The Scissors Effect: When Resize-Based Input Diversity Helps or Hurts Transfer Attacks

Anonymous authors
Paper under double-blind review

Abstract

Input Diversity (DI), which applies random resizing and padding at each attack iteration, is a near-default ingredient of transfer-based adversarial attacks because it is widely assumed to improve transferability. We show this assumption is regime-dependent and, for robustly trained surrogates, often reversed. Holding the attack fixed and varying only the surrogate, increasing the DI probability raises transfer success for standard surrogates but lowers it for robust ones: the two response curves separate like a pair of scissors, a pattern we call the *Scissors Effect*. The effect is strong and consistent on ImageNet: blind DI costs the robust source 10.3% attack success on average across CNN, ViT, Swin, and ConvNeXt targets, and the harm holds across ten attacks spanning 2018–2024; it is smaller on CIFAR-10 unless DI is made aggressive. A controlled robustness-strength sweep that fixes the ResNet-50 architecture and PGD-AT recipe and varies only the training ϵ shows the harm is graded rather than binary, crossing from beneficial to harmful already in the "little-robustness" regime. We trace it to gradient geometry: a resize/translation decomposition attributes roughly 67% of the harm to *resize*, and a direct source–target gradient-alignment measurement confirms the same resize operation improves alignment for standard surrogates but degrades it for robust ones. We summarize the regime with Local Gradient Consistency (LGC), a single input-space probe that cleanly separates the two surrogate types and tracks DI sensitivity at the regime level. A bias–variance theorem formalizes the mechanism qualitatively: a single crossover separates where DI helps from where its resize bias dominates. A training-free rule (CG-DI) that disables diversity when LGC is high then avoids the loss on robust surrogates while keeping DI's benefit on standard ones. We position the Scissors Effect as a DI-specific manifestation of the broader robustness–transferability trade-off, isolating which component of input diversity is responsible and when it should be disabled.

1 Introduction

Constructing transferable adversarial examples, perturbations that fool unseen target models, is a central tool for evaluating the safety of black-box deep learning systems. Since the discovery of transferability (Szegedy et al., 2014; Goodfellow et al., 2015; Papernot et al., 2016), many techniques have been proposed to enhance it. Among them, Input Diversity (DI) (Xie et al., 2019), which applies random resizing and padding at each attack iteration, has become a standard component of state-of-the-art attacks. Modern methods such as Admix (Wang et al., 2021), SSA (Long et al., 2022), and SIA (Wang et al., 2023a) aggressively stack input transformations under the working assumption that "more diversity is always better."

In this paper we show that this assumption is regime-dependent, and for robustly trained surrogates it is often reversed. We systematically study transferability from robust surrogates (*i.e.*, models trained with adversarial training (Madry et al., 2018) or related robustification) and uncover a phenomenon we call the *Scissors Effect* (Fig. 1). Holding the attack and target fixed and varying only the diversity probability p:

- For **standard** surrogates, increasing diversity ($p \to 1$) improves transferability, substantially on ImageNet and marginally on CIFAR-10, consistent with prior belief.

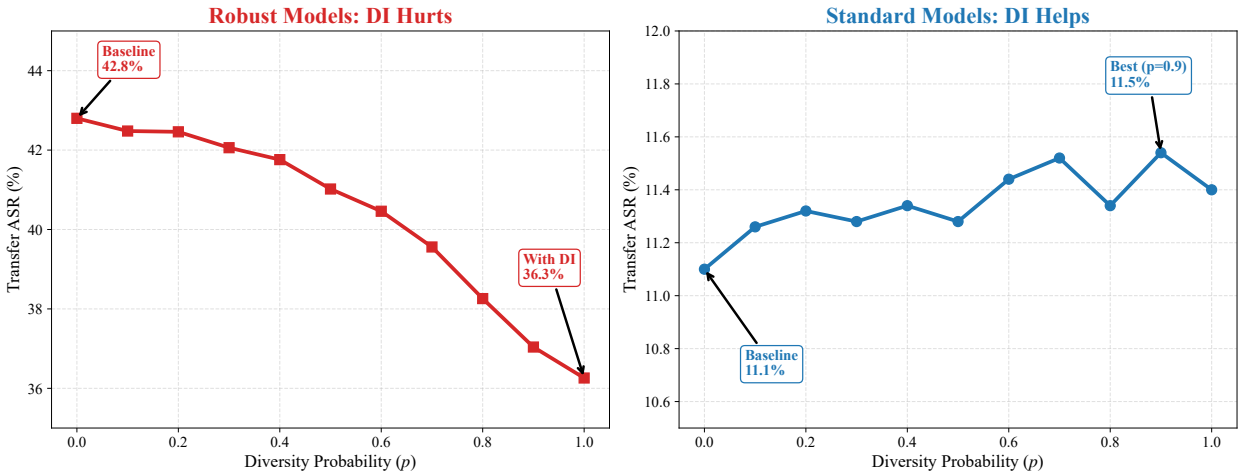

Figure 1: The Scissors Effect (CIFAR-10, `torchattacks` DI-FGSM, averaged over 5 RobustBench defended targets). Transfer ASR vs. diversity probability p depends on the surrogate type. *Left:* robust source (Engstrom; DI harms). *Right:* standard source (ResNet18; DI is neutral-to-beneficial).

- For **robust** surrogates, increasing diversity ($p \to 1$) often degrades transferability, in many cases performing worse than an attack with no diversity at all.

The two response curves separate in opposite directions as p increases, like the blades of a pair of scissors.

Why this matters, and what we claim precisely. Robust pretrained models are increasingly used as surrogates because of their stronger, perceptually aligned features (Salman et al., 2020). In that setting, leaving DI on by default can quietly weaken the attack: on ImageNet it costs 10.3% ASR on average for the robust source in our main setting. The effect is strong and consistent on ImageNet, but on 32×32 CIFAR-10 it is modest under default settings (∼2%) and only becomes large (6.6%) once DI is made aggressive, with no "safe" near-identity resize zone (Sec. 4.6). The harmful component is specifically the *resize* step of DI, not every transformation-based trick: a resize/translation decomposition attributes ∼67% of the harm to resize and shows translation to be nearly neutral. Throughout, our central claim is therefore the precise one: *resize-based input diversity helps standard surrogates but hurts sufficiently robust ones, and how much depends on the surrogate's gradient geometry.*

Relation to the robustness–transferability trade-off. That robustness changes a surrogate's gradient regime, and hence its transfer behavior, is itself known: Springer et al. (2021) show that "a little robustness" improves targeted transfer, and Zhang et al. (2024) explain how robustness strength trades off model smoothness against gradient similarity. Our contribution is not to re-discover that robustness matters, but to characterize its DI-specific manifestation: which component of input diversity is responsible (resize), how the effect emerges as robustness increases (a controlled ϵ-sweep, Sec. 3.4), and a single gradient-geometry probe that indicates when DI should be disabled.

Explanation. Our account starts from the geometry of input gradients. Intuitively, a model's gradient can be locally stable, pointing consistently in one direction under tiny input perturbations, or noisy, shifting under small perturbations. Standard models tend to produce noisy, high-frequency gradients, so DI helps by averaging out variance; robust models produce smoother, low-frequency gradients, so DI's resize step mainly perturbs a direction that is already useful. We capture this distinction with Local Gradient Consistency (LGC), a compact measure of the local gradient regime. LGC is useful chiefly as a coarse standard-vs-robust indicator; it also tracks DI sensitivity at the regime level, with Spearman $\rho = -0.75$ on ImageNet and Pearson $r = -0.87$ across the controlled ϵ-sweep. These correlations mainly reflect the regime gap rather than fine structure within it.

A simple, falsifiable consequence. This leads to a practical rule. We propose CG-DI (Consistency-Guided Diversity Input), a training-free guardrail that uses LGC to switch diversity on for standard-like surrogates and off for robust-like ones. CG-DI is not offered as a competitor to be tuned against baselines; it is a falsification test of LGC's predictive claim. A continuous p-sweep shows its binary choice $p \in \{0, 0.8\}$ sits within 1.5 percentage points of the LGC-predicted continuous optimum.

Our main contributions are:

1. We identify and verify the Scissors Effect: resize-based DI hurts transfer from robust surrogates across our CIFAR-10 and ImageNet settings, across diverse architectures (CNN, ViT, Swin, ConvNeXt, CLIP), and across ten attacks from 2018–2024.

2. We characterize how the effect depends on robustness through a controlled ϵ-sweep (fixed architecture and training recipe, varying only ϵ), showing the harm is graded rather than binary and crosses over already in the little-robustness regime, and we relate this to the broader robustness–transferability trade-off.

3. We explain the effect via gradient geometry: a resize/translation decomposition (resize ≈67% of harm), a direct source–target gradient-alignment measurement, a frequency analysis, and a bias–variance interpretation that we prove as a crossover theorem (DI helps iff LGC is below a threshold), all summarized by LGC.

4. We show LGC is useful beyond post-hoc explanation: it indicates DI sensitivity at the regime level, and a simple LGC-based rule (CG-DI) avoids the 10.3% average ASR loss caused by blind DI on robust surrogates while retaining most of its benefit on standard ones.

2 Related Work

2.1 Transfer-Based Attacks and Input Diversity

Adversarial attacks generate perturbations $x_{adv} = x + \delta$ that maximize a loss $\mathcal{L}(f(x_{adv}), y)$ subject to $\|\delta\|_\infty \leq \epsilon$. Basic gradient-based attacks such as FGSM and PGD (Goodfellow et al., 2015; Madry et al., 2018) achieve high white-box success but overfit to the surrogate's local curvature, limiting transferability. To address this, MI-FGSM (Dong et al., 2018) adds momentum to stabilize update directions, and NI-FGSM (Lin et al., 2020) couples Nesterov acceleration with scale invariance. A second, highly successful family augments the input at each iteration. Input Diversity (DI-FGSM) (Xie et al., 2019) applies, with probability p, a random *resize* (to a fraction r of the original side length, $r \in (0, 1]$; smaller r is more aggressive) followed by random padding (*i.e.*, a *translation*) back to the original size, effectively optimizing the expected loss over a distribution of transformed inputs. TI-FGSM (Dong et al., 2019) convolves the gradient with a translation-invariance kernel. Building on DI, Admix (Wang et al., 2021) mixes in other images, Spectrum Simulation (SSA) (Long et al., 2022) augments in the frequency domain, and SIA (Wang et al., 2023a), GRA (Zhu et al., 2023), PGN (Ge et al., 2023), BSR (Wang et al., 2024), and AdaMSI (Long et al., 2024) stack progressively richer structural transformations. Recent work further optimizes transferability via Bayesian priors (Fan et al., 2025), model quantization (Yang et al., 2024), or logit calibration with truncated feature mixing (Weng et al., 2025). Across this literature, input transformations are treated as broadly beneficial regardless of the surrogate's training regime; we show that this breaks down once the surrogate is robust, and isolate *resize* as the responsible component.

2.2 Input-Transformation Defenses

Input transformations also appear on the defense side: methods such as JPEG compression and resizing aim to purify adversarial perturbations (Guo et al., 2018). Paradoxically, attacking such defenses motivated making the attack itself transformation-robust, which is part of why DI and its descendants were adopted so widely. Our findings caution that this default is surrogate-dependent: when a robust surrogate is used to evaluate a defense, leaving DI on can understate attack strength.

2.3 Robustness, Gradient Geometry, and Transferability

Adversarial Training (AT) (Madry et al., 2018) and variants such as TRADES (Zhang et al., 2019) and ensemble AT (Tramèr et al., 2018) are the standard route to robustness. AT models learn perceptually aligned features and "interpretable" gradients (Ilyas et al., 2019; Salman et al., 2020). A line of work directly studies how robustness affects transferability. Springer et al. (2021) show that surrogates with a small amount of robustness yield better targeted transfer, and Zhang et al. (2024) give a more complete account: robustness strength trades off model *smoothness* against *gradient similarity*, and they conjecture that the data-distribution shift induced by stronger adversarial training degrades gradient similarity.

Our work operates in the same conceptual space, in that robustness changes surrogate gradient properties and those properties govern transfer behavior, but makes a different, complementary cut. (i) **Object of study.** Prior work explains *base* transferability as a function of robustness; we study the orthogonal axis of how a specific transfer-enhancing transformation (DI, and within it resize) interacts with that regime. In our controlled ϵ-sweep (Sec. 3.4) the base MI-FGSM transferability independently reproduces the "little-robustness" peak of Springer et al. (2021); Zhang et al. (2024), and the DI effect rides on top of it. (ii) **Metric.** LGC measures gradient stability under *input-space* perturbations at the scale of DI's transformation magnitude; it is closest to a localized form of gradient similarity, and is complementary to Zhang *et al.*'s parameter-space smoothness rather than a replacement. We therefore frame the Scissors Effect as a DI-specific specialization of the broader robustness–transferability trade-off, not a standalone re-explanation of it. While prior work (Wang et al., 2020; Xu et al., 2020) links neural networks to frequency-dependent behavior, we explicitly measure the frequency content and local consistency of gradients and tie them to the efficacy of input diversity.

3 The Scissors Effect: Discovery and Verification

3.1 Experimental Setup

Threat model. We study the standard black-box *transfer* setting: the attacker has white-box access to a *surrogate* (source) model but no access to the target: no parameters, no gradients, and no queries. The attacker crafts an L_∞-bounded perturbation on the surrogate and transfers it to the unseen target; success is measured by the target's misclassification. Attacks are untargeted unless stated otherwise (the targeted case is in Sec. E.1). We ask whether, given such a surrogate, adding input diversity to the surrogate-side optimization helps or hurts the transferred attack. The surrogate's training regime, which the attacker may not even know, is therefore the central variable.

Datasets. We use CIFAR-10 (5,000 test images) and ImageNet (5,000 validation images). Sample sizes vary by experiment and are stated in each table/figure caption; core analyses use N=500–1,000 with multiple seeds, while the CG-DI evaluation (Tab. 11) uses the full N=5,000.

Models. We consider Standard and Robust source (surrogate) models on each dataset. *CIFAR-10:* 5 standard (ResNet-18/50 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2015), DenseNet121 (Huang et al., 2017), ViT-B/16 (Dosovitskiy et al., 2021)) and 9 robust (Engstrom (Engstrom et al., 2019), Rice (Rice et al., 2020), Gowal (Gowal et al., 2020), Carmon (Carmon et al., 2019), Peng (Peng et al., 2023), Wang (Wang et al., 2023b), TRADES/Zhang (Zhang et al., 2019), Sehwag×2 (Sehwag et al., 2022)). *ImageNet:* 6 standard (ResNet50, ViT-B/16, DenseNet121, InceptionV3, Swin-B, ConvNeXt-B (Szegedy et al., 2016; Liu et al., 2021; 2022)) and 3 robust (Engstrom, Salman (Salman et al., 2020), Mo (Mo et al., 2022)). Target models are 5 RobustBench (Croce et al., 2021) defended models (Engstrom, Rice, Gowal, Carmon, Wang) for CIFAR-10, and diverse architectures (InceptionV3, ViT-B/16, Swin-B, ConvNeXt-B) for ImageNet. The full per-model architecture and training-recipe table is given in Tab. 16 (Sec. D). Our robust sources span two architecture families (ResNet-50 and ViT-B) and two training recipes (PGD-AT and ViT-AT), so the effect cannot be attributed to a single architecture. Throughout, "robust" denotes specifically L_∞ adversarially trained surrogates (PGD-AT, TRADES, and their variants); we do not test randomized-smoothing or L_2-robust models, whose gradient geometry could sit elsewhere on the LGC axis, and we scope our claims to the L_∞-AT family accordingly.

Table 1: Transfer ASR (%) on ImageNet (N=1,000, 5 seeds; DI-FGSM at p=1). MI-FGSM (p=0) is deterministic; DI-FGSM reported as mean±std. The Scissors Effect is consistent across target architectures.

