VITVIZ: ANALYZING ATTENTION REDISTRIBUTION UNDER ADVERSARIAL ATTACKS IN VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) have achieved state-of-the-art performance in image classification, yet they remain vulnerable to adversarial attacks. Since attention is a fundamental mechanism underlying these models, understanding how it behaves under adversarial perturbations is crucial. However, to the best of our knowledge, the literature lacks tools for systematically exploring attention under adversarial regimes. To address this gap, we introduce ViTViz, an open-source tool for exploring attention under adversarial attacks. ViTViz enables researchers to independently investigate key dimensions of attention—including the aggregation operator, anchoring class, layer, attention head, attack step, and perturbation budget—on any plain ViT classifier with a single class token, in the timm or the Hugging Face layout. For the benchmark's six models, it places each measurement at its percentile among images of the same model, attack, operator and outcome. Furthermore, we introduce an attention redistribution analysis based on three measures, each answering a different question: how far attention mass moved, whether the distribution changed at all, and whether attention spread or concentrated. As a proof of concept, we applied this analysis to 1,000 ImageNet images subjected to five attacks, using six ViT variants and two attention aggregation operators, one class-agnostic and one anchored on a class label. On the same images, the two operators disagree about whether a successful attack moves attention further than a failed one. How far attention moved therefore says something about an attack only when the operator and anchor that produced the measurement are reported.

1 INTRODUCTION

Reading a Vision Transformer's attention before and after an adversarial attack is one way to ask what the attack did (Wu et al., 2024a), and whether the explanation survived it (Hu et al., 2024). A ViT (Dosovitskiy et al., 2021) emits one attention matrix per head per layer. An operator usually condenses those matrices into the map a reader sees, one value per patch, and operators differ, among other things, on whether a class label enters the computation (Abnar & Zuidema, 2020; Chefer et al., 2021a;b).

Every map a reader sees comes from one choice of operator, anchoring class, layer and attack step, and it does not say whether the change it shows is large for that model, attack and outcome. Of the attention tools we surveyed, three read unperturbed input (Vig, 2019; Aflalo et al., 2022; Yeh et al., 2023) and one, for text, sets attention beside its adversarial counterpart (Correia & Lopes Cardoso, 2023); none does so for a ViT. Hu et al. (2024) score ViT maps under attack against foreground masks, not against whether the attack changed the prediction. Read against the outcome, two operators part ways: on the same images, a class-agnostic and a class-anchored one disagree on whether a successful attack moves attention further than a failed one (Section 6).

We therefore treat the operator as a factor of the measurement, and release ViTViz (Section 4), which attacks the image behind a map again and turns its operator, anchor, layer and step into settings (Fig. 1). At every step of the attack it computes both a class-agnostic rollout map and a Generic Attention-model Explainability (GAE) map (Chefer et al., 2021a), so the user chooses the reading after the run. A slider walks the attack step by step, and, for the six checkpoints of our

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

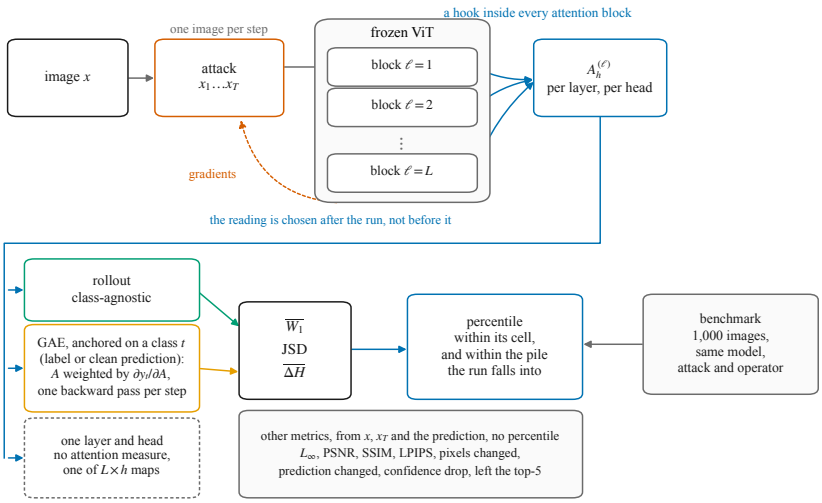


Figure 1: Where the attention comes from and what is done with it. The model is frozen, and a hook inside each attention block captures attention per layer and head. Rollout and the single layer-and-head view read that capture as it is; GAE weights it by the gradient of the anchor's logit. Only the three attention measures of the two aggregated readings are placed at a percentile.

benchmark, ViTViz places the clean-to-final change at its percentile among benchmark images of the same model, attack, operator and outcome (Fig. 2(c)).

We read the clean and attacked maps as distributions over patches and call the normalized transport distance between them the *displacement* of the attention, with the Jensen–Shannon divergence and the entropy change as secondary checks (Section 3). ViTViz's percentiles come from a benchmark that applies the same reading to six ViT variants, a thousand ImageNet images and five attacks at $\varepsilon = 4/255$. Each model under each attack forms a *cell* (Fig. 3), and the attack sorts its images into two *piles*: *flipped*, where it changed the prediction, and unflipped, where it did not. Comparing the two piles asks whether attention registers the attack's success at all. A cell separates them when a bootstrap interval on the ratio of their mean displacements excludes 1, where attention would have moved as far either way (Section 5).

On that benchmark, the attention reading determines the observed contrast. Of the twenty-four cells with enough unflipped images to measure, rollout separates the piles in eight and GAE, anchored on the ground-truth class, in twenty. The two reach the same verdict in only six (Table 2). On a checkpoint fine-tuned for traffic signs, under PGD (Madry et al., 2018), rollout finds more displacement in the flipped pile and GAE less, both intervals excluding 1 (Section 6.3). The readings also disagree about which images moved the most: across the estimable cells, their rankings of the same flipped images correlate at a median ρ of 0.27 (Section 6). On ViTViz's own gallery, the two place over a third of the 232 runs in opposite halves of their cells (Appendix L). Our contributions are ViTViz, which exposes that dependence on single images, and the evidence that conclusions about attention redistribution under attack depend on how attention is read, a dependence that can reverse the observed relation between displacement and attack success.

2 Related Work

This work is positioned at the intersection of two fields: attention interpretability in Vision Transformers and adversarial attacks on them. Work on attention interpretability asks whether the attention explains a prediction at all, a question first raised for recurrent models (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Reading a ViT's attention as a spatial map requires collapsing all per head per layer attention matrices into a single one, and the operators that do so differ on several aspects, among them, whether a class label enters the computation (Abnar & Zuidema, 2020; Chefer et al., 2021a;b; Jo et al., 2025; Joita & Ivanovska, 2026; Wu et al., 2024b; Bousselham et al., 2025).

Work on adversarial attacks, in turn, bounds a perturbation and maximizes the loss (Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Dong et al., 2018); its ViT-specific variants change how the gradient is formed rather than what is maximized (Wei et al., 2022; Zhang et al., 2023), or weight the gradient by the model's own attention rollout (Mahmood et al., 2021); and success is scored at the output, under protocols the field has standardized (Croce et al., 2021; Carlini et al., 2019). Where the fields meet, the instability of attention under perturbation is treated as a defect to be repaired (Hu et al., 2024), and maps are scored against foreground masks (Chefer et al., 2021b; Hu et al., 2024) rather than against the attack's outcome. Tools for inspecting attention directly read it on unperturbed input (Vig, 2019; Aflalo et al., 2022; Yeh et al., 2023), and the one that sets it beside its counterpart under attack does so for text (Correia & Lopes Cardoso, 2023); to the best of our knowledge, none does so for a ViT.

We treat the aggregation operator itself as a factor of the measurement, and read attention on populations split by whether the attack changed the prediction; ViTViz makes that comparison browsable case by case, so a practitioner may inspect how attention behaves along the attack.

3 Problem Formulation

For classification, a Vision Transformer (ViT) (Dosovitskiy et al., 2021) splits an image x into patches and prepends a learnable class token `[CLS]` to the N patch tokens. Attending over them, the ViT emits one attention matrix per head per layer. We write $A_h^{(l)} \in \mathbb{R}^{(N+1)\times(N+1)}$ for the post-softmax matrix at layer l and head h.

