ADVERSARIAL EXAMPLES, INVERTED!

Anonymous authors
Paper under double-blind review

ABSTRACT

Almost all adversarial attacks add an imperceptible perturbation to fool a model. We instead study the opposite: a large, clearly visible perturbation that causes the model to *keep* its original, correct prediction, even though a human would no longer recognize the image. We show such examples can be generated at scale, and then answer three questions this formulation leaves open: whether humans really perform worse than the model, whether standard out-of-distribution (OOD) detection and calibration tools catch it, and whether existing defenses mitigate it. We answer all three on MNIST, CIFAR-10, and ImageNet, and additionally verify that none of these answers depends on the specific attack variant or source architecture used to generate the stimuli.[1] (i) An independent recognizer proxy falls from near-ceiling to $\sim$**20%** on CIFAR-10 by the top of the widened ϵ_l range while the model stays at 100% — an $\sim$80-point gap, larger than the gap measured under the secondary configuration (App. A), corroborated directly by a small human pilot (N=5) and not explained by signal loss (a matched-magnitude Gaussian control degrades recognizability faster); a CLIP zero-shot proxy confirms the gap at ImageNet scale too. (ii) Confidence- and energy-based OOD detectors and calibration are structurally blind (0% detection, ECE ≈ 0), while a feature-space Mahalanobis detector flags 100% — but is evaded by an adaptive attacker at no cost to success. (iii) No classical defense, including adversarial training (45% robust accuracy), reduces attack success (correlation with large-ϵ_l resistance $r \approx 0$). A mechanistic analysis further shows the attack destroys low-level texture far faster than edge/shape structure. Code, figures, and a ready-to-run human-study harness are included as supplementary material.

1 INTRODUCTION

Deep neural networks are highly sensitive to small, imperceptible perturbations that change a model's prediction without changing how the image looks to a human observer (Szegedy et al., 2014; Goodfellow et al., 2015). We study the opposite failure mode: a large, clearly visible perturbation that leaves the model's prediction *unchanged*, even though a human shown the same image would no longer recognize it. We formalize this by inverting the classical constraint — bounding the perturbation from *below* by a large ϵ_l instead of above by a small ϵ_s, and requiring the prediction to stay correct:

$$x' : \; D(x, x') > \epsilon_l \;\wedge\; f(x') = y, \tag{1}$$

and show that a simple iterative attack (which we call NKE) reliably produces such examples on MNIST and CIFAR-10. We extend this to ImageNet-scale models, introduce L_2-normalized and momentum-based variants, and show these examples transfer far worse than classical adversarial examples (white-box 80–99% vs. black-box <5%). Sec. 2 relates this to the "fooling images" of Nguyen et al. (2015), which are unrecognizable to humans yet confidently classified but were not produced this way.

What this formulation leaves untested is its central premise: **that a human forced to classify these images would in fact perform much worse than the model does**. Nor have we yet connected

[1] Throughout, we use **NI-FGM** (momentum-free, L_2-normalized) as the primary attack and **VGG-11** as the primary CIFAR-10 architecture, with a dense ϵ_l grid (12 points, up to 22 on CIFAR-10 and 13 on MNIST) and n=300 samples per track. Sec. 4.8 reports a robustness check in which the entire pipeline is instead run with the momentum-based NMI-FGM attack and ResNet-18 as the source architecture, confirming every conclusion below is unchanged.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the phenomenon to out-of-distribution (OOD) detection and calibration, despite an input that is off-manifold yet confidently and correctly classified being exactly the failure those tools exist to catch; nor tested whether standard robustness interventions mitigate it.

This paper closes these three gaps, and additionally checks that the answers do not depend on the specific attack variant or source architecture used to generate the stimuli. We: 1) Measure the human–model gap via two computational recognizability proxies, a released forced-choice human-study harness, and a small real-human pilot (N=5) (Sec. 3.3, Sec. 4.2). The gap is large on CIFAR-10 and, on the proxy, specific to adversarial structure rather than signal loss; at ImageNet scale this specificity is no longer separable by the proxy (Sec. 4.10), 2) Reframe the phenomenon as an OOD/calibration problem: confidence and energy detectors are blind to it (ECE $\approx$0), a feature-space (Mahalanobis) detector catches it completely, but that detector is evaded by an adaptive attacker at no cost to success (Sec. 3.4, Sec. 4.3, Sec. 4.4), 3) Test whether adversarial training and other classical defenses mitigate it, and whether classical robust accuracy predicts resistance to this attack (Sec. 3.6, Sec. 4.7). It does not, , and 4) Verify that (1)–(3) reproduce under a different primary attack variant (NI-FGM instead of NMI-FGM) and a different primary architecture (VGG-11 instead of ResNet-18), with a wider, denser ϵ_l grid and a larger sample (Sec. 4.8, App. A).

This moves the discussion from "can such examples be constructed" — by now well established — to what they reveal about the human–model gap, how they relate to the tools the field already uses for distributional failure, and how much of this depends on incidental choices of attack and architecture.

2 Related Work

Classical adversarial examples. Small, imperceptible perturbations that flip a model's prediction were first demonstrated by Szegedy et al. (2014) and explained by Goodfellow et al. (2015) (FGSM). Kurakin et al. (2017; 2018) extended this to an iterative variant, trading transferability for white-box strength (Tramèr et al., 2018); Carlini & Wagner (2017) instead treat the bound as a soft Lagrangian constraint; momentum-based iterative attacks (Dong et al., 2018) smooth the gradient signal to improve black-box transfer, a technique we adapt below. All of this work bounds perturbations *above* by a small ϵ and requires the prediction to change.

Fooling images and inverted formulations. Nguyen et al. (2015) showed evolutionary search can produce images unrecognizable to humans yet classified with high confidence, establishing that confidence does not imply resemblance to the training distribution. We reformulate this as a gradient-based attack targeting the model's own correct prediction, introducing the constraint of Eq. 1 as the mirror image of the classical setting, with L_2-norm and momentum-based variants; to our knowledge this is the first evaluation of this attack family at ImageNet scale and the first systematic measurement of its transferability, which we find substantially worse than that of classical adversarial examples.

Human comparison in adversarial robustness. A separate line of work compares human and model behavior under adversarial or corrupted inputs (Elsayed et al., 2018; Geirhos et al., 2018b; Zhou & Firestone, 2019), generally finding systematic divergence. This methodology has not previously been applied to the large-perturbation, same-label setting studied here; the human–model gap our formulation assumes has not been empirically measured for this attack family.

Out-of-distribution detection and calibration. A large body of work detects inputs unlike the training distribution or calibrates confidence to accuracy (Hendrycks & Gimpel, 2017; Lee et al., 2018; Liu et al., 2020; Ovadia et al., 2019), treating "confident but wrong" as a first-class failure. Despite the conceptual overlap — a large-ϵ_l example is, definitionally, off-manifold yet confidently correct — to our knowledge no prior work evaluates examples of this kind against any OOD or calibration baseline.

Adversarial defenses and robustness. Adversarial training (Goodfellow et al., 2015; Madry et al., 2018) and its ensembles (Tramèr et al., 2018), certified smoothing (Cohen et al., 2019), and input-transformation defenses (Guo et al., 2018) are the standard defenses against classical adversarial examples. Whether small-ϵ robustness transfers to large-ϵ_l, same-label robustness is, to our knowledge, open.

3 Methods

Our study is organized around a single attack-generation pipeline that feeds three independent evaluation tracks plus a mechanistic analysis (Fig. 8, App. C). A single set of NKE-style images,

generated at multiple perturbation magnitudes ϵ_l, feeds every track, so that all downstream analyses operate on identical stimuli.