Source	Method	Avg	IncV3	ViT	Swin	ConvNeXt
Robust	MI-FGSM	**76.0**	**86.0**	**82.3**	**68.6**	**66.9**
	DI-FGSM	65.7±0.3	80.9±0.1	75.5±0.3	55.5±0.8	50.7±0.9
Standard	MI-FGSM	44.0	51.9	34.9	40.9	48.3
	DI-FGSM	**58.6±0.5**	**66.5±0.8**	**49.2±1.1**	**54.1±0.7**	**64.6±0.7**

Source–target overlap. Our ImageNet analyses deliberately use disjoint architecture families for source and target (*e.g.*, ResNet50 → ViT-B/Swin-B/ConvNeXt-B; Mo2022 ViT-B → ConvNeXt-B), so the effect is not an artifact of family similarity. We verify this directly in Sec. 3.5 with a same-family (RN50 → RN50) control. Ensemble surrogates are out of scope and left to future work.

Metric and attack. We report Attack Success Rate ASR = 1 − robust accuracy. Unless otherwise stated we use MI-FGSM (Dong et al., 2018) with T=10 iterations, step size α=2/255, and budget ϵ=8/255 (CIFAR-10) or ϵ=16/255 (ImageNet). Code for the experiments and for the symbolic/numerical verification of the theory is provided as anonymized supplementary material.

3.2 The Scissors Effect

We vary the input diversity probability $p \in [0, 1]$ and measure transfer ASR. Fig. 1 shows a clear divergence between surrogate types. On CIFAR-10 (N=1,000, 5 seeds, `torchattacks` MI-FGSM), the robust source (Engstrom) drops from 42.8% (p=0) to 36.3% (p=1), a 6.5% decline, while the standard source (ResNet18) stays around ∼11% across all p (peaking at 11.5% at p=0.9).

The effect amplifies on ImageNet (Tab. 1). For the robust source (Engstrom), DI degrades transfer to every target architecture: ViT (−6.8%), Swin (−13.1%), and ConvNeXt (−16.2%). For the standard source (ResNet50) DI yields large gains: ViT (+14.3%), Swin (+13.2%), ConvNeXt (+16.3%). On average, blind DI costs the robust source 10.3% ASR.

To keep the magnitude in perspective: robust sources transfer so well to begin with (avg. 76.0% vs. 44.0% for the standard source) that even after this harm they remain the stronger surrogate (65.7% with blind DI vs. the standard source's 58.6%). DI does not make a robust surrogate weak; it erodes the lead of the strongest surrogate. The consequence for defense evaluation is nonetheless real: leaving DI on by default underestimates the attack a robust surrogate can mount, and thus overestimates the target's robustness.

The pattern is not limited to CNN targets: ViTs and hybrid architectures (Swin, ConvNeXt) exhibit it too. One interpretation is that robustness aligns gradients with semantic directions that transfer across inductive biases, making DI's spatial distortion consistently harmful for robust surrogates; we offer this as a reading, not a measured claim, since our experiments establish the consistent *direction* across targets but not a shared-feature mechanism across architectures.

Scope on CIFAR-10. On CIFAR-10 with default DI (p=0.5, r=0.9) the absolute effect is modest (∼2%). Under more aggressive settings (p=0.8, r=0.6) it grows to 6.6% (Sec. 4.6). This resolution dependence, since 32×32 images leave little room for resize-induced frequency distortion, is important for interpreting our results: the *direction* of the Scissors Effect is consistent, but its *magnitude* depends on resolution and on how aggressive the resize is.

3.3 Generalization to Modern Attacks

Does the Scissors Effect extend beyond MI-FGSM? We evaluate 10 methods spanning 2018–2024: 3 classical momentum-based (MI-FGSM, NI-FGSM (Lin et al., 2020), VMI-FGSM (Wang & He, 2021)) and 7 modern input-transformation methods (Admix (Wang et al., 2021), SSA (Long et al., 2022), SIA (Wang et al.,

Table 2: Generalization to 10 attacks (Engstrom / ResNet50 → Swin-B; N=1,000, 5 seeds). DI harms the robust surrogate on all 10 and benefits the standard one on 9/10.

		Robust			**Standard**		
Method	Venue	Base	+DI	Δ	Base	+DI	Δ
MI-FGSM	CVPR'18	42.2	40.5	−1.7	23.9	24.1	+0.2
NI-FGSM	ICLR'20	30.7	30.0	−0.7	23.2	23.6	+0.4
VMI-FGSM	CVPR'21	42.6	42.0	−0.6	22.8	24.7	+1.9
Admix	ICCV'21	44.3	41.1	−3.2	25.6	27.3	+1.7
SSA	ECCV'22	41.8	26.2	**−15.6**	22.9	22.3	−0.6
SIA	ICCV'23	46.1	40.4	−5.6	25.2	26.3	+1.1
GRA	ICCV'23	37.2	36.6	−0.6	22.6	23.6	+1.0
PGN	NeurIPS'23	33.9	32.2	−1.8	21.8	23.2	+1.5
BSR	CVPR'24	41.2	36.1	−5.1	25.5	25.8	+0.4
AdaMSI	AAAI'24	43.1	41.3	−1.8	23.7	24.1	+0.4

2023a), GRA (Zhu et al., 2023), PGN (Ge et al., 2023), BSR (Wang et al., 2024), AdaMSI (Long et al., 2024)), implemented via TransferAttack (Trustworthy-AI-Group, 2023).[1]

Tab. 2 shows that DI degrades all 10 methods on the robust surrogate and benefits 9/10 on the standard one. Read as a directional sign test against a no-effect null, a one-sided binomial gives p=$2^{-10}\approx 0.001$ for the robust direction (10/10) and p=$11/2^{10}\approx 0.011$ for the standard direction (9/10), or $p \approx 2\times 10^{-5}$ pooled (19/20). We do not lean on the exact value: the ten attacks share a surrogate, target, and image set and overlap heavily in algorithm family (momentum and input-transform variants), so the effective number of independent observations is well below ten and we treat the result as descriptive directional consistency rather than a calibrated significance test. Per-row this matters: the seed std of each Δ is ∼0.4–0.8pp, so several of the smallest robust-side effects (NI −0.7, VMI −0.6, GRA −0.6) are within seed noise *individually*; what is robust is that all ten share the same sign, and the larger effects (SSA, SIA, BSR, Admix, PGN, MI, AdaMSI) each clear the noise on their own. SSA is the one apparent exception: it produces the largest harm on the robust surrogate (−15.6%) and the only standard-side effect that is not clearly positive (−0.6%, within seed noise). It is also the only frequency-domain method, and Sec. 4.4 shows this is a prediction of our theory rather than a counterexample. More generally, modern methods with rich built-in transforms (SIA, BSR) show larger DI harm on robust sources, as extra diversity compounds the distortion.

3.4 How the Effect Depends on Robustness: a Controlled ϵ-Sweep

The contrast above pits "standard" against "robust" as two clean categories, but the scientific question is not binary: *how does the DI effect evolve as surrogate robustness increases?* Multiple robust models differ along many axes at once (architecture, recipe, budget), so we isolate robustness strength with a controlled sweep. We fix the architecture (ResNet-50) and training recipe (PGD-AT) and vary only the training budget ϵ_{train}, using Salman *et al.*'s L_∞ checkpoints (Salman et al., 2020) for $\epsilon_{\text{train}} \in \{0.5, 1, 2, 4, 8\}/255$ plus the standard (ϵ=0) model. Target is Swin-B, N=500, 3 seeds, ϵ_{attack}=16/255.

Tab. 3 shows the harm is not a binary switch but emerges gradually: D crosses from +13.0% to negative at $\epsilon_{\text{train}} \in (0, 0.5]/255$, so DI becomes harmful already in the "little-robustness" regime, then deepens to a peak near ϵ_{train}=2/255 before tapering as overall transferability falls. The MI-FGSM curve itself peaks at $\epsilon_{\text{train}} \in [0.5, 1]/255$ (88.4%/86.0%), independently reproducing the little-robustness transferability peak of Springer et al. (2021); Zhang et al. (2024): our setup operates in the same regime they study, and the DI effect D is the orthogonal axis we contribute. The same sweep against a second target (ViT-B) gives

[1]Our core DI/MI-FGSM experiments (Tab. 1) use `torchattacks` (Kim, 2020); TransferAttack is used only for the modern attacks here. The two libraries differ in default iteration count, step size, and input pre-processing, so *absolute* ASR is not comparable across them: the same nominal Engstrom→Swin-B MI-FGSM reads 68.6% in Tab. 1 (torchattacks) but 42.2% here (TransferAttack), and the standard source 40.9% vs. 23.9%. Only *within*-table Δ columns, which hold the library and setup fixed, should be compared across rows; the Scissors direction is what each table establishes, and it is consistent in both.

Table 3: Controlled robustness-strength sweep (fixed ResNet-50 + PGD-AT, varying only ϵ_{train}; target Swin-B, N=500, 3 seeds). D is the DI effect (DI−MI). DI flips from beneficial to harmful at the smallest non-zero robustness; harm peaks near ϵ_{train}=2/255. The MI-FGSM column independently reproduces the little-robustness transferability peak of Springer et al. (2021); Zhang et al. (2024).

ϵ_{train} (×255)	MI-FGSM	DI-FGSM	D = DI−MI
0 (Standard)	41.2	54.2	**+13.0**
0.5	**88.4**	82.9	−5.5 (← crossover)
1	86.0	79.1	−6.9
2	78.4	65.9	**−12.5** (max harm)
4	62.2	51.3	−10.9
8	44.4	37.8	−6.6

the same qualitative picture (crossover in $[1, 2]/255$, peak harm −9.7% at ϵ_{train}=8; full numbers in Tab. 19, Sec. I). As we show in Sec. 4.6, the crossover location aligns with the LGC threshold used by CG-DI.

Across training methods, not just strength. The sweep isolates robustness strength at a fixed recipe (Salman ResNet-50, PGD-AT); to check the effect is not specific to that family, we run a cross-recipe panel of five robust ImageNet surrogates spanning four adversarial-training recipes and four architectures: PGD-AT (Engstrom, Salman; ResNet-50), ViT-aware AT (Mo2022; ViT-B), ARES adversarial training (ConvNeXt-B), and ConvStem AT (Singh; ViT-B). Each is attacked against two targets (Swin-B and ConvNeXt-B), with each surrogate's LGC measured alongside (N=500, 3 seeds; Tab. 20, Sec. J). DI harms every robust surrogate on all nine cross-family source–target pairs (D from −4.5 to −18.7), while the standard surrogate benefits (+13.0, +15.9); the *sign* of D tracks the LGC regime, not the recipe or architecture, with every LGC>0.92 pair harmed and both LGC<0.92 pairs benefiting. The CIFAR-10 robust pool additionally covers TRADES, semi-supervised, diffusion-augmented, and proxy-distribution training (Tab. 16), all high-LGC and all DI-harmed (Tab. 8). The Scissors Effect thus tracks the resulting gradient regime, not the particular objective that produced it. We read this as a replication of the *sign* and its LGC-tracking across recipes and a second target, not as a second controlled strength sweep (no public fixed-backbone ϵ-spectrum exists outside Salman's).

3.5 Source–Target Overlap: a Same-Family Control

To rule out that the Scissors Effect is driven by architecture-family similarity between source and target, we run a same-family control in which both source and target are ResNet-50 (3 seeds, N=500, target ASR on held-out ResNet-50 checkpoints). The direction is governed by the source regime, not by family overlap: a standard ResNet-50 source benefits from DI even against ResNet-50 robust targets (D=+1.3% vs. Engstrom-RN50, +1.1% vs. Salman ϵ=2 RN50), while a robust ResNet-50 source (Salman ϵ=2) is harmed by DI even against a standard ResNet-50 target (D= − 11.6%). These same-family numbers mirror the cross-family baseline (+13.0% standard, −12.5% robust against Swin-B), confirming the effect is a property of the surrogate's gradient geometry rather than of source–target similarity. Full numbers are in Tab. 21 (Sec. K).

4 Mechanism Analysis: Why Does This Happen?

Having established the Scissors Effect, we now investigate its mechanism. We propose the *Gradient Consistency Hypothesis*: the divergent effects of DI stem from differences in gradient geometry between standard and robust models.

4.1 The Gradient Consistency Hypothesis

Intuition. A model's input gradient can be either locally *stable*, pointing in essentially the same direction under tiny input perturbations, or *noisy*, shifting under small perturbations. Standard models, trained on clean data, tend to have noisy, high-frequency gradients (Ilyas et al., 2019; Wang et al., 2020). Robust

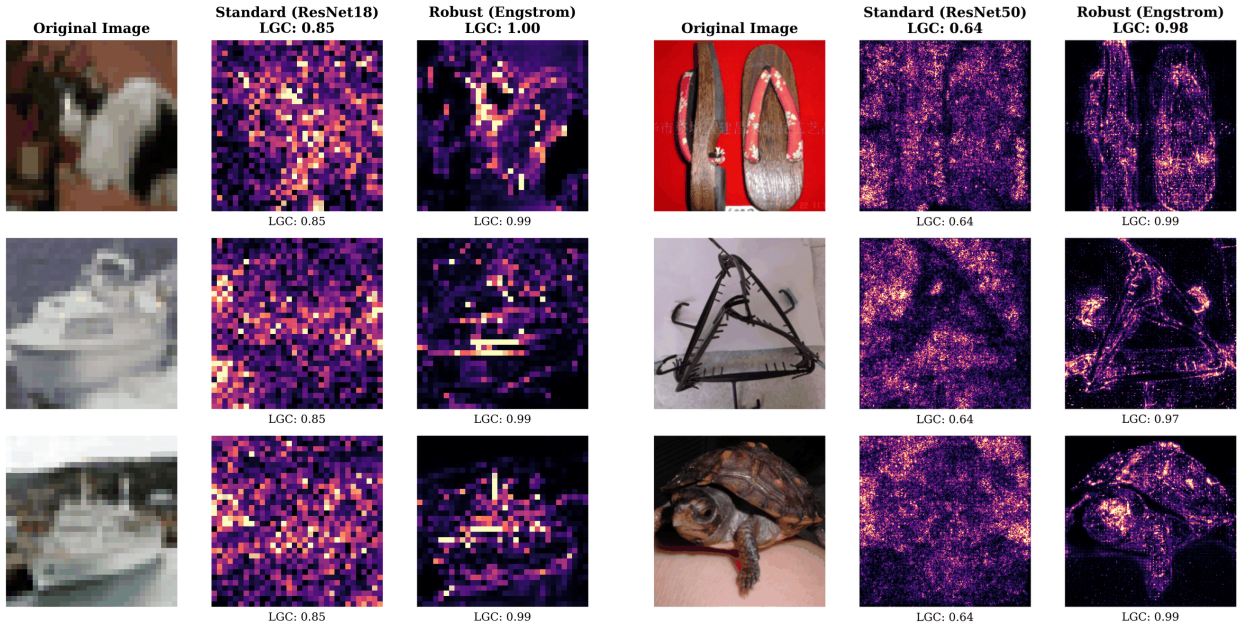

Figure 2: Gradient consistency across datasets. (*Left*) CIFAR-10: standard (ResNet18) gradients show scattered high-frequency noise; robust (Engstrom) gradients are spatially consistent. (*Right*) ImageNet: the same pattern holds at higher resolution.

models, forced to learn features stable under perturbation, have locally consistent gradients aligned with semantically meaningful directions. DI averages the gradient over randomly transformed views; this denoises a noisy gradient but perturbs an already-stable one.

We formalize stability through Local Gradient Consistency (LGC), the average cosine similarity between the gradient at a clean input x and at slightly perturbed neighbors $x'_k = x + \xi_k$:

$$\mathrm{LGC}(x) = \frac{1}{K} \sum_{k=1}^{K} \cos\big(\nabla_x \mathcal{L}(f(x), y),\ \nabla_{x'_k} \mathcal{L}(f(x'_k), y)\big), \tag{1}$$

where $\xi_k \sim \mathcal{U}(-\epsilon_{chk}, \epsilon_{chk})$ and ϵ_{chk}=1/255 (smaller than the attack step, probing the local landscape).

Empirical Validation. Fig. 2 visualizes gradient patterns. Robust models (*e.g.*, Engstrom) have LGC $\approx$ 0.98–1.00 with spatially smooth, semantically aligned gradients; standard models vary by architecture and dataset (ResNet18: 0.85 on CIFAR-10; ResNet50: 0.64 on ImageNet) and exhibit high-frequency gradient noise that benefits from DI's smoothing.

We further validate this in the frequency domain. Using the 2D FFT of the batch-averaged gradient, we compute a High-Frequency (HF) Ratio (inspired by Wang et al., 2020), the fraction of log-magnitude spectral content above 50% of the maximum spatial frequency (formal definition in Sec. F). Tab. 4 reports both metrics for 12 ImageNet models.