3.1 Two ways to collapse the attention matrices

Class-agnostic aggregation. Attention Rollout (Abnar & Zuidema, 2020) fuses the heads of each layer by an unweighted mean, averages the result with the identity (I) to carry the residual path, and propagates across depth by multiplication:

$$\hat{A}^{(l)} = \tfrac{1}{2}\left(\mathbb{E}_h\big(A_h^{(l)}\big) + I\right), \qquad \Phi^{(l)} = \hat{A}^{(l)}\,\Phi^{(l-1)}. \tag{1}$$

Here $\mathbb{E}_h(\cdot)$ is the mean across heads and $\hat{A}^{(l)}$ the fused matrix at layer l, whose rows the factor $\frac{1}{2}$ keeps summing to 1; the accumulator $\Phi^{(l)}$ starts at $\Phi^{(0)} = I$ and runs over the model's L layers.

The `[CLS]` row of $\Phi^{(L)}$, without its own entry and reshaped into the $G \times G$ grid ($G^2 = N$), is the attention map R of x under rollout, written R_{roll} where a distinction is needed. No class appears in Eq. (1).

Class-anchored explanation. Generic Attention-model Explainability (GAE) (Chefer et al., 2021a) weights the attention matrix elementwise by $\nabla A_h^{(l)} = \partial y_t / \partial A_h^{(l)}$, the gradient of the logit of a chosen class t, discarding negative contributions before averaging over heads:

$$\bar{A}^{(l)} = \mathbb{E}_h\left(\left(\nabla A_h^{(l)} \odot A_h^{(l)}\right)^{+}\right), \qquad \Phi^{(l)} = \Phi^{(l-1)} + \bar{A}^{(l)}\,\Phi^{(l-1)}, \tag{2}$$

where $\odot$ is the elementwise product and $(\cdot)^+$ sets negative entries to zero. From $\Phi^{(0)} = I$, the same readout gives R under GAE, written R_{gae} where a distinction is needed.

The class t is an *anchor*: it belongs to the reading, not to the model. A frozen model and a fixed image admit one rollout map, but as many GAE maps as there are classes.

Threat model. Every attack considered here is untargeted, succeeding by changing the model's prediction rather than by reaching a particular class, and white-box, with access to the weights and gradients of the model under attack. Given an image x, it produces x' with $\|x' - x\|_\infty \le \varepsilon$, and the same model and operator applied to x' give a second map, R'.

3.2 ATTENTION REDISTRIBUTION

We define *attention redistribution* as the change between R and R' read as distributions over the G^2 patches, each normalized to sum to 1. We read that change between distributions with three measures, each answering a different question.

$\overline{W_1}(R, R')$ is the earth mover's distance (Rubner et al., 2000) with Euclidean ground cost between patch positions, divided by the grid diameter $D = (G-1)\sqrt{2}$ so that it lies in $[0, 1]$. The ground cost makes it spatial, moving mass to a close patch being cheap and moving it across the image not, even where a divergence cannot tell the two apart (Fig. 7); it answers how far attention mass moved.

The Jensen–Shannon divergence (Lin, 1991) $\mathrm{JSD}(R, R')$, with base-2 logarithms, lies in $[0, 1]$ and has no ground cost, so nearby and distant moves count the same; it answers whether the distribution changed at all.

The entropy change $\overline{\Delta H} = \big(H(R') - H(R)\big) / \ln G^2$, with H the Shannon entropy (Shannon, 1948), lies in $[-1, 1]$ and answers whether attention spread or concentrated. It compares one number from each map, so very different maps can leave it near zero: four of the five synthetic pairs of Appendix D leave it at exactly zero.

4 VITVIZ TOOL

The operator is a factor of the measurement (Section 1), so we need an instrument for varying it. ViTViz starts from one image and follows the path of Fig. 1. The attack produces one image per step, and for each of them a hook inside every attention block captures $A_h^{(l)}$, from which ViTViz accumulates the rollout and GAE maps of Eqs. (1) and (2).

Both operators, one run. Every run computes the class-agnostic rollout map R_{roll} and the class-anchored R_{gae} of Section 3 at every step of the attack. The interface switches between them after the run, without running the model again (Fig. 2(b)). The anchor t of the GAE map, set before the run, can be the ground-truth class or the model's clean prediction, and the interface labels which one is in use.

One layer and one head. The `[CLS]` row of $A_h^{(l)}$ can also be read on its own, so that a single layer and head can be inspected under attack (Fig. 2(d)). ViTViz attaches none of the three attention measures to it.

Along the attack. A step slider walks the iterates (Fig. 2(a)), from the clean image to the attacked one, and the map animates between them under the selected operator.

A displacement read against its own pile. For $\overline{W_1}$, JSD and $\overline{\Delta H}$, ViTViz places the run's value at its percentile among the benchmark's images for the same model, attack and operator, and within the pile the run itself falls into (Fig. 2(c)). Pooling the two piles would compare a successful attack against a distribution that Table 2 shows to be shifted. A run that departs from the benchmark's budget, step size, step count, head fusion or anchor, or reads an uploaded image, still receives its percentile, with a note naming the departure; an uploaded checkpoint receives none.

Models. Besides the benchmark's six checkpoints, a user may supply a plain ViT classifier with a single class token and any number of classes, in the timm or Hugging Face `vit.*` layout; ViTViz declines distillation and register tokens, LayerScale, a final `fc_norm`, query and key normalisation, and blocks that leave the plain design, as in Swin.

5 BENCHMARK DESIGN

Models and images. We benchmark six Vision Transformers distributed through `timm` (Wightman, 2019), three capacities at two patch sizes each, all fine-tuned on ImageNet-1k (Appendix B gives the checkpoints). Patch-16 models run at 224×224 pixels and patch-32 models at 384×384, giving grids of $G = 14$ and $G = 12$, so patch size, input resolution—and, for ViT-S/16, the only one

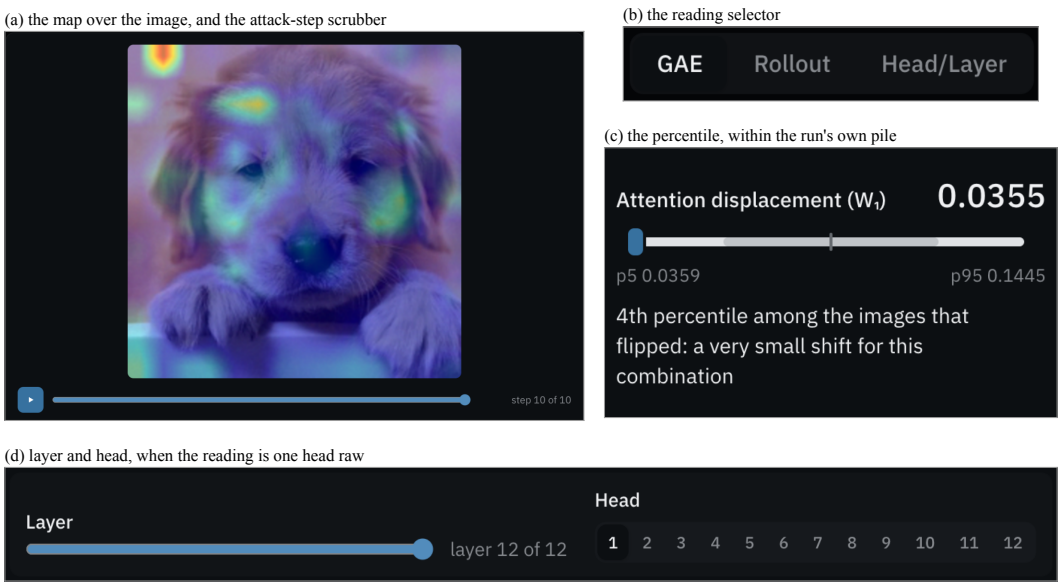


Figure 2: ViTViz after a run on one gallery image, ViT-B/16 under PGD. (a) The map at the current attack step, here the last, and the step slider. (b) The reading selector, rollout, GAE, or one layer and head, switched without running the model again. (c) $\overline{W_1}$ at its percentile among images of the same model, attack, operator and pile. (d) The layer and head controls.

not pretrained on ImageNet-21k, pretraining data—vary together. Every model reads the same $1{,}000$ images, one per ImageNet-1k class: 905 from the ILSVRC-2012 validation split (Russakovsky et al., 2015) and 95 from the full-ImageNet pool of Guillaumin et al. (2014), which are the ones that carry foreground masks (Appendix C). We read a seventh checkpoint, fine-tuned on a different domain, under the same protocol in Section 6.3.

Attacks. All attacks follow the threat model of Section 3 at $\varepsilon = 4/255$ and optimize against the clean prediction. We chose that budget from a probe over seven budgets (Appendix A), and holding all five attacks to one budget is what makes them comparable. FGSM (Goodfellow et al., 2015) takes a single step; PGD (Madry et al., 2018), MIM (Dong et al., 2018), SAGA (Mahmood et al., 2021) and TGR (Zhang et al., 2023) take ten, of size $\varepsilon/10$. SAGA weights the attacked ViT's gradient by that model's rollout map (Eq. (1)) and blends it with the gradient of an auxiliary CNN, and TGR, designed for transfer to a black-box target, runs here white-box; Appendix B gives every hyperparameter.