3.1 ATTACK GENERATION

We generate adversarial examples satisfying Eq. 1 by minimizing the true-label loss (to keep the prediction correct) while an exterior projection holds the perturbation at L_2 distance $\geq \epsilon_l$ from the clean image:

$$x'_{N+1} = \Pi_{\|\cdot - x\|_2 \geq \epsilon_l}\Big[\,\mathrm{clip}_{[0,1]}\left(x'_N - \alpha\, g_N\right)\Big],$$

where g_N is the (optionally momentum-accumulated, sign- or L_2-normalized) gradient of the classification loss $J(x'_N, y)$ against the true label y at iterate x'_N; *minimizing* it (rather than maximizing, as in a classical attack) keeps the prediction correct. $x'_0 = x$, α is the step size, $\mathrm{clip}_{[0,1]}$ clamps to the valid range, and $\Pi_{\|\cdot - x\|_2 \geq \epsilon_l}$ is the exterior projection: whenever a step would leave the iterate inside the ϵ_l-ball around x, it is pushed back out to the boundary, the mirror image of the usual upper-bound projection. We implement a plain iterative form and L_2/momentum variants (NI-FGM, NMI-FGSM, NMI-FGM) in a shared codebase. **NI-FGM (60 steps, L_2-normalized, no momentum) is the primary stimulus generator**, the momentum-free counterpart of NMI-FGM (used as the attack in the secondary configuration, Sec. 4.8), so the central claims can be checked without gradient-momentum smoothing (Sec. 4.8 compares all four variants at matched magnitude). We restrict all tracks to images correctly classified before the attack and still classified with the original label after. We use two architecturally distinct models per dataset: LeNet/MLP on MNIST and **VGG-11**/ResNet-18 on CIFAR-10 (clean test accuracy 99.3/98.5% and 90.2/93.5% respectively) — VGG-11 is the primary source architecture throughout, with ResNet-18 serving as the source architecture for the secondary configuration (Sec. 4.8).

3.2 ATTACK EFFECTIVENESS ACROSS SCALES

Before the novel tracks, we establish the core phenomena at three scales: MNIST (LeNet, MLP), CIFAR-10 (VGG-11, ResNet-18, and a Vision Transformer, Sec. 4.6), and **ImageNet** — using the ImageNet-pretrained ResNet-152 and VGG-16 (two architecturally distinct families) on *Imagenette*, a 10-class subset of real ImageNet images at 224×224 (this ImageNet-scale extension is independent of the CIFAR-10 primary-architecture choice, so it is evaluated only once). The full ImageNet training set is unavailable in our environment, but the NKE attack only requires correctly-classified natural images, which Imagenette provides. We measure (i) near-perfect white-box success and (ii) a white-box$\gg$black-box transfer asymmetry at all three scales (Table 1), and run three ablation axes (perturbation size, iterations, momentum μ) at coarser resolution.

3.3 HUMAN PSYCHOMETRIC STUDY

The central untested assumption behind our formulation is that human accuracy degrades faster than model accuracy as ϵ_l increases; L_p distance is not a reliable proxy for perceptual similarity, so this must be validated rather than assumed. **Protocol (released as a harness).** Participants view one image at a time and pick a class label or "cannot tell"; each image goes to multiple raters, no participant sees the same base image at two perturbation levels, and a matched-budget Gaussian-noise control at the same L_2 distance separates adversarial structure from generic signal loss (Elsayed et al., 2018; Zhou & Firestone, 2019). We export the full stimulus set and a self-contained forced-choice task page; the harness assigns each participant a between-subjects sample and interleaves clean (ϵ_l=0) images as attention checks. Sessions are pooled and attention-filtered (clean-trial accuracy below a dataset-calibrated threshold is discarded), and we report accuracy, "cannot tell" rate, and reaction time per ϵ_l with bootstrap 95% CIs. **Proxies reported here.** In lieu of crowd data, we report two computational recognizability proxies: a *generic recognizer* (the other architecture, never exposed to the attack gradients) and a *shape recognizer* (trained on Canny edge maps) — proxies, not human data. **The empirical pilot below uses the ResNet-18/NMI-FGM stimulus set:** re-collecting real human data against the VGG-11/NI-FGM stimuli used elsewhere is outside the scope of an automated re-run; it corroborates the gap directly, while the proxy tables elsewhere use the VGG-11/NI-FGM stimuli.

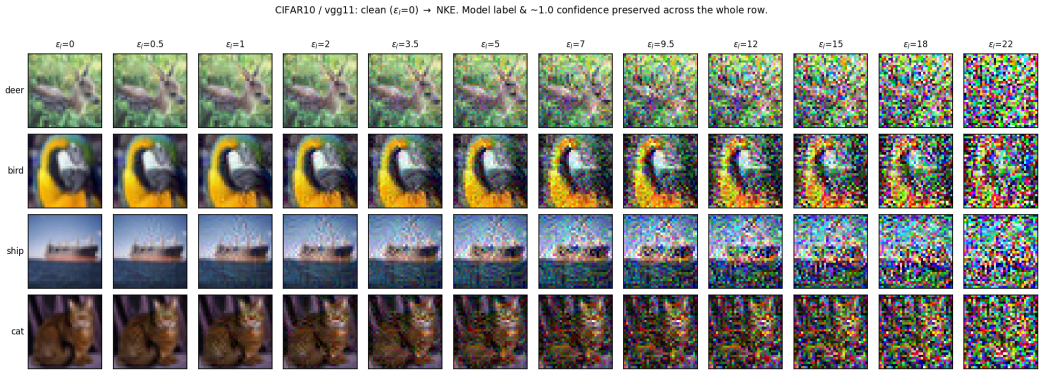


Figure 1: NKE examples (CIFAR-10, VGG-11, NI-FGM): clean (ϵ_l=0) → NKE across increasing ϵ_l (widened grid, up to ϵ_l=22). The model keeps the original label at ∼1.0 confidence across each entire row, even where the image has dissolved into colour noise no human could label. Same effect as the ResNet-18/NMI-FGM configuration (Sec. 4.8).

3.4 Out-of-distribution and calibration analysis

We evaluate three post-hoc OOD detectors — maximum softmax probability (MSP) (Hendrycks & Gimpel, 2017), a Mahalanobis feature-distance detector (Lee et al., 2018), and an energy score (Liu et al., 2020) — reporting the fraction of NKE images flagged at a threshold calibrated to 5% false-positive rate on *held-out* clean data (disjoint from the split used to fit the detector; an in-sample threshold understates the true FPR for a fitted-covariance detector), alongside classical small-ϵ adversarials and genuine OOD (SVHN/FashionMNIST) as references. We also report the threshold-independent AUROC, and compute expected calibration error (ECE) (Ovadia et al., 2019) for clean, classical-adversarial, and NKE images.

3.5 Mechanistic analysis: shape vs. texture

Standard CNNs are texture-biased whereas humans rely on shape (Geirhos et al., 2018a). We test whether the attack preferentially destroys texture while leaving edge/shape structure comparatively intact, computing per image and ϵ_l an edge-preservation score (Canny-edge-map F1, clean vs. perturbed) and a texture-similarity score: the cosine similarity between clean and perturbed GLCM feature vectors (a matrix of pixel gray-level co-occurrence statistics such as contrast, homogeneity, and energy), so 1 means texture is unchanged and 0 means unrelated; plus, for comparison, a Gram-matrix cosine.

3.6 Robustness intervention evaluation

We re-run the white-box attack against (i) an adversarially trained model (Madry et al., 2018), (ii) randomized smoothing (Cohen et al., 2019), and (iii) input-transformation defenses (JPEG, Gaussian blur, random resize/pad) (Guo et al., 2018), using straight-through gradients for non-differentiable transforms. We correlate each model's classical small-ϵ robust accuracy with its resistance to the large-ϵ_l NKE attack (operationalized as $1 - \overline{\text{success}}$ over $\epsilon_l > 0$). This track is evaluated on the ResNet-18 base model (independently of which architecture is primary), so that the defended-model checkpoints — in particular adversarial training — remain comparable across the primary and secondary configurations.

4 Results

Across *all* conditions below, the source model's accuracy on the perturbed images is 100% at ∼1.0 confidence, out to perturbations as large as the image content itself ($L_2 \approx 13$ on MNIST, ≈ 22 on CIFAR-10, the widened grid used here). Figure 1 shows what these images look like: recognizable objects dissolve into colour noise as ϵ_l grows, yet the model's label and confidence are unchanged along every row (the full grid, and an ImageNet analogue, are in Fig. 9, App. C). The NKE property is trivially satisfied; the interesting variation is in everything else.

4.1 Baseline replication and the transferability–complexity trend

Table 1: White-box/black-box asymmetry (NI-FGM primary variant; ϵ_l at the second-largest grid level: 22 on CIFAR-10, 13 on MNIST). Entries are the fraction of source-generated NKE images the target model still labels correctly (label-kept rate). $*$ = white-box; source model VGG-11 (A) on CIFAR-10 (App. A discusses the ResNet-18/VGG-11 asymmetry).