Standard models show HF Ratio $\approx$ 0.37–0.57 versus robust $\approx$ 0.17–0.35 ($\sim$1.7$\times$ lower), confirming the spectral gap. Mo2022 (ViT-B + AT) achieves the lowest HF ratio (0.17) and near-perfect LGC (1.00). Swin-B has low LGC (0.36) despite standard training, likely because shifted-window attention creates local gradient discontinuities, yet CG-DI still correctly assigns p=0.8 since $0.36 \ll \tau$.

4.2 Direct Mechanism Evidence: Gradient Alignment Under DI

LGC and the frequency analysis characterize the surrogate in isolation. To test the mechanism directly, namely that the same DI operation moves a standard surrogate's gradient *toward* transferable target di-

Table 4: Gradient geometry of 12 ImageNet models. Standard models exhibit higher HF content and lower LGC than robust models.

Model	Type	HF Ratio	LGC
ResNet18	Standard	0.48	0.83
VGG16	Standard	0.50	0.84
ResNet50	Standard	0.55	0.64
DenseNet121	Standard	0.40	0.92
InceptionV3	Standard	0.39	0.81
ViT-B/16	Standard	0.37	0.90
Swin-B	Standard	0.55	0.36
ConvNeXt-B	Standard	0.57	0.87
Engstrom	Robust	0.31	0.98
Salman2020	Robust	0.35	0.98
Mo2022	Robust	**0.17**	**1.00**
CLIP (Radford et al., 2021)	VLM	0.47	0.72

Table 5: Direct mechanism evidence: change in source–target sign-alignment induced by DI (N=500, 3 seeds, 20 EOT). The same DI operation improves alignment for standard surrogates and degrades it for robust ones; at aggressive r=0.6 the robust effect is a true reversal. Magnitude column is the median $|g_{\text{DI}}|/|g_{\text{no-DI}}|$: DI shrinks the noisy standard gradient $\sim$4$\times$ (variance reduction) but leaves the robust gradient essentially unchanged (nothing to denoise).

Source type	D_{sign} (r=0.9)	D_{sign} (r=0.6)	median $|g_{\text{DI}}|/|g_{\text{no-DI}}|$ (r=0.6)
Standard (ResNet50)	+0.0054	+0.0055	0.24
Robust (Engstrom, Salman, Mo2022)	−0.0004	**−0.0010**	1.27

rections but a robust surrogate's *away*, we measure source–target gradient alignment with and without DI (N=500 ImageNet images, 3 seeds, 20 Expectation-over-Transformation (EOT) (Athalye et al., 2018) samples, 4 sources $\times$ 3 ImageNet targets, on clean-correct subsets). The attack-relevant metric is the change in sign-alignment $D_{\text{sign}} := \Delta\cos\big(\text{sign}(g_{\text{src}}), \text{sign}(g_{\text{tgt}})\big)$ induced by DI, which directly tracks the sign-update step of (MI/DI)-FGSM.

Tab. 5 confirms the asymmetry. All 9 standard source–target pairs have $D_{\text{sign}} > 0$ (per-pair effect 7–15$\times$ SEM; strongest ResNet50$\rightarrow$ConvNeXt +0.0086, $\sim$15 SEM). The three robust sources, which span two architectures (ResNet50, ViT-B) and two recipes (PGD-AT, ViT-AT), show no improvement at default DI and actively degrade sign-alignment at aggressive r=0.6 (Mo2022$\rightarrow$ViT-B −0.0033, $\sim$5σ; Engstrom$\rightarrow$ConvNeXt −0.0011). Cross-seed CV is 6.4%. The magnitude column makes the variance-reduction story concrete and is the stronger of the two pieces of evidence here: DI shrinks the standard gradient $\sim$4$\times$ (removing high-frequency noise) but leaves robust gradients essentially unchanged (1.0–1.4$\times$), because they are already low-frequency and have no noise for DI's resize low-pass filter to remove. The sign-alignment column needs a careful reading. The robust-side shifts (−0.0004 to −0.0010) are small in absolute cosine terms; we read them as *directional* evidence (the sign is consistently negative, 5–15$\times$ SEM per pair) for the claim that the same DI operation pushes robust gradients the wrong way, not as a quantitative account of the 10–16% ASR swing. The quantitative driver is the asymmetric variance reduction (the magnitude column): DI's averaging buys a large variance gain on noisy standard gradients and almost none on robust ones, while paying the same resize bias on both. This is the attack-step mechanism behind the Scissors Effect, complementing the causal decomposition below. We also tested where this signal stops. The sign-alignment signal is directional at the *population* level but is not a per-image mediator: the specific images DI flips do not have a distinguishable sign-alignment shift from those it leaves alone (Sec. P.4). The mechanism is distributional, which is why we rest the quantitative claim on the variance-reduction asymmetry rather than on the sub-percent mean alignment change.

Table 6: Transform decomposition (target: Swin-B; N=500, 3 seeds, DI at p=1). Columns are measured against this experiment's own no-transform base (robust 67.6%). Resize causes ∼67% of DI harm on the robust surrogate; translation is nearly neutral.

Surrogate	Base	+Resize	+Translation	Full DI
ResNet50	42.0	46.5	52.7	54.8
Engstrom	67.6	55.7	67.4	49.7
		Δ vs. Baseline		
ResNet50	—	+4.5	+10.7	+12.8
Engstrom	—	**−11.9**	−0.2	−17.9

4.3 Component Decomposition: Resize vs. Translation

To isolate the harmful part of DI, we decompose it into *resize* (which alters spatial frequency content via interpolation) and *translation* (which shifts the image while preserving frequency content).

Tab. 6 reveals a clear asymmetry. For the robust surrogate, resize alone causes −11.9%, about 67% of Full DI's −17.9%, while random translation (EOT-style) barely affects it (−0.2%).[2] Because each column uses this table's own no-transform base, the Full-DI figure here (−17.9) runs larger than the N=1,000 headline of Tab. 1 (−13.1); the gap is sampling plus the self-contained DI configuration, and the table isolates the component split rather than matching that magnitude. The bilinear/bicubic interpolation in resize acts as a low-pass filter: for standard models with noisy gradients this smoothing helps, but for robust models whose gradients are already low-frequency and precisely aligned it introduces bias without variance reduction. An interpolation-mode ablation (Sec. P.3) confirms the *direction* of the effect is invariant across bilinear, bicubic, and antialiased resize, while its *magnitude* scales with the filter's low-pass strength (bicubic, a gentler rolloff, roughly halves the harm): the dependence the low-pass-bias account predicts, and the opposite of what a sign-agnostic interpolation artifact would produce. This is also where the frequency analysis (Tab. 4 and Sec. F) connects to the theory of Sec. 4.5: the spectra are what establish that resize is a genuine low-pass contraction (so the theorem's bias factors $c, g_r < 1$), whereas a small centered translation acts as $R \approx I$ (so $c, g_r \approx 1$ and no bias). The frequency story thus explains why the bias term is large for resize and negligible for translation, rather than standing on its own.

4.4 SSA: A Frequency-Domain Regime Shift, Not a Counterexample

SSA is the one apparent exception in Tab. 2: it produces the largest DI harm on the robust surrogate (−15.6%) and is the only method whose standard-side DI effect is not clearly positive (−0.6%). Because all 9 other modern methods are spatial-domain while SSA augments in the *frequency* domain (Gaussian noise plus DCT spectral scaling), this is exactly what our theory predicts. Modeled as a low-pass spectral pre-filter, SSA raises the surrogate's *effective* gradient consistency and shifts it rightward along the crossover of Proposition 1; this is the content of Corollary 1. The consequences match the data on both sides:

- **Robust source:** already past the crossover, SSA pushes it deeper into the harmful regime, and the resize bias compounds in the shared low-frequency band, giving the largest harm of any method.
- **Standard source:** SSA lifts its effective LGC up to near the crossover $\tau^\star$, where the DI effect is small and *sign-labile*. This reconciles an otherwise puzzling discrepancy: under the TransferAttack default config the standard-side effect is −0.6% (within seed noise), whereas under a stronger SSA ensemble (Tab. 7 below) it is +7.8%. Both are small near-crossover values on opposite sides of $\tau^\star$, consistent with Corollary 1. This is a *post-hoc reconciliation*: Corollary 1 carries a free strength parameter for the SSA pre-filter, so the model can accommodate either sign of a near-zero standard-side effect; we present it as a coherent account of the observed signs, not as a sign predicted in

[2] This differs from TI-FGSM's kernel-based translation invariance, which aggressively smooths gradients and causes larger harm (−18.4%); see Sec. P.2.

Table 7: SSA × transform decomposition (MI-FGSM + 20-sample SSA ensemble, target Swin-B). On both ImageNet and CIFAR-10, resize is the dominant harm on the robust surrogate (matching Tab. 6) and translation is nearly neutral. ASR (%).

Dataset	Surrogate	SSA-only	+Resize	+Translation	+Full DI
ImageNet	Engstrom (Rob)	54.1	**44.5** (−9.6)	51.7 (−2.4)	47.5 (−6.6)
	ResNet50 (Std)	73.8	79.8 (+6.0)	79.3 (+5.5)	81.6 (+7.8)
CIFAR-10	Engstrom (Rob)	23.2	**20.3** (−2.9)	23.7 (+0.5)	20.8 (−2.5)
	Standard (Std)	12.1	13.5 (+1.5)	12.5 (+0.5)	13.7 (+1.6)

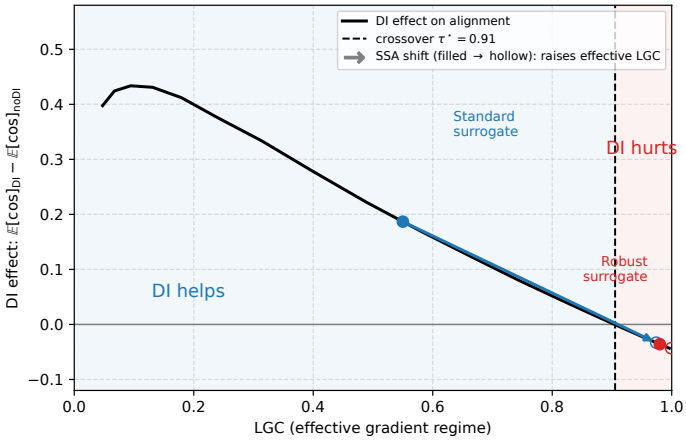

Figure 3: Why SSA is the exception (Corollary 1). The DI effect on alignment (black curve) crosses zero at $\tau^\star$. As a low-pass pre-filter, SSA raises a surrogate's *effective* LGC (arrows): a standard surrogate is pushed from the DI-helpful region up to near $\tau^\star$ (small, sign-labile effect), while a robust surrogate is pushed deeper into the DI-harmful region. Markers are the Monte-Carlo of Sec. B.

advance (Fig. 3). The robust-side consequence, larger and resize-driven harm, is the falsifiable part, and it is borne out on both datasets (Tab. 7).

We confirm the resize-driven robust harm with an SSA-based transform decomposition analogous to Tab. 6 (target Swin-B, N=500, 3 seeds; the SSA base here is MI-FGSM with a 20-sample SSA ensemble, which is stronger than the TransferAttack default and so reaches higher absolute ASR).

Tab. 7 confirms that on the robust surrogate, SSA+resize alone causes the dominant harm (−9.6% on ImageNet; −2.9% on CIFAR-10, where it even exceeds full DI's −2.5%), while SSA+translation is approximately neutral, the same resize-driven mechanism as Tab. 6. The interpretation, made precise by Corollary 1, is that DI's resize low-pass filter shares spectral support with SSA's DCT-based augmentation, so on the robust surrogate's low-frequency gradient the two compound. In plain terms (Fig. 3): SSA pre-sharpens the gradient the way robustness does, nudging any surrogate toward the "DI-hurts" end of the axis, a long push for robust models and just over the line for standard ones. It is a frequency-domain *regime shift*, not a counterexample.

4.5 A Bias–Variance Perspective

The effect can be read through a bias–variance lens. DI averages gradients over transformed views, reducing the variance of the gradient estimate at the cost of bias (the averaged direction may deviate from the original optimum). For standard models (low LGC), the estimate has high variance from noisy, high-frequency loss landscapes, so $\mathbb{E}_T[\nabla\mathcal{L}(T(x))]$ acts as beneficial variance reduction. For robust models (high LGC), the landscape is smooth with intrinsically low gradient variance ($\mathrm{Var}[\nabla\mathcal{L}] \approx 0$), so variance reduction

is unnecessary; meanwhile resize shifts the spatial frequency content and introduces systematic bias that degrades an already well-aligned direction. LGC thus proxies the tradeoff: high LGC indicates low variance where avoiding bias matters; low LGC indicates high variance where averaging helps.

This intuition can be made precise. Modeling the surrogate gradient as signal plus isotropic noise and resize-DI as a symmetric contraction R with m-fold averaging (assumptions in Sec. B), the alignment between the attack direction and the transferable target direction admits a closed form, and DI helps exactly below a threshold on the gradient regime.

Proposition 1 (Scissors crossover; proof in Sec. B)**.** *Under Assumptions 1 to 3, let* $\rho = \|\mu\|^2/(n\sigma^2)$ *be the gradient signal-to-noise ratio, so that* $\mathrm{LGC} = \rho/(1+\rho)$*. Write* $c = \cos(R\mu, \mu) \in (0, 1]$*,* $g_r = \|R\mu\|^2/\|\mu\|^2$*, and* $\kappa = \mathrm{tr}(R^2)/(nm)$*, and assume* $\kappa/g_r < c^2 < 1$*. Then resize-based DI strictly improves the signal-to-RMS alignment if and only if* $\mathrm{LGC} < \tau^\star$*, and strictly degrades it if and only if* $\mathrm{LGC} > \tau^\star$*, where*

$$\tau^\star = \frac{\rho^\star}{1+\rho^\star}, \qquad \rho^\star = \frac{c^2 - \kappa/g_r}{1 - c^2}. \tag{2}$$

In particular, as $\mathrm{LGC} \to 1$ *(a fully robust surrogate), DI necessarily hurts.*

The proposition is a *surrogate-side* bias–variance crossover: a single threshold in the gradient regime separates where averaging (DI) helps from where its resize bias dominates. The bias enters only through $c, g_r < 1$, *i.e.* through $R \neq I$, the resize component; a small, centered translation corresponds to $R \approx I$, leaving the bias factor near unity and matching the near-neutral translation in Tab. 6. The bridge from this surrogate-side statement to *transfer* is Assumption 3, which takes the target's vulnerable direction to be aligned with the surrogate signal μ. We do not prove that assumption; it is the standard robust-features-transfer premise (Salman et al., 2020) and is what our gradient-alignment measurement (Tab. 5) tests directly. In fact the theorem does not even require exact alignment: Proposition 2 generalizes it to an *arbitrary* target direction u and shows DI still necessarily hurts as $\mathrm{LGC} \to 1$ precisely when resize reduces alignment with u ($\gamma_R < \gamma_\mu$), a condition needing no architectural match and matching the sign of the D_{sign} that Sec. 4.2 measures. The proposition therefore makes the bias–variance intuition explicit and testable, though its two parts carry different weight. The fully robust limit ($\mathrm{LGC} \to 1$, i.e. $\sigma \to 0$) is close to definitional: with no gradient variance left to reduce, any direction-altering contraction can only hurt. The non-trivial content is therefore the existence and uniqueness of the interior crossover, whose location the theorem does not fix. What the proposition does not do is settle, from first principles, whether a surrogate's gradient aligns with a target's transferable directions, which is an empirical matter (Sec. 4.2). All algebra is symbolically verified, and a numerical simulation (Fig. 4) exhibits the crossover near $\mathrm{LGC} \approx 0.9$ for the illustrative constants used there (Sec. B). Estimating c, g_r, κ on 13 real surrogates keeps the model instantiable ($g_r \leq 1$ and the crossover condition holds for all), and the resulting per-surrogate threshold separates the two regimes and predicts the DI direction correctly on all eight surrogates we ran as sources; it does not, however, pin a single threshold, so CG-DI's τ stays empirical (Sec. B.5).