Reading the maps. We anchor GAE on the ground-truth class, the same class for the clean and the attacked map, so the two readings differ in operator, anchor and map normalisation at once; the random-anchor control below separates the anchor and Appendix H the normalisation. Normalizing each map to sum to 1 (Section 3) does not cancel the difference in scale, so we compare operators through ratios and never through levels.

Cells and the ratio test. A cell is one entry of the six-model by five-attack table, thirty in all, read under each operator on the images the model classifies correctly without attack. An image is flipped when the attack changes the model's prediction, regardless of the label, and the flipped and unflipped images of a cell form its two piles. Each cell reports the mean displacement of its flipped pile divided by that of its unflipped pile, so a ratio of 1 means attention moved as far when the attack changed the prediction as when it did not. The interval on the ratio is a 95% percentile bootstrap (Efron & Tibshirani, 1993) over 10^4 resamples, each pile resampled independently, and a cell *separates* the piles when its interval excludes 1. The same test runs on JSD and on the absolute entropy change. Figure 3 walks one cell through the test. We call a cell *estimable* when its unflipped pile holds at least 20 images, since with fewer the interval runs too narrow: a flat cell cut to five unflipped images separates in a fifth of draws, and cut to twenty in a tenth (Appendix N). We count

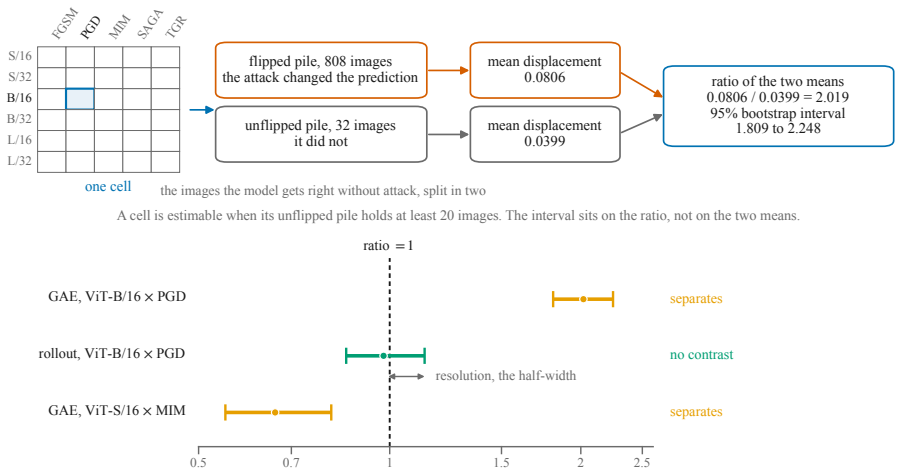


Figure 3: One cell, from the table to the verdict. Top: ViT-B/16 under PGD, read under GAE: the images the model classifies correctly split into a flipped and an unflipped pile, and the cell reports the ratio of their mean displacements with a bootstrap interval. Bottom: three cells against a ratio of 1 on a logarithmic axis: the same cell separates its piles under GAE and not under rollout, and ViT-S/16 under MIM separates below 1.

Table 1: The four readings in this paper and the question each one answers.

Reading	Question it answers	Where
Rollout (no anchor)	What moves without class information?	Section 6
GAE, ground-truth class	What moves with the correct class?	Section 6
GAE, clean prediction	Does the identity of the prediction matter?	Appendix I
GAE, random class	Is the contrast specific to class information?	Appendix J

no cell below that. A cell whose interval contains 1 reports no contrast, which means something only if the cell could have found one, so we also give its *resolution*: the half-width of the interval, the smallest departure from 1 it could have registered. No ratio is pooled across architectures, since a pooled ratio can land on the other side of 1 from the cells inside it (Section 6).

Controls. Table 1 lists the four readings and the question each answers. The random-anchor control draws a class uniformly from the thousand, never the true one, and holds it for both maps: if the contrast rests on class information it should vanish, and if it survives, the gradient weighting is doing the work. It runs on thirteen of the twenty cells that separate under the ground-truth anchor, with SAGA as a positive control (Appendix J). Within each model we also decompose the displacement over images and attacks, and over both operators at once, and report each factor's share η^2 (Appendix B).

6 Results

6.1 One image at a time

The same attack, read twice. The five panels are the view ViTViz puts on a single image (Section 4), filled here from the benchmark's stored maps so that this section and Table 2 rest on the same run. In the first two rows of Fig. 4 the class-agnostic map is hard to tell apart from one side to the other, while the class-anchored map is visibly reorganised; in the third, both operators place the image high in their own distribution.

Disagreement is the rule, not the exhibit. We rank the 808 flipped images of this cell by displacement under each operator, and the two orderings correlate at $\rho = 0.173$. Across the twenty-

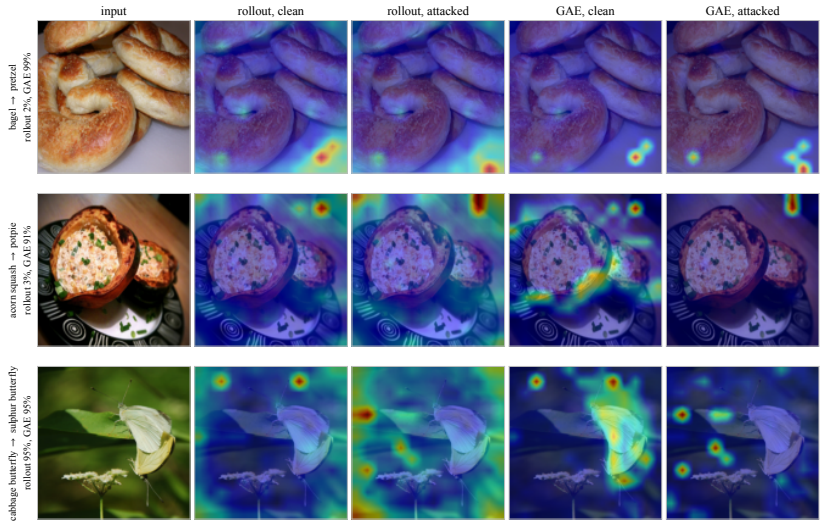


Figure 4: Three images ViT-B/16 classifies correctly and PGD flips at $\varepsilon = 4/255$, read before and after the attack under rollout and under GAE anchored on the ground-truth class. Percentages rank each image's $\overline{W_1}$ among the 808 flipped images of this cell, within its operator. The first two rows are the largest rank disagreements; the third is the closest agreement within the top decile of both. Colour scales are shared within an operator, not across operators.

Table 2: Ratio of mean $\overline{W_1}$ displacement between attacks that changed the prediction and attacks that did not, per model and attack, under both operators, on images the model classifies correctly without attack. **Bold**: the 95% bootstrap interval on the ratio excludes 1 (intervals not shown). Grey: fewer than 20 non-flipped images, excluded from every count; **–**: none. Ratios compare across the two panels; levels do not.

	FGSM	PGD	MIM	SAGA	TGR
Attention rollout (class-agnostic)					
ViT-S/16	0.972	0.788	0.864	**1.381**	1.042
ViT-S/32	**1.106**	–	0.560	**2.446**	1.043
ViT-B/16	1.028	0.978	1.200	0.970	1.009
ViT-B/32	1.020	0.717	0.938	**2.727**	**0.939**
ViT-L/16	0.989	1.040	1.027	**2.072**	1.010
ViT-L/32	1.008	1.101	1.097	**2.419**	**1.177**
GAE, ground-truth anchor (class-informed)					
ViT-S/16	**0.704**	0.707	**0.660**	0.881	**0.833**
ViT-S/32	0.885	–	1.206	**1.563**	**0.916**
ViT-B/16	**1.362**	**2.019**	1.699	**1.903**	**1.362**
ViT-B/32	**1.097**	0.875	1.156	**3.140**	0.982
ViT-L/16	**1.072**	**1.422**	**1.220**	**1.641**	1.042
ViT-L/32	**1.137**	**1.155**	**1.205**	**2.190**	**1.140**

four estimable cells the correlation has a median of 0.270 and never exceeds 0.461: the readings do not merely rescale the displacement, they disagree about which images moved the most. One level up the disagreement holds: in Table 2 this cell has a ratio of 0.978 under rollout, which does not separate its piles, and 2.019 under GAE, which does.

