16	97	72	32	5	0	0	3
DeiT-S	97	81	64	9	1	0	4
Swin-B	98	70	43	21	5	0	5
HRNet-W48	95	70	44	10	2	0	4
ConvMixer-768	96	82	61	26	11	1	5
ViT-B/CLIP	98	82	67	31	14	2	6
ConvNeXtV2-L	98	95	83	42	13	1	6
EVA-Giant/CLIP	97	74	36	4	1	0	3
DINOv3-7B	98	89	78	51	33	8	8
R50-AT (Engstrom)	17	0	0	0	0	0	*none*
R50-AT (Salman)	9	0	0	0	0	0	*none*
WRN-50-AT	11	0	0	–	–	–	*none*

The white-box frontier is architecture-dependent. Table 3 gives the single-model frontier at the fixed budget. At K=1 every standard backbone is essentially fully attackable (92–98%), so top-1 targeted universality discriminates nothing. The K-axis does: $K^\star$, the largest K still reaching 20%, ranges from 3 for ViT-B/16 and EVA-Giant/CLIP to 8 for DINOv3-7B. Capacity and pretraining appear to buy *depth*—ConvNeXtV2-L, ViT-B/CLIP and DINOv3-7B sustain the longest ordered lists—but the relationship is not clean, since EVA-Giant/CLIP is foundation-scale and CLIP-pretrained yet shares the shallowest frontier. We therefore describe the frontier's range rather than fit a law to it.

Adversarial training suppresses the phenomenon rather than shifting it. The three adversarially-trained backbones (lower block of Table 3) are fooled on only 9–17% of images at K=1, against 92–98% for standard models, fall below 1% at K=2 and to 0.0% at K=3, so none has a $K^\star$. The decomposition locates the failure in *membership*: at K=2 the target pair enters the top-2 on at most 1.2% of images for every objective, against 39–81% on standard ResNet-50, and on those few set successes CW-K orders only 21–58% (pooled over seeds), against 96% on standard ResNet-50. Under the attacks and budget tested this is suppression, not removal (Appendix C.4).

Table 4: **Defences remove long ordered lists first, and ensembling buys robustness.** Per-model ordered ASR% of CW-K under one operating point per defence; 5,000 held-out `val` images, $\|\delta\|_2 \leq 24.34$, mean over seeds (and over 11 backbones for the single-model row); *single UAP → 3 unseen*: each of the R50, D121, ViT-B and DeiT-S UAPs scored on the other three. All objectives: Table 10.

UAP	K	none	JPEG 70	JPEG 50	resize-pad	blur σ=0.95	grayscale
single-model UAP	1	94.6	38.8	19.1	14.2	24.1	29.4
	2	77.7	12.6	2.8	3.3	5.3	10.3
	3	53.5	1.9	0.2	0.7	1.5	2.0
4-model UAP	1	90.1	60.0	42.8	33.6	46.6	40.0
	2	66.2	20.9	9.6	7.8	13.0	11.9
	3	30.9	0.5	0.1	0.5	0.6	1.1
18-model UAP	1	91.3	74.7	63.9	62.5	57.6	56.9
	2	61.6	26.7	17.0	21.2	11.6	20.0
	3	29.3	2.9	1.0	4.5	1.7	5.5
18-model UAP → 7 unseen	1	62.8	43.6	34.5	50.8	40.5	36.1
single UAP → 3 unseen	1	4.0	1.8	1.2	1.7	1.1	0.5

White-box depth is not transfer resistance. The seven held-out models, attacked both ways, separate the two axes: DINOv3-7B is the deepest white-box target ($K^\star$=8) and the most transfer-resistant (29.0%), while HRNet-W48 is shallow ($K^\star$=4) and the most transfer-susceptible (77.7%). Transfer susceptibility instead follows the pretraining objective — supervised ImageNet models are fooled most, language- or masked-image-pretrained ones less, the self-supervised foundation-scale model least — so the models at the top of the deployment stack are the ones a shared perturbation reaches least well (Table 18; Appendix C.4).

Resistance from the input side: defences remove long lists first. A deployed defender can also transform the input, so we re-score every learned perturbation — every configuration, objective and K — under seven input corruptions swept over a magnitude ladder (Table 4 fixes one operating point per defence; Table 10 gives all). The perturbations are learned without augmentation, so this measures a *non-adaptive* attacker; one optimising through the defence (Athalye & Sutskever, 2017) would be expected to keep more. Two regularities hold for every objective. *Fragility grows*

with K: JPEG at quality 70 leaves the 18-model CW-K perturbation 82% of its success at $K{=}1$ ($91.3 \to 74.7$), 43% at $K{=}2$ and 10% at $K{=}3$. *Ensembling buys robustness*: under the same JPEG, single-model perturbations keep 38.8% at $K{=}1$ against 74.7% for the 18-model one, and under resize-pad 14.2% against 62.5%; on the seven unseen models the 18-model perturbation keeps 50.8 of its 62.8% under resize-pad, while single-model perturbations barely transfer at all (4.0% at $K{=}1$ for CW-K, undefended). These defences are cheap: JPEG-70, resize-pad and blur at $\sigma{=}0.95$ cost 2–5 points of clean top-1 accuracy on the attacked models and 1–3 on the held-out ones; only strong blur ($\sigma{=}1.84$, 15 points) is expensive.

What the defences break: membership, not ordering. Re-scoring every 18-model perturbation under each defence on both criteria (Table 5) shows which factor gives way. At $K{=}2$, JPEG, blur and resize-pad remove 49–75% of the set term on the attacked models but keep 88–97% of $P(\text{order} \mid \text{set})$; at $K{=}3$ they keep 8–19% of membership and still 86–91% of the ordering term. When the targets survive a defence, they survive in order. The one exception is grayscale, which also costs about a third of the ordering term, so colour carries part of the ordering signal. The same holds on the seven held-out models, where resize-pad is the weakest defence — it keeps 59% and 48% of membership at $K{=}2$ and 3, against 37% and 17% on the attacked models — so the cheapest defence for the defender is also the one that does least against a transferred perturbation. Transfer (§5.1), input defences and adversarial training therefore fail the attack the same way: each removes membership, and under transfer and input defences ordering survives.

Table 5: **Defences break membership, not ordering.** Share (%) of the undefended set rate and of $P(\text{o}|\text{s})$ that survives each defence, for the 18-model perturbations; mean over objectives and seeds, 5,000 held-out `val` images, $\|\delta\|_2 \le 24.34$. $K{=}4, 5$: Table 12.

	attacked (18)					unseen (7)				
	$K{=}1$	$K{=}2$		$K{=}3$		$K{=}1$	$K{=}2$		$K{=}3$	
defence	set	set	$P(\text{o}\|\text{s})$	set	$P(\text{o}\|\text{s})$	set	set	$P(\text{o}\|\text{s})$	set	$P(\text{o}\|\text{s})$
JPEG 70	81	51	94	19	91	71	40	106	14	90
JPEG 50	69	35	92	8	89	58	26	110	7	77
resize-pad	65	37	88	17	86	79	59	100	48	85
blur $\sigma{=}0.95$	61	25	97	10	88	65	30	110	11	85
grayscale	60	44	64	23	63	57	48	77	43	67

Resistance from the budget side. Halving the budget on the four-model ensemble (Appendix C.5) is uneven across objectives: at $K{=}1$ the joint rate retains 0.47 and 0.43 of its value for KL and CW-K but 0.08 and 0.01 for QP and RisingAttacK: QP loses *agreement* before per-model strength, RisingAttacK loses both. The collapse of QP at $K{=}1$ is an interaction with ensembling, not a property of the loss: on single models its retention is indistinguishable from KL's and CW-K's.

6 Set-then-Order: Combining What the Diagnosis Separates

The diagnosis locates the strengths of the benchmark objectives in different factors of one product. KL attains the highest set rate on every backbone and the highest cross-model agreement in all but one ensemble cell, but the weakest ordering; CW-K has the strongest ordering but gives up membership (§§5.1–5.2). Ordered success is their product (Eq. 3), so an attack that obtained KL's set term and CW-K's order-conditional term would exceed both. We propose the simplest attack that does this.

Method. Set-then-Order (StO) optimises the shared perturbation with the membership objective first and the ordering objective second,

$$\mathcal{L}_{\text{StO}}(\delta;\, e) \;=\; \begin{cases} \mathcal{L}_{\text{KL}}(\delta) & e < T_s, \\ \mathcal{L}_{\text{CW-K}}(\delta) & e \ge T_s, \end{cases} \tag{7}$$

where e is the epoch and both losses are averaged over images and attacked models as in §4. Phase one places the targets in the top-K, on the same images across models; phase two arranges targets that are already present. Both terms are functions of the same logits, so StO costs one forward pass per model per step and runs for the same 50 epochs as each single objective: any gain is not bought with compute. The alternative that does *not* follow from the diagnosis is a fixed blend $(1-\alpha)\mathcal{L}_{\text{KL}} + \alpha\mathcal{L}_{\text{CW-K}}$, which we run as a control.

What the diagnosis predicts. If the two phases supplied the two factors perfectly, StO would reach the ceiling $\Pr_{\text{KL}}[S] \cdot \Pr_{\text{CW-K}}[O \mid S]$. This is an upper bound, not a forecast, since nothing in the hinges preserves the membership KL buys in phase one; what decides the gain is how much of KL's membership lead StO still holds at the end, which we report for every setting.

Selection without test leakage. We chose T_s (and the blend weight α of the control) on held-out images of odd ground-truth class, on the four-model ensemble at $K \in \{2, 3\}$, and *froze* $T_s{=}35$ of 50 epochs for every setting in Table 6. Every number there is on the even half, and each StO run is

compared only with baselines trained on the same number of GPUs, since the batch is per process (Appendix C.6).

Where StO helps, and why. Table 6 tests the frozen schedule in ten settings. StO is never meaningfully worse than CW-K — the largest loss is 1.2 points — and in two settings it is much better: on the four-model ensemble at $K{=}3$ it adds 6.7 points ($4/4$ seeds), and on ViT-B/16 it adds 38.1 points ($4/4$ seeds), exceeding the ceiling. What separates these from the settings where StO is flat is not the size of KL's membership lead but how much of it survives the switch: StO keeps 75% of the lead on ViT-B and 49% on the four-model ensemble, but 11–33% on Swin-B, ConvNeXtV2-L and the 18-model ensemble, and none on ResNet-50. Swin-B shows the distinction most sharply: KL out-places CW-K by 25 points there, the largest lead after ViT-B, yet StO keeps 11% of it and trails CW-K on four of five seeds. The ceiling, which assumes full retention, is reached only where retention is high. The $K{=}2$ settings are flat for a simpler reason: CW-K already orders most of what it places, so there is little for a membership phase to add; at half the budget both constituents are weak ($\leq 18\%$ at $K{=}2$, $\leq 3.3\%$ at $K{=}3$).

Table 6: **Set-then-Order across settings.** Per-model ordered ASR% on the *even* ground-truth-class half of the held-out re-scoring set; the schedule (KL for 35 of 50 epochs, then CW-K) was chosen once, on the odd half of the 4-model $K{\in}\{2,3\}$ cells, and frozen everywhere else. Δ: StO minus CW-K, paired by seed (seeds won). *ceiling*: the gain if StO kept KL's set rate and CW-K's ordering. *lead*: KL's set-rate lead over CW-K; *kept*: the share of that lead StO retains after the switch. Each StO run is compared only with baselines trained at the same world size (10 runs excluded on that ground).

setting	K	KL	CW-K	StO	Δ vs CW-K	ceiling	lead	kept
4-model	3	21.8	32.6	39.3	+6.7 (4/4)	+10.7	+13.5	49%
4-model	2	49.7	68.8	67.6	−1.2 (1/4)	+4.1	+4.4	-5%
4-model, $\ell_2{\leq}12.17$	3	3.3	2.5	2.4	+0.0 (2/5)	+3.9	+7.2	0%
4-model, $\ell_2{\leq}12.17$	2	16.6	17.9	16.8	−1.1 (1/5)	+4.6	+5.6	18%
ViT-B/16	3	33.0	23.1	61.3	+38.1 (4/4)	+34.2	+49.1	75%
Swin-B	3	37.3	41.1	43.5	+2.4 (1/5)	+18.6	+24.6	11%
ResNet-50	3	21.2	57.5	57.3	−0.2 (1/4)	+9.5	+10.9	-22%
ConvNeXtV2-L	3	51.3	81.4	83.5	+2.1 (4/5)	+7.0	+7.3	33%
18-model	3	15.2	28.3	29.5	+1.2 (5/5)	+6.6	+8.6	25%
18-model	2	42.5	60.2	60.6	+0.4 (3/5)	+4.6	+5.1	33%
design variants (same protocol)								
optimizer reset, 4-model	3	21.8	32.6	39.4	+6.8 (4/4)	+10.7	+13.5	54%
optimizer reset, Swin-B	3	37.3	41.1	40.4	−0.7 (1/5)	+18.6	+24.6	-1%
fixed mix $\alpha{=}0.3$, 4-model	3	21.8	32.6	31.3	−1.3 (1/4)	+10.7	+13.5	90%

Design variants. A fixed blend $0.7\,\mathrm{KL} + 0.3\,\mathrm{CW\text{-}K}$ keeps 90% of KL's membership lead but loses CW-K's ordering and ends 1.3 points below CW-K: sequencing, not mixing, combines the two factors. Resetting the AdamW moments at the switch removes a transient dip in training success but leaves the final result unchanged (+6.8 against +6.7 on the four-model ensemble), so the transient is not what limits retention (Appendix C.6).