4.6 From Explanation to Indication: LGC as a Regime Identifier

Beyond the binary direction, does LGC quantify DI sensitivity? We find a clear two-tier picture. LGC is a reliable **coarse** regime indicator: standard surrogates cluster near $\mathrm{LGC} \approx 0.64$ and robust ones above 0.95, a clean separation that is all CG-DI uses. Whether LGC also gives *fine-grained* resolution within a regime is a weaker claim we do not make: the correlation analysis below shows LGC saturates across the robust region, so we read the quantitative correlations as *suggestive* of this regime picture rather than a per-surrogate predictor of harm.

Effect-size identifiability. We first note why a naive correlation fails on CIFAR-10. Define the *identifiability margin* $\Delta_m = \mathrm{ASR}(p^*) - \mathrm{ASR}(p_{\mathrm{second}})$. Under standard DI ($r$=0.9) all 10 CIFAR-10 surrogates have $\Delta_m < 1\%$ (Sec. H), so the argmax is noise-dominated, an effect-size limitation of 32×32 images rather than a failure of LGC. Reducing the resize rate recovers identifiable effect sizes (Tab. 8). Two features stand out. First, there is *no "safe" near-identity resize*: even r=0.97 (a single-pixel shrink on 32×32 images) already costs −2.5 to −2.9% on the robust surrogates, confirming that any interpolation perturbs the already-aligned robust gradient. Second, the harm is roughly flat for mild resize ($r \geq 0.9$) and then amplifies monotonically

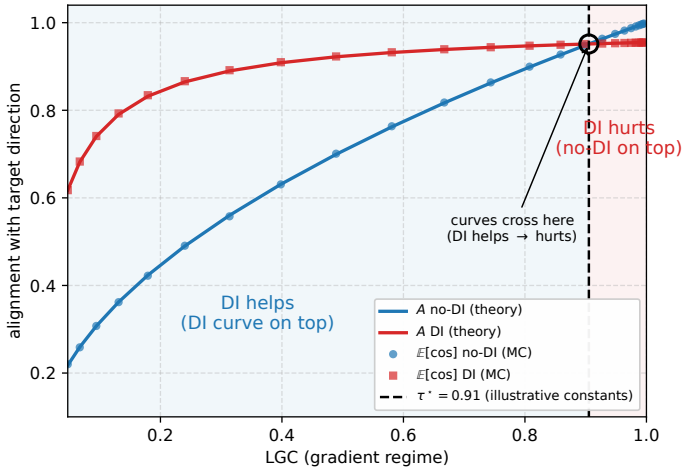

Figure 4: The theorem, visualized. As the gradient regime sweeps from standard (low LGC) to robust (high LGC), the DI effect on alignment changes sign at a single crossover, and DI necessarily hurts as LGC $\to 1$ (this qualitative behaviour is what the theorem guarantees). The crossover location (LGC $\approx$ 0.9 here) depends on the constants c, g_r, κ, which are illustrative values chosen for the simulation, not measured from a real surrogate; we do not claim the theory predicts CG-DI's τ=0.92. Both the A-metric (lines) and the measured $\mathbb{E}[\cos]$ (markers) cross at the same point. Setup in Sec. B.

Table 8: DI aggressiveness vs. harm (CIFAR-10, p=0.8; ΔASR vs. p=0, N=1,000, 5 seeds, target Engstrom). Even near-identity resize (r=0.97) harms robust surrogates; harm amplifies monotonically as r decreases.

Surrogate	r=0.97	r=0.95	r=0.9	r=0.8	r=0.7	r=0.6
Engstrom	−2.9	−2.7	−2.9	−3.9	−4.8	**−6.6**
Rice	−2.5	−2.2	−2.7	−3.0	−4.5	−5.2

Table 9: LGC correlation with DI sensitivity across three settings.

Setting	Quantity vs. LGC	n	Coefficient	p-value
CIFAR-10	ΔASR (harm)	14	Spearman −0.41	0.146
ImageNet	optimal p^*	9	Spearman −0.75	0.019
ϵ-sweep	DI effect D	6	Pearson −0.87	0.025

as the resize grows more aggressive, reaching −6.6% (Engstrom) at r=0.6. This rules out the intuition that the effect appears only under extreme resize, while explaining why it is modest under the default r=0.9.

Evidence: primary and supporting. The robust core of the empirical case is *directional*: DI hurts 10/10 robust pairs and helps 9/10 standard ones (Tabs. 1 and 2; one-sided binomial $p\approx$0.001 for the robust direction and $p\approx$0.011 for the standard direction, read as directional consistency given the attacks are not independent; see Sec. 3.3). This is the claim we stand on, and it is what CG-DI acts on. The rank correlations (Fig. 5 and Tab. 9) are consistent with it but we read them only as *supporting*: on ImageNet, LGC vs. optimal p^* gives Spearman $\rho = -0.75$ (p=0.019, n=9), and across the controlled ϵ-spectrum LGC vs. the DI effect D gives Pearson $r = -0.87$ (p=0.025, n=6). Two caveats temper these. First, the coefficients are driven by the *coarse* standard-vs-robust separation, not by fine structure: within the robust region LGC saturates at 0.98–0.99 (Tab. 22) while the harm D still varies roughly two-fold and non-monotonically (peaking at ϵ_{train}=2/255, Tab. 3), so LGC does not resolve harm within the robust regime. Second, n=6 is small and assumes linearity, so we do not lean on it. On CIFAR-10, where the correlation uses an expanded pool of 14 surrogates (the margin analysis above uses the 10 core models), the direction holds but the coefficient is weak (Spearman $\rho = -0.41$, p=0.146), consistent with the limited effect size.

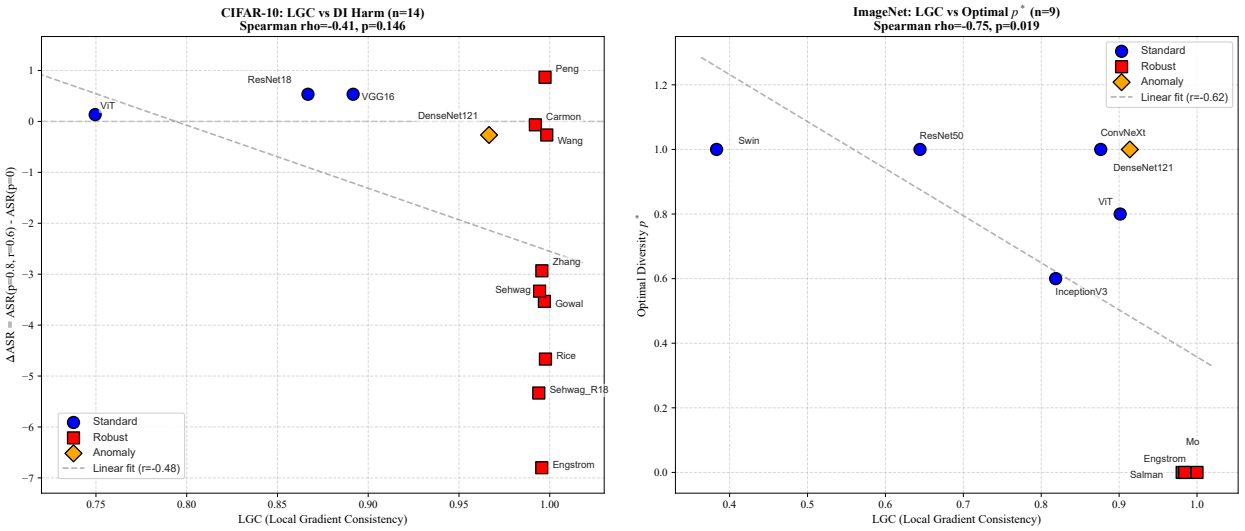

Figure 5: LGC correlation analysis. (*Left*) CIFAR-10 (n=14): LGC vs. DI-induced harm (ΔASR); Spearman $\rho = -0.41$, p=0.146. (*Right*) ImageNet (n=9): LGC vs. optimal diversity p^*; $\rho = -0.75$, p=0.019. Higher LGC consistently associates with less benefit from diversity.

In practice this supports a simple a-priori check from a single gradient-consistency measurement: a surrogate with LGC > 0.95 should be treated as robust and, as a rule of thumb, suffers $\sim$3% ASR degradation under DI (p=0.8, r=0.9), amplifying to $\sim$5–7% under aggressive settings ($r \leq 0.6$), which warrants disabling diversity. We frame this as a regime-level guideline rather than a precise per-surrogate prediction of harm.

5 Application: Consistency-Guided Diversity Input

As a direct, falsifiable consequence of our findings, we propose **CG-DI** (Consistency-Guided Diversity Input), a training-free rule that uses LGC to choose the diversity setting per surrogate. CG-DI is not a new attack submitted to win a hyper-parameter benchmark, but a falsification test of LGC's predictive claim. If a parameter-free, LGC-only switch recovers the right regime decision, that is evidence LGC's utility extends beyond post-hoc explanation; if it failed, it would refute the claim. Its deliberate simplicity is the point.

One might ask why a gradient probe is needed at all when an attacker usually knows whether their own surrogate was adversarially trained. The answer is the case where they do not: third-party, fine-tuned, or distilled checkpoints carry training metadata that is unverifiable, and a standard-trained backbone can present a robust-like gradient regime (the DenseNet anomaly of Sec. 5.3). CG-DI keys on input-space gradient geometry, which is directly measurable from the checkpoint, rather than on a provenance label that may be wrong or absent. When the label is trustworthy, CG-DI agrees with it; its value is precisely when it is not.

5.1 Algorithm

CG-DI detects the gradient regime with a lightweight LGC probe and makes a binary decision: enable diversity (p=0.8) for standard-like models, disable it (p=0) for robust-like ones. Alg. 1 is a two-step process: (1) probe LGC with K local gradient queries; (2) compare against a threshold τ.

Why binary suffices: a continuous-p sweep. One might worry a binary $p \in \{0, 0.8\}$ is a crude heuristic. It is not: it is the near-optimal projection of a continuous, LGC-predicted curve. We sweep $p \in \{0, 0.1, \ldots, 1.0\}$ for three sources against Swin-B (N=500, 3 seeds; Tab. 10). The ASR(p) curves are monotone with a *sign-flipped slope* between low-LGC and high-LGC sources, so LGC predicts the slope direction, and CG-DI's binary choice lands within 1.5 percentage points of the continuous optimum (exact

Algorithm 1 CG-DI: Consistency-Guided Diversity Input

Require: Surrogate f, input x, label y, threshold τ=0.92

1: $g \leftarrow \nabla_x \mathcal{L}(f(x), y)$; $S \leftarrow 0$
2: **for** $k = 1$ to K=5 **do**
3: $\quad x'_k \leftarrow x + \mathcal{U}(-1/255, 1/255)$
4: $\quad S \leftarrow S + \cos(g, \nabla_{x'_k} \mathcal{L}(f(x'_k), y))$
5: **end for**
6: $p^* \leftarrow 0$ if $S/K > \tau$ else 0.8
7: **return** Attack with diversity p^*

Table 10: Continuous-p sweep (target Swin-B, N=500, 3 seeds). ASR(p) is monotone with a slope whose sign LGC predicts. CG-DI's binary projection $p \in \{0, 0.8\}$ tracks the continuous optimum p^* within 1.5pp (exact for robust sources). ASR (%).

Source	LGC	p=0	p=0.5	p=1.0	p^*	CG-DI (p^*_{bin})	gap
Standard RN50	0.65	41.2	51.9	**54.2**	1.0	0.8→52.7	+1.5
Salman ϵ=2	0.99	**78.4**	73.7	65.9	0.0	0→78.4	**0.0**
Engstrom ϵ=4	0.98	**67.6**	64.3	55.3	0.0	0→67.6	**0.0**

Table 11: CG-DI on CIFAR-10 (N=5,000, 5 seeds). CG-DI adapts p for DI-FGSM (default p=0.5), matching the stronger setting per surrogate.

Source	MI-FGSM	DI-FGSM	CG-DI
Robust	**43.64**	41.09±0.23	**43.64**
Standard	12.98	12.90±0.13	12.91±0.13

for both robust sources). The binary rule is thus the projection of the LGC-predicted curve, not an arbitrary heuristic.

Computational cost. These K=5 extra gradient queries are on the *surrogate*, which the attacker controls and runs locally; CG-DI adds zero target-model queries (a transfer attack never queries the target while crafting), so it costs nothing against a query-limited black-box budget, the constrained resource in API attacks. The only cost is surrogate compute: in per-image mode CG-DI adds ∼75% wall-clock overhead on a T=10 attack (Tab. 26); and since LGC is consistent within a model (Sec. G), pre-computing it once on a small calibration batch (N=50) and reusing the binary decision amortizes this to ∼5% (and to essentially zero across many attacks with the same surrogate).

5.2 Main Results

CG-DI is a guardrail, so the right baseline to beat is *the better of* MI-FGSM (p=0) and DI-FGSM: it should match the stronger one per surrogate rather than lose to blind DI on robust sources.

CIFAR-10. Tab. 11: for Engstrom, CG-DI sets p=0 and matches MI-FGSM (43.64%), avoiding DI-FGSM's degradation (41.09%); for ResNet18, DI is neutral on this dataset and CG-DI adapts correctly.

ImageNet. On ImageNet (Tab. 1; per-target breakdown in Sec. N), CG-DI prevents the 10.3 ± 0.3% average ASR degradation that blind DI causes on the robust source: for Engstrom it sets p=0 and matches MI-FGSM's 76.0%; for ResNet50 it enables p=0.8 and captures the majority of DI's gains. On the robust surrogates we evaluate, CG-DI matches MI-FGSM and so does not lose to blind DI, which resolves the apparent "CG-DI is worse" reading of an enhancer-style comparison that does not apply to a guardrail.

Table 12: DenseNet anomaly on CIFAR-10: CG-DI avoids harm from blind DI (p=0.8) where a metadata rule would misroute.

Model	LGC	Blind (p=0.8)	CG-DI
ResNet18	0.85	13.0±0.1%	12.9±0.1%
DenseNet121	**0.96**	12.8±0.1%	**12.9±0.0%**
Engstrom	0.99	36.2±0.5%	**43.6±0.0%**

Table 13: Small perturbation (ϵ=4/255, ImageNet, DI-FGSM at p=1). Effect persists.

Source	MI-FGSM	DI-FGSM	CG-DI
Robust	**34.9±0.0%**	32.0±0.4%	**34.9±0.0%**
Standard	32.5±0.0%	**47.1±0.5%**	46.0±0.7%

5.3 Ablations

Threshold sensitivity. We ablate $\tau \in [0.80, 0.98]$ on ImageNet (target Swin-B, N=500, 3 seeds; Sec. P.6). The robust source (Engstrom) holds ∼67.1% across $\tau \in [0.80, 0.96]$, dropping only at the extreme τ=0.98; the standard source (ResNet50) plateaus at ∼53.3% for $\tau \geq 0.86$. This gives a broad safe zone $\tau \in [0.86, 0.96]$ with our default τ=0.92 centered in it. As long as a substantial LGC gap separates source types, the exact τ has little practical impact.

The DenseNet anomaly. DenseNet121, though standard-trained, has unusually high LGC on CIFAR-10 (0.96; 0.92 on ImageNet per Tab. 4) due to dense-connectivity gradient smoothing. A naive metadata rule ("standard → enable DI") would set p=0.8, while LGC correctly sets p=0. This is why CG-DI keys on input-space gradient geometry rather than training metadata, which is unverifiable for fine-tuned, distilled, or third-party checkpoints (Tab. 12).

Smaller perturbation budget. The Scissors Effect persists at ϵ=4/255 (Tab. 13), supporting that it is a property of gradient geometry rather than of a specific budget.

When does LGC alone mislead? A corner case. Across the controlled ϵ-sweep, CG-DI classifies 7/8 robust sources correctly. The single miss is Salman ϵ=8/255 *ResNet-50* (LGC 0.80, so CG-DI would enable DI). We attribute this to architecture capacity, not an LGC failure: at the same training ϵ=8/255, Mo2022's *ViT-B* yields LGC 1.00 (correct), and Salman *et al.*'s own clean accuracy for this checkpoint drops to 54.5% (a 21-point fall from ϵ=0), a documented capacity strain; RobustBench and Mo et al. (2022) likewise reserve higher-capacity backbones for $\epsilon \geq 8/255$. We report this case in Sec. L.