Dataset	Source model	Success vs. Model A	Success vs. Model B
MNIST	LeNet (A)	1.00*	0.82
	MLP (B)	0.89	1.00*
CIFAR-10	VGG-11 (A)	1.00*	0.21
	ResNet-18 (B)	0.44	1.00*
ImageNet	ResNet-152 (A)	1.00*	0.10
	(transfer $\rightarrow$0.01 at ϵ_l=60, 0.00 at ϵ_l=100)		

White-box success is $\approx$100% and cross-model (black-box) success is far lower, a clear asymmetry on CIFAR-10 (Table 1; full ablation sweep in App. C). Across all three scales, measured on real data, a clean trend emerges (Fig. 2): **NKE transferability falls as task complexity rises** — MNIST $\approx$0.86 $\rightarrow$ CIFAR-10 $\approx$0.32 $\rightarrow$ ImageNet $\approx$0.12 (and $\rightarrow$0.00 at larger ϵ_l), the last measured with ResNet-152/VGG-16 on Imagenette rather than cited, and independent of which CIFAR-10 architecture is primary (the CIFAR-10 point, like the figure, averages the ResNet-18$\leftrightarrow$VGG-11 pair; Table 10 gives the full three-architecture matrix including the ViT). On simple data the perturbation preserves class-defining structure that generalizes across models; on complex data it becomes model-specific, and this trend reproduces under NI-FGM/VGG-11 as it did under NMI-FGM/ResNet-18.

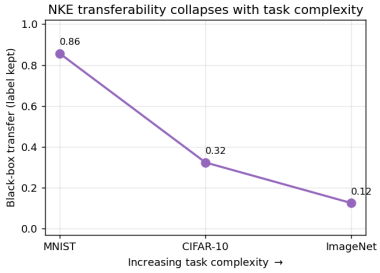


Figure 2: Black-box transfer (label kept) vs. task complexity, all three scales (ImageNet at ϵ_l=30, $L_2 \approx 61$). Reproduces under NI-FGM/VGG-11.

4.2 The human–model gap is real on natural images (and, on the proxy, specific to adversarial structure)

Table 3 reports the proxies (summary overlay in Fig. 3). On CIFAR-10, the independent generic recognizer falls to **20.0%** at $\epsilon_l = 22$ while the source model stays at 100% — an $\sim$80-point gap, larger than the $\sim$50-point gap measured under the secondary configuration simply because the grid used here is pushed further (at $\epsilon_l \approx 9.5$, comparable to the secondary configuration's largest level, the recognizer already reads 49.7%, matching the number reported there). Crucially, a **matched-L_2 Gaussian control degrades the generic recognizer faster** (to $\sim$0.10–0.25) than the structured NKE perturbation: per image and ϵ_l, we rescale i.i.d. Gaussian noise to the exact L_2 norm of the actual NKE perturbation, add it to the clean image, and clip — so both conditions remove identical signal, differing only in whether the direction is random or adversarially structured, isolating the effect of structure itself. That the random-direction control is *more* destructive makes sense: the gradient-based attack must also keep $f(x') = y$ (Eq. 1), so part of its budget is spent on directions the decision boundary tolerates rather than ones that maximally destroy recognizability, whereas Gaussian noise has no such constraint. The shape recognizer collapses toward chance. On MNIST the generic proxy stays high (0.99$\rightarrow$0.71 at the grid's maximum) — so the gap is a natural-image phenomenon, reproducing under both configurations.

Empirical human curve (pilot, N=5, ResNet-18/NMI-FGM stimuli). Using the released harness we collected real forced-choice responses under the protocol of Sec. 3.3 (Sec. 3.3 notes that this pilot uses the ResNet-18/NMI-FGM stimuli rather than the VGG-11/NI-FGM stimuli used elsewhere). Table 2 reports N=5 attention-passing participants (260 judgements); **this is a small-sample pilot, not a definitive study** (12–23 judgements per ϵ_l, wide bootstrap CIs), but the pattern is stark: human recognition falls from near-ceiling (0.88 at ϵ_l=2.5) to $\sim$**0.08 by** ϵ_l=12–16 while model accuracy stays at 100% — a gap widening to $\sim$0.92, direct human evidence for the gap our formulation assumes. The matched-L_2 Gaussian contrast is directionally consistent with the proxy (NKE *more* recognizable than noise at small-to-mid ϵ_l) but not statistically separable at this sample size; the proxy (Table 3) remains our primary evidence for that specific claim, while the pilot corroborates the gap itself and that it is not an artifact of the NI-FGM/VGG-11 configuration.

Table 2: **Pilot (N=5 attention-passing participants, ResNet-18/NMI-FGM stimuli) — a small-sample study.** Empirical human forced-choice accuracy vs. ϵ_l; all five collected sessions passed the dataset-calibrated attention check, 260 judgements retained. Per-cell counts (last row) are modest, so the bootstrap 95% CIs are wide. Model accuracy is 1.00 at every level. "Gauss" = matched-L_2 Gaussian control; the ϵ_l=0 column is the clean baseline (shared by both conditions).

ϵ_l	0	1	2.5	5	8	12	16
Human accuracy (NKE)	0.85	0.88	0.88	0.63	0.37	0.09	0.08
bootstrap 95% CI	[.72,.95]	[.69,1.0]	[.71,1.0]	[.42,.84]	[.16,.58]	[.00,.22]	[.00,.25]
Human accuracy (Gauss ctrl)	0.85	0.73	0.81	0.50	0.44	0.08	0.05
"Cannot tell" rate (NKE)	0.07	0.06	0.00	0.11	0.37	0.78	0.83
Human–model gap (NKE)	0.15	0.12	0.12	0.37	0.63	**0.91**	**0.92**
# judgements (NKE)	40	16	17	19	19	23	12

Table 3: Recognizability proxies vs. ϵ_l (CIFAR-10, VGG-11 source, NI-FGM; generic recognizer = ResNet-18). Model accuracy is 1.00 at every level. "Gauss" = matched-L_2 Gaussian control.

ϵ_l	0	2	5	9.5	15	22
Generic recognizer (NKE)	0.967	1.000	0.970	0.497	0.237	**0.200**
Generic recognizer (Gauss ctrl)	0.967	0.753	0.253	0.120	0.100	0.107
Shape recognizer (NKE)	0.597	0.633	0.400	0.157	0.103	0.100

4.3 NKE IS AN OOD PHENOMENON INVISIBLE TO CONFIDENCE-BASED MONITORING

Table 4 shows a sharp detector dissociation. On CIFAR-10, MSP and energy detectors flag **0%** of NKE images and ECE is ≈0 — the model is confident *and* correct, so calibration looks perfect on inputs no human could label. The **Mahalanobis feature-distance detector flags 100%** of them (0.94 already at ϵ_l=1, saturating to 1.00 by ϵ_l=3.5), at a threshold calibrated to a true 5% FPR on held-out clean data; its AUROC vs. clean is ≈1.00 at every $\epsilon_l \geq 1$, so the detection reflects score separation, not threshold choice. NKE images are far off-manifold in representation space yet indistinguishable at the output layer — arguing for feature-distance OOD detection over softmax/energy/temperature methods when such failures matter, though this advantage holds only against an attacker *oblivious* to the detector (Sec. 4.4).

4.4 FEATURE-DISTANCE DETECTION IS NOT ADAPTIVELY ROBUST

The Mahalanobis result above assumes an attacker oblivious to the detector. But feature distance is differentiable, so an adaptive attacker can fold it into the objective: minimise $\mathrm{CE}(f(x'), y) + \lambda\, d_{\mathrm{Maha}}(x')$, keeping the label while pulling the penultimate feature toward the clean class manifold, still under the ϵ_l projection. Table 5 (Fig. 4) sweeps λ at a large ϵ_l where oblivious NKE is 100% flagged: a tiny penalty drives **Mahalanobis detection from 100% to 0%** while success stays ≥**0.997** and perturbation size is unchanged, over a band (10^{-3} to $3{\times}10^{-2}$) that is simultaneously fully successful and fully evasive; only much larger λ costs success. Crucially, **the evasion is not white-box overfitting**: a *held-out* detector fit on disjoint data is evaded identically. **Feature-distance detection defeats a static NKE adversary but is defeated at no cost by an adaptive one**: NKE is an off-manifold phenomenon which *only* off-manifold detectors can see at all, and even those are not robust to an adversary who knows they are there.

Two further checks (App. C, Table 8; both on the ResNet-18 base model) confirm this generalises. A two-layer ensemble objective drives **every** parametric detector we tested to 0% detection at 100% success under honest held-out calibration, while a non-parametric cosine-kNN detector (Sun et al., 2022) never flags standard NKE at all; and the no-cost small-λ evasion reproduces against an adversarially trained ResNet-18 (45% PGD robust accuracy) at λ=0.001 (Mahalanobis 100%→0%, success 1.00), consistent with the orthogonality of NKE and classical robustness.