7 Limitations

We observe five aspects that bound what we claim. *(i) Budget.* The corpus is at one energy budget; the half-budget replication covers the four-model ensemble and two single models but not $N{=}18$, and two points do not determine a curve. *(ii)–(iii) Resistance.* The white-box/transfer dissociation rests on seven models, and every transfer number comes from one conventionally supervised training ensemble. *(iv) Set-then-Order* uses one schedule frozen across settings with four or five seeds each, so only its large gains are individually decisive; we explain where it helps by retention of membership, not why retention varies across backbones, and claim it as the attack the diagnosis implies, not a new loss. *(v)* The ordering account applies where set successes exist, not to the collapse at large K. Appendix D states each in full.

8 Conclusion

We posed ordered top-K universal perturbations. A benchmark of four objectives over 3,180 UAPs shows that one perturbation imposes an exact ordered list of 3 to 8 classes on a single model but 3 on every ensemble, transfers to unseen models, and that no objective wins everywhere. Decomposing ordered success into membership, ordering given membership, and cross-model agreement explains why: the loss sets the ordering term, and each failure is one factor collapsing — membership under transfer, input defences and adversarial training, agreement at large ensembles and small budgets. The same decomposition prescribes Set-then-Order, which at matched compute beats the hinge loss by up to 38 points where its first-phase membership survives, and stays within 1.2 points elsewhere. A single ordered success rate is therefore the wrong instrument for benchmarking this threat, and defending against it needs only one factor broken.

REPRODUCIBILITY STATEMENT

Every table and figure is generated by a script from saved results rather than transcribed by hand; we will release the code upon publication. The supplementary material contains an HTML viewer of the learned perturbations, with each one's per-model metrics and, where evaluated, its success under each input defence. To fit the 100 MB limit on supplementary material, the viewer holds the 3,336 perturbations that the paper's tables and figures draw on, stored as lossy WebP images; the full set of 3,509 learned perturbations, stored losslessly, exceeds that limit and will be released with the code. The protocol is stated in full in Appendix B: the model roster (all public `timm` or published robust checkpoints, Appendix E), the training images and their filter, the seeded uniform target sampler, the single ℓ_2 budget $\|\delta\|_2 \leq 24.34$, the 50-epoch schedule, and the fact that no data augmentation is used when learning a perturbation. Every reported number uses the final-epoch perturbation, never the best epoch, and every table names the held-out image set it is scored on (the 5,000-image re-scoring set or the full `val` split). The universal lifting of RisingAttacK is specified in §4 and Appendix C.1. The re-scoring pass stores per-image, per-model success indicators with an alignment fingerprint, so every agreement, majority and joint statistic can be recomputed without re-learning a perturbation. Set-then-Order's schedule was selected on images of odd ground-truth class and every Set-then-Order number is reported on the even half, with each run compared only to baselines trained on the same number of GPUs (§6); the per-process batch and the world sizes used are listed in Appendix B. Results are reported over up to five seeds with dispersion. Two limits on exact reproduction: the training code does not force deterministic cuDNN kernels, so runs with convolutional members agree with a re-run statistically rather than bit for bit; and the held-out-model decomposition under defences is computed with TF32 matrix multiplication, whose retention ratios divide by an undefended baseline computed in the same pass.

ETHICS STATEMENT

This paper studies an attack, so it is dual-use by construction. We judge the balance to favour publication. The capability we measure is already implied by published per-image ordered attacks and by universal perturbations; what we add is a measurement of how far it extends across images and models, and, equally, where it fails. Much of the paper is therefore directly useful to defenders: it shows that adversarial training suppresses the attack's ability to place the target classes, that standard input transformations remove long ordered lists first, and which models a shared perturbation reaches least well. Our perturbations are learned under an ℓ_2 energy budget and are visible on inspection (Fig. 1), and we evaluate a non-adaptive attacker against defences. We do not attack any deployed system. The work uses only the public ImageNet-1k dataset and publicly released models, involves no human subjects, no new data collection and no personally identifying information. We will release code for research use upon publication, and no perturbation tuned to a specific production system. We declare no conflicts of interest.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper: to edit, condense and restructure prose, and to help construct LaTeX tables and figure scaffolding. We also used AI coding assistants to pull training and evaluation logs from our compute servers and to write the scripts that aggregate saved results into the paper's tables and figures, and to build the HTML viewer of learned perturbations in the supplementary material. The authors reviewed all AI-assisted text and code and checked every claim and number against the underlying results. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. *CoRR*, abs/1707.07397, 2017. URL `http://arxiv.org/abs/1707.07397`.

Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016. URL http://arxiv.org/abs/1608.04644.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4312–4321, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Dong_Evading_Defenses_to_Transferable_Adversarial_Examples_by_Translation-Invariant_Attacks_CVPR_2019_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2755–2764, 2017.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2016.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Thomas Paniagua, Ryan Grainger, and Tianfu Wu. Quadattack$k$: A quadratic programming approach to learning ordered top-k adversarial attacks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Thomas Paniagua, Chinmay Savadikar, and Tianfu Wu. Adversarial perturbations are formed by iteratively learning linear combinations of the right singular vectors of the adversarial jacobian. In *The Forty-second International Conference on Machine Learning, ICML*, 2025. URL `https://openreview.net/pdf?id=3q6T4hQ4Dy`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5636–5643, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, pp. 516–533. Springer, 2022.

Asher Trockman and J Zico Kolter. Patches are all you need? *Transactions on Machine Learning Research*, 2023.

Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. Geometry-inspired top-k adversarial perturbations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3398–3407, 2022.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Juanjuan Weng, Zhiming Luo, Zhun Zhong, Dazhen Lin, and Shaozi Li. Exploring non-target knowledge for improving ensemble universal adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2768–2775, 2023.

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

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2730–2739, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14521–14530, 2020a.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14521–14530, 2020b.

Zekun Zhang and Tianfu Wu. Learning ordered top-k adversarial attacks via adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 776–777, 2020.

A Extended Related Work

Universal adversarial perturbations. That adversarial examples transfer across models trained with different hyperparameters or data was already observed by Szegedy et al. (2013), and Liu et al. (2016) studied transfer across architectures explicitly, distinguishing targeted from untargeted attacks and introducing the ensemble-based construction whose spirit our multi-model optimisation shares. The universal setting—one perturbation for all images—was established by Moosavi-Dezfooli et al. (2017) and Shafahi et al. (2020) for classification and by Hendrik Metzen et al. (2017) for segmentation. Subsequent work sharpened what a UAP may specify: Zhang et al. (2020a) identified shared class-specific structure across universal perturbations, and Benz et al. (2020) constrained a UAP to move one source class to one sink class while leaving others intact. This literature is almost entirely untargeted or top-1 targeted, and reports targeted universality as difficult to obtain at all (Liu et al., 2016; Zhang et al., 2020b; Weng et al., 2023).

Ordered top-K attacks. Dictating the exact content *and* order of the top-K predictions was posed by Adversarial Distillation (Zhang & Wu, 2020), which designs a target distribution over classes and minimises a KL divergence to it, and by QuadAttacK (Paniagua et al., 2023), which solves a quadratic program in the feature embedding space so that the ordered-logit inequalities are enforced explicitly. RisingAttacK (Paniagua et al., 2025) attains the strongest per-image results in this family, building perturbations from the right singular vectors of the adversarial Jacobian, and the ordered extension of the Carlini–Wagner hinge (Carlini & Wagner, 2016) supplies a fourth objective with explicit consecutive-pair ordering terms. All four are single-model, single-image methods. Three we use unchanged, as instruments spanning a range of ordering explicitness. RisingAttacK is an inner solver rather than a loss, so it has no per-example objective to average over images and models, and we lift it to the universal setting by stacking the per-model Jacobians into one system and treating the distance to the solver's solution as the outer objective (Appendix C.1). We do not contribute a new loss: the finding is that this design axis, and not the attacked architecture, is what determines the ordering term of Eq. 3, and Set-then-Order (§6) sequences two existing losses rather than adding a new one.

The top-K objective has also been taken in a universal direction before: Tursynbek et al. (2022) construct top-K perturbations from a geometry-inspired construction. That work is *untargeted*—it displaces the true class out of the top-K rather than installing a prescribed list—so it does not face

the ordering constraint that Eq. 3 factors. We accordingly scope our setting as the *ordered, targeted, multi-model* regime, and make no claim of priority over top-K perturbations as such.

The two literatures therefore leave a gap on the diagonal. Ordered top-K work is targeted and ordered but per-image and single-model; universal top-K work is multi-image but untargeted and unordered. Neither asks whether one fixed perturbation can impose the *same prescribed ranking* on many images and many models at once, which is the setting this paper studies and, to our knowledge, the first time it is posed.

What prior work measures, and what it cannot see. Across both literatures, success is reported as a single rate: the fraction of inputs on which the attack achieves its objective. For an ordered top-K attack that rate is the product of Eq. 3, and for a multi-model attack it is the intersection of Eq. 4. To our knowledge no prior study reports the factors separately, and consequently none can distinguish an attack that places the target classes but misorders them from one that never places them, or an attack that is weak on every model from one that is strong on each model but on disjoint image sets. Section 5.2 shows the consequence concretely: an objective that is the strongest we measure on both constituents of universality appears, under a single ordered rate, to collapse entirely.

Why perturbations transfer. A substantial line of work improves transfer by modifying the optimisation— input diversity (Xie et al., 2019), momentum (Dong et al., 2018), translation invariance (Dong et al., 2019), and ensemble training (Liu et al., 2016)—generally evaluated by whether the transferred success rate rises. Our $\bar{\phi}$ (Eq. 5) is a measurement of the quantity those methods implicitly target: the extent to which per-model successes co-occur on the same inputs. Reporting it separately from per-model strength is what allows §5.2 to separate how much of the gap between per-model and joint success is compounding of per-model strength and how much is genuine cross-model dependence, and to show that under transfer to unseen models the agreement structure is nearly objective-independent, so that a method which improves transfer there must be improving per-model strength instead.

Robustness and the models being attacked. Adversarial training remains the reference defence, and we attack the standard publicly released robust ImageNet backbones (Engstrom et al., 2019; Salman et al., 2020) catalogued by RobustBench (Croce et al., 2020). Separately, the modern model landscape has moved toward language-supervised and self-supervised pretraining at scale (Radford et al., 2021; Fang et al., 2023; Siméoni et al., 2025), which prior UAP work— evaluated almost exclusively on convolutional ImageNet classifiers—does not cover. Section 5.3 attacks both families under one protocol, and finds that they resist in different ways: adversarial training suppresses ordered top-K universality by removing membership, whereas foundation-scale pretraining leaves the model trivially attackable in the white box while making it the least reachable target of a shared perturbation.

Ranked outputs as the object of attack. The prevailing evaluation convention in both literatures above is top-1: an attack succeeds if the argmax changes, or becomes the target. For systems whose downstream behaviour is driven by a ranked list rather than a single label, that convention understates the attack surface in two directions. It credits no additional capability to an adversary who can dictate positions two through K, and it gives no failure signal when the argmax is untouched but the remainder of the ranking has been rewritten. Ordered top-K attacks make both explicit by construction, which is the reason we treat the ordering constraint as a first-class axis of difficulty rather than as a stricter success threshold.