6 Discussion

Implications. Many modern attacks (*e.g.*, Admix, SSA, SIA) stack transformations expecting extra diversity to help. Our results show this is unreliable once the surrogate is robust: a standard recipe can become weaker simply because the source model changed. The point matters most for *defense evaluation*: if DI is left on by default, attacks from robust surrogates can look weaker than they are, inflating robustness estimates.

Relation to the robustness–transferability trade-off. The Scissors Effect is best read as a DI-specific specialization of the trade-off studied by Springer et al. (2021) and Zhang et al. (2024). Their account explains how robustness reshapes *base* transferability through smoothness and gradient similarity; we isolate how one transfer-enhancing transformation (resize within DI) interacts with that regime, identify the responsible component, and give a single input-space probe (LGC) that indicates when to disable it. Our controlled sweep (Sec. 3.4) connects the two: base MI-FGSM transferability reproduces their little-robustness peak, and the DI effect rides on top as an orthogonal axis.

Limitations. (1) The CIFAR-10 effect is modest at default p=0.5 (∼2%) due to 32×32 resolution, amplifying to 6.6% only under aggressive DI (p=0.8, r=0.6); our central claim is therefore scoped to resize-based DI

and is strongest on ImageNet. (2) ImageNet uses three robust surrogates spanning two architectures and two recipes; CIFAR-10 additionally covers TRADES, but stronger coverage is limited by public checkpoint availability. (3) SSA interferes with DI on both source types via spectral overlap; CG-DI's gradient-geometry-only logic does not detect such method-specific interference. (4) A binary $p \in \{0, 0.8\}$ is a guardrail, not an optimizer; the continuous-p sweep shows it is within 1.5pp of optimal, but it can still miss the exact value. (5) Ensemble surrogates mixing gradient regimes are unexplored. (6) LGC can mislead when capacity, not robustness, drives low consistency (the Salman ϵ=8 ResNet-50 corner case).

On theory. The crossover itself is now proved: under an additive signal–noise gradient model with resize as a symmetric contraction, Proposition 1 (proved in Sec. B, with machine-checked algebra) shows DI helps iff LGC $< \tau^\star$ and necessarily hurts as LGC $\to 1$, and Corollary 1 extends this to frequency-domain methods such as SSA. What remains open is a quantitative model of SSA's DCT-band interference and a data-dependent estimate of the constants (c, g_r, κ) for a given surrogate.

Future work. Two directions seem natural. First, CG-DI could be made *method-aware*, recognizing cases like SSA where spectral operations interact with DI. Second, it would be worth testing whether the effect appears more broadly in multimodal systems; our preliminary CLIP results (Sec. M) suggest LGC remains informative beyond standard supervised vision models. We also leave continuous p-selection, ensemble surrogates, and a tighter link between LGC and loss-surface curvature to future work.

7 Conclusion

We revisited an assumption largely taken for granted in transfer attacks: that more input diversity always helps. It does not. For robustly trained surrogates, resize-based DI can substantially hurt transferability, producing the opposite trend from standard models, a mismatch we call the Scissors Effect. The scope is specific: the effect is strong and consistent on ImageNet, smaller on CIFAR-10 unless DI is aggressive, and driven specifically by the resize component, with translation nearly neutral. A controlled robustness-strength sweep shows the harm is graded rather than binary and crosses over already in the little-robustness regime, placing the phenomenon within the broader robustness–transferability trade-off as a DI-specific specialization of it.

LGC gives a compact description of the underlying difference in gradient regime, validated by a direct source–target gradient-alignment measurement, a resize/translation decomposition, frequency analysis, and a bias–variance account that we sharpen, under a signal-plus-noise gradient model, into a crossover theorem: DI improves the attack direction exactly when LGC is below a threshold, and necessarily hurts as LGC approaches one. LGC serves mainly as a coarse standard-vs-robust indicator that also tracks DI sensitivity at the regime level ($\rho = -0.75$, p=0.019 on ImageNet; $r = -0.87$, p=0.025 across the controlled sweep), reflecting the regime gap rather than fine structure within it. CG-DI shows this is actionable: by selecting the diversity setting from the gradient regime, it avoids the 10.3% average ASR drop caused by blind DI on robust surrogates while retaining most of its benefit on standard ones, and a continuous-p sweep confirms its binary choice is the projection of an LGC-predicted curve. As robust pretrained models become more common as surrogates, the practical takeaway is simple: DI should not be a default-on component, and attack design should account for the surrogate's gradient geometry.

Broader Impact Statement

This work studies when a standard component of transfer-based adversarial attacks (input diversity) helps or hurts. Like all research on adversarial transferability, the techniques could in principle be used to attack deployed black-box models. However, our central message is cautionary and primarily benefits *defenders* and *evaluators*: we show that leaving input diversity on by default can make attacks from robust surrogates look weaker than they actually are, leading to over-optimistic robustness estimates. Reporting this effect, and providing a cheap gradient-geometry probe to detect it, helps practitioners avoid systematically underestimating the strength of transfer attacks during defense evaluation. We use only publicly available models and datasets and introduce no new attack capability beyond a guideline for setting an existing hyperparameter.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, pp. 284–293. PMLR, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy S. Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: A standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4312–4321, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Mingyuan Fan, Cen Chen, Wenmeng Zhou, and Yinggui Wang. Transferable adversarial examples with bayesian approach. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security (ASIACCS)*, pp. 517–529, 2025.

Zhijin Ge, Hongying Liu, Xiaosen Wang, Fanhua Shang, and Yuanyuan Liu. Boosting adversarial transferability by achieving flat local maxima. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 70141–70161, 2023.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Hoki Kim. Torchattacks: A PyTorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2020.

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 (ICCV)*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022.

Sheng Long, Wei Tao, Shuohao Li, Jun Lei, and Jun Zhang. On the convergence of an adaptive momentum method for adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14132–14140, 2024.

Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision (ECCV)*, pp. 549–566. Springer, 2022.

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

Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 18599–18611. Curran Associates, Inc., 2022.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

ShengYun Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, Mansi Phute, Jason Martin, and Duen Horng Chau. Robust principles: Architectural design principles for adversarially robust CNNs. In *34th British Machine Vision Conference (BMVC)*. BMVA, 2023.

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

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, pp. 8093–8104. PMLR, 2020.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3533–3545, 2020.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=WVXONNVBBkV`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Jacob M. Springer, Melanie Mitchell, and Garrett T. Kenyon. A little robustness goes a long way: Leveraging robust features for targeted transfer attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9759–9773, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Trustworthy-AI-Group. TransferAttack: A PyTorch framework for adversarial transferability, 2023. URL `https://github.com/Trustworthy-AI-Group/TransferAttack`.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8684–8694, 2020.

Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24336–24346, 2024.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1924–1933, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16158–16167, 2021.

Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4607–4619, 2023a.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263. PMLR, 2023b.

Juanjuan Weng, Zhiming Luo, and Shaozi Li. Improving transferable targeted adversarial attack via normalized logit calibration and truncated feature mixing. *IEEE Transactions on Information Forensics and Security*, 2025. doi: 10.1109/TIFS.2025.3563820.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2730–2739, 2019.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Yulong Yang, Chenhao Lin, Qian Li, Zhengyu Zhao, Haoran Fan, Dawei Zhou, Nannan Wang, Tongliang Liu, and Chao Shen. Quantization aware attack: Enhancing transferable adversarial attacks by model quantization. *IEEE Transactions on Information Forensics and Security*, 19:3265–3278, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pp. 7472–7482. PMLR, 2019.

Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, Wei Wan, and Hai Jin. Why does little robustness help? a further step towards understanding adversarial transferability. In *IEEE Symposium on Security and Privacy (SP)*, pp. 3365–3384, 2024.

Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4741–4750, 2023.

A Appendix

Code and data. Code to reproduce the experiments, together with the symbolic and numerical verification of the theory, is provided as anonymized supplementary material with this submission. Section, figure, and table numbers in this appendix continue the main-paper numbering.

B A Bias–Variance Theorem for the Scissors Effect

We make the bias–variance account of Sec. 4.5 precise and prove Proposition 1. The model abstracts the attack as choosing a perturbation direction from the surrogate gradient and transferring it to a target whose vulnerable direction is aligned with the surrogate signal (consistent with the observation that robust features transfer (Salman et al., 2020)).

B.1 Model and assumptions

Assumption 1 (Additive gradient model)**.** At a fixed input x, the surrogate input-gradient is $g = \mu + \eta$, where $\mu \in \mathbb{R}^n$ is a deterministic signal and η is zero-mean noise with $\mathrm{Cov}(\eta) = \sigma^2 I_n$. The gradient regime is summarized by the signal-to-noise ratio $\rho := \|\mu\|^2/(n\sigma^2)$.

Assumption 2 (Resize-based DI)**.** Input diversity replaces g by $\bar{g} = \frac{1}{m}\sum_{i=1}^{m} g_i$ with $g_i = R\mu + \zeta_i$ i.i.d., where $R \in \mathbb{R}^{n\times n}$ is the resize operator (symmetric, $0 \preceq R \preceq I$) and ζ_i are zero-mean with $\mathrm{Cov}(\zeta_i) = \sigma^2 R^2$. Write $c = \cos(R\mu, \mu) \in (0, 1]$, $g_r = \|R\mu\|^2/\|\mu\|^2 \in (0, 1]$, and $\kappa = \mathrm{tr}(R^2)/(nm) \in (0, 1]$.

Assumption 3 (Transfer figure of merit)**.** The transferable target direction is proportional to μ, and attack quality is the signal-to-RMS alignment $A(d) = \langle \mathbb{E}[d], \hat{\mu}\rangle/\sqrt{\mathbb{E}\|d\|^2}$, with $\hat{\mu} = \mu/\|\mu\|$, evaluated at $d = g$ (no DI) or $d = \bar{g}$ (DI).

The metric A uses only the first two moments of d, so it is exact: no high-dimensional concentration is invoked. It lower-bounds the linearized expected transfer gain $\langle \mathbb{E}[d/\|d\|], \hat{\mu}\rangle$ and, empirically, tracks the measured $\mathbb{E}[\cos(d, \mu)]$ almost exactly (Fig. 4 below).

B.2 Lemmas and theorem

Lemma 1 (LGC–SNR identity)**.** *Under Assumption 1, for independent draws g, g',*

$$\mathrm{LGC} \;:=\; \frac{\mathbb{E}\langle g, g'\rangle}{\sqrt{\mathbb{E}\|g\|^2\,\mathbb{E}\|g'\|^2}} \;=\; \frac{\rho}{1+\rho}, \tag{3}$$

which is strictly increasing in ρ.

Proof. Independence and $\mathbb{E}\eta = 0$ give $\mathbb{E}\langle g, g'\rangle = \langle \mathbb{E}g, \mathbb{E}g'\rangle = \|\mu\|^2$ and $\mathbb{E}\|g\|^2 = \|\mu\|^2 + \mathrm{tr}(\sigma^2 I_n) = \|\mu\|^2 + n\sigma^2$. Dividing through by $n\sigma^2$ and using $\rho = \|\mu\|^2/(n\sigma^2)$ gives Eq. (3).

Lemma 2 (Closed forms)**.** *Under Assumptions 1 to 3,*

$$A_{\mathrm{noDI}} = \sqrt{\frac{\rho}{1+\rho}} = \sqrt{\mathrm{LGC}}, \qquad A_{\mathrm{DI}} = \frac{c}{\sqrt{1+\kappa/(g_r\rho)}}. \tag{4}$$

Proof. For the baseline, $\mathbb{E}[g] = \mu$ gives $\langle \mathbb{E}[g], \hat{\mu}\rangle = \|\mu\|$ and $\mathbb{E}\|g\|^2 = \|\mu\|^2 + n\sigma^2$, so

$$A_{\mathrm{noDI}} = \frac{\|\mu\|}{\sqrt{\|\mu\|^2 + n\sigma^2}} = \sqrt{\frac{\rho}{1+\rho}}.$$

For DI, $\mathbb{E}[\bar{g}] = R\mu$ gives $\langle R\mu, \hat{\mu}\rangle = \mu^\top R\mu/\|\mu\| = c\sqrt{g_r}\,\|\mu\|$, while $\mathbb{E}\|\bar{g}\|^2 = \|R\mu\|^2 + \frac{\sigma^2}{m}\operatorname{tr}(R^2) = g_r\|\mu\|^2 + n\sigma^2\kappa$. Hence

$$A_{\mathrm{DI}} = \frac{c\sqrt{g_r}\,\|\mu\|}{\sqrt{g_r\|\mu\|^2 + n\sigma^2\kappa}} = \frac{c}{\sqrt{1+\kappa/(g_r\rho)}}.$$

Theorem 1 (Scissors crossover)**.** *Under Assumptions 1 to 3, assume $\kappa/g_r < c^2 < 1$, and define*

$$\varphi(\rho) := \left(\frac{A_{\mathrm{DI}}}{A_{\mathrm{noDI}}}\right)^2 = c^2\,\frac{\rho+1}{\rho+\kappa/g_r}. \tag{5}$$

Then φ is strictly decreasing, with $\varphi(0^+) = c^2 g_r/\kappa > 1$ and $\varphi(\infty) = c^2 < 1$, so it has a unique root $\rho^\star > 0$ given by

$$\rho^\star = \frac{c^2 - \kappa/g_r}{1-c^2}. \tag{6}$$

Consequently $A_{\mathrm{DI}} > A_{\mathrm{noDI}}$ iff $\rho < \rho^\star$, and $A_{\mathrm{DI}} < A_{\mathrm{noDI}}$ iff $\rho > \rho^\star$. Equivalently, by Lemma 1, DI helps iff $\mathrm{LGC} < \tau^\star$ *and hurts iff* $\mathrm{LGC} > \tau^\star$*, where $\tau^\star = \rho^\star/(1+\rho^\star)$. In particular, as* $\mathrm{LGC} \to 1$*, DI necessarily hurts.*

Proof. By Lemma 2, $\varphi(\rho) = c^2(1+1/\rho)/(1+\kappa/(g_r\rho))$, which simplifies to Eq. (5). Differentiating,

$$\varphi'(\rho) = c^2\,\frac{(\rho+\kappa/g_r) - (\rho+1)}{(\rho+\kappa/g_r)^2} = c^2\,\frac{\kappa/g_r - 1}{(\rho+\kappa/g_r)^2} < 0,$$

since $\kappa/g_r < 1$; hence φ is strictly decreasing. The limits are immediate, and $\varphi(0^+) = c^2 g_r/\kappa > 1 \Leftrightarrow \kappa/g_r < c^2$. Being continuous and strictly decreasing from above 1 to below 1, φ attains the value 1 at a unique $\rho^\star$; solving $c^2(\rho+1) = \rho + \kappa/g_r$ yields Eq. (6). The LGC statement then follows from the strict monotonicity of $\mathrm{LGC}(\rho)$ in Lemma 1.

Assumption 3 takes the target's vulnerable direction to be μ itself, which is a strong premise. The next proposition shows the theorem does not actually need it: it survives for an *arbitrary* target direction, under a weaker condition that Sec. 4.2 measures directly.

Proposition 2 (General target direction; relaxing Assumption 3)**.** *Let the target's vulnerable direction be an arbitrary unit vector u (not assumed equal to μ), and write $\gamma_\mu = |\cos(\mu, u)|$ and $\gamma_R = |\cos(R\mu, u)|$, so the figure of merit uses $\langle \mathbb{E}[d], u\rangle$ in place of $\langle \mathbb{E}[d], \hat{\mu}\rangle$. Then*

$$\varphi(\rho) = \Gamma\,\frac{g_r(\rho+1)}{g_r\rho+\kappa}, \qquad \Gamma := \frac{\gamma_R^2}{\gamma_\mu^2}, \tag{7}$$

which is still strictly decreasing (as $\kappa < g_r$), with $\varphi(\infty) = \Gamma$. Hence DI necessarily hurts as $\mathrm{LGC} \to 1$ *if and only if $\Gamma < 1$, i.e. $\gamma_R < \gamma_\mu$: resize moves the surrogate signal away from the target's vulnerable direction. Theorem 1 is the special case $u = \mu$ (γ_μ=1, γ_R=c, Γ=c^2).*

Proof. With the target u, $A_{\mathrm{noDI}}^2 = \gamma_\mu^2\,\rho/(\rho+1)$ and $A_{\mathrm{DI}}^2 = \gamma_R^2\,g_r\rho/(g_r\rho+\kappa)$, whose ratio is Eq. (7). The factor Γ does not depend on ρ, so φ' has the same sign as in Theorem 1 ($\propto \kappa - g_r < 0$), and $\varphi(\infty) = \Gamma$, $\varphi(0^+) = \Gamma g_r/\kappa$. All steps are machine-verified (Sec. B.4).