6.2 One cell at a time

The attention reading determines whether the contrast registers. Twenty-four of the thirty model and attack combinations leave a failed-attack pile large enough to estimate, and in Table 2 we give, for each of them, the ratio between the mean $\overline{W_1}$ of successful and of failed attacks. Eight

Table 3: Ratio of mean $\overline{W_1}$ displacement between attacks that changed the prediction and attacks that did not, on a ViT-B/16 fine-tuned on GTSRB at $\varepsilon = 4/255$, under both operators, on the 403 of 430 images it classifies correctly without attack. **Bold**: the 95% bootstrap interval on the ratio excludes 1 (intervals not shown). Ratios compare across the two columns; levels do not.

	flipped / unflipped	rollout	GAE
FGSM	194 / 209	1.039	**1.094**
PGD	364 / 39	**1.109**	**0.840**
MIM	358 / 45	**1.304**	1.070
TGR	125 / 278	1.009	**1.102**

of the twenty-four intervals exclude 1 under rollout and twenty do under GAE anchored on the ground-truth class; Appendix E plots each cell under both operators.

The class-agnostic reading is not blind. Rollout separates the piles in five of the six SAGA cells, with ratios from 1.381 to 2.727, though SAGA weights its gradient by the rollout map itself (Section 5), and three of the cells it separates are ones GAE leaves flat. Where rollout reports no contrast, its interval was still narrow enough to have caught one, with median resolution 0.076 across those sixteen cells (Appendix F). What rollout does not report is therefore not what it cannot see.

The direction of the contrast follows capacity. Direction is not uniform across the twenty cells GAE separates: four have the ratio below 1, all four on a Small backbone, and in ViT-S/16 under MIM it is the successful attacks that displace least, at a ratio of 0.660. The split runs wider than those four cells, since the change of operator moves every Small cell down and fourteen of seventeen Base and Large cells up; it holds under $\overline{W_1}$ and does not replicate under JSD or $\overline{\Delta H}$, so we report it without explaining it. Pooling the six models hides the split (Appendix F).

What the contrast does not license. Class identity is not established as the active ingredient: drawing the anchor at random rather than from the label leaves the contrast standing in ViT-L/32 under FGSM, at 1.175 against 1.137 for the true class. Part of what GAE registers is its anchor's own probability: compared only among images that end with the same probability on the true class, the five cells both operators separate keep a median 64% of their log-ratio under GAE and 92% under rollout (Appendix J). Normalisation, one of the three differences between the readings, closes 20% of the gap between their mean displacements in ViT-B/16 and 37% in ViT-L/32 (Appendix H). Appendix F gives these qualifications in full.

Which attack was run explains little. Across the six models, the choice of attack accounts for between 3% and 24% of the variance in $\overline{W_1}$, 9% on average, against 42% to 63% for the identity of the image, between 1.8 and 19.7 times more. Over both readings at once, the image accounts for 32%, the attack for 7%, and the operator's interaction with the image for 17% (Appendix B): the reading turns on which image it reads more than twice as much as the attack explains by itself. Whether an attack appears to leave a trace in attention is not determined by the attack alone.

Step by step. Averaged over images, displacement grows at every step of an iterating attack, but under PGD and SAGA almost no single trajectory does: attention falls partway back toward the clean map, which the mean curve hides, and the separation of Table 2 is there from the first step or not at all (Appendix G).

6.3 One checkpoint from another domain

Every cell so far reads ImageNet images with a model fine-tuned on ImageNet-1k, so the operator dependence could be a property of that pairing. We ran the protocol of Section 5 on a ViT-B/16 fine-tuned to the 43 classes of the German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al., 2012), at the same $\varepsilon = 4/255$ and without SAGA; Appendix K gives the checkpoint, the images and why SAGA is absent. The change does not isolate the domain: images, checkpoint, provenance and class count change together.

Under $\overline{W_1}$ the two operators disagree in all four cells of Table 3: three by presence, and PGD by direction. Rollout puts the ratio between PGD's flipped and unflipped piles at 1.109, interval 1.007 to 1.226; GAE, anchored on the ground-truth class, puts it at 0.840, interval 0.730 to 0.967. The class-agnostic reading says attention moved further on the images the attack flipped; the class-anchored reading says it moved less. Both intervals exclude 1: this is two signals, opposed, over the same images sorted into the same piles. The same reversal appears with the operator held fixed: ViT-B/16 under PGD, read under GAE, gives 2.019 on ImageNet (Table 2) against 0.840 here.

No estimable cell of Table 2 does this: five of its cells have both operators separating and all five point the same way, and raising the floor only removes cells, never reverses one. The reversal repeats under the absolute entropy change (Appendix K), so it is not an artefact of one measure, and we do not establish why the class-anchored reading inverts here.

7 Discussion

The aggregation operator is a measurement decision, not a presentation detail. A displacement read under one operator does not say what another would register, and neither of the two counts of Table 2 makes its operator the reference. Reports should name the operator and compare operators by ratio, which normalisation barely moves (Appendix H), and not by level. And since within each model the attack explains less of the displacement than the image does (Section 6), we want an effect size beside any ordering of attacks. ViTViz exposes the operator as a setting the user chooses (Section 4). Its percentiles are approximate: the reference comes from the benchmark's GPU runs and ViTViz runs on CPU, where the same image lands a median of 10 to 13 percentile points away under PGD and of 4 or fewer under the other attacks (Appendix L).

Limitations. Attack success, not sample size, limits what can be measured: PGD and MIM flip 91.5% to 100% of clean-correct predictions, so 6 of their 12 cells fall below the floor of 20 unflipped images, and more images would not lift them. The floor does not decide the result: at 30, GAE still separates more than twice as many cells as rollout, 16 of 20 against 7. Seventy-four of the thousand images lie in the split the models were fine-tuned on, and dropping them moves the counts from 8 and 20 to 6 and 21. The single budget measures TGR before it has finished gaining; rerun at eight budgets on two models, the operator dependence holds at every one (Appendix A). The evidence comes from one configuration: two operators, untargeted white-box attacks, patch size travelling with resolution, and one checkpoint from another domain. The variance shares are η^2, which overstates them; after the ω^2 correction the image still explains 1.4 to 16.8 times more than the attack. Appendix M gives each in full.

8 Conclusion

We asked what a Vision Transformer's attention map says about an attack once we treat the operator behind it as a choice and read its change against images of the same model, attack and outcome. The reading changes the verdicts: rollout and GAE agree on six of the twenty-four estimable cells, and each separates cells the other leaves flat (Table 2); on a checkpoint fine-tuned for traffic signs they read one attack in opposite directions (Section 6.3). Image by image the operators largely disagree on which images moved.

ViTViz takes this reading to single images, the user's own included, computing both operators in one run and placing each displacement at its percentile within its own cell (Section 4). We conclude that a displacement says something about an attack only when it is read against its own cell, with its operator and anchor named.

AI use statement

Large language models assisted this work in four ways, each checked by the authors. *Writing*: Claude (Anthropic) drafted part of the prose from the authors' outlines and decisions, and ChatGPT (OpenAI) gave a round of reviewer-style comments on a draft. *Retrieval and discovery*: Elicit and Claude helped find some of the papers we read; the authors read every cited paper and checked every

citation against its source. *Research ideation and execution*: Claude helped build both the analysis and interface code. The research questions, the design and every decision about what the paper claims are the authors'. *Other purposes*: Claude launched and monitored runs on GPU, checked every number in the text against the analysis outputs, formatted the source and helped prepare the anonymized supplementary material.

ETHICS STATEMENT

The attacks we study are published, and we run them white-box against released checkpoints to measure how attention maps register them; the paper introduces no new attack and no way to make one more effective. Attacks on traffic-sign classifiers bear on safety-critical systems, and our case study reads one publicly released checkpoint to report how its explanations change, not to target a deployed system. The ImageNet images are used under the ImageNet terms of access for non-commercial research; the traffic-sign images are drawn from the public GTSRB test set (Stallkamp et al., 2012), and the traffic-sign checkpoint is released under Apache-2.0. No personal data is collected and no human subjects are involved.

REPRODUCIBILITY STATEMENT

The supplementary material holds the anonymized code of ViTViz and of the benchmark: the interface; the sweep runner with its configurations; the per-image results of both sweeps and of the traffic-sign case; and scripts that regenerate Tables 2, 3 and 10, the budget analysis of Appendix A from its per-image data, the median correlation between the two operators, and the verdicts without the seventy-four benchmark images that fall in the ImageNet-1k training split, which it lists. The appendices give the checkpoints, the attack settings, the budget probe and the statistical protocol (Appendices A and B). The traffic-sign checkpoint is `bazyl/gtsrb-model` on the Hugging Face Hub, with its configuration at revision `95f6c95` and weights whose SHA-256 begins `0d2c52a1`. The timm release the sweeps ran under was not recorded; rerunning 500 benchmark images at $\varepsilon = 4/255$ under both operators, on the device of the original runs, reproduces all but one of $10{,}000$ prediction outcomes and 99% of displacements to within 0.23%.