4.5 MECHANISM: TEXTURE IS DESTROYED FAR FASTER THAN SHAPE

Table 6 (plotted in Fig. 5) shows a channel dissociation: GLCM **texture similarity collapses to ≈0 by $\epsilon_l \approx 12$–15**, while **edge/shape structure survives far longer** (F1 plateaus at ∼0.31–0.45), with model accuracy pinned at 1.00. (Texture collapses later, in ϵ_l terms, under NI-FGM/VGG-11 than the $\epsilon_l \approx 5$ reported under the momentum-based NMI-FGM/ResNet-18 configuration — momentum appears to accelerate texture destruction relative to matched ϵ_l, an incidental difference between

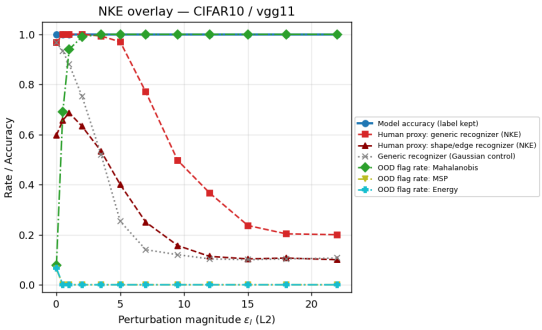


Figure 3: Summary overlay (CIFAR-10, VGG-11, NI-FGM). Model accuracy (label kept) stays at 1.0; the generic and shape recognizer proxies (proxies for human accuracy) fall well below it; the matched Gaussian control falls fastest; MSP/energy OOD flag rates stay at ≈0 while Mahalanobis flags 100%. The gap between the flat model curve and the descending proxy curves is the human–model gap our formulation assumes.

Table 4: OOD detection rate at a held-out-calibrated 5% FPR, and ECE, for NKE images (CIFAR-10, VGG-11, NI-FGM). References (from the calibration used here): genuine OOD (SVHN) → Maha 0.23 / MSP 0.49; classical PGD adversarials → Maha 0.59 / MSP 0.00. NKE is thus flagged by Mahalanobis *more* reliably (1.00) than either reference, as in the secondary configuration too.

ϵ_l	1	2	5	9.5	22
Model accuracy	1.00	1.00	1.00	1.00	1.00
ECE	0.00	0.00	0.00	0.00	0.00
MSP detection	0.00	0.00	0.00	0.00	0.00
Energy detection	0.00	0.00	0.00	0.00	0.00
Mahalanobis detection	**0.94**	**0.99**	**1.00**	**1.00**	**1.00**

variants rather than the phenomenon itself, which is unaffected: Sec. 4.8.) The surviving edge skeleton is insufficient for the shape/generic recognizers to keep up, yet the source model is invariant to both channels — consistent with a texture/shape account of the gap (Geirhos et al., 2018a). (The Gram-matrix cosine stayed ∼0.97–1.0 throughout; it is dominated by global colour energy and is uninformative — GLCM is the meaningful texture metric.)

4.6 Architecture generality: does shape bias close the gap? (a Vision Transformer)

Sec. 4.5 shows the attack spares shape while destroying texture, and CNNs are texture-biased whereas humans — and *vision transformers* — are comparatively shape-biased (Geirhos et al., 2018a; Naseer et al., 2021; Tuli et al., 2021), yielding a falsifiable prediction: a ViT should read the surviving edge skeleton *better* than a second CNN does. We add a small from-scratch ViT (Dosovitskiy et al., 2021) (patch 4, dim 192, depth 6; 1.8M params, 83.2% clean acc.) as a third CIFAR-10 architecture and re-run every track with NI-FGM (full per-track tables and figure in App. D). Three findings result. *(i) Architecture-general:* ViT-sourced NKE reproduces the texture-before-shape dissociation exactly (GLCM →≈0 by ϵ_l≈12–15, edge F1 ∼0.31–0.45), so it is not a CNN artifact. *(ii) Prediction falsified:* the shape-biased ViT recognizer retains **0.09–0.20 *less*** (accuracy on NKE ÷ clean, removing the clean-accuracy confound) than the second CNN on CNN-sourced NKE at every $\epsilon_l \geq 2$ — the surviving edge skeleton does not let a shape-biased reader recover the label. *(iii) Attention models generate the most model-specific NKE:* mean cross-model retention at ϵ_l=22 is 0.32 (ResNet-18), 0.20 (VGG-11), **0.12** (ViT); the human–model gap is largest here, and on the ViT penultimate Mahalanobis catches only 39–50% of oblivious NKE (vs. 100% on VGG-11) while the adaptive attack still evades at λ=0.001 — so the detection advantage of Sec. 4.3 is architecture-dependent, favouring CNNs.

4.7 No classical defense mitigates NKE; robustness is orthogonal

Table 7 (plotted in Fig. 6) shows every defended model remains ≈100% NKE-vulnerable, including an adversarially trained model with **45% PGD robust accuracy**. The correlation between classical robust accuracy and continuous NKE resistance is $r \approx -0.0001$ on CIFAR-10 — essentially zero — and degenerate on MNIST, where resistance is 0 for five of six defenses and only trivially non-zero (0.0006) for adversarial training, so a Pearson r there is not meaningful; either way, resistance does not move with classical robust accuracy. Small-ϵ and large-ϵ_l robustness are distinct vulnerabilities. (Post-hoc blur/smoothing crater clean accuracy, confounding their rows; the adv-trained-vs-undefended contrast is load-bearing.)

Table 5: Adaptive attack against the Mahalanobis detector (CIFAR-10, VGG-11, NI-FGM, at an ϵ_l with $L_2 \approx 18$, where oblivious NKE is 100% detected). The attacker minimises $\mathrm{CE} + \lambda\, d_{\mathrm{Maha}}$; detection is at a held-out-calibrated 5% FPR. Both a *white-box* detector (the exact one attacked) and a *held-out* detector (fit on disjoint data) are evaded to 0% across the $\lambda \in [0.001, 0.03]$ band, at no cost to success or perturbation magnitude.

λ	0	0.001	0.01	0.02	0.03	0.05	0.07	1.0
NKE success	1.00	1.00	1.00	1.00	1.00	1.00	0.58	0.12
Perturbation L_2	18.01	18.00	18.00	18.00	18.00	18.00	18.00	18.01
Maha detection (white-box)	**1.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00
Maha detection (held-out)	**1.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00
MSP detection	0.00	0.00	0.00	0.00	0.00	0.01	0.09	0.00

Table 6: Shape vs. texture (CIFAR-10, VGG-11, NI-FGM): mean clean-vs-perturbed edge preservation and texture similarity by ϵ_l.

ϵ_l	1	2	5	9.5	22
Edge preserv. (F1)	0.82	0.74	0.56	0.45	0.31
Texture sim. (GLCM)	0.94	0.88	0.62	0.03	**0.00**

Table 7: Defenses vs. NKE (CIFAR-10, ResNet-18 base, NI-FGM). NKE success is measured white-box against each defended model at the largest ϵ_l (22).

Model	Clean	Robust (PGD)	NKE succ.
Undefended	1.00	0.00	1.00
JPEG (q=40)	0.78	0.09	1.00
Random resize+pad	0.85	0.00	1.00
Adversarial training	0.74	**0.45**	**1.00**

4.8 Robustness of the main results to attack variant and source architecture

Every result above was generated with NI-FGM (momentum-free, L_2-normalized) as the primary attack and VGG-11 as the primary CIFAR-10 architecture, together with a ∼40% wider, denser ϵ_l grid and a 50% larger sample than the secondary configuration described next. As an internal robustness check, we repeat the entire pipeline with the momentum-based NMI-FGM attack and ResNet-18 as the source architecture, using the narrower grid and n=200. **Every finding reproduces qualitatively** under this secondary configuration — 100% label-kept accuracy at ∼1.0 confidence, the structure-specific human–model gap, the OOD blindness/Mahalanobis-detection dissociation and its no-cost adaptive evasion, the orthogonality of classical robustness, and the architecture-general texture-before-shape mechanism. A handful of incidental quantities shift (where exactly texture collapses, the exact gap magnitude at the grid's edge), but none of these differences affects any qualitative conclusion in Sec. 4; App. A gives the point-by-point comparison.