Remark 1 (Why this defuses the "strong assumption" concern). The exact-alignment premise "target $\propto \mu$" is replaced by the weaker condition $\gamma_R < \gamma_\mu$: a robust surrogate is harmed by DI whenever resize reduces the signal's alignment with the target direction. This requires no architectural match between surrogate and target. Its sign is the empirical counterpart of what Sec. 4.2 measures: the robust sources there show $D_{\text{sign}} < 0$ (DI reduces source–target sign-alignment), matching $\gamma_R < \gamma_\mu$. So the assumption the theorem really needs is not assumed but observed.

Remark 2 (Why resize and not translation). The bias factor enters Theorem 1 only through $c, g_r < 1$, *i.e.* through $R \neq I$. Resize is a genuine low-pass contraction ($R \prec I$), giving $c, g_r < 1$ and a finite crossover. A small, centered random translation, by contrast, averages to $R \approx I$ (shifts cancel in expectation), so $c, g_r \approx 1$ and the bias factor ≈ 1: DI's translation component is then near-neutral, matching the resize/translation decomposition in Tab. 6. The theorem also reproduces the monotone ϵ-sweep (Sec. 3.4): increasing adversarial-training strength raises ρ (hence LGC) through the single threshold $\tau^\star$.

B.3 Corollary: a frequency-domain pre-method (SSA) shifts the regime

The same theory explains the one apparent exception in Tab. 2, SSA. SSA augments the input with Gaussian noise and a random DCT spectral down-scaling before the gradient is taken; to first order this is a symmetric low-pass operator applied to the gradient, with its own ensemble averaging. We model such a method as a pre-operator that conditions the gradient before the DI step.

Corollary 1 (Pre-method regime shift)**.** *Let a method M apply, before the DI step, a symmetric PSD operator S ($0 \preceq S \preceq I$) to the gradient with m_S-fold averaging, so the M-conditioned gradient is $g_M = S\mu + \xi$ with $\mathbb{E}\xi = 0$ and $\mathrm{Cov}(\xi) = \sigma^2 S^2/m_S$. Then its signal-to-noise ratio is*

$$\rho_{\text{eff}} = \frac{\|S\mu\|^2}{\sigma^2 \operatorname{tr}(S^2)/m_S} = \rho \cdot \frac{s_r}{\kappa_S}, \qquad s_r = \frac{\|S\mu\|^2}{\|\mu\|^2}, \quad \kappa_S = \frac{\operatorname{tr}(S^2)}{n\, m_S}, \tag{8}$$

and the effective consistency is $\mathrm{LGC}_{\text{eff}} = \rho_{\text{eff}}/(1+\rho_{\text{eff}})$. If S removes more noise than signal, $\kappa_S < s_r$, then $\mathrm{LGC}_{\text{eff}} > \mathrm{LGC}$: M moves the surrogate rightward along the crossover axis of Theorem 1. Applying Theorem 1 to g_M, DI on top of M helps iff $\mathrm{LGC}_{\text{eff}} < \tau_S^\star$. Hence (i) a surrogate already in the robust region ($\mathrm{LGC} > \tau^\star$) is pushed further in and DI's harm increases; (ii) a standard surrogate ($\mathrm{LGC} < \tau^\star$) is pushed toward $\tau_S^\star$, where the DI effect is small and its sign is sensitive to the strength of M.

Proof. For independent draws,

$$\mathbb{E}\langle g_M, g_M' \rangle = \|S\mu\|^2, \qquad \mathbb{E}\|g_M\|^2 = \|S\mu\|^2 + \tfrac{\sigma^2}{m_S} \operatorname{tr}(S^2),$$

which give ρ_{eff} in Eq. (8) and, by Lemma 1 applied to g_M, $\mathrm{LGC}_{\text{eff}} = \rho_{\text{eff}}/(1+\rho_{\text{eff}})$. The ratio $\rho_{\text{eff}}/\rho = s_r/\kappa_S$ is immediate, and since $x \mapsto x/(1+x)$ is increasing, $\mathrm{LGC}_{\text{eff}} > \mathrm{LGC} \Leftrightarrow s_r > \kappa_S$. The two consequences are Theorem 1 applied to g_M with $\rho \to \rho_{\text{eff}}$.

Remark 3 (SSA matches the data on both sides). SSA instantiates S as a DCT low-pass that preserves the low-frequency band where robust gradients concentrate while attenuating broadband noise (and averages over its ensemble), so $\kappa_S < s_r$ and the shift is rightward. This predicts both empirical signatures. On a robust surrogate the harm is the largest of any method (Tab. 2) and is resize-driven (Tab. 7), because the resize bias of Theorem 1 now acts on the SSA-narrowed, even-lower-frequency signal. On a standard surrogate, SSA raises the effective LGC to near $\tau^\star$, so the DI effect is small and its sign flips with the SSA strength, which explains the -0.6% of the default config (Tab. 2) versus the $+7.8\%$ of the stronger-ensemble config (Tab. 7). Fig. 3 (main text) visualizes the shift.

B.4 Machine-checked verification

All algebraic steps above (the moments of Lemmas 1 and 2, the sign of φ', the limits, the root $\rho^\star$, the pre-method identity $\rho_{\text{eff}} = \rho\, s_r/\kappa_S$ of Corollary 1, and the general-target ratio $\varphi = \Gamma\, g_r(\rho+1)/(g_r\rho+\kappa)$ of Proposition 2) are verified symbolically with a computer-algebra system; the verification is released with our code. As an independent check, we run a Monte-Carlo simulation with a DCT-based low-pass resize

operator R (n=64, m=10 EOT samples) and a signal μ with mixed low/high-frequency content. These constants (c=0.95, g_r=0.81, κ=0.056) are illustrative choices for the simulation, not measured from a real surrogate, and they place the crossover at $\tau^\star$=0.905; a different choice would move it (*e.g.*, to 0.7 or 0.99). What the theorem guarantees is the *qualitative* behaviour, not this number. With these constants, both the theorem's A-metric and the directly measured $\mathbb{E}[\cos(d, \mu)]$ flip from DI-helpful to DI-harmful at the same point, LGC $\approx$ 0.90 (Fig. 4); we note this happens to be close to CG-DI's empirically chosen τ=0.92, but we do not claim the theory predicts that value. The same simulation adds an SSA-like low-pass pre-operator S (s_r=0.79, κ_S=0.026, m_S=20, so $\rho_{\text{eff}}/\rho \approx 31\times$) and confirms Corollary 1: it lifts a standard surrogate's effective LGC from 0.55 to 0.974 (past $\tau^\star$), flipping DI's effect from clearly helpful (+0.19) to near-zero/harmful (−0.02), while a robust surrogate stays harmful, reproducing the sign behaviour of Tabs. 2 and 7. The crossover (Fig. 4) and the SSA regime shift (Fig. 3) are visualized in the main text.

B.5 Estimating the constants on real surrogates

We go a step further and estimate (c, g_r, κ) directly from real surrogates, taking R to be the deterministic resize-only low-pass (downsample by r=0.9 then upsample; the translation component is modeled separately as $R \approx I$, so it is excluded here). We denoise the signal μ by averaging the input-gradient over K=5 small-noise probes, set $c = \cos(R\mu, \mu)$, $g_r = \|R\mu\|^2/\|\mu\|^2$, and estimate $\kappa = \text{tr}(R^2)/(nm)$ by Hutchinson probing (N=500 ImageNet images, m=10). Tab. 14 reports the result for 13 surrogates.

Table 14: Theorem constants (c, g_r, κ) estimated on real surrogates (resize-only R, ImageNet, N=500). LGC is the canonical value of Tabs. 4 and 22. "Pred." is the theorem's prediction (DI helps iff LGC $< \tau^\star$); "Src." marks the surrogates whose DI direction we measured as a source. The instantiability check ($g_r \leq 1$, $\kappa/g_r < c^2$) and the 8/8 source-side accuracy are discussed in the text.

Source	Type	c	g_r	κ/g_r	$\tau^\star$	LGC	Pred.	Src.
ResNet50	Std	0.86	0.38	0.068	0.72	0.64	help	✓
ViT-B/16	Std	0.95	0.63	0.041	0.89	0.90	hurt	—
Swin-B	Std	0.83	0.36	0.072	0.67	0.36	help	—
ConvNeXt-B	Std	0.83	0.35	0.075	0.65	0.87	hurt	—
DenseNet121	Std	0.91	0.56	0.046	0.81	0.92	hurt	—
InceptionV3	Std	0.93	0.51	0.050	0.86	0.81	help	—
Engstrom	Rob	0.84	0.43	0.061	0.68	0.98	hurt	✓
Salman ϵ=0.5	Rob	0.93	0.62	0.042	0.87	0.99	hurt	✓
Salman ϵ=1	Rob	0.87	0.50	0.051	0.74	0.99	hurt	✓
Salman ϵ=2	Rob	0.78	0.36	0.071	0.58	0.99	hurt	✓
Salman ϵ=4	Rob	0.76	0.31	0.084	0.54	0.98	hurt	✓
Salman ϵ=8	Rob	0.78	0.31	0.085	0.57	0.80	hurt	✓
Mo2022	Rob	0.97	0.76	0.034	0.94	1.00	hurt	✓

Two things follow. (i) The model is not vacuous on real data: $g_r \leq 1$ and $\kappa/g_r < c^2$ for all 13 surrogates, so the crossover exists. (ii) On the eight surrogates whose source-side DI direction we actually measured, ResNet50 (DI helps) and the seven robust sources (DI hurts), the rule "DI helps iff LGC $< \tau^\star$" is correct 8/8, and the independently estimated $\tau^\star$ falls below every robust LGC and above the clearly-standard ones, i.e. it separates the two regimes. The robust predictions are close to the near-definitional LGC $\to$ 1 limit; the one robust surrogate not near that limit is Salman ϵ=8, whose capacity strain lowers its LGC to 0.80, and it is still predicted correctly ($0.80 > \tau^\star$=0.57), and is in fact where a single fixed threshold fails while the per-surrogate $\tau^\star$ does not (Sec. L). For the five standard models we did not run as sources we make no claim; the theory predicts harm for the three with unusually high LGC (ViT-B 0.90, ConvNeXt 0.87, DenseNet 0.92, the last being the documented anomaly of Tab. 12), consistent with their atypically consistent gradients, but we lack source-side ground truth for them. Two caveats temper everything: μ is estimated with only K=5 probes and is noisier for the high-variance standard gradients, so the standard-side $\tau^\star$ carry more uncertainty (which is also why the borderline cases sit on the standard side); and $\tau^\star$ varies across surrogates (0.54–0.94), so this still does not yield a single predicted threshold, and CG-DI's τ=0.92 remains an empirical

choice (Sec. P.6). The qualitative crossover and the LGC $\rightarrow 1$ limit are the operationalization-independent guarantees.

C Experimental Hyperparameters

All attacks are untargeted and optimized under the L_∞ norm; the standard configuration follows common transfer-attack practice (Tab. 15).

Table 15: Hyperparameter settings for attacks.

Parameter	CIFAR-10	ImageNet
Perturbation budget (ϵ)	8/255	16/255
Step size (α)	2/255	2/255
Iterations (T)	10	10
LGC probe (ϵ_{chk})	1/255	1/255
LGC samples (K)	5	5

Baseline method settings. MI-FGSM: momentum decay μ=1.0. DI-FGSM: transformation probability p=0.5 (default) unless swept; resize factor r=0.9 (images resized to a random size in $[\lfloor 224 \times 0.9 \rfloor, 224]$ pixels then zero-padded back; for CIFAR-10, 32 $\rightarrow$ 35 pixels). TI-FGSM: Gaussian kernel k=15 (ImageNet), k=5 (CIFAR-10).

Compute. All experiments ran on a server with an NVIDIA RTX 4090 (24 GB) and an Intel Xeon Platinum 8352V (16 vCPU) @ 2.10 GHz with 120 GB RAM, Python 3.10, CUDA 11.8.

D Model Architectures and Training Recipes

Tab. 16 lists every surrogate, its architecture, and its training recipe. All CIFAR-10 robust models are L_∞, ϵ=8/255 RobustBench checkpoints. The robust sources used in the main analyses span two architecture families (ResNet-50 and ViT-B) and several recipes (PGD-AT, TRADES, semi-supervised, diffusion-augmented, ViT-aware AT), so the Scissors Effect cannot be attributed to any single architecture or training procedure.

E Additional Experimental Details

All modern attacks were implemented via TransferAttack (Trustworthy-AI-Group, 2023) following their official codebases, with default hyperparameters except the DI probability p.

E.1 Targeted Attack Generalization

We evaluate targeted MI-FGSM (Dong et al., 2018) on ImageNet (N=1,000, 5 seeds, ϵ=16/255) with randomly assigned target classes (Tab. 17).

Despite the inherently low absolute ASR of targeted cross-architecture transfer, the direction is consistent: DI helps the standard surrogate ($\sim +0.6\%$) and harms the robust one (-1.2% avg, $\sim -32\%$ relative). The Scissors Effect is thus independent of attack objective (untargeted vs. targeted).

E.2 Unified Y-Axis Visualization

Fig. 1 uses separate Y-axes to highlight direction; Fig. 6 shows the same data on a shared scale.

F Spectral Visualization of Gradient Geometry

We analyze the frequency spectra of input gradients via 2D FFT (Fig. 7).

Table 16: Surrogate architectures and training recipes. Robust accuracy (RA) for CIFAR-10 robust models is the RobustBench AutoAttack value.

Dataset	Model	Architecture	Recipe	RA (%)
CIFAR-10 (Std)	ResNet18	ResNet-18	natural	—
	ResNet50	ResNet-50	natural	—
	VGG16	VGG-16	natural	—
	DenseNet121	DenseNet-121	natural	—
	ViT-B/16	ViT-B/16	natural	—
CIFAR-10 (Rob)	Engstrom	ResNet-50	PGD-AT	49.3
	Rice	WRN-34-20	PGD-AT (early stop)	53.4
	Gowal	WRN-70-16	PGD-AT (extra data)	57.1
	Carmon	WRN-28-10	semi-supervised (RST)	59.5
	Wang2023	WRN-28-10	diffusion data	67.3
	Zhang (TRADES)	WRN-34-10	TRADES	44.8
	Sehwag	WRN-34-10	proxy distribution	60.3
	Sehwag-R18	ResNet-18	proxy distribution	55.5
ImageNet (Std)	ResNet50	ResNet-50	natural	—
	ViT-B/16	ViT-B/16	natural	—
	DenseNet121	DenseNet-121	natural	—
	InceptionV3	InceptionV3	natural	—
	Swin-B	Swin-B	natural	—
	ConvNeXt-B	ConvNeXt-Base	natural	—
ImageNet (Rob)	Engstrom	ResNet-50	PGD-AT, ϵ=4/255	—
	Salman (×5)	ResNet-50	PGD-AT, $\epsilon \in \{0.5, 1, 2, 4, 8\}/255$	—
	Mo2022	ViT-B	ViT-aware AT, ϵ=8/255	—

Table 17: Targeted attack transfer ASR (%). DI harms robust surrogates even in the targeted setting.

Surrogate	Target	Naked	+DI	Δ (abs)	Δ (rel)
ResNet50 (Std)	Swin-B	0.2%	0.7%	+0.5%	+250%
ResNet50 (Std)	IncV3	0.6%	1.2%	+0.6%	+100%
Engstrom (Rob)	Swin-B	2.7%	1.8%	**−0.9%**	−33%
Engstrom (Rob)	IncV3	5.2%	3.6%	**−1.6%**	−31%

High-frequency ratio. From the 2D FFT $G(f)$ we form the log-magnitude spectrum $S(f) = \log(1+|G(f)|)$, take its radially-averaged profile $P(r)$, and define

$$R_{HF} = \frac{\sum_{r>r_{\max}/2} P(r)}{\sum_r P(r)}, \tag{9}$$

where r is the radial spatial frequency and $r_{\max}$ the maximum resolvable frequency. The log transform prevents dominant low-frequency components from masking the high-frequency signal.