Attacks that iterate do not reproduce exactly across devices; Appendix L measures by how much and explains why.

REFERENCES

Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.385. URL https://doi.org/10.18653/v1/2020.acl-main.385.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers, 2022. URL https://arxiv.org/abs/2203.17247.

Walid Bousselham, Angie W. Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 1–10. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01891. URL https://doi.org/10.1109/ICCV51701.2025.01891.

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

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 387–396.

IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00045. URL https://doi.org/10.1109/ICCV48922.2021.00045.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 782–791. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.00084. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html.

Pedro Gonçalo Correia and Henrique Lopes Cardoso. Towards explaining shortcut learning through attention visualization and adversarial attacks. In Lazaros Iliadis, Ilias Maglogiannis, Serafin Alonso, Chrisina Jayne, and Elias Pimenidis (eds.), *Engineering Applications of Neural Networks*, pp. 558–569, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34204-2.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark, 2021. URL https://arxiv.org/abs/2010.09670.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum, 2018. URL https://arxiv.org/abs/1710.06081.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, December 2014. ISSN 0920-5691. doi: 10.1007/s11263-014-0713-9.

Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. Improving interpretation faithfulness for vision transformers, 2024. URL https://arxiv.org/abs/2311.17983.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1357. URL https://doi.org/10.18653/v1/n19-1357.

Sehyeong Jo, Gangjae Jang, and Haesol Park. GMAR: gradient-driven multi-head attention rollout for vision transformer interpretability. In *IEEE International Conference on Image Processing, ICIP 2025, Anchorage, AK, USA, September 14-17, 2025*, pp. 582–587. IEEE, 2025. doi: 10.1109/ICIP55913.2025.11084559. URL https://doi.org/10.1109/ICIP55913.2025.11084559.

Sohambhai Joita and Tatyana Ivanovska. GMAR++: efficient gradient enhanced attention rollout. In Antonino Furnari and Petia Radeva (eds.), *Proceedings of the 21st International Conference on Computer Vision Theory and Applications, VISAPP 2026, Marbella, Spain, March 9-11, 2026, Volume 1*, pp. 13–18. SCITEPRESS, 2026. doi: 10.5220/0014215300004084. URL https://doi.org/10.5220/0014215300004084.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2020. URL https://arxiv.org/abs/1912.11370.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2017. URL https://arxiv.org/abs/1611.01236.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples, 2021. URL https://arxiv.org/abs/2104.02610.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos, 2024.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Ayon Sen, Xiaojin Zhu, Erin Marshall, and Robert Nowak. Popular imperceptibility measures in visual adversarial attacks are far from human perception. In *Decision and Game Theory for Security (GameSec)*, volume 12513 of *Lecture Notes in Computer Science*, 2020.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Jesse Vig. A multiscale visualization of attention in the transformer model. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://aclanthology.org/P19-3007/.

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers, 2022. URL https://arxiv.org/abs/2109.04176.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1002. URL https://doi.org/10.18653/v1/D19-1002.

Ross Wightman. PyTorch image models, 2019.

Jialin Wu, Kaikai Pan, Yanjiao Chen, Jiangyi Deng, Shengyuan Pang, and Wenyuan Xu. Protego: detecting adversarial examples for vision transformers via intrinsic capabilities. In *IEEE International Conference on Metaverse Computing, Networking, and Applications, MetaCom*, pp. 105–112. IEEE, 2024a. doi: 10.1109/MetaCom62920.2024.00029.

Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 10926–10935. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.01039. URL https://doi.org/10.1109/CVPR52733.2024.01039.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. Attentionviz: A global view of transformer attention, 2023. URL https://arxiv.org/abs/2305.03210.

Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization, 2023. URL https://arxiv.org/abs/2303.15754.

A Choosing the perturbation budget

A benchmark that compares attacks at one budget has to say why that budget. We ran a probe before the sweeps: four models, ViT-S/16, ViT-S/32, ViT-B/16 and ViT-B/32, crossed with the five attacks and seven budgets from $2/255$ to $16/255$, on 200 images, for $28{,}000$ runs. It asked two questions of the data and took a third from the literature.

Where the iterative attacks stop gaining. Success rises steeply to $4/255$ and then flattens (Fig. 5). Doubling the budget to $8/255$ buys every attack but TGR between 1.2 and 3.0 percentage points of attack success rate, against the 2.7 to 15.5 points they gain going from $2/255$ to $4/255$ (Table 4). A budget inside the flat region is one where a further relaxation would not change which attack looks strongest, which is the property a comparison needs.

Where the perturbation stops being quiet. At $4/255$ the median structural similarity between the clean and attacked image stays above 0.92 for every attack, from 0.930 under FGSM to 0.986 under SAGA. At $8/255$ FGSM falls to 0.791 and TGR to 0.824. The single-step attack degrades the image fastest because it spends its whole budget in one direction.

What the perceptual measures can and cannot carry. We report structural similarity as a description of each attack's perceptual profile and not as the budget itself, following Sen et al. (2020), who measure how far each popular imperceptibility measure is from human judgement and find ℓ_∞ the closest of those tested and structural similarity the furthest by three orders of magnitude. There is no citable threshold at which structural similarity becomes perceptible, so matching attacks on it would mean matching them on the worst available ruler; matching them on the best one collapses back to a fixed ℓ_∞ budget. That is the ground on which $\varepsilon = 4/255$ sits, and it is also the budget RobustBench (Croce et al., 2021) standardises for ℓ_∞ on ImageNet.

A criterion we retired. An earlier version of this argument also compared how far apart the five attacks sit at each budget, on the reasoning that a budget which spreads them apart is more informative. It does spread them more at $4/255$ than at $8/255$, 0.329 against 0.198 in attack success rate. We do not rely on it, for two reasons. Most of that spread is TGR sitting below the other four

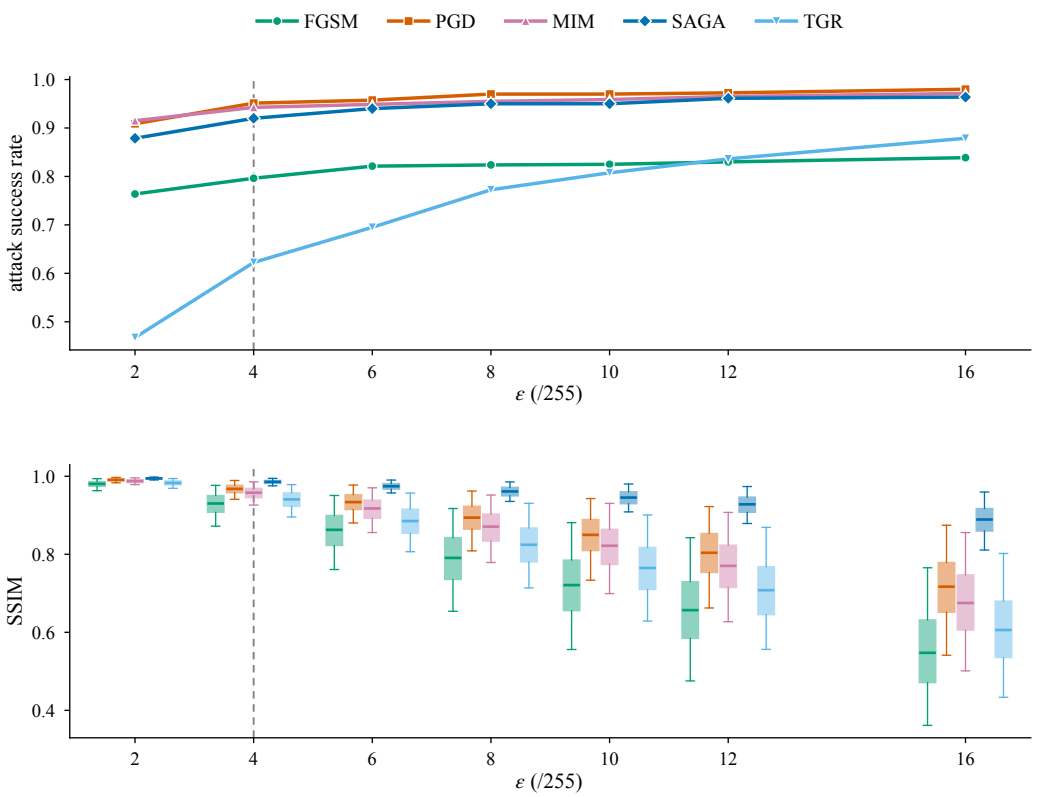


Figure 5: The probe behind $\varepsilon = 4/255$: seven budgets, four models (ViT-S/16, ViT-S/32, ViT-B/16 and ViT-B/32), 200 images per cell; the dashed rule marks the chosen budget. *Above*, attack success rate, averaged over the four models. *Below*, the structural similarity of all 800 images, as median, interquartile range and whiskers at the fifth and ninety-fifth percentiles, the five boxes of each budget in the order of the legend. Structural similarity describes each attack's perceptual profile; it is not the quantity being bounded.

rather than the five standing apart: dropping TGR leaves 0.155 of the 0.329, so the criterion would credit the budget for one attack's failure to converge. And the same spread runs from 0.255 to 0.410 depending only on which models enter the average. It is a property of the attacks and the models, not of the budget.