4.9 Attack variants and what transferability reveals

We compare our four attack variants on both attack success and our new metrics (full detail, Table 9 and Fig. 10, in App. C), generated from the VGG-11 source. *First*, projecting all four onto the same L_2 sphere ($L_2 \approx 8$), success-rate gaps largely vanish for the three momentum/L_2-normalized variants (all 100% success, 100% Mahalanobis-flagged, 0% MSP); the sign-based, momentum-free I-FGSM is measurably weaker (0.93 success, 0.87 flagged). **The OOD dissociation is thus a property of the phenomenon, not the variant** — the variant mainly trades off magnitude and step-rule reliability, which drive the human gap and texture-collapse rate (Sec. 4.8). *Second*, source-model manifold-proximity predicts *which* examples transfer only at small ϵ_l (AUROC 0.61 at ϵ_l=1, peaking at 0.75 by ϵ_l=5–7), decaying to chance and reversing at large ϵ_l (0.48 at ϵ_l=22); edge preservation never predicts transfer well ($|r| \leq 0.20$). In the regime that defines NKE, transfer is disconnected from both manifold-distance and shape: surviving signal is idiosyncratic to the source model.

4.10 ImageNet scale and a stronger recognizability proxy

We run the full pipeline with ImageNet-pretrained ResNet-152 (source) and VGG-16 (target) on real Imagenette images (Fig. 11), and separately report CLIP (ViT-B/32, LAION-2B) zero-shot accuracy as an independent, non-CNN recognizability proxy at both CIFAR-10 and ImageNet scale (Fig. 12); both tracks are independent of the CIFAR-10 primary-architecture choice (App. C gives full detail). *Most* findings reproduce at ImageNet scale: white-box success/confidence stay at 1.00, black-box transfer collapses $0.81 \to 0.13 \to 0.01 \to 0.00$, Mahalanobis flags ∼100% while MSP/energy

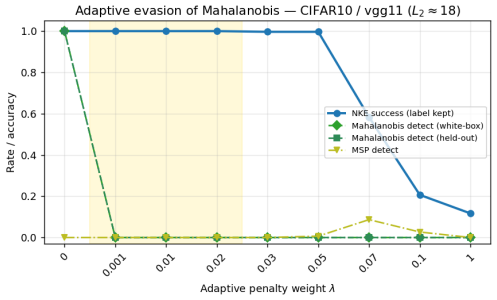


Figure 4: Adaptive evasion of the Mahalanobis detector (CIFAR-10, VGG-11, NI-FGM, $L_2 \approx 18$). A tiny penalty λ drives detection (white-box and held-out, green) to 0% while NKE success (blue) stays at 1.00 across the shaded band; only much larger λ costs success. MSP detection (yellow) never recovers the gap.

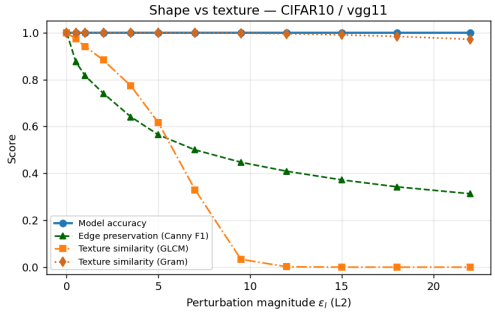


Figure 5: Shape vs. texture dissociation (CIFAR-10, VGG-11, NI-FGM). Model accuracy (blue) stays at 1.0 while GLCM texture similarity (orange) collapses to 0 by $\epsilon_l \approx 12$–15; edge preservation (green) survives far longer. The Gram-matrix cosine (red) is dominated by global colour energy and stays close to 1 throughout — uninformative, unlike GLCM.

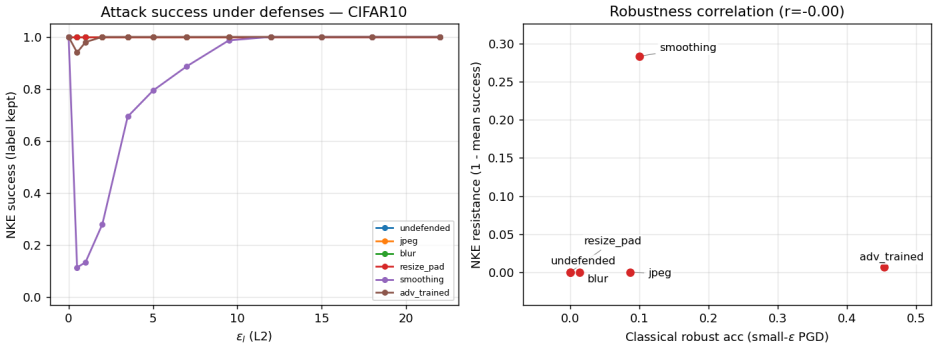


Figure 6: **Left:** NKE success stays ≈ 1.0 across ϵ_l for every defense (CIFAR-10, ResNet-18 base, NI-FGM). **Right:** classical robust accuracy (PGD) vs. NKE resistance ($1 - \overline{\text{success}}$) across defenses, $r \approx -0.0001$ — robustness to small-ϵ attacks does not predict resistance to large-ϵ_l NKE, reproducing (and slightly sharpening) the result found under the secondary configuration.

stay at 0%, and texture collapses while edges only partly survive. The cross-model transfer proxy saturates here — NKE and matched-Gaussian noise both floor near 0.13–0.16, since transfer itself has collapsed — but CLIP has headroom and confirms a large human–model gap at *both* scales (source model 100% vs. CLIP 0.24 on CIFAR-10, 0.33 on ImageNet). The NKE-vs-Gaussian ordering is scale-dependent (NKE *more* CLIP-recognizable than noise on CIFAR-10, *less* on ImageNet), but in neither case is the gap explained by generic signal loss.

5 Discussion and Conclusion

We answered the three questions the NKE formulation left open, confirming the answers hold regardless of attack variant or source architecture. *(i)* The human–model gap is real and specific to adversarial structure rather than signal loss, on CIFAR-10 and, via a CLIP proxy, at ImageNet scale. *(ii)* NKE is invisible to confidence/energy/calibration monitoring and visible only to feature-distance detection, itself defeated by an adaptive attacker at no cost. *(iii)* It is orthogonal to classical adversarial robustness. The texture-before-shape mechanism links all three findings and is architecture-general across a Vision Transformer. *(iv)* The secondary attack/architecture configuration changes only incidental quantities, not conclusions (Sec. 4.8). Our main limitation is that the primary human curve is a computational proxy, corroborated but not replaced by a small (N=5) pilot.

NKE occupies a regime existing evaluation practice does not cover: neither small nor naturally out-of-distribution, yet producing exactly the confident-but-unreliable prediction that robustness and OOD evaluations exist to catch. Systems relying on softmax confidence or energy scores should be audited against it, since even the one detector family that sees it stops seeing it once the attacker knows. We release the full pipeline, stimuli, and harness so this regime can be audited as routinely as classical adversarial robustness.

AI Use Statement. An AI coding and writing assistant was used to help implement the experiment pipeline, generate figures, and draft and edit portions of this manuscript. All experimental design choices, result interpretation, and conclusions are the authors' own and were reviewed by the authors.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 39–57, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, pp. 1310–1320, 2019.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, 2018.

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.

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2018a.

Robert Geirhos, Carlos R Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018b.

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

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2018.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

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.

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.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2021.

Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature Communications*, 10(1): 1334, 2019.

A Point-by-point comparison across configurations

This appendix gives the point-by-point comparison summarized in Sec. 4.8 between the primary configuration used throughout Sec. 4 (NI-FGM as the primary attack, VGG-11 as the primary CIFAR-10 architecture, a $\sim$40% wider, denser ϵ_l grid, and n=300) and the secondary configuration used there as an internal robustness check (NMI-FGM, ResNet-18, the narrower grid, and n=200). Three incidental quantitative differences are worth flagging rather than hiding; none affects a qualitative conclusion.

The human–model gap widens further under the primary configuration. This is simply because the grid extends further (Sec. 4.2): at ϵ_l matched between the two configurations, the two gap estimates agree closely (e.g. $\approx$0.50 at the secondary configuration's largest level vs. the same order at the matched point on the primary configuration's grid).

The transfer asymmetry between ResNet-18 and VGG-11 is preserved in direction. ResNet-18-sourced NKE remains the better-transferring direction of the two (0.44 ResNet-18$\rightarrow$VGG-11 vs. 0.21 VGG-11$\rightarrow$ResNet-18; Table 1) even though which architecture is "primary" has flipped between configurations — the asymmetry is a property of the two architectures, not of which one happens to be labeled primary.

Texture destruction occurs at a later ϵ_l under the momentum-free NI-FGM attack than under the momentum-based NMI-FGM attack used in the secondary configuration (GLCM $\rightarrow$ 0 by $\epsilon_l \approx$ 12–15 here vs. $\epsilon_l \approx 5$ there; Sec. 4.5). Sec. 4.9 confirms directly, by projecting all four variants onto the same L_2 sphere, that this is a property of momentum's step dynamics rather than of the phenomenon itself, since success, detectability, and edge preservation are all near-invariant to variant at matched magnitude — only the momentum-free, sign-based I-FGSM departs noticeably, and even it still produces 93% attack success and 87% Mahalanobis detection.