The radial profile (Fig. 8) shows a ∼1.7× HF gap (standard mean ≈ 0.48 vs. robust ≈ 0.28), explaining why DI's resize, which corrupts low-frequency signals via interpolation, is particularly harmful for robust models. Among standard models, attention-based architectures (ViT-B/16: R_{HF}=0.37) produce smoother gradients than CNNs (ResNet50: 0.55), consistent with their higher LGC.

G Per-Image LGC Computation

Although Alg. 1 defines LGC per image, its variability depends on model type. Robust models have very low per-image variance (std < 0.02 over 1,000 images); standard models vary more widely (*e.g.*, ResNet50, model-level mean ≈ 0.64), but their mean stays well below τ=0.92. Since the model-level mean is a reliable

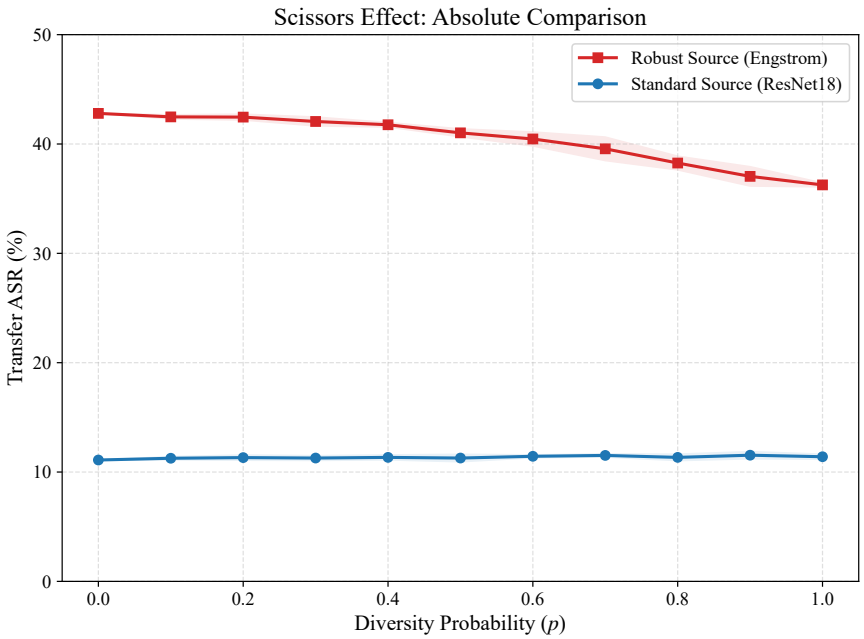

Figure 6: Scissors Effect with unified Y-axis (same data as Fig. 1). The robust surrogate has higher absolute ASR across all p, while the divergent trends (robust ↓, standard ↑) remain evident.

regime indicator, one can estimate the binary decision on a small calibration batch (N=50), amortizing the K=5 queries across all attack iterations.

H Optimal p^* Identifiability Analysis

We measure the ASR gap between adjacent p values (the margin) as a notion of signal strength for strategy selection (Fig. 9 and Tab. 18).

Table 18: Identifiability summary: p^* margin statistics.

Dataset	Models	Margin $\geq 1\%$	Mean margin
CIFAR-10	10	0	0.20%
ImageNet	9	3	0.96%

I Controlled ϵ-Sweep: Second Target (ViT-B)

Tab. 19 repeats the controlled ϵ-sweep of Sec. 3.4 against a second target (ViT-B), confirming the same qualitative picture: DI flips from beneficial to harmful in the little-robustness regime and the harm deepens with robustness.

J Cross-Recipe Replication Panel

Tab. 20 gives the cross-recipe panel of Sec. 3.4: five robust ImageNet surrogates spanning four adversarial-training recipes (PGD-AT, ViT-aware AT, ARES adversarial training, ConvStem AT) and four architectures (ResNet-50, ViT-B, ConvNeXt-B, ViT-B-ConvStem), each attacked against two targets (Swin-B and ConvNeXt-B), with the per-source LGC measured alongside (N=500, 3 seeds, ϵ_{attack}=16/255). DI harms every robust surrogate on all nine cross-family pairs and benefits the standard surrogate on both; the sign of

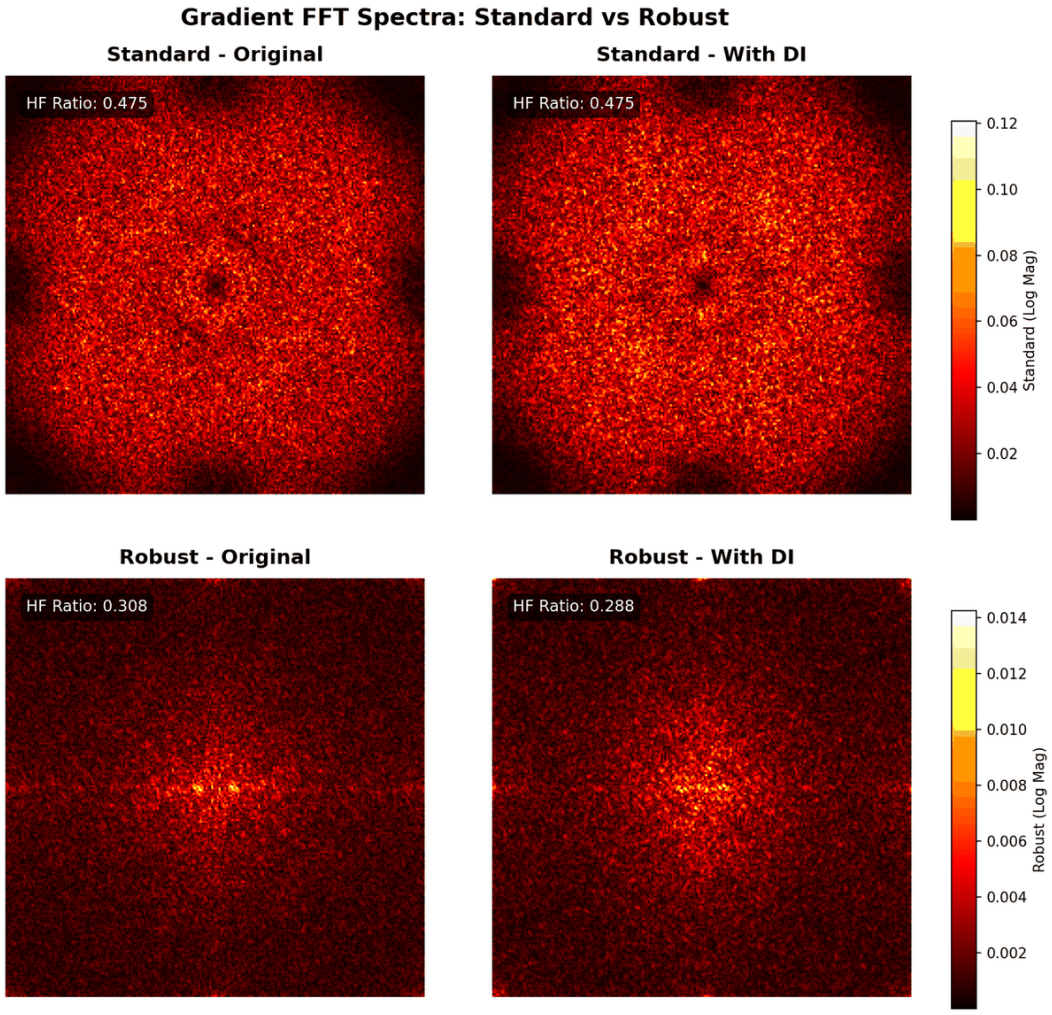

Figure 7: 2D FFT spectra of gradient maps. (*Top row*) Standard model (ResNet18): diffuse, broadband content. (*Bottom row*) Robust model (Engstrom): concentrated low-frequency energy. Left/right columns compare original vs. DI-augmented inputs.

Table 19: Controlled ϵ-sweep, target ViT-B (N=500, 3 seeds, ϵ_{attack}=16/255). Same crossover and monotone harm as the Swin-B target in Tab. 3.

ϵ_{train} ($\times 255$)	MI-FGSM	DI-FGSM	D = DI−MI
0 (Standard)	34.4	50.4	+16.0
0.5	**91.8**	91.1	−0.7
1	90.0	90.1	+0.1
2	87.8	82.1	−5.7
4	78.8	71.4	−7.4
8	62.6	52.9	**−9.7**

D follows the LGC regime, not the recipe or architecture. This establishes that the Scissors direction replicates beyond the Salman ResNet-50 / PGD-AT family and to a second target; it is a replication of the sign, not a second controlled strength sweep (no public fixed-backbone ϵ-spectrum exists outside Salman's). The one source–target pair sharing an architecture family (ARES ConvNeXt-B source against the ConvNeXt-B target) is marked and excluded from the nine cross-family count.

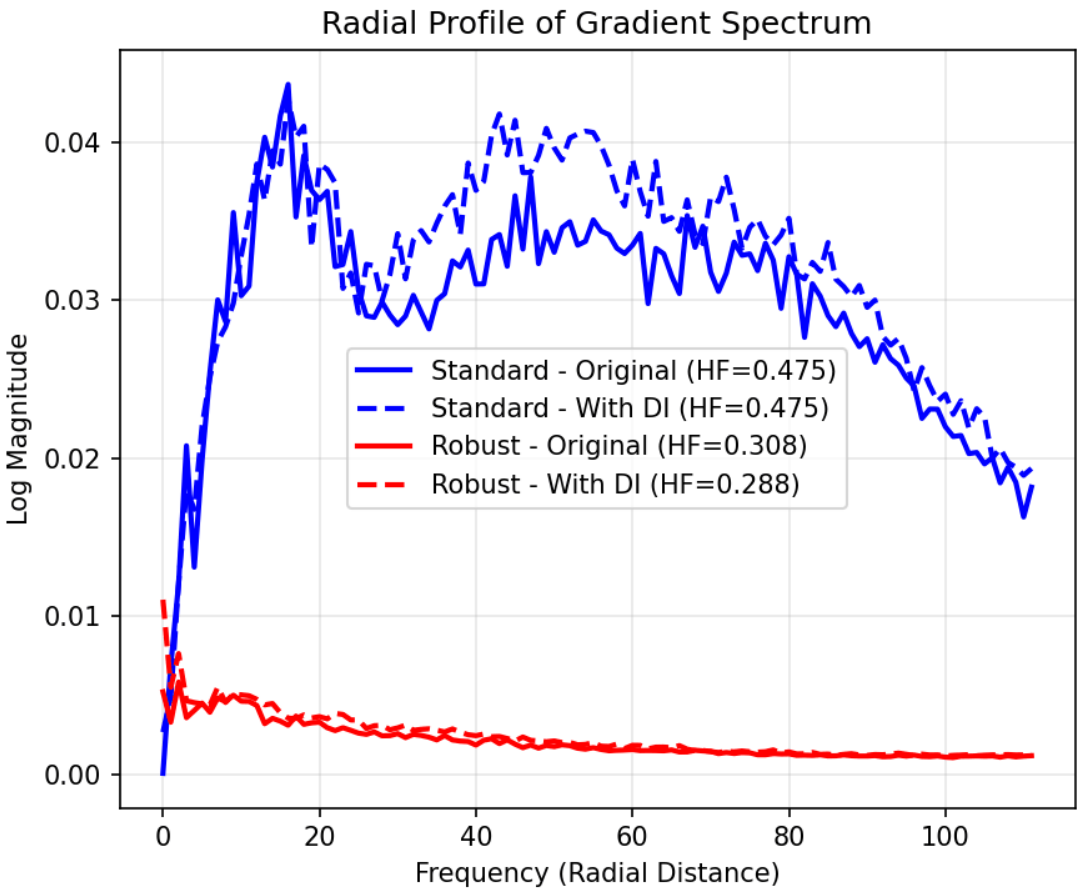

Figure 8: Radial spectral profile. Standard models (*blue*) have higher log-magnitude at high frequencies; robust models (*red*) concentrate at low frequencies. Shaded regions show standard deviation across samples.

Table 20: Cross-recipe replication panel (N=500, 3 seeds, target Swin-B and ConvNeXt-B). D = DI−MI (%). DI harms all five robust surrogates across four recipes and four architectures on both targets; the standard surrogate benefits; the sign tracks LGC. † same-family pair (excluded from the cross-family count).

Source	Recipe	Arch	LGC	D (Swin-B)	D (ConvNeXt-B)
Standard RN50	—	RN50	0.63	**+13.0**	**+15.9**
Engstrom	PGD-AT	RN50	0.98	−12.3	−18.7
Salman	PGD-AT	RN50	0.98	−10.9	−15.2
Mo2022	ViT-AT	ViT-B	1.00	−6.1	−6.7
ARES	ARES-AT	ConvNeXt-B	1.00	−5.5	−7.1†
Singh	ConvStem-AT	ViT-B-CS	1.00	−4.5	−5.3

K Source–Target Overlap: Same-Family Control

Tab. 21 gives the full numbers for the same-family (RN50 → RN50) control of Sec. 3.5. The DI effect's sign is set by the source regime, not by source–target family overlap: a standard source benefits from DI even against robust RN50 targets, while a robust source is harmed even against a standard RN50 target.

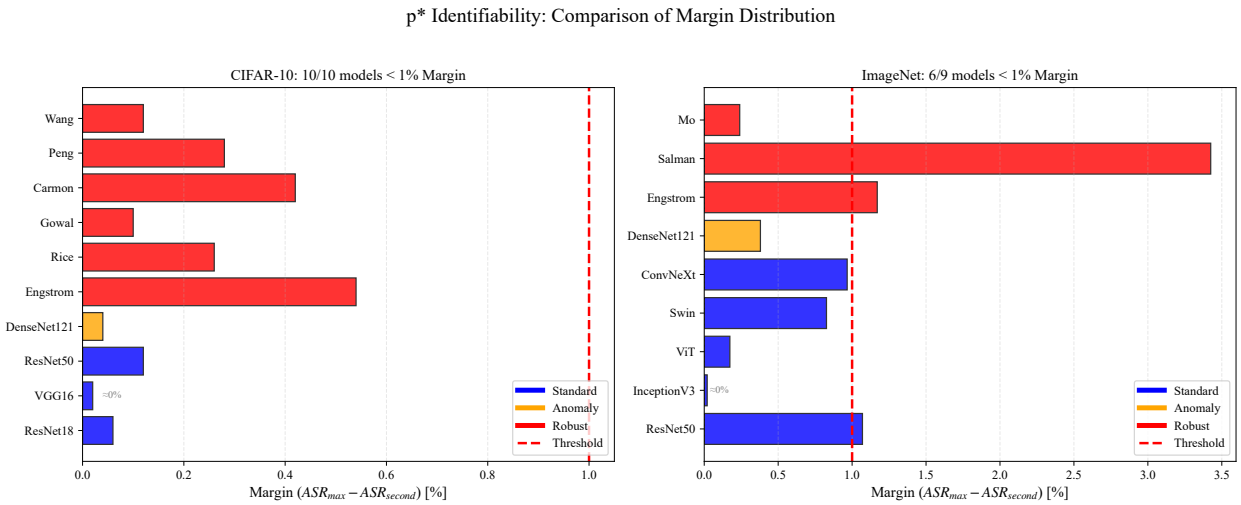

Figure 9: Identifiability margin distribution. (*Left*) CIFAR-10, standard DI (r=0.9): all 10 models have margin < 1%. (*Right*) ImageNet: 6/9 below 1%; only 3 robust surrogates exceed it.

Table 21: Same-family control (N=500, 3 seeds). D = DI−MI. Cross-family baselines (against Swin-B) repeated for reference.

Source	Target	Relation	MI	DI	D
Standard RN50	Swin-B	cross-family (baseline)	41.2	54.2	+13.0
Standard RN50	Engstrom RN50	same-family (robust tgt)	37.4	38.7	+1.3
Standard RN50	Salman ϵ=2 RN50	same-family (robust tgt)	34.0	35.1	+1.1
Salman ϵ=2 RN50	Swin-B	cross-family (baseline)	78.4	65.9	−12.5
Salman ϵ=2 RN50	Standard RN50	same-family (std tgt)	88.8	77.2	**−11.6**

L LGC Across the Robustness Spectrum, and a Corner Case

We compute LGC on 8 robust ImageNet sources spanning the controlled ϵ-spectrum plus Engstrom and Mo2022 (clean-correct subsets, K=5, σ=1/255; Tab. 22). LGC tracks DI sensitivity quantitatively: against the DI effect D from Tab. 3, Pearson $r(\text{LGC}, D) = -0.87$, p=0.025 (n=6; Fig. 10). CG-DI classifies 7/8 sources correctly.