The budget holds on the larger models. The two Large models were not in the probe, so we repeated it on them, $14{,}000$ runs on the same images and grid. Compared against the model of the same patch size rather than against the average over all six, which mixes two regimes, ViT-L/16 and ViT-B/16 gain within a point of each other from $4/255$ to $8/255$, $+3.0$ against $+2.0$ under PGD and identically, $+2.5$ and $+5.0$, under MIM and SAGA. Median structural similarity at $4/255$ runs from 0.935 to 0.988 on the two. The Large models sit in the same regime as the rest.

TGR is not in the flat region, and this qualifies every TGR cell. TGR gains 15.0 points from $4/255$ to $8/255$ on the probe's four models, and between 11.0 and 17.5 across all six taken one at a time, against -1.0 to $+5.0$ for the other four attacks over the same twenty-four cells. Its curve is still climbing where theirs have converged. Fixing the budget is still the comparison we want, since it holds the perturbation allowance equal, but a reader should take every TGR result in this paper as an attack measured before it has finished gaining while the others are measured after. TGR was designed to transfer to a black-box target and is run white-box here, and the two facts compound.

The contrast across budgets. The benchmark reads the contrast at one budget. To see whether the dependence on the operator is a property of that choice, we reran ViT-B/16 and ViT-B/32 on the first

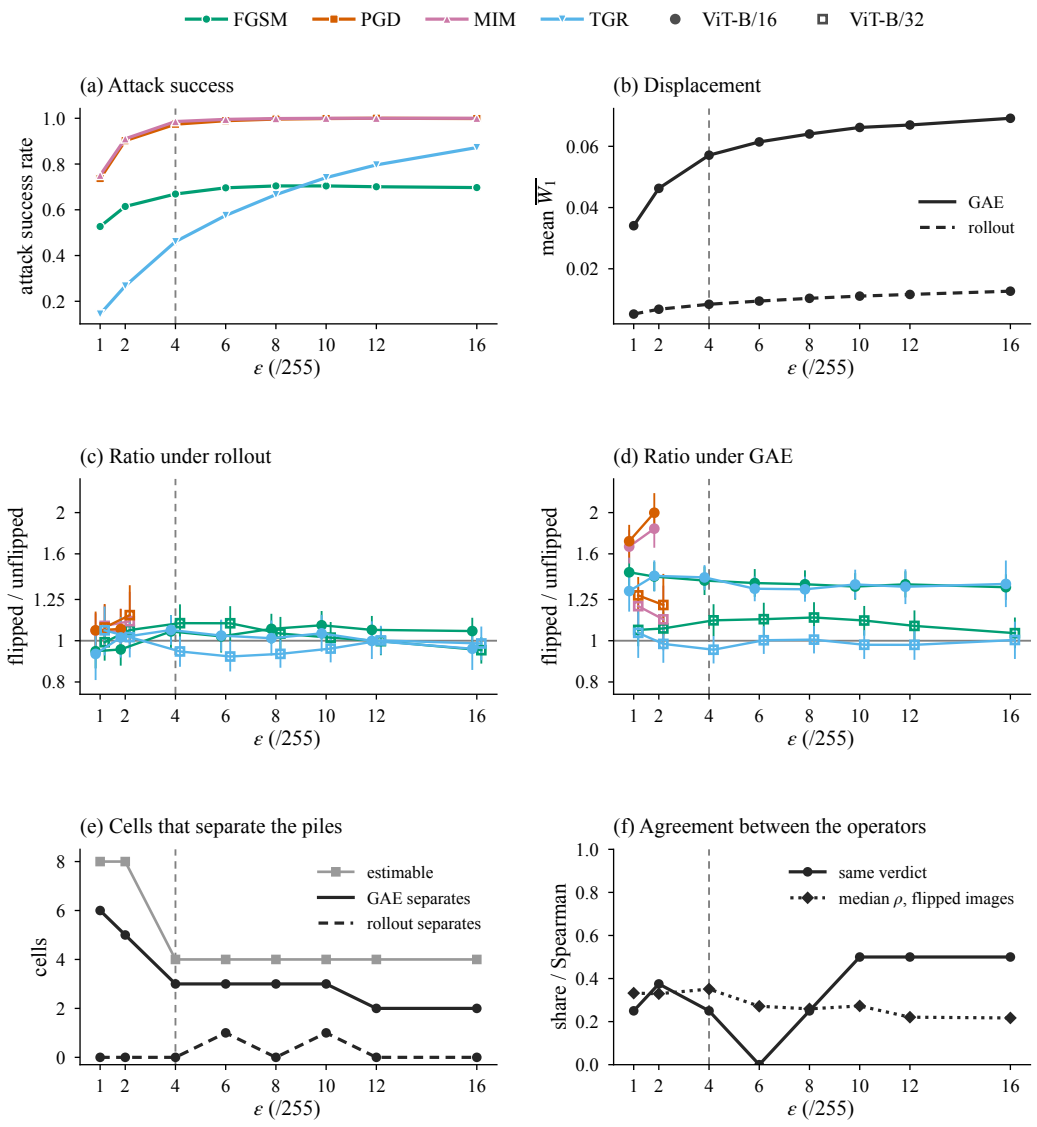


Figure 6: The contrast across budgets: ViT-B/16 and ViT-B/32 on the first 500 benchmark images, under FGSM, PGD, MIM and TGR and both operators; the dashed rule marks the benchmark's budget. (a) Attack success rate, averaged over the two models. (b) Mean $\overline{W_1}$ under each operator. (c, d) The ratio of mean displacements between flipped and unflipped images, with its 95% bootstrap interval, in each cell with at least 20 unflipped images; PGD and MIM appear only at $1/255$ and $2/255$, because beyond that they leave fewer than 20 unflipped images per cell. (e) How many of those cells each operator separates. (f) The share of them where the operators reach the same verdict, and the median Spearman correlation between their rankings of the same flipped images.

500 benchmark images under FGSM, PGD, MIM and TGR, with both operators, at eight budgets from $1/255$ to $16/255$ (Fig. 6); at $4/255$ the rerun reproduces the benchmark image by image. At every budget GAE separates the piles in two to six of the estimable cells and rollout in at most one, the operators reach the same verdict in at most three, and their rankings of the same flipped images correlate at a median ρ between 0.22 and 0.35. Beyond $2/255$, PGD and MIM flip nearly every image: their cells keep at most nineteen unflipped images at $4/255$ and at most one from $10/255$ on, so they have no unflipped pile to compare against, and the estimable cells fall from eight to four. SAGA, whose auxiliary network makes it by far the costliest attack, is left out.

	ASR at $4/255$	ASR at $8/255$	Δ (points)	median SSIM at $4/255$
FGSM	0.796	0.824	+2.7	0.930
PGD	0.951	0.970	+1.9	0.968
MIM	0.942	0.955	+1.2	0.958
SAGA	0.920	0.950	+3.0	0.986
TGR	0.623	0.772	**+15.0**	0.941

Table 4: The probe behind $\varepsilon = 4/255$, averaged over ViT-S/16, ViT-S/32, ViT-B/16 and ViT-B/32, 200 images per cell: attack success rate at the chosen budget and at twice it, the gain between them, and the median structural similarity between clean and attacked image at the chosen budget.

B BENCHMARK DETAILS

Checkpoints. The checkpoints are the ones distributed through `timm` (Wightman, 2019): ViT-S/16, ViT-S/32, ViT-B/32 and ViT-L/16 from the AugReg release (Steiner et al., 2022), ViT-L/32 from the original release (Dosovitskiy et al., 2021), and ViT-B/16 a later re-fine-tuning of the AugReg ImageNet-21k weights that `timm` distributes as `augreg2`; all six are fine-tuned on ImageNet-1k, and all but ViT-S/16 are pretrained on ImageNet-21k first.