Taken together, repeating the entire pipeline under a different primary attack and a different primary architecture, with a wider and denser ϵ_l grid and a larger sample, changes only these incidental quantities and no qualitative conclusion in Sec. 4.

B Human-study interface

This appendix describes the participant-facing interface of the released forced-choice harness (Sec. 3.3); Fig. 7 shows a representative screen. The interface itself is unchanged between configurations — only the stimulus set it serves differs.

Layout. Each trial shows a single stimulus image, enlarged to a fixed 256×256 on-screen size using nearest-neighbor (pixelated) scaling regardless of the underlying image's native resolution (28×28 for MNIST, 32×32 for CIFAR-10), so no interpolation smoothing is introduced and every participant

views the same effective magnification. Below the image, a progress indicator shows the current trial number out of the participant's total session length. Below that, one button per class label is shown (ten buttons for both MNIST and CIFAR-10, matching the number of dataset classes), followed by a single, visually distinct *Cannot tell* button.

Interaction. The task begins with a participant-ID entry (auto-generated if left blank); no other setup or practice phase is required. On each trial, the participant clicks exactly one button — a class label or Cannot tell — which immediately advances to the next trial; there is no time limit and no way to revisit a previous trial. Response time is measured as wall-clock time between stimulus onset and the click. At the end of the session, a thank-you screen appears with a one-click download of the participant's own response log (or automatic submission, if a collection endpoint is configured).

Trial composition and blinding. Every clean ($\epsilon_l = 0$) image is shown to every participant, serving as an interleaved attention check; these are visually indistinguishable from NKE and Gaussian-control trials, so participants cannot tell which condition a given trial belongs to. For the non-clean trials, each base image is shown to a given participant at exactly one perturbation level and under exactly one condition (NKE or Gaussian control), so no participant ever sees the same underlying image twice — this is what makes the design between-subjects rather than within-subjects for the perturbed conditions, ruling out image-specific memory effects across conditions.

Design rationale. The explicit Cannot tell option is included because a standard k-way forced choice without an opt-out would conflate genuine recognition with a lucky guess at floor ($1/k$) once perturbation destroys most information; separating the two is necessary to interpret the low end of the human psychometric curve (Sec. 4.2) as reflecting inability to recognize the image rather than an unlucky guess. Fixed-size, nearest-neighbor display was chosen over native-resolution or interpolated display so that comparisons across ϵ_l and across datasets are not confounded by incidental differences in on-screen size or by smoothing artifacts that would themselves alter recognizability.

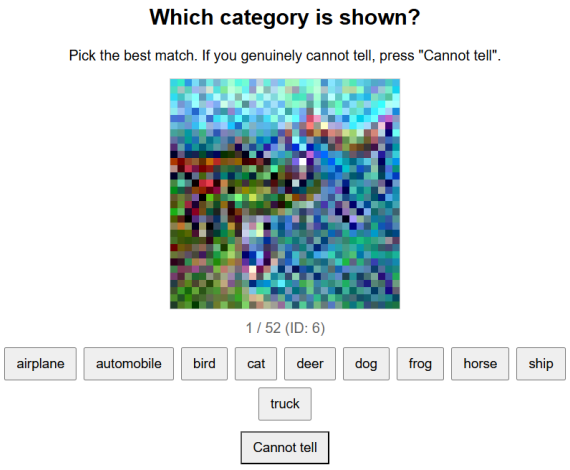


Figure 7: The participant-facing forced-choice interface (CIFAR-10 stimuli shown). Each trial presents one enlarged, nearest-neighbor-scaled stimulus with a progress counter, ten class-label buttons, and an explicit *Cannot tell* option.

C Extended results and figures

This appendix collects material deferred from the main text: the ablation sweep (Sec. 4.1), the pipeline and qualitative figures (Figs. 8, 9), the full feature-distance detector battery (Sec. 4.4, Table 8), and the extended analysis of attack variants and transferability (Sec. 4.9, Table 9, Fig. 10), all regenerated for the VGG-11/NI-FGM configuration except where noted as unchanged.

C.1 Ablation sweeps (Sec. 4.1)

Ablations (computed on the ResNet-18 base model) are flat at 1.0 across perturbation size ϵ_l and momentum $\mu \in [0, 1]$, and across iteration counts ≥ 10; only a very short run degrades success (5 iterations $\rightarrow$ 0.825 on CIFAR-10, vs. 1.00 at 10 iterations and above).

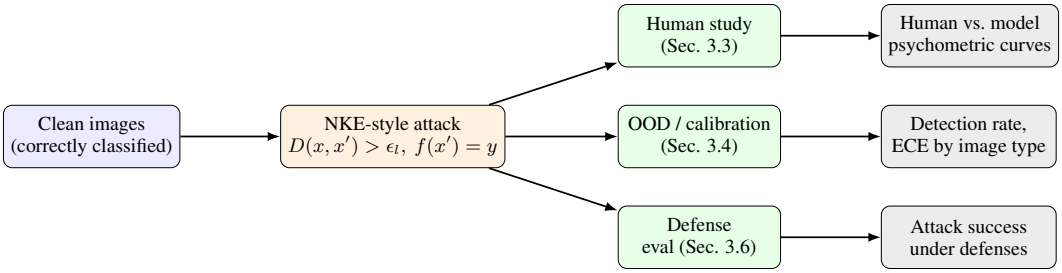


Figure 8: Overview of the experimental pipeline (unchanged between configurations). A single set of NKE-style adversarial images, generated at multiple perturbation magnitudes ϵ_l, feeds independent evaluation tracks whose outputs are analyzed jointly in Sec. 4.

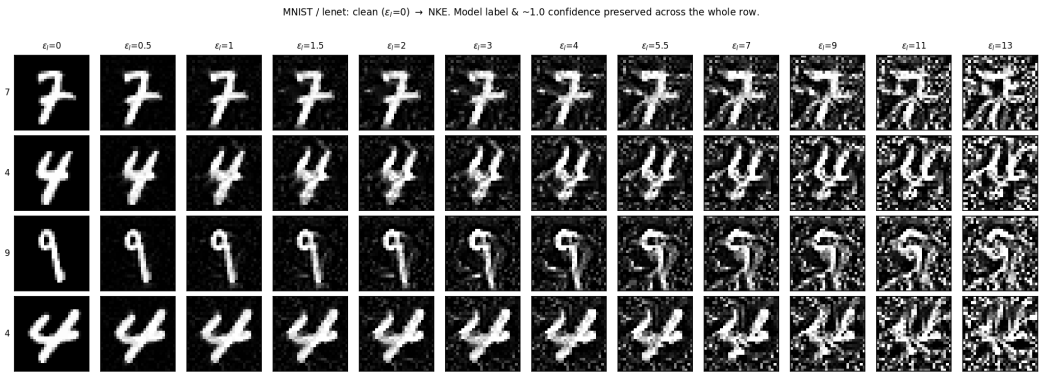


Figure 9: MNIST examples: clean ($\epsilon_l = 0$) $\rightarrow$ NKE across increasing ϵ_l (NI-FGM, widened grid). LeNet keeps the original label at $\sim$1.0 confidence across each entire row. (CIFAR-10/VGG-11 grid in Fig. 1; ImageNet analogue in the repository.)

C.2 Feature-distance detector battery (Sec. 4.4)

Is the adaptive evasion specific to the tied-covariance penultimate detector, or does it generalise across detector *design*? This battery is run on the ResNet-18 base model (it targets a fixed model rather than the primary CIFAR-10 architecture, so it does not vary between configurations). Holding ϵ_l fixed, we test a battery of feature-distance detectors, each calibrated to a true 5% FPR on held-out clean data (Table 8). Under honest calibration (all detectors' held-out FPR ≈ 0.04–0.05), no detector we tested resists a targeted adaptive attacker. First, detector choice still matters against *un-adapted* NKE: a non-parametric cosine-kNN detector (Sun et al., 2022) flags **0%** of standard NKE (which retains high cosine similarity to training features), so only whitened-distance (Mahalanobis-type) detectors catch NKE at all. Second, evasion generalises across design: penalising the tied-covariance penultimate detector alone already drives the *per-class* (untied) detector on the same features to 0% as well, and a two-layer ensemble objective drives **every** parametric detector — tied and untied, both layers — to 0% at 100% success. **No feature-distance detector we tested resists an adaptive attacker that can name and differentiate it**; the negative result of Table 5 is a property of feature-distance detection in general, not of the particular metric, and reproduces under NI-FGM as it did under NMI-FGM.