Table 22: LGC across the robustness spectrum (8 robust ImageNet sources). CG-DI is correct on 7/8. The single miss (Salman ϵ=8 ResNet-50) is an architecture-capacity corner case, not an LGC failure (see text).

Source	Arch	ϵ_{train}	LGC	CG-DI
Standard	RN50	0	0.65	p=0.8 ✓
Salman	RN50	0.5	0.99	p=0 ✓
Salman	RN50	1	0.99	p=0 ✓
Salman	RN50	2	0.99	p=0 ✓
Salman	RN50	4	0.98	p=0 ✓
Engstrom	RN50	4	0.98	p=0 ✓
Mo2022	ViT-B	8	**1.00**	p=0 ✓
Salman	RN50	8	**0.80**	p=0.8 (corner case)

The corner case. Salman ϵ=8/255 ResNet-50 has LGC 0.80, so CG-DI would (incorrectly) enable DI. We attribute this to architecture capacity, not an LGC failure, on three independent grounds. (i) At the same training ϵ=8/255, Mo2022's ViT-B yields LGC 1.00 and CG-DI is correct; the only varying factor is

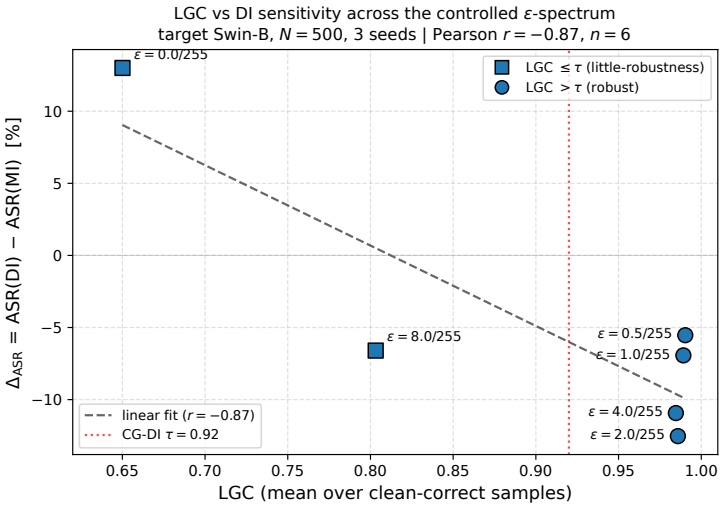

Figure 10: LGC vs. the DI effect D across the controlled ϵ-spectrum (Pearson $r = -0.87$, p=0.025, n=6). The CG-DI threshold τ=0.92 is marked.

the backbone. (ii) Salman *et al.*'s own reported clean accuracy for this checkpoint falls from 76.1% (ϵ=0) to 54.5% (ϵ=8), a documented 21-point capacity strain. (iii) RobustBench and Mo et al. (2022) reserve higher-capacity backbones (WRN-70-16, ViT-B) for $\epsilon \geq 8/255$, using ResNet-50 only for $\epsilon \leq 4/255$. We report the case rather than tuning it away.

It is worth being precise about what fails here, because the per-surrogate theory of Sec. B.5 sharpens it. The capacity strain lowers this checkpoint's LGC (from the ≈0.98 typical of robust models to 0.80). What that breaks is CG-DI's *single fixed* threshold τ=0.92: $0.80 < 0.92$, so CG-DI misroutes it as standard. It is not a failure of LGC or of the theorem: the same capacity strain also lowers this surrogate's crossover threshold (its resize geometry gives $\tau^\star$=0.57), so the per-surrogate test LGC(0.80) $> \tau^\star(0.57)$ still correctly predicts harm (Tab. 14). In other words, a per-surrogate threshold is robust to a capacity-induced LGC drop in a way a one-size-fits-all τ is not, which is exactly why CG-DI treats τ as a deliberately conservative guardrail rather than a precise predictor.

M Multimodal Surrogate: CLIP Transfer

We attack from CLIP ViT-B/32 (Radford et al., 2021) to robust CIFAR-10 targets (Tab. 23). CLIP's gradients are "standard-like" (LGC $\approx$ 0.72; Tab. 4), likely because contrastive pretraining does not induce AT-level gradient smoothness, so CG-DI enables moderate diversity (p=0.8) and reaches near-oracle performance.

Table 23: CLIP ViT-B/32 → robust targets. CG-DI selects a near-optimal p=0.8.

Target	p=0	p=0.5	p=1.0	CG-DI
Engstrom	13.9%	14.1%	14.5%	**14.3%**
Rice	16.1%	16.5%	16.6%	**16.6%**
Wang-WRN	8.4%	8.9%	9.0%	**9.0%**

N Extended ImageNet Architecture Breakdown

Tab. 24 gives the per-target CG-DI results underlying the averaged ImageNet metrics in Tab. 1, plus an extra target (ResNet50) for the robust source.

Table 24: ImageNet per-architecture breakdown. CG-DI adapts to each source type.

Source	Target	p=0	p=1	CG-DI	Δ
ResNet50	IncV3	51.9	66.5	**66.1**	+14.6
	ViT	34.9	49.2	**48.1**	+14.3
	Swin	40.9	54.1	**53.5**	+13.2
	ConvNeXt	48.3	64.6	**64.9**	+16.3
Engstrom	IncV3	**86.0**	80.9	**86.0**	-5.1
	ResNet50	**81.3**	68.7	**81.3**	-12.6
	ViT	**82.3**	75.5	**82.3**	-6.8
	Swin	**68.6**	55.5	**68.6**	-13.1
	ConvNeXt	**66.9**	50.7	**66.9**	-16.2

O Model Preprocessing Details

Tab. 25 documents input sizes and normalization for ImageNet targets. All use standard ImageNet normalization; InceptionV3 requires 299×299, others 224×224. CIFAR-10 RobustBench models use native 32×32 inputs; torchvision architectures used for cross-architecture experiments are upscaled to 224×224.

Table 25: Model preprocessing configuration for ImageNet targets.

Model	Input size	Norm mean	Norm std
ResNet50	224×224	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
InceptionV3	**299×299**	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
ViT-B/16	224×224	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
Swin-B	224×224	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
ConvNeXt-B	224×224	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
Engstrom (Robust)	224×224	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)

P Additional Analysis

P.1 Computational Cost of LGC

Tab. 26 benchmarks LGC (K=5) against a 10-step MI-FGSM attack. Per-batch LGC adds ∼76% overhead, reduced to ∼5% with pre-computation on a small calibration set (LGC is stable across set sizes; Tab. 27).

Table 26: Runtime overhead (batch size 32).

Component	Time (s)	Overhead
LGC (K=5)	0.239	—
Attack (T=10)	0.316	+75.6%
Amortized (N=1000)		
Pre-comp (50 imgs)	0.5	—
Attack	10.1	**+5.0%**

P.2 Impact of Translation (TI-FGSM) on Robust Models

We compare a translation-only variant (TI-FGSM, kernel 5) to standard DI and the naked baseline (Tab. 28). Kernel-based translation invariance degrades robust-model attacks by −18.4%, even more than standard DI (−6.9%), because the convolution kernel over-smooths already-consistent robust gradients. This differs from the mild random (EOT) translation used in the main-paper decomposition (Tab. 6), and reinforces that robust gradients are highly sensitive to spatial structure.

Table 27: Stability of model-level LGC estimation across calibration set sizes.

Calibration set size	Estimated LGC
N=50	0.9815 ± 0.0091
N=1000	0.9812 ± 0.0098

Table 28: Input transformations on a robust surrogate (Engstrom → Swin-B, N=500).

Method	Transform	ASR (%)
CG-DI (p=0)	none (naked)	**67.6**
DI-FGSM (p=0.5)	resize + translation	60.7
TI-FGSM	translation only (kernel)	49.2

P.3 Interpolation Mode Ablation (Ruling Out Artifacts)

To rule out that the effect is tied to the bilinear interpolation in DI's resize, we ablate three modes, namely bilinear (default), bicubic, and antialiased, while sweeping $p \in \{0, 0.3, 0.5, 0.7, 1.0\}$, using `torchattacks` (Kim, 2020) DIFGSM and modifying only the interpolation call (3 seeds, ImageNet, N=1,000; Tab. 29).

Table 29: Interpolation mode ablation. Average transfer ASR (%) across 4 targets; Δ is relative to p=0 (44.0% Standard, 76.0% Robust).

Source	Interp mode	$p=0$	$p=0.3$	$p=0.5$	$p=0.7$	$p=1.0$
Standard	Bilinear	44.0	$52.8^{+8.8}$	$55.6^{+11.6}$	$57.2^{+13.2}$	$58.9^{+14.9}$
	Bicubic	44.0	$51.7^{+7.7}$	$54.2^{+10.2}$	$56.0^{+12.0}$	$57.3^{+13.3}$
	Antialias	44.0	$52.3^{+8.3}$	$55.3^{+11.3}$	$57.5^{+13.5}$	$59.2^{+15.2}$
Robust	Bilinear	76.0	$73.7^{-2.3}$	$71.8^{-4.2}$	$69.8^{-6.2}$	$65.6^{-10.4}$
	Bicubic	76.0	$75.1^{-0.9}$	$74.6^{-1.4}$	$73.7^{-2.3}$	$70.9^{-5.1}$
	Antialias	76.0	$73.7^{-2.3}$	$71.8^{-4.2}$	$69.4^{-6.6}$	$65.0^{-11.0}$

All three modes show the same qualitative pattern (DI helps Standard +13–15%, harms Robust −5 to −11%): the *direction* is invariant across all three interpolation filters. The *magnitude*, however, is not: bicubic roughly halves the robust harm (−5.1% vs. bilinear −10.4%), while standard surrogates are largely insensitive to interpolation choice (<2% variation). We read this as support for, not against, the gradient-geometry account. A pure interpolation artifact would have no reason to keep a fixed sign across filters; what we instead see is a sign that never flips and a magnitude that scales with how aggressively the filter removes low-frequency content: bicubic has a gentler low-pass rolloff than bilinear, so on a robust surrogate's low-frequency gradient it injects less resize bias and does less harm. That is exactly the dependence the low-pass-bias mechanism (Sec. 4.5) predicts: the harm is set by the resize operator's contraction strength, not by an interpolation idiosyncrasy. The Scissors Effect is thus a property of gradient geometry interacting with a low-pass resize, with magnitude tunable by the filter and direction fixed.

P.4 A Per-Image Negative Control: Sign-Alignment Does Not Mediate Individual Flips

The sign-alignment evidence of Sec. 4.2 is a *population* statement: averaged over images, DI moves a robust surrogate's gradient the wrong way and a standard one's the right way. A natural follow-up question is whether it also acts as a *per-image* mediator: do the specific images whose transfer DI breaks have a larger negative sign-alignment shift than the images it leaves alone? We tested this directly and report a clean negative result, because it bounds what the sign-alignment column should be read to claim.

For each clean-correct image (Engstrom→Swin-B, N=500, 311 correct on both models, 3 seeds, ϵ=16/255) we paired (i) the per-image sign-alignment shift d_{sign} induced by DI (EOT= 10, r=0.9, the same metric as

Sec. 4.2) with (ii) the per-image transfer outcome under an actual attack (MI-FGSM vs. DI-FGSM at p=1). The attack side reproduces the Scissors Effect at the outcome level (net transfer change −14.9% robust, +16.1% standard; DI flips 18.5% of robust images from success to failure). But d_{sign} does not distinguish the flipped images: those DI breaks have mean $d_{\text{sign}}=-0.0001$, statistically indistinguishable from the unchanged images (+0.0001; Mann–Whitney p=0.43), the correlation between d_{sign} and a signed outcome (+1 helped, −1 hurt) is null (r=0.02, p=0.48), and P(DI hurts) is flat across d_{sign} quintiles (0.17–0.21).

We read this straightforwardly: the clean-point sign-alignment shift is a reliable *directional, population-level* signal but not a per-image cause of which individual attack transfers, since the transferred outcome depends on the full ten-step attack trajectory rather than the initial gradient alignment. This is why the main text leads the asymmetric *variance reduction* (the ∼4× magnitude column of Tab. 5) as the quantitative driver and treats d_{sign} as directional evidence only (Sec. 4.2). The Scissors mechanism is *distributional*, a property of the surrogate's gradient regime as a whole, and we do not claim it reduces to a per-image mediator. We include this negative control rather than omit it precisely because the per-image test is the first one a skeptical reader would run.

P.5 Remark on High Baseline ASR

The high baseline ASR of the robust source (Engstrom) against standard targets (avg 76.0%) is a documented phenomenon: adversarially trained models learn perceptually aligned, transferable features (Salman et al., 2020), and robust models are known to be strong surrogates for transfer attacks (Springer et al., 2021). Our values use unmodified RobustBench (Croce et al., 2021) checkpoints.

P.6 Threshold Sensitivity Analysis

We ablate $\tau \in [0.80, 0.98]$ on ImageNet (target Swin-B, N=500, 3 seeds; Fig. 11) and identify a safe zone $\tau \in [0.86, 0.96]$. The robust source (Engstrom) holds ∼67.1% across $\tau \in [0.80, 0.96]$, dropping only at τ=0.98 (64.7%); the standard source (ResNet50) rises from 51.5% (τ=0.80) to a ∼53.3% plateau for $\tau \geq 0.86$. Because CG-DI computes LGC per image (or on a calibration batch), the model-level mean (≈ 0.64 for ResNet50) does not place all images below τ; as τ decreases, more individual images exceed it and are assigned p=0, explaining the gradual standard-source transition.

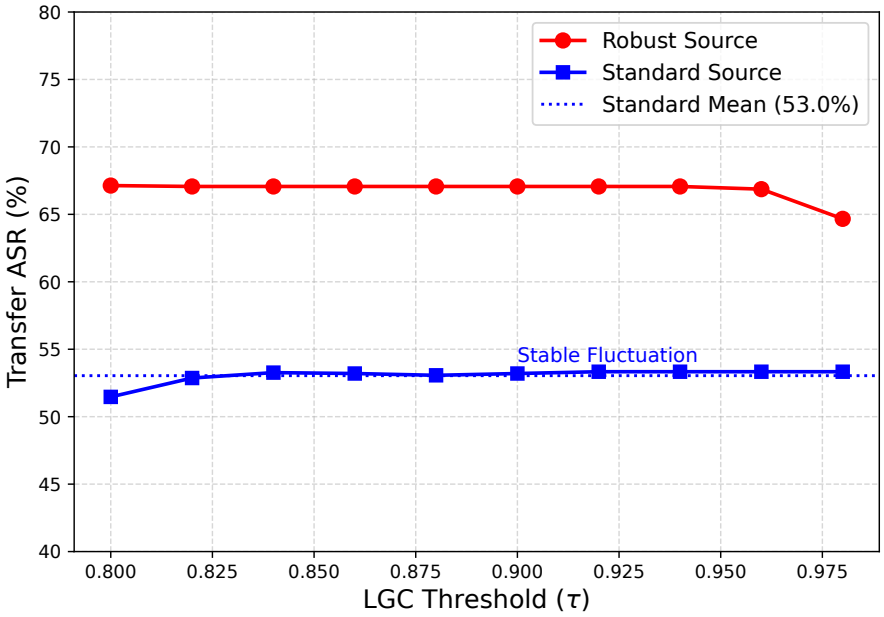

Figure 11: Threshold sensitivity on ImageNet (target Swin-B, N=500, 3 seeds). Transfer ASR vs. the LGC threshold τ for both source types. The robust source (Engstrom, *red*) is stable at ∼67% across $\tau \in [0.80, 0.96]$, dropping only at the extreme τ=0.98 (64.7%); the standard source (ResNet50, *blue*) rises to a ∼53.3% plateau for $\tau \geq 0.86$. Both source types are simultaneously stable in the safe zone $\tau \in [0.86, 0.96]$, with our default τ=0.92 centered in it.