B Measurement Protocol

Models. We attack 28 distinct ImageNet-1k networks arranged into 19 training configurations. The largest is an **18-model ensemble** chosen to span the major architecture families and capacities— 13 convolutional (ResNet-18/34/50 ($\times$2 recipes)/101, DenseNet-121/161/169/201, HRNet-W18, ConvNeXt-T/S/B), 4 transformer (DeiT-S, DeiT3-S, DeiT3-M, ViT-B/16), and 1 MLP-style (Mixer-B/16)—so that a single δ must satisfy a broad function class rather than one inductive bias. Nested subsets of this ensemble give the ensemble-size axis: $N = 1$ (R50, D121, ViT-B, DeiT-S individually), $N = 2$ (R50+D121, ViT-B+DeiT-S), $N = 4$ (all four), and $N = 18$. Transfer is measured on **7 held-out models** never seen during UAP learning: Swin-B, HRNet-W48, ConvMixer-768, ViT-B/CLIP, EVA-Giant/CLIP, ConvNeXtV2-L, and DINOv3-7B—a set chosen to span scale and pretraining objective (supervised, vision–language, and self-supervised). The same seven are additionally attacked *as white-box single-model targets*, which is what lets §5.3 compare white-box depth against

transfer resistance on the same networks. Finally, three adversarially-trained backbones (two robust ResNet-50s and a robust WideResNet-50) are attacked as single models. The full roster is in Appendix E.

Data. UAPs are learned from 5000 ImageNet-1k `train` images, five per class, restricted to images correctly classified by at least 11 of a 15-model reference ensemble so that the attack is not credited for images the models already get wrong. *This threshold is a filter on images; it is not a success criterion*—success is always measured over the full attacked model set, as defined in §3. All reported numbers are on held-out `val` images, unseen both during UAP learning and during the networks' own training, drawn from two sets that we name per table: a *re-scoring set* of 5 images per class (5000 images, class-ordered; it keeps the five images of class t_1, so at $K{=}1$ a correctly classified one is a free success, at most 0.1%), used for every decomposition and agreement table, Fig. 2 and the augmentation table; and the *full* split minus the images of class t_1 (49,950 images), used for the single-model K-frontier (Tables 3 and 17), which needs K up to 10. On the 1,515 runs scored on both, per-model ordered ASR agrees to 0.7 percentage points on average (median 0.3), so the two sets are interchangeable at the precision we report.

Budget, and what it does and does not bound. The benchmark is run at one fixed ℓ_2 budget, $\|\delta\|_2 \leq 24.34$. Fixing it is deliberate: the location of the K-frontier moves with the budget, so numbers gathered at different budgets are not comparable, and a study that varies budget implicitly along with K and N cannot attribute an effect to any of them. Fixing it makes all 3,180 benchmark runs mutually comparable. Budget dependence is measured separately, by a controlled half-budget replication ($\|\delta\|_2 \leq 12.17$: 65 runs on the four-model ensemble and 120 on two single models; Appendix C.5), and its limits are stated in §7.

The value 24.34 is the ℓ_2 radius of a fully saturated $\ell_\infty = 16/255$ ball at this input resolution, i.e. the two describe the same total perturbation energy. **They do not describe the same constraint**, and we are explicit about this because it is easy to misread. An ℓ_2-only constraint permits the optimiser to concentrate energy in few pixels, and it does. Measured on the perturbation the models actually receive, $\mathrm{Clamp}(x+\delta) - x$, over 1,000 held-out `val` images (one per class) for all 3,180 final-epoch UAPs, clipping keeps the ℓ_2 norm just inside the budget (mean 23.3; no run's mean above 23.6) while the mean ℓ_∞ is $103/255$ and the largest on any image $223/255$, with 99.9% of runs averaging above $16/255$. energy budget, not a per-pixel imperceptibility guarantee, and the perturbations are visible on inspection. We report ℓ_∞ alongside ℓ_2 throughout rather than describing the budget in ℓ_∞-equivalent terms. An ℓ_∞-projected variant of this attack is materially weaker: enabling a per-step ℓ_∞ projection makes it, not the ℓ_2 budget, the binding constraint and costs 27.6 ASR points on average across paired runs.

Targets. Target lists T are sampled uniformly at random, not chosen for semantic coherence, since coherent lists are markedly easier and would flatter the attack. We discard images whose ground-truth label equals t_1. We do not otherwise filter T against benign predictions; instead we measured, post hoc, how often a benign ordered top-K prediction already equals T on the evaluated images: at most 0.03% per model at $K{=}1$ (0.05% for the union over the four-model ensemble) and zero at every $K \geq 2$, across 32 backbones and all five seeds. No run starts materially pre-satisfied.

Attack objectives as instruments. We use four existing objectives, chosen because they span a range in *how explicitly the loss encodes order*—the design axis Eq. 3 predicts should matter:

- **CW-K** — hardest ordering. For $K > 1$ the loss is $\sum_{i=1}^{K-1} \max(z_{t_{i+1}} - z_{t_i}, 0) + \max(\max_{c \notin T} z_c - z_{t_K}, 0)$: $K{-}1$ explicit consecutive-pair hinges plus a boundary hinge.
- **QP** (AllQuadAttacK, Paniagua et al., 2023) and **RisingAttacK** — ordering appears as explicit inequality constraints, solved as a quadratic program and as a sequential QP over a Jacobian subspace respectively.
- **KL** (AllKLAttacK, after Zhang & Wu, 2020) — softest ordering. It constructs a target distribution in which the targets receive descending logits separated by a fixed margin (0.3) and minimises the KL divergence to it. Order is thus expressed as a small preference inside a matched distribution rather than as a constraint.

None of these objectives is contributed here. Three are used exactly as published, held fixed as instruments so that the ordering term of Eq. 3 can be read off as a function of loss design. The fourth could not be used as published in this setting, for the structural reason given next.

Lifting RisingAttacK. The construction is given in Appendix C.1; the constrained system it stacks is the same ordered top-K system as the QP objective (`constraint_builder`), and the inner solver runs to the same tolerance and iteration cap as the published per-image attack.

Optimization. Each UAP is optimised for 50 epochs. **No data augmentation is used when learning a UAP**: the training loader applies only a deterministic resize (no random crop, flip or colour transform), and no transformation is applied to $x + \delta$ during optimisation (no expectation over transformations (Athalye & Sutskever, 2017)). The input corruptions of §5.3 and Appendix G are applied only when a learned perturbation is evaluated. For the 18-model ensemble a mini-batch of 4 of the 18 models is sampled per gradient step, which keeps the step cost independent of N; joint success is nonetheless always evaluated over all 18. Sampling models stochastically within an ensemble attack is a standard implementation device rather than a contribution of ours, and we treat it as one. Training is data-parallel with 8 images per process, so the global batch is 8 times the number of GPUs. That number is 8 or 10 for nearly every run and fixed within most configurations; a few cells mix 6 and 8 (single ResNet-50, DenseNet-121, ViT-B/16, DeiT-S) or 9 and 10 (the adversarially trained models), i.e. global batches of 48–80. Every Set-then-Order comparison pairs runs with the same number of GPUs (§6). Each (configuration, objective, K) cell is repeated over up to 5 random seeds, and we report mean $\pm$ standard deviation across seeds throughout. **We report the perturbation at the final epoch, never the epoch with the highest held-out success.** The latter is a selection made on the evaluation set, and it is not small: across the 1,515 runs where both are available it adds 3.2 percentage points to per-model ordered ASR on average, 4.3 at $K \in \{2, 3\}$, 10.8 on DINOv3-7B, and up to 46 on individual runs whose success oscillates from epoch to epoch (Appendix I).

Instrumentation. To measure the quantities in Eq. 5, each learned δ is scored in a dedicated pass, without augmentation, and for every (model, image) pair we record both indicators $\mathbf{1}[S_\Theta(x)]$ and $\mathbf{1}[O_\Theta(x)]$ computed from the *same* top-K logits. Storing the $N \times 5000$ indicator matrices rather than their means makes joint success over any model subset, and the pairwise correlations $\bar{\phi}$, computable offline without re-learning anything. Because indicator vectors from different runs are only comparable index-by-index, each matrix is stored with the ground-truth label sequence of its evaluation loader as an alignment fingerprint, and the aggregation refuses to combine runs whose fingerprints disagree. We scored 1,516 runs this way; all fingerprints matched, and the joint rates recomputed from the indicator matrices reproduce the independently accumulated scalars to within 10^{-6} percentage points.

Scale. The study comprises **3,180 completed UAPs** over the four objectives and 19 configurations: $K \in \{1, \ldots, 10\}$ for the single models and the two- and four-model ensembles, $K \leq 5$ for the 18-model ensemble, typically five seeds per cell. Table 17 gives the number of runs per configuration and marks the cells with fewer seeds; one further configuration, a separate ResNet-50 control with three objectives at $K \leq 3$, is not reported. All runs are at the single fixed budget and the single protocol above (5 images per class, 50 epochs). Only runs that completed all 50 epochs enter the study.

C Extended Diagnosis and Method Details

C.1 Lifting RisingAttacK to a Universal Perturbation

RisingAttacK, the strongest per-image ordered top-K attack, is not a loss. It is an inner sequential-QP solver that, for a *single* image and a *single* model, builds a perturbation inside the subspace spanned by the leading right singular vectors of that model's adversarial Jacobian. There is no per-example objective to average, and running it per image would return $|\mathcal{D}|$ perturbations rather than one. We lift it in two steps. First, the per-model Jacobians and their compressed logit vectors are *stacked* into one constrained system and the SVD is taken of the stacked operator, so the search subspace is one the attacked models *share* rather than any single model's own. Second, the converged inner iterate $x^\star$ becomes an outer objective on the shared perturbation: we minimise $\|x^\star - \mathrm{Clamp}(x + \delta)\|_2^2$, whose gradient pulls δ toward the solution the solver found. The stacked system is the same ordered top-K constraint system as the QP objective, and the inner solver runs to the same tolerance and iteration cap as the published per-image attack (Appendix B). We claim the lifting preserves the method, not that it is the best way to universalize it.

C.2 ORDERING (EXTENDS §5.1)

A controlled pair. On the 18-model ensemble at $K{=}2$, the no-order control PGD-CE and KL are a controlled pair. Their set rates are indistinguishable (73.0 ± 6.2 against 73.2 ± 5.3), so they place the target classes equally often; they differ by eleven points in $P(\mathrm{o}|\mathrm{s})$ (47.8 against 58.8), and the only relevant difference between them is that KL's target distribution carries a 0.3 logit margin between consecutive targets while PGD-CE's is flat. Holding membership fixed and varying only whether the loss says anything about order therefore moves the ordering term and nothing else, which is the cleanest form of the loss-design to behaviour link §5.1 claims. That the two objectives tie on membership at all is itself informative: two unrelated formulations saturating at the same 73% is what we would expect if membership at this budget is limited by the budget rather than by the loss (§C.5).

Varying the architecture. Table 19 tests whether that link survives when the architecture is varied instead of the ensemble. At $K{=}3$ across eleven single-model backbones spanning CNNs, transformers, and CLIP-, masked-image- and self-supervised-pretrained networks, the pattern is unanimous: **KL attains the highest set rate on 11 of 11 backbones and the lowest ordering rate on 11 of 11**, and CW-K attains the highest ordering rate on 10 of 11. Averaged over the eleven, KL leads on set membership by roughly twenty points (81 vs 61 for CW-K) while trailing on ordering by more than forty (42 vs 84). We report these as counts rather than as a fitted relationship: eleven backbones establish that the ordering is consistent, not the functional form of any trend. The single exception is EVA-Giant/CLIP, where QP (74) and RisingAttacK (72) edge out CW-K (70) on $P(\mathrm{o}|\mathrm{s})$. Its CW-K set rate is 20 ± 30 across seeds, i.e. the objective failed on some seeds entirely, so the conditional ratio there is estimated from very few successes and we do not read anything into the reversal.

The ensemble frontier. On the full held-out `val` split, the best objective's per-model ordered ASR at $K{=}3$ and $K{=}4$ is 41.4 and 18.6% on R50+D121, 47.9 and 14.9% on ViT-B+DeiT-S, 31.2 and 8.7% on the four-model ensemble, and 29.8 and 7.1% on the 18-model ensemble (final epoch, mean over seeds), so $K^\star{=}3$ for every ensemble (§4), against 3–8 for single models (Table 3).

Hence "strongest attack" is metric-dependent. Because the two factors move in opposite directions across objectives, there is no ordering-free answer to which attack is best. On the 18-model ensemble at $K{=}2$, KL has the highest joint *set* rate of any objective (36.4%) and the lowest joint *ordered* rate (0.2%). A benchmark reporting one ordered ASR cannot distinguish a strong set attacker, a strong ordering attacker, and an objective mediocre at both—and, as §6 shows, it would also hide that the first two can be combined.