C.3 Extended analysis of attack variants and transferability (Sec. 4.9)

Variants at matched magnitude. The four variants show success-rate gaps (*e.g.* NMI-FGSM > NI-FGM), but those attacks also differ in the perturbation magnitude they induce. Projecting all

Table 8: Detection of adaptive NKE by a battery of feature-distance detectors, each at a *held-out-calibrated* 5% FPR (CIFAR-10, ResNet-18, NI-FGM, $L_2 \approx 18$; adapt rows at λ=0.01). Row 1 confirms honest calibration (clean FPR ≈0.04–0.05). Attacking only the tied-penultimate detector also evades the untied one; a two-layer ensemble objective evades every parametric detector. Cosine-kNN never flags NKE at all.

Scenario	Success	tied-penult	untied-penult	Maha layer-3	cosine-kNN
Clean false-positive rate	—	0.04	0.04	0.05	0.05
Standard NKE (no adaptation)	1.00	1.00	0.97	1.00	0.00
Adapt vs. tied-penult	1.00	**0.00**	**0.00**	0.94	0.36
Adapt vs. tied ensemble (penult + layer-3)	1.00	**0.00**	**0.00**	**0.00**	0.01

four onto the *same* L_2 sphere ($L_2 \approx 8$ on CIFAR-10, VGG-11 source), so they differ only in step rule and momentum, the reported differences largely vanish for the three L_2-normalized/momentum variants: NI-FGM (the primary attack), NMI-FGSM, and NMI-FGM all reach 100% success and are flagged 100% by Mahalanobis and 0% by MSP, with nearly identical edge preservation (Table 9). The sign-based, momentum-free I-FGSM is measurably weaker at matched magnitude (0.93 success, 0.87 Mahalanobis-flagged, 0.06 MSP-flagged) and also the most human-destructive of the four (generic proxy 0.38, vs. 0.68–0.83 for the others). The variant mainly trades off perturbation magnitude and step-rule reliability, and magnitude in turn drives the human gap and the rate of texture collapse (Sec. 4.8).

Table 9: Attack variants at matched $L_2 \approx 8$ (CIFAR-10, VGG-11 source, reference recognizer ResNet-18). Success and detectability are near-invariant to variant among the three L_2/momentum variants; I-FGSM (sign-based, no momentum) is the outlier; the human proxy varies more.

Variant	Success	Maha detect	MSP detect	Human proxy	Edge preservation
I-FGSM (L_∞ step)	0.93	0.87	0.06	0.38	0.49
NI-FGM (L_2 step, **primary**)	1.00	1.00	0.00	0.68	0.48
NMI-FGSM (mom.+sign)	1.00	1.00	0.00	0.83	0.47
NMI-FGM (mom.+L_2)	1.00	1.00	0.00	0.71	0.47

What the transferability collapse reveals. Black-box transfer collapses (<5% on ImageNet, Sec. 4.10) but this does not explain *which* examples transfer. We test whether source-model manifold-proximity (negative Mahalanobis distance) or edge preservation predicts transfer to a second architecture (ResNet-18→VGG-11, NI-FGM). At *small* perturbations, proximity predicts transfer above chance (AUROC 0.61 at $\epsilon_l = 1$, rising to ≈0.75 by $\epsilon_l = 5$–7): the examples that transfer are those still near the manifold. But this predictivity **decays through chance and reverses at large** ϵ_l (AUROC 0.48 at $\epsilon_l = 22$), and edge preservation never predicts transfer well ($|r| \leq 0.20$). Thus in the large-ϵ_l regime that defines NKE, transfer becomes disconnected from both manifold-distance and shape.

C.4 ImageNet scale and CLIP proxy: full detail (Sec. 4.10)

The second-CNN transfer proxy used on CIFAR-10 conflates "a human would recognize this" with "this transfers to another CNN"; at ImageNet the NKE transfer collapse drives the proxy to its floor, so NKE (0.125 at $\epsilon_l = 30$) and the matched-Gaussian control (0.155) become indistinguishable in Fig. 11 — not because the gap is signal loss, but because the proxy has no headroom left to show it. CLIP (ViT-B/32, LAION-2B; Fig. 12), never trained on these tasks and not the attack's transfer target, has ample headroom instead: it falls to 0.24 (CIFAR-10) and 0.33 (ImageNet) at the largest ϵ_l while the source model stays at 100%. The NKE-vs-Gaussian-control ordering is scale-dependent: on CIFAR-10, NKE remains *more* CLIP-recognizable than matched noise (*e.g.* $\epsilon_l = 8$: 0.66 vs. 0.17), replicating the "not signal loss" effect on a non-CNN proxy; on ImageNet the ordering *reverses* ($\epsilon_l = 100$: 0.33 vs. 0.61) — the perturbation that keeps ResNet-152 confident actively destroys broad semantic content *faster* than equal-energy noise, i.e. the examples are adversarial in the human direction too at this scale.

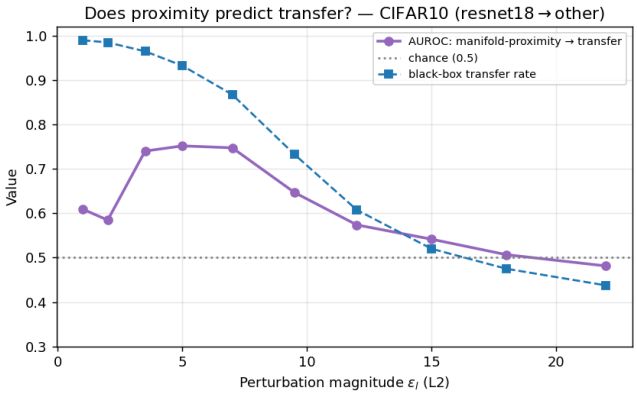


Figure 10: Manifold-proximity predicts transfer only at small-to-mid ϵ_l (ResNet-18→VGG-11, NI-FGM); its predictive power (AUROC, purple) rises then decays through chance and reverses as ϵ_l grows, tracking the black-box transfer rate (blue). In the NKE regime, transfer is not explained by proximity or shape.

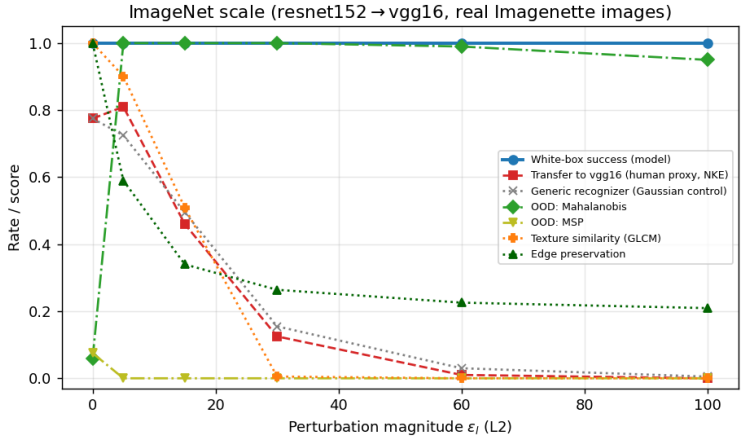


Figure 11: ImageNet scale (ResNet-152→VGG-16, real Imagenette images; independent of the CIFAR-10 primary-architecture choice). White-box success/confidence flat at 1.0; transfer (= generic-recognizer proxy) collapses to 0; Mahalanobis detects ∼100% while MSP stays at 0; texture is destroyed while edges partially survive. Note the NKE-transfer (red) and Gaussian-control (gray) curves nearly coincide — unlike CIFAR-10, the cross-model proxy cannot separate structured from random degradation here (see Sec. 4.10).

D VISION TRANSFORMER SHAPE-BIAS PROBE: FULL RESULTS

This appendix reports the per-track numbers underlying Sec. 4.6. The ViT is a from-scratch model (patch size 4 ⇒ 64 tokens, embedding dim 192, depth 6, 3 attention heads, MLP ratio 2; 1.8M parameters), trained for 70 epochs with AdamW (lr 5×10^{-4}, weight decay 0.05, 5-epoch linear warmup, label smoothing 0.1) under the same crop+flip augmentation as the CNNs, reaching **83.2%** clean test accuracy. It is registered as a third CIFAR-10 architecture and passed through the identical pipeline (attack, stimulus set, and every evaluation track) as VGG-11 and ResNet-18. All attacks use NI-FGM at 60 steps as in the main text; n=300 correctly-classified, class-stratified images per source.