Attack hyperparameters. MIM and TGR use momentum decay 1.0. TGR scales its gradients by 0.25 on attention and on the MLP and by 0.75 on query, key and value. SAGA blends the attacked ViT's rollout-weighted gradient with the gradient of an auxiliary Big Transfer network, BiT-M-R152x4 (Kolesnikov et al., 2020), under the weights its authors report for ImageNet, 0.99 and 0.01. TGR, designed to transfer from a surrogate to a black-box target, runs here white-box against the model it attacks.

The ratio test. The interval sits on the ratio and not on the two means, since overlapping intervals on separate means are not a test of the difference between them. No ratio is pooled across architectures: pooling returns one number, weighted toward the cells contributing the most images, that can land on the other side of 1 from the cells inside it.

Variance decomposition. To ask how much of the displacement the image explains and how much the attack does, we decompose it within each model over the $1{,}000$ images and five attacks, $\mathrm{SS}_{\text{total}} = \mathrm{SS}_{\text{image}} + \mathrm{SS}_{\text{attack}} + \mathrm{SS}_{\text{residual}}$, and report each factor's share $\eta^2_f = \mathrm{SS}_f/\mathrm{SS}_{\text{total}}$, under GAE. A second decomposition adds the operator as a third factor over both readings of the same runs, with each operator's displacements divided by their own mean. Without that division an operator main effect would measure the difference in units and nothing else (Appendix H); with it, the operator shows itself only through its interactions. With one observation per cell, the highest interaction falls into the residual in both, and the shares are η^2, whose bias Appendix M quantifies.

C THE SAME MAPS READ AGAINST FOREGROUND MASKS

The body of this paper scores an attention map against the attack's outcome. The literature more often scores it against a ground-truth foreground mask, asking whether the map lands on the object (Chefer et al., 2021b; Hu et al., 2024), and Section 2 notes that difference. This appendix runs the second ruler over the same maps, for the reader who wants the two side by side. It carries no claim of the body.

Two quantities. *fg-IoU* is the single-class intersection over union between the map, thresholded at its own mean, and the mask; *mAP* ranks the map's values against the mask and needs no threshold. Both are reported as the change the attack produces, Δ = adversarial − clean, because the clean levels differ between the two operators and their levels are therefore not comparable: on GTSRB the clean fg-IoU is 0.358 under rollout and 0.460 under GAE. Only the changes may be set beside each other.

ImageNet. Of the $1{,}000$ ImageNet images, 95 carry the human foreground masks released by Guillaumin et al. (2014); after the clean-correct cut each cell keeps between 59 and 84 of them.

Averaged over the thirty cells, the attack moves fg-IoU by -0.004 under rollout and -0.024 under GAE, and the sign is not uniform: the per-cell change runs from -0.019 to $+0.012$ under rollout and from -0.188 to $+0.048$ under GAE. Nine of the thirty rollout cells and thirteen of the thirty GAE cells have an interval clear of zero (Table 5). Paired within image, the difference between the two operators separates in eight of the thirty cells, six of them with GAE losing more overlap.

GTSRB. On the case study's checkpoint we annotated 43 signs with SAM 2 (Ravi et al., 2024) and kept the 41 the model classifies correctly without attack. Here the two operators do not merely differ in size. Under rollout the overlap with the sign *rises* after the attack in all four cells, from $+0.028$ to $+0.072$, every interval clear of zero; under GAE no cell separates at all. The mAP readings follow the same pattern.

What this does not say. It does not say that attention leaves the object. The mean change on ImageNet is a hundredth of an fg-IoU point under rollout, the direction reverses between cells, and on GTSRB the change is positive. What survives is comparative and narrow: the two operators disagree about what the attack does to foreground overlap, in the same direction of disagreement the body reports for displacement, and a reader who scores maps against masks would reach different conclusions from a reader who scores them against the attack's outcome depending only on which aggregation was used.

One exclusion is worth naming. We also computed a foreground entropy ratio, but an offline recomputation of it from the stored maps does not reproduce the sweep's own values to the precision fg-IoU and mAP do, so it is not reported here.

D How the three measures behave

Section 3 reads the change from R to R' along three measures, and this appendix shows what each one is sensitive to. The maps in Fig. 7 are synthetic rather than model output: each clean and attacked pair is constructed to isolate one kind of change. What holds by construction is where the zeros fall. Four of the five pairs leave the multiset of map values unchanged, so $\overline{\Delta H}$ is exactly zero in all four, although $\overline{W_1}$ runs from 0.054 to 0.530 across them. Two pairs isolate the distinctions of Section 3: the adjacent shift has the JSD of transport with a tenth of its $\overline{W_1}$, and the small transport has the $\overline{W_1}$ of diffusion with none of its entropy change. We solve the transport behind $\overline{W_1}$ exactly, by network simplex in POT (Flamary et al., 2021), not by entropic regularisation, so no solver parameter enters its value.

The three definitions in full. Dividing $\overline{W_1}$ by D, the largest ground cost the grid admits, bounds it in $[0, 1]$. JSD treats patches as unordered categories: moving mass to a nearby patch and moving it across the image count the same. $\overline{\Delta H}$ compares a number taken from each map rather than the maps themselves, so it can take small values even for very different distributions; its divisor $\ln G^2$ is the largest entropy a G^2-patch grid admits.

Why it is worth showing. Because the three are read together in the body and disagree there. Section 6 reports a split by capacity that holds under $\overline{W_1}$ and does not replicate under JSD or $\overline{\Delta H}$. Entropy change is a weak instrument for it: its cell means stay below 0.0014 in magnitude under rollout, and under GAE reach 0.0353 but change direction with the model. A reader who takes the three as interchangeable measures of the same thing will read those disagreements as noise. They are not, and the five pairs tell the three apart pairwise: entropy change is blind to what the other two report in every row but diffusion, and the last row reverses the relation between the remaining two, which the first four leave in one order.

E Every estimable cell under both operators

We plot every estimable cell of Table 2 against itself under the two operators, with only the operator differing between the two coordinates (Fig. 8).

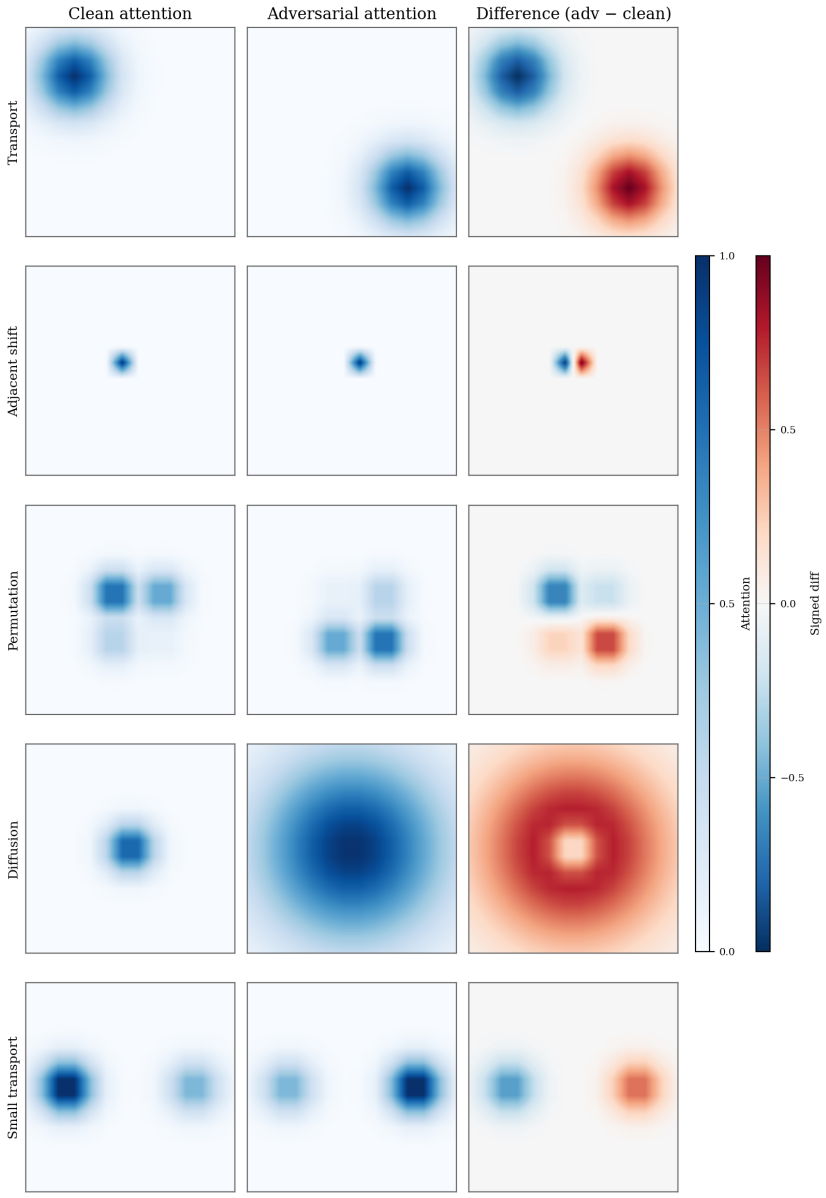


Figure 7: Five synthetic pairs on a $G = 14$ grid, read as $\overline{W_1}$, JSD, $\overline{\Delta H}$. *Transport*, a peak moved across the grid: 0.530, 0.996, 0.000. *Adjacent shift*, a peak about one patch wide moved to the neighbouring patch: 0.054, 0.964, -0.000. *Permutation*, the same spatial cluster with its amplitudes permuted: 0.104, 0.271, -0.000. *Diffusion*, a peak spreading at a fixed centre: 0.163, 0.566, $+0.432$. *Small transport*, thirty-eight percent of the mass moved between two distant blobs: 0.163, 0.104, -0.000. The rows share one colour scale. The measures are in the normalisation of Section 3: their zeros compare across measures, their levels do not.