Where ordering stops being the binding constraint. The decomposition has a boundary, and we state it explicitly because it bounds the claim. At $K \geq 4$ on the 18-model ensemble the *joint* set rate is itself approximately zero for every objective—the perturbation is not failing to arrange the targets, it is failing to place them in the top-K at all. In that regime $P(\mathrm{o}|\mathrm{s})$ is estimated from a vanishing denominator and carries no information about ordering (our estimator returns it as undefined below a minimum-support threshold; Appendix I). The ordering account of difficulty therefore applies at $K \in \{2, 3\}$ for the large ensemble, and further out in K for single models and small ensembles, where set successes remain plentiful (Table 19). Ordering explains the near-frontier behaviour; it does not explain the collapse itself.

C.3 AGREEMENT (EXTENDS §5.2)

Isolating agreement from ordering. At $K \geq 2$ the statistic r computed on ordered success conflates the two factors: a low joint-ordered rate can reflect disagreement between models or merely the ordering cost of §5.1. Table 20 therefore recomputes the same quantities on *set* success, which divides the ordering factor of Eq. 3 out by construction and is comparable across K. The result is a second, independent decay: r_{set} falls monotonically with K in every one of the sixteen (config, objective) cells—for the 18-model ensemble, from 0.82 to 0.50 to 0.19 for KL, and from 0.54 to 0.18 to 0.005 for QP. **Lengthening the target list costs cross-model agreement in its own right**, independently of any difficulty in arranging the targets. That is not visible in any ordered success rate, and it is a distinct mechanism from the ordering cost of §5.1.

Compounding versus dependence. Substituting each 18-model run's eighteen per-model rates into $\prod_\Theta \widehat{\Pr}[O_\Theta]$ at $K{=}1$ and averaging over seeds gives independence-predicted joint rates of 23.5% (KL), 22.0% (CW-K), 16.7% (RisingAttacK) and 5.5% (QP): an 18.0-point spread, against the

observed 30.9 (74.9, 68.9, 62.2, 44.0%; Table 2). The remaining amplification is dependence, which D measures (§5.2).

What $\bar{\phi}$ does and does not explain. The pairwise correlation $\bar{\phi}$ recovers the r ordering exactly at $N{=}4$ and $N{=}18$—the regimes where r actually separates objectives—but not everywhere, and we state the exceptions. At $N{=}2$ on R50+D121, RisingAttacK has the highest $\bar{\phi}$ (0.664) yet only the third-highest r, because at $N{=}2$ the surviving fraction is compressed near 1 and small differences in per-model level dominate. Under transfer, $\bar{\phi}$ is nearly constant across objectives (0.376–0.397) while r varies from 0.19 to 0.32; there the agreement *structure* among unseen models is essentially objective-independent, and the differences in joint transfer trace back to per-model strength instead. Agreement and strength therefore do not stand in a fixed relationship—which of the two governs depends on the regime, and the honest general statement is the one the trend in Table 2 supports: agreement's share of the outcome grows with the size and diversity of the model set, and is dominant by $N{=}18$ on attacked models.

C.4 Resistance (extends §5.3)

Adversarial training, in full. The lower block of Table 3 is qualitatively different. The three adversarially-trained backbones are already hard at $K{=}1$ (9–17% versus 92–98% for standard models) and are eliminated by $K{=}2$: R50-AT (Engstrom), R50-AT (Salman) and WRN-50-AT all fall below 1%, and all three are at 0.0% by $K{=}3$. None of the three reaches 20% at any K, so $K^\star$ is undefined for all of them. This is not the same effect as a shallower frontier. Every standard model retains a usable ordered top-K attack over several values of K; the adversarially-trained ones support no ordered list at all, and only a weak single-target one. Among the interventions we measure, adversarial training is the only one that suppresses universal ordered top-K attackability to this degree rather than merely reducing it — though under the attacks and budget tested, suppression is what we observe, not removal. The decomposition locates the failure: adversarial training removes *membership*. At $K{=}2$ the target pair enters the top-2 on at most 1.2% of images for every objective on all three robust backbones, against 39–81% on standard ResNet-50 (QP 39.2, RisingAttacK 41.0, CW-K 80.6, KL 81.0), so almost nothing is left to order. Most runs have fewer than the 25 set successes our estimator requires (Appendix I), so we pool over seeds: CW-K then orders only 21–58% of its few set successes on the three robust backbones, against 96% on standard ResNet-50.

White-box depth is not transfer resistance. The seven models attacked both ways separate the two axes, and the endpoints are opposite: DINOv3-7B is the deepest white-box target in the study ($K^\star{=}8$) and simultaneously the most transfer-resistant (29.0%), while HRNet-W48 has one of the shallowest frontiers ($K^\star{=}4$) and is the most transfer-susceptible (77.7%). Susceptibility to a *shared* perturbation is therefore governed by something other than how far a model's own logit geometry can be pushed. We resist converting this into a trend: Spearman's $\rho = -0.44$ is not meaningful at $n{=}7$ and the middle of the range is not monotone (Table 18, Appendix H).

What transfer resistance does track. Ordering the seven by transfer susceptibility recovers a recognisable gradient: conventionally supervised ImageNet models are fooled most (HRNet-W48 77.7, ConvMixer-768 70.5, Swin-B 69.3), models pretrained with language or masked-image supervision less (ViT-B/CLIP 62.8, ConvNeXtV2-L 60.1, EVA-Giant/CLIP 40.1), and the self-supervised foundation-scale model least (DINOv3-7B 29.0). The pretraining objective, not the architecture family, organises this ordering—ConvMixer and HRNet share neither architecture nor scale and sit together; scale alone does not, since ConvNeXtV2-L and EVA-Giant/CLIP are both large and sit twenty points apart. This is the sense in which our results speak to the concern that motivates the ordered top-K threat model: the models being deployed at the top of the stack, trained with language supervision or at foundation scale, are the ones a shared perturbation reaches least well, even though each is trivially attackable on its own.

Defence tiers. Across objectives the defences fall in the same four tiers, from flip and colour jitter (nearly free) to blur, so a defence's effect is a property of the corruption rather than of the attack (Appendix G).

Table 7: **Halving the budget costs cross-model agreement before it costs per-image strength.** Four-model ensemble, ordered ASR (%) on 5,000 held-out `val` images, final epoch, mean±std over five seeds. Only the budget differs between the two halves (24.34 and 12.17 are saturated ℓ_∞=16/255 and 8/255 balls). At K=1 the per-model rate retains 0.65, 0.64, 0.42 and 0.07 for KL, CW-K, QP and RisingAttacK, but the *joint* rate only 0.47, 0.43, 0.08 and 0.01. The split is ensembling, not loss design (Table 8). –: not run at the half budget.

		per-model ASR		joint ASR (all 4)	
K	objective	24.34	12.17	24.34	12.17
1	QP	71.7±6.2	29.9±4.1	40.2±15.1	3.4±2.9
	Rising	54.9±16.2	3.8±4.0	33.0±19.8	0.3±0.5
	KL	91.3±2.6	59.7±3.0	83.9±4.8	39.3±4.7
	CW-K	90.1±3.1	58.0±3.6	80.7±5.8	35.1±5.2
2	QP	41.1±16.7	4.2±0.4	7.7±7.9	0.0±0.0
	Rising	21.0±10.7	0.0±0.1	4.8±3.4	0.0±0.0
	KL	48.8±6.7	17.3±6.7	13.4±4.0	2.1±1.8
	CW-K	66.2±11.2	18.6±11.7	43.5±15.4	4.8±5.8
3	QP	18.0±7.0	1.6±2.9	0.0±0.0	0.0±0.0
	Rising	3.2±4.5	0.0	0.1±0.2	0.0
	KL	21.2±5.0	3.5±1.8	1.6±1.2	0.0±0.0
	CW-K	30.9±8.4	2.6±3.1	8.8±5.1	0.2±0.4
4	QP	3.5±3.3	–	0.0±0.0	–
	Rising	0.4±0.5	–	0.0±0.0	–
	KL	6.3±2.5	0.6±0.4	0.0±0.0	0.0±0.0
	CW-K	8.4±5.6	0.3±0.4	0.5±0.8	0.0±0.0
5	QP	0.9±0.8	–	0.0±0.0	–
	Rising	0.1±0.1	–	0.0±0.0	–
	KL	1.4±0.9	0.1	0.0±0.0	0.0
	CW-K	1.6±1.3	0.0	0.0±0.0	0.0

C.5 Resistance from the Budget Side: Halving ϵ

The budget itself is the third way to take the capability away, and unlike a corruption it does not act uniformly on the four objectives. We repeat the four-model ensemble at half the energy budget — $\|\delta\|_2 \leq 12.17$, the radius of a saturated ℓ_∞=8/255 ball — holding every other field fixed, so the budget is the only variable (65 runs; Table 7). Read across objectives, the cost is very unequal: at K=1 the per-model rate retains 0.65 and 0.64 of its value for KL and CW-K, 0.42 for QP, and 0.07 for RisingAttacK, which the halved budget removes as a universal ordered attack outright.

The more informative comparison is with the joint rate, because **agreement gives way before per-image strength does**. The same halving leaves joint success at 0.47 and 0.43 for KL and CW-K but at 0.08 for QP and 0.01 for RisingAttacK—so the agreement ratio r of Eq. 5 falls from 0.56 to 0.11 for QP and from 0.60 to 0.09 for RisingAttacK, while KL and CW-K fall only from 0.92 and 0.90 to 0.66 and 0.60. Tightening the budget therefore widens the spread in r across objectives in the same direction that enlarging the ensemble does (§5.2): both make agreement, rather than per-model power, the term that decides universal success. A geometric objective at half budget still fools each model on a respectable fraction of images; it stops fooling them on the *same* images.

This is an interaction, not a property of the loss. The obvious reading—that distribution-shaping objectives tolerate a tight budget and geometric ones do not—is available only because both measurements are made on an ensemble. Repeating the halving on two *single* models separates the explanations (Table 8): at N=1 QP retains 0.84 and 0.78 of its K=1 success, indistinguishable from KL (0.78, 0.77) and CW-K (0.82, 0.80), so its collapse to 0.42 is something ensembling does to it, not the budget. Only RisingAttacK is budget-sensitive as a loss ($\approx$ 0.50 at N=1, 0.07 at N=4). Going from one model to four costs KL and CW-K 0.12 and 0.17 of retention but QP 0.40 and RisingAttacK 0.43: at K=1, neither a tight budget nor a large model set removes the geometric objectives alone, only the conjunction. The interaction account is a K=1 statement. At K=2 and 3 the half budget costs every objective on single models as well, with no consistent ordering among

Table 8: **The geometric objectives' budget collapse is a property of ensembling, not of the loss.** Ordered ASR (%) at full and half budget and the retention ratio $12.17/24.34$, final epoch, five seeds per cell; $N{=}1$ on the full held-out `val` split, $N{=}4$ on the 5,000-image re-scoring set. At $K{=}1$ and $N{=}1$ three of the four objectives retain 0.77–0.84 and are indistinguishable; only RisingAttacK is budget-sensitive on its own (≈ 0.50). Ensembling then costs KL and CW-K 0.12 and 0.17 of retention but QP 0.40 and RisingAttacK 0.43: neither a tight budget nor a large model set breaks them alone.

		R50 ($N{=}1$)			ViT-B ($N{=}1$)			4-mix
K	objective	24.34	12.17	ret.	24.34	12.17	ret.	ret. ($N{=}4$)
1	QP	59.1	49.9	0.84	73.9	57.9	0.78	**0.42**
	Rising	64.3	32.4	**0.50**	73.7	36.0	**0.49**	**0.07**
	KL	92.1	71.4	0.78	97.1	75.0	0.77	0.65
	CW-K	92.0	75.3	0.82	94.3	75.4	0.80	0.64
2	QP	32.3	19.1	0.59	49.6	18.0	0.36	0.10
	Rising	32.6	7.8	0.24	43.5	9.9	0.23	0.00
	KL	49.0	25.3	0.52	58.5	26.5	0.45	0.36
	CW-K	77.8	40.9	0.53	72.2	26.7	0.37	0.28
3	QP	20.3	10.2	0.50	27.5	3.9	0.14	0.09
	Rising	20.9	1.3	0.06	20.7	0.2	0.01	0.00
	KL	21.7	9.7	0.45	32.3	6.8	0.21	0.17
	CW-K	55.5	13.3	0.24	26.2	0.7	0.03	0.08

them (on ViT-B at $K{=}3$ retention ranges from 0.01 for RisingAttacK and 0.03 for CW-K to 0.21 for KL; Table 8).