Cross-model transferability. Table 10 extends the black-box matrix (Table 1) to all three CIFAR-10 architectures. White-box success is 1.00 on the diagonal; every off-diagonal (black-box) entry is low, and the ViT row (NKE *generated by* the ViT) is the lowest of all.

Retention-normalised cross-recognizer analysis. Table 11 gives the retention (accuracy on NKE ÷ accuracy on that source's clean images, i.e. normalised per recognizer to remove the clean-accuracy

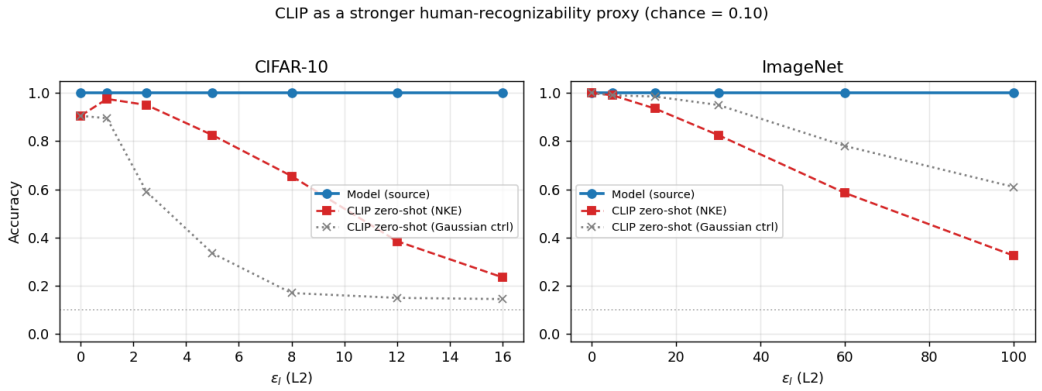


Figure 12: CLIP zero-shot accuracy (a non-CNN, broadly-trained recognizability proxy) vs. ϵ_l. Model stays at 1.0; CLIP falls far below at both scales (gap confirmed). CIFAR-10: NKE $>$ Gaussian control (not signal loss). ImageNet: NKE $<$ Gaussian control (perturbation is *more* destructive to semantics than noise). Chance $= 0.10$.

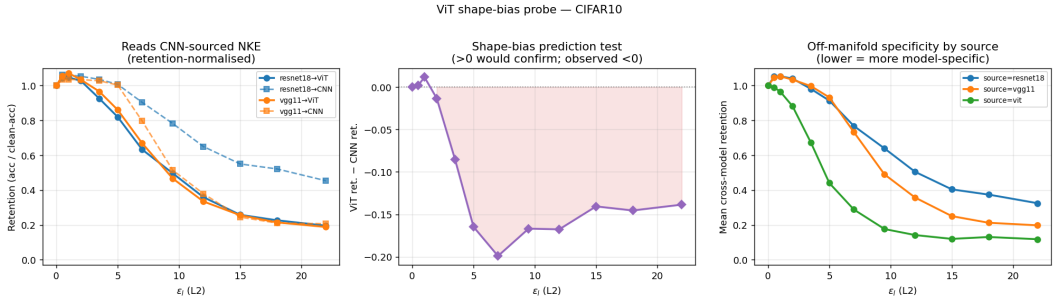


Figure 13: ViT shape-bias probe (CIFAR-10, NI-FGM), retention-normalised. **Left:** retention of CNN-sourced NKE by the ViT vs. the other CNN recognizer. **Middle:** the shape-bias prediction test — ViT-minus-CNN retention is *negative* throughout (past ϵ_l=2), so the shape-biased model is *not* a better reader (prediction falsified). **Right:** off-manifold specificity by source — NKE generated by the ViT (green) transfers worst and collapses earliest, i.e. attention models produce the most model-specific perturbations.

confound) that underlies Fig. 13. The shape-bias contrast (last two rows) is negative from $\epsilon_l \geq 2$ onward: the ViT recognizer is not a better reader of CNN-sourced NKE. The source-specificity block shows ViT-sourced NKE collapsing earliest and lowest.

Mechanism (ViT source). Table 12 is the ViT analogue of Table 6: the same texture-before-shape dissociation, with model accuracy pinned at 1.00.

Human proxies (ViT source). Table 13 is the ViT analogue of Table 3; the generic recognizer is ResNet-18. Note that the NKE curve nearly coincides with the matched-Gaussian control by $\epsilon_l \geq 9.5$ (unlike the VGG-11 source in Table 3), i.e. ViT NKE is almost as cross-model-destructive as noise.

OOD detection (ViT source). Table 14 is the ViT analogue of Table 4. The confidence/energy blindness reproduces, but Mahalanobis detection on the ViT's global CLS token is far weaker than on the VGG-11 penultimate (peaking at 0.42 rather than 1.00).

Adaptive attack (ViT source). Table 15 is the ViT analogue of Table 5. Oblivious NKE is only 40% detected here (weaker feature-distance signal), and a tiny penalty (λ=0.001) drives both white-box and held-out detectors to $\approx$0 while success stays high; success then declines faster with λ than on the CNN, so the no-cost evasion band is narrower but present.

Table 10: CIFAR-10 cross-model label-kept matrix at ϵ_l=18 (NI-FGM; fraction of source-generated NKE the target still labels correctly). Rows: source; columns: target. $*$ = white-box.

Source ↓ / Target →	ResNet-18	VGG-11	ViT
ResNet-18	1.00*	0.44	0.22
VGG-11	0.21	1.00*	0.17
ViT	0.11	0.17	1.00*

Table 11: Retention (acc on NKE / acc on clean) on CIFAR-10, NI-FGM. *Top:* shape-bias contrast — mean (ViT-recognizer retention − other-CNN-recognizer retention) for CNN-sourced NKE; negative means the shape-biased ViT reads CNN NKE *worse*. *Bottom:* source specificity — mean cross-model retention of the NKE each source generates (lower = more model-specific).

ϵ_l	2	5	9.5	15	22
Shape-bias prediction test (CNN-sourced NKE)					
ViT − CNN retention (contrast)	−0.013	−0.165	−0.167	−0.141	−0.139
Source specificity (mean cross-model retention by source)					
ResNet-18 source	1.04	0.91	0.64	0.40	0.32
VGG-11 source	1.03	0.93	0.49	0.25	0.20
ViT source	**0.88**	**0.44**	**0.18**	**0.12**	**0.12**

Table 12: Shape vs. texture for ViT-sourced NKE (CIFAR-10, NI-FGM). Model accuracy is 1.00 at every ϵ_l.

ϵ_l	1	2	5	9.5	22
Edge preservation (Canny F1)	0.83	0.76	0.57	0.45	0.31
Texture similarity (GLCM)	0.96	0.91	0.64	0.04	0.00

Table 13: Recognizability proxies vs. ϵ_l for ViT-sourced NKE (CIFAR-10, NI-FGM; generic recognizer = ResNet-18). Model accuracy is 1.00 at every level.

ϵ_l	0	2	5	9.5	15	22
Generic recognizer (NKE)	0.96	0.82	0.27	0.10	0.10	0.11
Generic recognizer (Gauss ctrl)	0.96	0.84	0.24	0.11	0.10	0.09
Shape recognizer (NKE)	0.68	0.60	0.30	0.15	0.13	0.10

Table 14: OOD detection rate at a held-out-calibrated 5% FPR, and ECE, for ViT-sourced NKE (CIFAR-10, NI-FGM). Model accuracy 1.00 throughout.

ϵ_l	1	2	5	9.5	22
ECE	0.03	0.03	0.03	0.03	0.04
MSP detection	0.00	0.00	0.00	0.00	0.00
Energy detection	0.00	0.00	0.00	0.00	0.00
Mahalanobis detection	0.05	0.13	0.32	0.37	0.39

Table 15: Adaptive attack against the Mahalanobis detector on the ViT (CIFAR-10, NI-FGM, $L_2 \approx 18$). Detection at a held-out-calibrated 5% FPR.

λ	0	0.001	0.01	0.02	0.03	1.0
NKE success	1.00	0.99	0.88	0.69	0.34	0.13
Perturbation L_2	18.01	17.99	17.99	17.99	18.00	18.00
Maha detection (white-box)	0.40	0.00	0.00	0.00	0.00	0.00
Maha detection (held-out)	0.43	0.01	0.00	0.00	0.00	0.00
MSP detection	0.00	0.02	0.06	0.09	0.25	0.00