The ordering term behaves differently again. At $K{=}2$, $P(\text{order} \mid \text{set})$ falls with the budget for QP (85% to 63%), RisingAttacK (76% to 57%) and CW-K (93% to 84%) but is unchanged for KL (63% to 62%): the objective with the softest ordering signal is the one whose ordering is budget-invariant. The ranking among objectives is therefore a statement about an operating point—budget as well as metric—not a property of the objective. Two points on one configuration do not determine a curve, and we do not claim the exchange rate is constant.

C.6 SET-THEN-ORDER (EXTENDS §6)

What the diagnosis predicts, and what it does not. If the two phases supplied the two factors perfectly, StO would reach the ceiling $\Pr_{\text{KL}}[S] \cdot \Pr_{\text{CW-K}}[O \mid S]$, and its gain over CW-K would be CW-K's ordering rate applied to KL's membership *lead*. That ceiling is an upper bound, not a forecast: it assumes the membership KL buys in phase one survives phase two, which nothing in the losses guarantees, since the hinges act on the targets' relative logits and are indifferent to how the rest of the distribution is shaped. The quantity that decides the gain is therefore how much of KL's lead StO still holds at the end, which we report for every setting.

Selection without test leakage. T_s and α are hyperparameters, so we chose them on held-out images whose ground-truth class index is odd, in a first round on the four-model ensemble at $K \in \{2, 3\}$ (switch at epoch 25 or 35; fixed blends $\alpha \in \{0.3, 0.5\}$), and fixed the split before any result was inspected. The selected $T_s{=}35$ of 50 epochs is then *frozen* for every setting in Table 6, and every number there is on the even half, the single-objective baselines included. Because the batch is per process, each StO run is compared only with baselines trained on the same number of GPUs; the selection round, run on a different number of GPUs, is used for selection only.

Design variants. Two alternatives fail in ways that support the account. A fixed blend $0.7\,\text{KL} + 0.3\,\text{CW-K}$, run at matched batch, keeps nearly all of KL's membership (90% of the lead) but loses CW-K's ordering, and ends below CW-K (-1.3 points): sequencing, not mixing, is what combines the two factors. Resetting the optimizer's moment estimates at the switch removes a transient: without it, training success on Swin-B halves right after the switch ($41 \rightarrow 20\%$) while the new loss rises, consistent with AdamW moments calibrated on the smaller KL gradients; with it, success rises through the switch ($41 \rightarrow 55\%$). But the reset leaves the final result unchanged ($+6.8$ against $+6.7$ on the four-model ensemble; still no gain on Swin-B), so the transient is not what limits retention.

D LIMITATIONS IN FULL

Five bound what we can claim. *(i) Partial budget coverage.* The 3,180-run corpus is at a single energy budget, $\|\delta\|_2 \leq 24.34$, which is what makes those runs comparable. The halved-budget replication of the four-model ensemble and of two single models shows that the objective ranking moves with the budget and that the movement is an ensembling effect (Appendix C.5), but two points do not determine a curve, and we do not vary the budget at $N{=}18$, where agreement dominates and the effect would be largest. *(ii) Seven models for the dissociation.* Spearman's $\rho = -0.44$ at $n{=}7$ is not significant and EVA-Giant/CLIP is non-monotone; we claim the axes dissociate, not a relationship between them. *(iii) One training ensemble.* All transfer numbers come from perturbations trained on eighteen conventionally supervised ImageNet classifiers, so what we report as foundation-scale resistance is confounded with distance from that ensemble. *(iv) Set-then-Order is tested, not tuned, per setting.* Its schedule was selected once and frozen, so Table 6 measures transfer of one schedule rather than a per-setting optimum; with four or five seeds per setting, only the large gains are individually decisive. We explain *where* StO helps by retention of the first phase's membership, but we do not yet explain *why* retention is high on ViT-B/16 and low on Swin-B. StO combines two existing losses, and we claim it as the attack the diagnosis implies, not as a new loss. *(v) The ordering account has a range.* At $K \geq 4$ on the large ensemble the set term is itself near zero, so $P(\text{order} \mid \text{set})$ carries no information; the ordering account explains near-frontier behaviour, not the collapse.

E MODEL ROSTER

The 18-model training ensemble. Chosen to span architecture family and capacity, so that a single perturbation must satisfy a broad function class. All are `timm` (Wightman, 2019) checkpoints trained on ImageNet-1k (Russakovsky et al., 2015).

- **Convolutional (13).** ResNet (He et al., 2016): `resnet18.a1_in1k`, `resnet34.a1_in1k`, `resnet50.a1_in1k`, `resnet50.b1k_in1k`, `resnet101.a1_in1k`; DenseNet (Huang et al., 2017): `densenet121.ra_in1k`, `densenet161.tv_in1k`, `densenet169.tv_in1k`, `densenet201.tv_in1k`; HRNet (Wang et al., 2020): `hrnet_w18.ms_in1k`; ConvNeXt (Liu et al., 2022): `convnext_tiny.fb_in1k`, `convnext_small.fb_in1k`, `convnext_base.fb_in1k`
- **Transformer (4).** DeiT (Touvron et al., 2021): `deit_small_patch16_224.fb_in1k`; DeiT3 (Touvron et al., 2022): `deit3_small_patch16_224.fb_in1k`, `deit3_medium_patch16_224.fb_in1k`; ViT (Dosovitskiy et al., 2020): `vit_base_patch16_224.augreg_in1k`
- **MLP-style (1).** MLP-Mixer (Tolstikhin et al., 2021): `mixer_b16_224.goog_in21k_ft_in1k`

The ensemble-size axis uses nested subsets: $N{=}1$ is each of `resnet50.a1_in1k`, `densenet121.ra_in1k`, `vit_base_patch16_224.augreg_in1k` and `deit_small_patch16_224.fb_in1k` individually; $N{=}2$ is R50+D121 and ViT-B+DeiT-S; $N{=}4$ is all four. Every member of the axis is a subset of the 18, so membership is nested; note that composition, architectural diversity and per-model optimisation exposure still change along the axis, so N is not the only quantity varying.

The 7 held-out models. Never seen during UAP learning; attacked both as transfer targets of the 18-model perturbation and as white-box single-model targets. `swin_base_patch4_window7_224.ms_in1k` (Liu et al., 2021), `hrnet_w48_ssld.paddle_in1k` (Wang et al., 2020), `convmixer_768_32.in1k` (Trockman & Kolter, 2023), `vit_base_patch16_clip_224.openai_ft_in12k_in1k` (Radford et al., 2021), `eva_giant_patch14_224.clip_ft_in1k` (Fang et al., 2023), `convnextv2_large.fcmae_ft_in22k_in1k` (Woo et al., 2023), and `dinov3_vit7b16_imagenet` (Siméoni et al., 2025).

The 3 adversarially-trained models. Two robust ResNet-50s (Engstrom et al., 2019; Salman et al., 2020) and a robust WideResNet-50, as catalogued by RobustBench (Croce et al., 2020).

Table 9: **The quadratic program dominates the cost.** Median seconds per epoch, from training-log timestamps. QP costs 7–10× CW-K and KL and 2–3× RisingAttacK at every ensemble size. The 18-model row samples 4 of 18 models per gradient step, so its cost is not 18× the single-model cost.

ensemble size N	CW-K	Rising	QP	KL
1	18	50	122	18
2	15	46	107	14
4	25	72	182	25
18	86	280	831	86

F Computational Cost

Table 9 gives median seconds per epoch, measured from training-log timestamps. The quadratic program is the dominant cost at every ensemble size—831 s per epoch on the 18-model ensemble against 86 s for the hinge and distribution-matching objectives (7–10×) and 280 s for RisingAttacK (2–3×)—because each step solves a differentiable QP whose constraint matrix grows with the label set. This is the practical reason the study's grid is what it is: extending K or the ensemble is cheap for the hinge and distribution-matching objectives, moderate for RisingAttacK and expensive for QP, which dominates the cost of any future budget or ensemble sweep. The 18-model row samples 4 of 18 models per gradient step, so its cost is not 18× the single-model cost.

G Robustness Under Corruption

Table 13 repeats the decomposition for every configuration and every K in the corpus, and Table 14 gives it per defence at $K{=}2$. The pattern of Table 5 is not specific to the 18-model ensemble: at $K{=}2$ the defences keep 0–38% of the set term on every standard backbone and ensemble, but 61–88% of $P(\mathrm{o}|\mathrm{s})$. Single-model perturbations lose more membership than the 18-model one (1–50% of the $K{=}1$ set rate kept, against 67%), and those learned on Swin-B and ConvNeXtV2-L are almost entirely removed (about 1% kept). Grayscale costs the most ordering on 9 of the 11 standard backbones and on every ensemble. On the adversarially trained backbones, and at $K \geq 4$, the undefended set rates are too small for stable ratios.

Resistance from the input side: defender augmentations. The main-text results are evaluated on clean inputs. A deployed defender can also simply transform the input, so we re-score each 18-model perturbation ($K{=}1$–5) under seven common corruptions, each swept over a magnitude ladder, and report the ASR retained relative to the un-augmented case (Table 15). Two results. First, all four objectives place the seven defences in the *same four tiers*: horizontal flip and colour jitter are nearly free (95–98% retained), Gaussian noise and grayscale cost moderately (78–83%), JPEG and resize-pad more (57–69%; the two swap places within the tier for KL), and Gaussian blur is the strongest defence for every objective, roughly halving the attack (47–50%). These tiers are a $K{=}1$ result. Every defence removes a growing share as K grows, and JPEG and resize-pad overtake blur from $K{=}2$: at $K{=}5$ they leave 1–9% of the attack against 24–28% for blur (Table 15). A defence's effectiveness is therefore a property of the corruption rather than of the attack it faces, which is what makes the ordering useful to a defender: it does not need to be tuned to an assumed attacker. Second, retention differs across objectives by at most seven points in every column except resize-pad (twelve) while their un-augmented joint rates differ by thirty (Table 2), so the $K{=}1$ method ranking of §5.2 survives every corruption and the agreement result is not an artefact of evaluating on pristine inputs.

These ratios are averages over each corruption's magnitude ladder at $K{=}1$, where even the strongest corruption leaves roughly half the attack intact. That does not carry to longer lists: at $K \geq 2$ the same corruptions remove most of the attack (Table 4), because they break membership while leaving ordering largely intact (Tables 5 and 12).

The ordering also survives transfer. Re-scoring the same perturbations on the seven held-out models, under the same ladder, reproduces the same tiers with the four objectives again within four points of one another except under resize-pad (Table 16); a defender need not know which attack, or which attacked ensemble, it is facing. One defence does not carry over: resize-pad retains 57–69% of the attack on the models it was fitted to but 72–83% on unseen ones, so its strength is specific to those models, whereas JPEG and blur transfer intact.

Table 10: **Table 4 for all four objectives: single-model and four-model UAPs.** Per-model ordered ASR%, same operating points and protocol as Table 4, every K from 1 to 5 that was run; –: not evaluated. The single-model and four-model UAPs were also run at K=6–10, where every defended cell rounds to 0.0%.

UAP	K	objective	none	JPEG 70	JPEG 50	resize-pad	blur σ=0.95	grayscale
single-model UAP	1	CW-K	94.6	38.8	19.1	14.2	24.1	29.4
		Rising	82.8	23.8	11.2	9.5	15.4	21.2
		QP	83.2	26.6	12.8	9.4	14.1	19.4
		KL	96.7	44.6	23.0	21.0	29.4	39.6
	2	CW-K	77.7	12.6	2.8	3.3	5.3	10.3
		Rising	61.2	7.0	2.0	1.2	2.0	4.1
		QP	66.8	9.4	2.3	1.6	2.7	5.3
		KL	61.9	14.3	4.2	2.9	5.5	7.8
	3	CW-K	53.5	1.9	0.2	0.7	1.5	2.0
		Rising	40.3	1.3	0.2	0.2	0.4	0.4
		QP	42.8	2.0	0.3	0.2	0.6	0.9
		KL	34.7	3.6	0.5	0.5	1.0	2.5
	4	CW-K	31.9	0.2	0.0	0.1	0.4	0.5
		Rising	23.9	0.2	0.0	0.1	0.2	0.2
		QP	24.4	0.3	0.1	0.1	0.2	0.2
		KL	17.4	0.7	0.1	0.1	0.3	0.9
	5	CW-K	17.5	0.0	0.0	0.0	0.2	0.2
		Rising	14.9	0.0	0.0	0.0	0.0	0.0
		QP	14.4	0.0	0.0	0.0	0.1	0.1
		KL	8.0	0.1	0.0	0.0	0.0	0.1
4-model UAP	1	CW-K	90.1	60.0	42.8	33.6	46.6	40.0
		Rising	51.9	19.4	11.8	8.9	10.6	13.3
		QP	71.7	31.8	15.0	12.0	18.6	22.0
		KL	91.3	63.0	48.7	38.1	50.6	41.0
	2	CW-K	66.2	20.9	9.6	7.8	13.0	11.9
		Rising	18.5	2.7	1.1	0.8	1.2	2.1
		QP	41.1	10.5	2.9	1.9	6.8	5.3
		KL	48.8	20.5	11.7	8.6	12.7	8.1
	3	CW-K	30.9	0.5	0.1	0.5	0.6	1.1
		Rising	2.6	0.1	0.0	0.0	0.0	0.0
		QP	18.0	0.3	0.0	0.1	0.4	0.3
		KL	21.2	3.3	0.8	1.2	1.6	1.4
	4	CW-K	8.4	0.0	0.0	0.0	0.0	0.0
		Rising	0.4	0.0	0.0	0.0	0.0	0.0
		QP	3.5	0.0	0.0	0.0	0.0	0.1
		KL	6.3	0.3	0.0	0.1	0.1	0.2
	5	CW-K	1.5	0.0	0.0	0.0	0.0	0.0
		Rising	0.1	0.0	0.0	0.0	0.0	0.0
		QP	0.9	0.0	0.0	0.0	0.0	0.0
		KL	1.4	0.0	0.0	0.0	0.0	0.0

Table 15 scores each perturbation under corruption on the models it was optimised against. Table 16 repeats the measurement on the seven held-out models, using the same corruption ladder and the same final-epoch perturbations, so the two differ only in whether the scored model was attacked. The defence tiers are unchanged and the four objectives again agree to within four points per column except resize-pad, which is the transfer version of the claim in Appendix C.4: the ordering belongs to the corruption, not to the attack. Resize-pad is the one defence that does not carry over, and the un-augmented transfer rates in the *base* column (47–68%) are high enough that none of these ratios rests on a handful of successes.

H Supporting Tables

Table 19 is the per-backbone version of the ordering result of §5.1: across all eleven single-model backbones the distribution-matching objective attains the highest set rate and the lowest ordering rate, which is the claim the main text makes in aggregate. Table 20 recomputes the agreement statistics of

Table 11: **Table 4 for all four objectives: the 18-model UAP and transfer.** Per-model ordered ASR%, same operating points and protocol as Table 4, every K from 1 to 5 that was run; –: not evaluated. The 18-model UAP was run to K=5; on the unseen models the robustness evaluation covers K=1 (and K=2 for KL).

UAP	K	objective	none	JPEG 70	JPEG 50	resize-pad	blur σ=0.95	grayscale
18-model UAP	1	CW-K	91.3	74.7	63.9	62.5	57.6	56.9
		Rising	88.5	70.6	59.8	53.6	54.9	49.9
		QP	81.8	63.9	52.4	49.4	44.6	46.6
		KL	91.7	77.6	68.2	66.1	60.1	58.5
	2	CW-K	61.6	26.7	17.0	21.2	11.6	20.0
		Rising	54.9	27.1	18.5	19.2	12.4	15.5
		QP	42.2	20.2	13.5	11.8	11.4	13.6
		KL	43.4	23.8	17.0	19.1	12.0	15.7
	3	CW-K	29.3	2.9	1.0	4.5	1.7	5.5
		Rising	21.9	3.5	1.6	3.3	1.4	3.4
		QP	14.3	1.8	0.7	1.2	1.4	2.2
		KL	15.7	3.2	1.3	3.6	1.4	3.3
	4	CW-K	7.0	0.1	0.0	0.2	0.0	0.3
		Rising	4.1	0.1	0.0	0.2	0.0	0.3
		QP	3.0	0.1	0.0	0.0	0.1	0.1
		KL	4.6	0.4	0.1	0.5	0.2	0.5
	5	CW-K	0.9	0.0	0.0	0.0	0.0	0.0
		Rising	0.9	0.0	0.0	0.0	0.0	0.0
		QP	0.6	0.0	0.0	0.0	0.0	0.0
		KL	1.0	0.0	0.0	0.0	0.0	0.0
18-model UAP → 7 unseen	1	CW-K	62.8	43.6	34.5	50.8	40.5	36.1
		Rising	56.5	40.4	32.8	42.9	38.0	30.3
		QP	46.7	32.3	26.7	34.7	28.4	27.2
		KL	68.1	50.3	41.2	57.7	46.9	40.7
	2	KL	15.3	6.2	4.1	10.3	4.1	6.6
single UAP → 3 unseen	1	CW-K	4.0	1.8	1.2	1.7	1.1	0.5
		Rising	1.4	0.9	0.6	0.9	0.4	0.4
		QP	0.8	0.5	0.3	0.5	0.3	0.2
		KL	5.3	2.9	1.9	2.3	1.5	1.1
	2	CW-K	0.0	0.0	0.0	0.0	0.0	0.0
		Rising	0.0	0.0	0.0	0.0	0.0	0.0
		QP	0.0	0.0	0.0	0.0	0.0	0.0
		KL	0.2	0.0	0.0	0.0	0.0	0.0
	3	CW-K	0.0	0.0	0.0	0.0	0.0	0.0
		Rising	0.0	0.0	0.0	0.0	0.0	0.0
		QP	0.0	0.0	0.0	0.0	0.0	0.0
		KL	0.0	0.0	0.0	0.0	0.0	0.0
	4	CW-K	0.0	0.0	0.0	0.0	0.0	0.0
		Rising	0.0	0.0	0.0	0.0	0.0	0.0
		QP	0.0	0.0	0.0	0.0	0.0	0.0
		KL	0.0	0.0	0.0	0.0	0.0	0.0
	5	CW-K	0.0	0.0	0.0	0.0	0.0	0.0
		Rising	0.0	0.0	0.0	0.0	0.0	0.0
		QP	0.0	0.0	0.0	0.0	0.0	0.0
		KL	0.0	0.0	0.0	0.0	0.0	0.0

§5.2 on *set* success rather than ordered success, which divides the ordering factor out by construction and shows that cross-model agreement decays with K in its own right.

Table 18 places white-box depth and transfer resistance side by side for the seven models attacked both ways; §5.3 states its two endpoints and Appendix C.4 the Spearman value.

Table 17 extends Table 3 to every K from 1 to 10 and to the adversarially-trained backbones and the ensembles, with the number of completed runs behind each row.

Table 21 scores the same perturbations under four multi-model criteria. Joint success is the strictest; when the ensemble stands in for uncertainty about which model is deployed, the per-model and worst-model rates are the relevant ones, and a majority-vote system sits in between. KL's collapse at K=2 on the 18-model ensemble (§5.2) is specific to the joint criterion: a majority of the eighteen models emit the exact pair on 40.4% of images, against 0.2% for all eighteen; for CW-K the two are 63.8% and 15.2%.

Table 12: **Table 5 at every K.** Share (%) of the undefended set rate / of $P(\mathrm{o}|\mathrm{s})$ that survives each defence, 18-model perturbations, mean over the four objectives. "–": too few set successes to estimate $P(\mathrm{o}|\mathrm{s})$; values above 100 are sampling noise on few successes.

	K	JPEG 70	JPEG 50	resize-pad	blur σ=0.95	grayscale
attacked (18)	1	81 / 100	69 / 100	65 / 100	61 / 100	60 / 100
	2	51 / 94	35 / 92	37 / 88	25 / 97	44 / 64
	3	19 / 91	8 / 89	17 / 86	10 / 88	23 / 63
	4	9 / 54	3 / 41	8 / 81	3 / 81	9 / 68
	5	3 / 32	1 / 74	2 / 131	1 / 56	3 / 82
unseen (7)	1	71 / 100	58 / 100	79 / 100	65 / 100	57 / 100
	2	40 / 106	26 / 110	59 / 100	30 / 110	48 / 77
	3	14 / 90	7 / 77	48 / 85	11 / 85	43 / 67
	4	4 / 33	0 / 36	22 / 75	3 / 44	5 / 115
	5	0 / –	0 / –	7 / 0	0 / –	2 / 0

Table 13: **What the defences break, for every configuration.** Share (%) of the undefended set rate / of $P(\mathrm{o}|\mathrm{s})$ that survives, averaged over the five defences of Table 5 and the four objectives; attacked models (upper block) and held-out models (lower block: each single-model UAP on the other three 4-model members, the 18-model UAP on the seven unseen models). 5,000 held-out `val` images, $\|\delta\|_2 \leq 24.34$. At K=1 only the set term is defined. –: the undefended set rate is below 0.5% for every objective, or $P(\mathrm{o}|\mathrm{s})$ has fewer than 25 set successes; values above 100 are sampling noise on few successes. Per defence at K=2: Table 14.

UAP	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
ResNet-50	26	6 / 86	1 / 80	1 / 41	0 / 12	0 / 52	0 / –	0 / –	0 / –	0 / –
DenseNet-121	43	19 / 87	8 / 77	6 / 61	2 / 41	1 / 28	1 / 9	1 / 0	1 / –	0 / –
ViT-B/16	25	9 / 75	2 / 77	1 / 26	1 / 70	1 / 20	0 / –	0 / –	0 / –	0 / –
DeiT-S	16	10 / 61	4 / 44	2 / 34	0 / 16	0 / 0	0 / 0	0 / –	0 / –	0 / –
Swin-B	1	0 / 88	0 / –	0 / –	0 / –	0 / –	0 / –	0 / –	0 / –	0 / –
HRNet-W48	50	23 / 82	6 / 57	2 / 38	1 / 14	1 / 8	1 / 30	0 / –	0 / –	0 / –
ConvMixer-768	19	8 / 72	3 / 59	1 / 40	0 / 17	0 / 0	0 / –	0 / –	0 / –	0 / –
ViT-B/CLIP	28	5 / 68	1 / 46	0 / 17	0 / 0	0 / –	0 / –	0 / –	0 / –	0 / –
ConvNeXtV2-L	1	1 / 66	0 / 56	0 / 9	0 / –	0 / –	0 / –	0 / –	0 / –	0 / –
EVA-Giant/CLIP	25	12 / 67	2 / 42	1 / 54	0 / 0	0 / –	0 / –	0 / –	0 / –	0 / –
DINOv3-7B	39	19 / 64	10 / 71	7 / 53	4 / 42	2 / 28	0 / 86	0 / 48	0 / 0	0 / 0
R50-AT (Engstrom)	41	19 / 120	–	–	–	–	–	–	–	–
R50-AT (Salman)	47	–	–	–	–	–	–	–	–	–
WRN-50-AT	42	33 / 92	–	–	–	–	–	–	–	–
R50+D121	45	24 / 82	8 / 73	4 / 53	2 / 45	0 / 27	1 / 19	0 / 8	0 / –	0 / –
ViT-B+DeiT-S	22	15 / 65	4 / 75	1 / 47	0 / 33	0 / 0	0 / 0	0 / –	0 / –	0 / –
4-model	39	20 / 82	5 / 72	1 / 61	1 / 46	0 / –	0 / –	–	–	–
18-model	67	38 / 87	15 / 83	6 / 65	2 / 74	–	–	–	–	–
ResNet-50 → 3	21	–	–	–	–	–	–	–	–	–
DenseNet-121 → 3	47	–	–	–	–	–	–	–	–	–
ViT-B/16 → 3	21	4 / 33	–	–	–	–	–	–	–	–
DeiT-S → 3	19	4 / 83	–	–	–	–	–	–	–	–
18-model → 7	66	41 / 100	25 / 81	9 / 75	–	–	–	–	–	–

I Estimator Details

Why a strength-controlled dependence statistic. Under full independence with equal marginals $p_\Theta \equiv \Pr[O_\Theta] = p$ for every model, the joint rate is p^N and so $r = p^{N-1}$: retention falls with N and with attack *weakness* even when the models are perfectly independent, which confounds the quantity of interest with per-model strength. The statistic D of Eq. 6 instead places the observed joint rate on the interval its own marginals allow: its lower endpoint $\prod_\Theta \widehat{\Pr}[O_\Theta]$ is the independent joint rate and its upper endpoint $\min_\Theta \widehat{\Pr}[O_\Theta]$ is the largest joint rate those marginals permit (every model's successes nested inside the weakest model's). So $D = 0$ means independence, $D = 1$ maximal co-occurrence, and—unlike r—two objectives with different per-model rates are comparable in D. We report r as well because it is the quantity an attacker experiences (what fraction of per-model success is actually usable), and $\bar{\phi}$ because it is pairwise and therefore insensitive to N; where the three disagree we say so.

Table 14: **Table 13 per defence, at K=2.** Share (%) of the undefended set rate / of $P(\mathrm{o}|\mathrm{s})$ that survives each defence, mean over the four objectives; same conventions.

UAP	JPEG 70	JPEG 50	resize-pad	blur σ=0.95	grayscale
ResNet-50	12 / 93	2 / 70	6 / 90	1 / 81	11 / 101
DenseNet-121	37 / 98	16 / 92	5 / 83	7 / 88	28 / 76
ViT-B/16	27 / 84	7 / 73	1 / 67	8 / 82	1 / 64
DeiT-S	29 / 69	9 / 58	1 / 77	6 / 64	6 / 42
Swin-B	0 / –	0 / –	0 / 113	0 / 99	1 / 77
HRNet-W48	48 / 92	16 / 86	9 / 82	7 / 94	35 / 57
ConvMixer-768	17 / 74	3 / 73	3 / 81	4 / 84	15 / 48
ViT-B/CLIP	7 / 76	1 / 82	1 / 65	3 / 65	12 / 52
ConvNeXtV2-L	0 / –	0 / –	2 / 67	0 / 89	1 / 65
EVA-Giant/CLIP	23 / 82	5 / 67	3 / 60	3 / 68	27 / 56
DINOv3-7B	8 / 62	0 / 43	10 / 63	37 / 85	38 / 57
R50-AT (Engstrom)	43 / 150	36 / 149	1 / –	4 / –	10 / 30
R50-AT (Salman)	–	–	–	–	–
WRN-50-AT	54 / 116	42 / 122	2 / –	0 / –	64 / 37
R50+D121	41 / 88	20 / 86	16 / 82	4 / 94	40 / 58
ViT-B+DeiT-S	31 / 70	10 / 52	4 / 74	18 / 80	10 / 48
4-model	31 / 87	15 / 85	11 / 84	19 / 94	24 / 60
18-model	51 / 94	35 / 92	37 / 88	25 / 97	44 / 64
ResNet-50 $\rightarrow$ 3	–	–	–	–	–
DenseNet-121 $\rightarrow$ 3	–	–	–	–	–
ViT-B/16 $\rightarrow$ 3	13 / 33	0 / –	2 / –	5 / –	2 / –
DeiT-S $\rightarrow$ 3	6 / 49	2 / –	2 / –	10 / 117	0 / –
18-model $\rightarrow$ 7	40 / 106	26 / 110	59 / 100	30 / 110	48 / 77

Minimum support for the order-conditional term. $P(\mathrm{order} \mid \mathrm{set})$ is estimated per model as $\widehat{\Pr}[O_\Theta]/\widehat{\Pr}[S_\Theta]$ and averaged over models. The ratio is unstable when a model has almost no set successes—a single set-success that happens to be correctly ordered yields 100%—which occurs on the adversarially-trained backbones at $K \geq 3$, where the set rate is approximately zero. We therefore require at least 25 set-successes out of the 5000 evaluation images (0.5%) for a model to contribute, and report the term as undefined otherwise. Without this guard the adversarially-trained rows report spurious near-perfect ordering computed from one or two images.

Why not the independence-excess ratio. The natural way to express dependence is $\widehat{\Pr}[J_\mathcal{M}]/\prod_\Theta \widehat{\Pr}[O_\Theta]$. At N=18 the denominator is a product of eighteen probabilities; it underflows, and the ratio exceeds 10^7 in cells whose joint rate is a fraction of a percent, while being undefined wherever any single model contributes zero. The statistics of Eq. 5 are used instead: r is bounded in $[0, 1]$ and $\bar{\phi}$ is pairwise, so neither degrades as N grows.

Alignment of success indicators. Per-image success vectors from different runs are comparable only if they index the same images in the same order. Each stored indicator matrix carries the ground-truth label sequence of its evaluation loader as a fingerprint, and the aggregation refuses to combine runs whose fingerprints differ. Across the 1,516 re-scored runs every fingerprint matched (5 images per class, 1000 classes, in class order), and joint rates recomputed from the indicator matrices reproduce the independently accumulated scalar rates to within 10^{-6} percentage points.

Completion criterion. A run enters the study only if it completed all 50 epochs under the protocol of Appendix B (5 images per class). Incomplete runs are excluded rather than reported at whatever epoch they reached.

Checkpoint. Every number is the perturbation at the final epoch. We also record, for each run, the epoch with the highest held-out success rate, but do not report it, because it is selected on the evaluation set. On the 1,515 runs for which both are available the best-epoch value exceeds the final-epoch value by 3.2 percentage points on average (median 1.1), by 4.3 at $K \in \{2, 3\}$, by 10.8 on DINOv3-7B, and by up to 46 on single runs of the QP and RisingAttacK objectives, whose held-out

Table 15: **The defence ordering is a property of the corruption, not of the attack.** 18-model ensemble, $K{=}1,\ldots,5$, 5,000 held-out `val` images. Cells are ASR retained relative to the un-augmented perturbation (%), averaged over the magnitude ladder and seeds. All four objectives place the seven defences in the same four tiers (flip/colour, noise/grayscale, JPEG/resize-pad, blur) at $K{=}1$; at larger K the retained share falls for every defence, and at $K \geq 4$ it rests on a small undefended rate. Single eval config (n5-a6c093); mixing configs would compare different ladders.

K	objective	h-flip	colour	noise	gray	JPEG	resize-pad	blur
1	CW-K	97	96	83	81	65	65	49
	Rising	96	95	83	78	64	58	49
	QP	95	95	82	78	61	57	47
	KL	98	96	83	82	68	69	50
2	CW-K	89	90	70	66	36	32	36
	Rising	88	89	72	63	39	32	37
	QP	86	89	72	64	37	26	37
	KL	92	91	72	68	43	41	39
3	CW-K	79	84	62	58	16	15	31
	Rising	77	83	64	57	19	13	31
	QP	78	83	64	56	17	8	32
	KL	86	86	65	60	22	21	33
4	CW-K	67	78	52	52	6	3	27
	Rising	68	80	54	53	8	4	28
	QP	68	79	59	51	11	2	29
	KL	77	81	60	55	14	10	29
5	CW-K	57	74	40	50	2	1	24
	Rising	62	77	52	51	6	2	26
	QP	61	74	55	50	9	1	28
	KL	70	78	53	51	7	5	27

Table 16: **The defence ordering survives transfer, but resize-pad does not.** The 18model perturbation scored on the 7 *held-out* models it never saw, $K{=}1$, 5,000 `val` images, final epoch. Cells are ASR retained relative to each run's own un-augmented *transfer* ASR (%), averaged over the magnitude ladder and seeds; *base* is that un-augmented rate. The four objectives agree within a few points per column, as on attacked models (Table 15). The exception is **resize-pad**: strong on the attacked models (≈65% retained), weak under transfer (72–83%), so its effect is specific to the models fitted against. Single eval config (n5-fe60f9); no run fell below the 5% baseline guard.

objective	n	base	h-flip	colour	noise	gray	JPEG	resize-pad	blur
QP	5	47	99	96	81	78	56	72	49
Rising	5	56	100	96	82	76	57	74	51
KL	5	68	100	96	81	79	59	83	51
CW-K	5	63	100	96	81	78	55	79	50

success oscillates from epoch to epoch on the large backbones; CW-K, KL, and the ensembles are stable. For 665 runs in the high-K tail and on the adversarially-trained backbones, per-epoch validation stops once the attack has failed, so only the best-epoch value is available; since the final-epoch value cannot exceed it, we use it as an upper bound for those runs (599 of them are exactly 0, 64 are below 0.5 points, and the remaining two are 1.4 and 4.1).

Table 17: **The full K-frontier.** Per-model ordered ASR%, best over the four objectives, final epoch, full held-out `val` split, $\|\delta\|_2 \leq 24.34$, for every configuration and K in the corpus. *runs*: completed UAPs over all K (a full cell is 20: four objectives $\times$ five seeds); †: fewer than 20 runs at that K; –: not run. The 18-model ensemble was run to K=5.

configuration	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	$K^\star$	runs
ResNet-50	92	78	55	32	13	6	1	0	0	0	4	208
DenseNet-121	97	74	52	31	13	5	2	1†	1†	0†	4	172
ViT-B/16	97	72	32	16	5	1†	0†	0†	0†	0†	3	145
DeiT-S	97	81	64	26	9	3	1†	0†	0†	0†	4	155
Swin-B	98	70	43	43	21	4	5	6	3	0	5	230
HRNet-W48	95	70	44	22	10	5	2	0	0	0	4	200
ConvMixer-768	96	82	61	38	26	19	11	5	3	1	5	200
ViT-B/CLIP	98	82	67	49	31	23	14	10	6	2	6	200
ConvNeXtV2-L	98	95	83	62	42	25	13	6	4	1	6	200
EVA-Giant/CLIP	97	74	36	18	4	2	1	0	0	0	3	200
DINOv3-7B	98	89	78	61	51	46	33†	20†	19†	8†	8	184
R50-AT (Engstrom)	17	0	0	0	0	0	0	0	0	0	*none*	200
R50-AT (Salman)	9	0†	0†	0†	0†	0†	0†	0†	0†	0†	*none*	130
WRN-50-AT	11	0†	0†	0†	–	–	–	0†	0†	–	*none*	44
R50+D121	93	71	41	19	9	3	2	1	0	0	3	205
ViT-B+DeiT-S	96	76	48	15	3	1	0	0	0	0	3	205
4-model	92	67	31	9	2†	0†	0†	0†	0†	0†	3	158
18-model	92	62	30†	7	1	–	–	–	–	–	3	99

Table 18: **White-box depth and transfer resistance are dissociated.** The seven held-out models, each attacked *both* as a white-box target and by the 18-model UAP it never saw. $K^\star$ and white-box ASR are the final values of Table 3; transfer ASR is that UAP at K=1 on the 5,000-image re-scoring set, mean over objectives. Spearman $\rho = -0.44$ ($n = 7$): at this sample size we report a dissociation, not a trend.

held-out model	white-box $K^\star$	white-box ASR (K=1)	transfer ASR (K=1)
HRNet-W48	4	95	77.7
ConvMixer-768	5	96	70.5
Swin-B	5	98	69.3
ViT-B/CLIP	6	98	62.8
ConvNeXtV2-L	6	98	60.1
EVA-Giant/CLIP	3	97	40.1
DINOv3-7B	8	98	29.0

Table 19: **The objective ranking on ordering is consistent across the tested architectures.** Single-model attacks at $K = 2, \ldots, 5$ (every K re-scored on the 5,000-image held-out set), set ASR% and $P(\text{o}|\text{s})$%; –: $P(\text{o}|\text{s})$ undefined (fewer than 25 set successes). At K=3, across all eleven backbones the distribution-matching objective (KL) attains the *highest* set rate and the *lowest* ordering rate; at K=2 the same holds on all eleven and CW-K orders best on all eleven, and at K=4, 5, where set rates fall, KL still places best on 10 of 11 but CW-K orders best on only 6 and 7. The four objectives differ in more than how explicitly they encode order — KL's fixed 0.3 target-logit margin is one alternative explanation we do not separate — so this is a consistent ranking across architectures, not an isolated mechanism.

		CW-K		Rising		QP		KL	
K	model	set	$P(\text{o}\|\text{s})$	set	$P(\text{o}\|\text{s})$	set	$P(\text{o}\|\text{s})$	set	$P(\text{o}\|\text{s})$
2	ResNet-50	81 ± 5	96 ± 2	41 ± 11	77 ± 11	39 ± 20	80 ± 8	81 ± 4	60 ± 6
	DenseNet-121	77 ± 5	94 ± 2	70 ± 11	91 ± 5	70 ± 9	90 ± 6	86 ± 4	65 ± 7
	ViT-B/16	75 ± 13	95 ± 1	51 ± 13	84 ± 5	56 ± 16	87 ± 9	91 ± 2	64 ± 6
	DeiT-S	84 ± 5	95 ± 3	51 ± 21	87 ± 7	77 ± 10	92 ± 6	92 ± 4	78 ± 12
	Swin-B	76 ± 10	90 ± 9	53 ± 28	78 ± 16	66 ± 13	88 ± 5	91 ± 6	70 ± 12
	HRNet-W48	82 ± 9	94 ± 4	82 ± 8	94 ± 5	84 ± 7	93 ± 6	91 ± 3	64 ± 8
	ConvMixer-768	85 ± 4	96 ± 3	82 ± 4	95 ± 2	80 ± 8	95 ± 3	92 ± 3	66 ± 6
	ViT-B/CLIP	85 ± 5	97 ± 2	77 ± 3	94 ± 2	78 ± 10	95 ± 5	89 ± 4	64 ± 5
	ConvNeXtV2-L	95 ± 2	99 ± 1	81 ± 6	91 ± 4	90 ± 1	99 ± 2	95 ± 2	81 ± 4
	EVA-Giant/CLIP	64 ± 37	95 ± 4	80 ± 10	91 ± 3	79 ± 11	92 ± 10	87 ± 4	71 ± 7
	DINOv3-7B	90 ± 3	97 ± 1	54 ± 7	96 ± 3	71 ± 13	91 ± 6	95 ± 2	69 ± 6
	mean	**81**	**95**	**66**	**89**	**72**	**91**	**90**	**68**
3	ResNet-50	64 ± 6	86 ± 4	34 ± 13	58 ± 15	33 ± 12	57 ± 14	76 ± 3	28 ± 3
	DenseNet-121	60 ± 4	84 ± 4	42 ± 15	72 ± 11	59 ± 8	73 ± 5	76 ± 4	30 ± 6
	ViT-B/16	36 ± 8	72 ± 10	34 ± 8	60 ± 21	41 ± 10	67 ± 19	82 ± 3	39 ± 7
	DeiT-S	73 ± 13	87 ± 4	36 ± 10	62 ± 8	60 ± 13	77 ± 16	83 ± 9	51 ± 7
	Swin-B	56 ± 6	75 ± 10	39 ± 22	54 ± 20	48 ± 17	61 ± 19	80 ± 7	47 ± 4
	HRNet-W48	57 ± 11	80 ± 12	67 ± 9	77 ± 5	68 ± 11	75 ± 12	78 ± 7	32 ± 9
	ConvMixer-768	68 ± 7	88 ± 3	69 ± 9	84 ± 5	63 ± 9	82 ± 9	84 ± 6	43 ± 8
	ViT-B/CLIP	73 ± 6	91 ± 2	70 ± 7	88 ± 4	67 ± 8	87 ± 8	84 ± 5	38 ± 12
	ConvNeXtV2-L	86 ± 4	95 ± 4	78 ± 8	86 ± 9	73 ± 11	93 ± 6	93 ± 3	55 ± 12
	EVA-Giant/CLIP	20 ± 30	70 ± 12	49 ± 16	72 ± 14	39 ± 28	74 ± 16	68 ± 17	52 ± 10
	DINOv3-7B	82 ± 5	93 ± 2	39 ± 11	77 ± 10	56 ± 16	76 ± 15	92 ± 1	47 ± 7
	mean	**61**	**84**	**51**	**72**	**55**	**75**	**81**	**42**
4	ResNet-50	43 ± 7	73 ± 5	28 ± 10	49 ± 12	20 ± 9	35 ± 11	68 ± 3	16 ± 1
	DenseNet-121	45 ± 9	65 ± 12	39 ± 6	50 ± 10	50 ± 13	58 ± 9	68 ± 7	14 ± 3
	ViT-B/16	11 ± 6	25 ± 15	22 ± 7	34 ± 10	17 ± 14	33 ± 7	67 ± 7	23 ± 5
	DeiT-S	37 ± 21	58 ± 27	15 ± 17	25 ± 24	41 ± 15	65 ± 20	71 ± 12	30 ± 7
	Swin-B	58 ± 16	70 ± 16	44 ± 22	40 ± 21	26 ± 18	39 ± 22	69 ± 13	24 ± 6
	HRNet-W48	28 ± 5	49 ± 14	49 ± 9	56 ± 12	41 ± 12	50 ± 16	58 ± 15	18 ± 5
	ConvMixer-768	53 ± 9	70 ± 9	56 ± 8	67 ± 8	56 ± 10	67 ± 13	77 ± 9	19 ± 3
	ViT-B/CLIP	60 ± 7	78 ± 5	60 ± 8	80 ± 4	52 ± 9	59 ± 11	77 ± 5	31 ± 11
	ConvNeXtV2-L	72 ± 9	83 ± 10	59 ± 11	62 ± 19	63 ± 14	77 ± 19	91 ± 2	36 ± 5
	EVA-Giant/CLIP	1 ± 2	28	31 ± 27	45 ± 31	18 ± 18	38 ± 31	22 ± 30	30 ± 22
	DINOv3-7B	72 ± 2	82 ± 7	35 ± 14	63 ± 15	56 ± 10	61 ± 8	89 ± 1	28 ± 7
	mean	**44**	**62**	**40**	**52**	**40**	**53**	**69**	**25**
5	ResNet-50	26 ± 6	48 ± 9	17 ± 8	27 ± 11	20 ± 12	22 ± 12	57 ± 3	7 ± 2
	DenseNet-121	31 ± 10	37 ± 6	25 ± 8	30 ± 9	34 ± 5	36 ± 7	56 ± 7	8 ± 3
	ViT-B/16	5 ± 4	11 ± 5	10 ± 5	16 ± 9	8 ± 5	13 ± 7	41 ± 19	10 ± 5
	DeiT-S	19 ± 13	40 ± 23	9 ± 10	14 ± 8	20 ± 13	39 ± 17	57 ± 13	14 ± 4
	Swin-B	22 ± 13	37 ± 23	44 ± 21	37 ± 27	23 ± 9	33 ± 15	65 ± 13	10 ± 4
	HRNet-W48	9 ± 7	22 ± 12	39 ± 7	34 ± 7	23 ± 8	30 ± 10	46 ± 15	7 ± 4
	ConvMixer-768	40 ± 9	52 ± 11	45 ± 5	53 ± 9	49 ± 8	51 ± 12	69 ± 5	8 ± 4
	ViT-B/CLIP	46 ± 7	62 ± 11	50 ± 6	62 ± 7	42 ± 15	45 ± 17	70 ± 5	16 ± 11
	ConvNeXtV2-L	57 ± 5	70 ± 21	58 ± 12	54 ± 12	55 ± 14	68 ± 15	86 ± 3	24 ± 5
	EVA-Giant/CLIP	0 ± 0	–	18 ± 12	18 ± 10	9 ± 9	29 ± 20	9 ± 19	27 ± 22
	DINOv3-7B	67 ± 3	74 ± 6	23 ± 11	37 ± 17	47 ± 11	56 ± 11	86 ± 2	18 ± 6
	mean	**29**	**45**	**31**	**35**	**30**	**38**	**58**	**13**

Table 20: **Agreement, isolated from ordering.** The same intersection statistics computed on *set* success, which removes the ordering factor of Eq. 3 by construction and so is comparable across K. r_{set} falls with K in every cell, and with N in most: a shared perturbation loses agreement as the target list lengthens, independently of any difficulty in arranging it.

		$K=1$		$K=2$		$K=3$		$K=4$		$K=5$	
config	objective	r_{set}	$\bar{\phi}_{\text{set}}$	r_{set}	$\bar{\phi}_{\text{set}}$	r_{set}	$\bar{\phi}_{\text{set}}$	r_{set}	$\bar{\phi}_{\text{set}}$	r_{set}	$\bar{\phi}_{\text{set}}$
R50+D121 (N=2)	CW-K	0.970	0.638	0.881	0.574	0.778	0.589	0.599	0.500	0.339	0.356
	Rising	0.953	0.664	0.880	0.585	0.782	0.576	0.664	0.509	0.549	0.432
	QP	0.889	0.465	0.781	0.467	0.467	0.343	0.195	0.225	0.106	0.169
	KL	0.972	0.609	0.930	0.634	0.876	0.635	0.788	0.595	0.727	0.580
ViT-B+DeiT-S (N=2)	CW-K	0.957	0.479	0.918	0.591	0.816	0.564	0.503	0.350	0.344	0.174
	Rising	0.931	0.464	0.828	0.464	0.748	0.467	0.511	0.318	0.443	0.254
	QP	0.915	0.390	0.841	0.500	0.738	0.481	0.461	0.278	0.317	0.213
	KL	0.982	0.599	0.967	0.686	0.895	0.635	0.813	0.587	0.590	0.412
4-model (N=4)	CW-K	0.896	0.533	0.724	0.538	0.427	0.500	0.180	0.361	0.004	0.145
	Rising	0.601	0.493	0.377	0.424	0.157	0.174	0.019	0.103	0.000	0.024
	QP	0.561	0.368	0.227	0.323	0.004	0.226	0.000	0.167	0.000	0.133
	KL	0.919	0.573	0.810	0.585	0.639	0.584	0.420	0.496	0.086	0.306
18-model (N=18)	CW-K	0.755	0.510	0.391	0.531	0.106	0.458	0.001	0.244	0.000	0.124
	Rising	0.702	0.495	0.357	0.516	0.086	0.452	0.004	0.329	0.001	0.229
	QP	0.538	0.460	0.176	0.434	0.005	0.294	0.000	0.121	0.000	0.094
	KL	0.817	0.549	0.497	0.564	0.188	0.516	0.024	0.415	0.000	0.237

Table 21: **Ordered success under four multi-model criteria.** The same perturbations and per-image indicators as Tables 1 and 2, scored as the mean over models (*per-model*), the hardest model (*worst*), more than half of the models fooled on the same image (*majority*; equal to *joint* at N=2), and all models (*joint*), for every K re-scored. 5,000 held-out `val` images, $\|\delta\|_2 \leq 24.34$, final epoch, mean over seeds.

		per-model				worst				majority				joint			
UAP	K	CW-K	Ris.	QP	KL	CW-K	Ris.	QP	KL	CW-K	Ris.	QP	KL	CW-K	Ris.	QP	KL
R50+D121	1	92.0	85.5	81.9	93.0	90.5	84.3	73.7	91.9	89.3	81.5	72.8	90.5	89.3	81.5	72.8	90.5
	2	70.2	62.0	56.8	50.8	62.9	60.3	43.7	48.8	59.9	52.2	40.7	32.9	59.9	52.2	40.7	32.9
	3	40.8	37.2	28.1	21.5	29.7	32.0	12.9	20.1	26.5	25.3	11.2	7.6	26.5	25.3	11.2	7.6
	4	17.3	18.2	13.2	7.4	8.7	12.6	1.6	7.2	6.8	8.0	1.2	1.4	6.8	8.0	1.2	1.4
	5	5.2	8.4	6.9	3.2	0.9	4.3	0.3	2.7	0.6	2.2	0.2	0.3	0.6	2.2	0.2	0.3
ViT-B+DeiT-S	1	91.9	87.0	87.4	95.7	90.6	85.0	82.3	95.1	87.9	80.9	80.0	94.0	87.9	80.9	80.0	94.0
	2	75.0	55.1	57.1	61.1	73.8	51.6	48.3	55.8	66.8	43.7	42.3	43.2	66.8	43.7	42.3	43.2
	3	47.0	31.0	37.8	31.3	43.6	27.7	33.1	27.7	34.6	19.2	24.4	15.7	34.6	19.2	24.4	15.7
	4	14.7	7.0	9.0	13.2	9.6	5.4	7.2	11.4	5.8	1.7	2.5	3.8	5.8	1.7	2.5	3.8
	5	3.3	3.1	2.3	3.5	1.6	1.9	1.6	3.0	0.8	0.6	0.3	0.4	0.8	0.6	0.3	0.4
4-model	1	90.1	54.9	71.7	91.3	87.6	47.6	46.1	89.0	89.7	48.7	67.8	90.3	80.7	33.0	40.2	83.9
	2	66.2	21.0	41.1	48.8	53.4	14.9	9.9	40.1	62.1	12.9	27.7	37.6	43.5	4.8	7.7	13.4
	3	30.9	3.2	18.0	21.2	15.3	1.5	0.1	15.2	22.3	0.7	2.3	8.8	8.8	0.1	0.0	1.6
	4	8.4	0.4	3.5	6.3	2.1	0.0	0.0	4.0	2.8	0.0	0.0	0.7	0.5	0.0	0.0	0.0
	5	1.6	0.1	0.9	1.4	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18-model	1	91.3	88.5	81.8	91.7	85.2	80.6	68.4	86.5	93.0	90.4	84.3	92.7	68.9	62.2	44.0	74.9
	2	61.6	54.9	42.2	43.4	41.7	36.3	19.4	27.7	63.8	55.8	38.3	40.4	15.2	10.1	2.3	0.2
	3	29.3	21.9	14.3	15.7	8.9	8.8	0.8	6.3	24.0	14.9	3.3	2.9	0.4	0.1	0.0	0.0
	4	7.0	4.1	3.0	4.6	0.2	0.9	0.0	0.8	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	5	0.9	0.9	0.6	1.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18-model → 7 unseen	1	62.8	56.5	46.7	68.1	29.3	23.9	18.4	31.1	68.9	61.2	48.2	75.8	18.9	13.8	8.8	21.7
	2	17.9	16.5	9.9	13.5	4.2	3.3	1.5	1.2	8.3	7.2	2.4	4.8	0.1	0.0	0.0	0.0
	3	2.9	1.9	1.1	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0