Table 5: Change in foreground overlap after the attack, Δ fg-IoU = adversarial − clean, per model and attack, under both aggregation operators. fg-IoU is the single-class intersection over union between the map, thresholded at its mean, and the ground-truth foreground mask, in the convention of Chefer et al. (2021b). Restricted to images the model classifies correctly without attack, which is why n falls below the number of masked images and varies by cell. Confidence intervals from a 10^4-resample bootstrap on the mean of the paired difference; **bold** marks an interval that excludes zero. Levels are not comparable across operators, only changes: the clean fg-IoU differs between them.

		n	Δ fg-IoU (rollout)	Δ fg-IoU (GAE)
ImageNet (95 masked images; $n = 59 - -84$ per cell after the clean-correct cut)				
ViT-B/16	FGSM	73	**-0.014**	**-0.085**
ViT-B/16	MIM	73	**-0.018**	**-0.188**
ViT-B/16	PGD	73	-0.012	**-0.178**
ViT-B/16	SAGA	73	**-0.019**	**-0.184**
ViT-B/16	TGR	73	+0.005	**-0.082**
ViT-B/32	FGSM	64	+0.004	**-0.022**
ViT-B/32	MIM	64	-0.008	**-0.023**
ViT-B/32	PGD	64	-0.007	-0.017
ViT-B/32	SAGA	64	-0.007	**-0.025**
ViT-B/32	TGR	64	**+0.012**	+0.002
ViT-L/16	FGSM	84	**+0.005**	+0.012
ViT-L/16	MIM	84	-0.001	+0.010
ViT-L/16	PGD	84	-0.003	+0.009
ViT-L/16	SAGA	84	**-0.004**	+0.001
ViT-L/16	TGR	84	**+0.011**	**+0.020**
ViT-L/32	FGSM	64	-0.015	**+0.048**
ViT-L/32	MIM	64	**-0.017**	+0.011
ViT-L/32	PGD	64	-0.001	+0.008
ViT-L/32	SAGA	64	+0.001	+0.004
ViT-L/32	TGR	64	-0.017	**+0.038**
ViT-S/16	FGSM	65	+0.007	+0.012
ViT-S/16	MIM	65	-0.003	-0.027
ViT-S/16	PGD	65	-0.008	-0.035
ViT-S/16	SAGA	65	-0.007	**-0.038**
ViT-S/16	TGR	65	+0.005	**+0.031**
ViT-S/32	FGSM	59	**+0.011**	+0.006
ViT-S/32	MIM	59	+0.001	+0.007
ViT-S/32	PGD	59	-0.011	-0.010
ViT-S/32	SAGA	59	-0.009	-0.009
ViT-S/32	TGR	59	+0.002	-0.006
GTSRB (43 masked images; $n = 41$ after the clean-correct cut)				
GTSRB	FGSM	41	**+0.068**	+0.002
GTSRB	MIM	41	**+0.031**	-0.003
GTSRB	PGD	41	**+0.028**	+0.007
GTSRB	TGR	41	**+0.072**	+0.006

F FURTHER RESULTS

What rollout separates. Of the three cells rollout separates and GAE leaves flat, only one is a SAGA cell: the others are ViT-S/32 under FGSM and ViT-B/32 under TGR. Across the sixteen cells where rollout reports no contrast, the median resolution of 0.076 means that in half of them a ratio 8% away from 1 would have excluded 1.

Pooling hides the capacity split. Under FGSM the pooled ratio is 1.141, interval 1.106 to 1.178, against 0.704 for ViT-S/16 alone, interval 0.652 to 0.761, both clear of 1 and on opposite sides of it.

What the contrast does not license, in full. We attach further qualifications to the eight against twenty. Class identity is not established as the active ingredient: drawing the anchor at random rather than from the label leaves the contrast standing in ViT-L/32 under FGSM, at 1.175 against 1.137 for

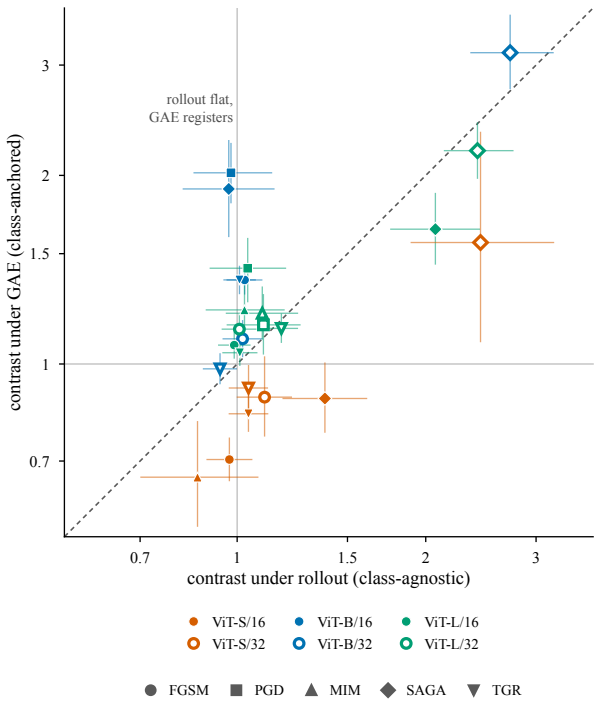


Figure 8: The twenty-four estimable cells of Table 2, each placed by its ratio under rollout and under GAE anchored on the ground-truth class, with bootstrap intervals on both axes. Colour gives capacity, fill gives patch size and shape gives the attack. The axes are logarithmic, so a ratio and its reciprocal sit equally far from 1, and the distance from the dashed diagonal is the disagreement between the operators. A point whose horizontal interval contains 1 and whose vertical interval does not is a contrast only GAE registers.

the true class, though under SAGA in the same model the random anchor moves the ratio from 2.190 to 2.418, reproducing rollout's 2.419 as a positive control should; and whether the anchor matters at all follows patch size rather than capacity: swapping it changes the reading in six of seven patch-16 cells against two of six patch-32 cells, with median absolute change 0.290 against 0.107. Neither pattern is explained here. Part of what GAE registers is its anchor's own probability: compared only among images that end with the same probability on the true class, the five cells both operators separate keep a median 64% of their log-ratio under GAE and 92% under rollout, though the Small cells below 1 stay below it (Appendix J). Normalisation, one of the three differences between the readings (Section 5), accounts for part of the gap: rerunning GAE with rollout's scaling closes 20% of the difference between their mean displacements in ViT-B/16 and 37% in ViT-L/32 (Appendix H).

The operator's share, model by model. Over both readings at once, the operator's interaction with the image accounts for 17% of the variance, between 14% and 19% across the six models. Orderings are correspondingly fragile: the two halves of Table 2 agree on only six of the twenty-four estimable cells, and each operator separates cells the other leaves flat.

Trajectories in detail. Under PGD the random start alone covers 12% to 35% of the terminal displacement, and since the five steps after the fifth still add 9% to 69%, the attack has not run out of room when it stops. The excursions of single trajectories are not rounding. MIM and TGR are the quieter pair, non-decreasing in up to 51.5% and 86.4% of images with median largest drops below 17%. The number of iterating cells that separate is six at the first step and seven at the tenth, and six cells separate at no step at all. In the four SAGA cells that separate from the first step, the ratio gains 46% to 105% by the tenth, while the two TGR cells that separate move 2% and 16%: where the piles are apart from the first step, they end further apart. For PGD the sum behind path efficiency begins with the jump from R_0 to $\tilde{R}_0$. The low efficiency is what the jagged individual curves are

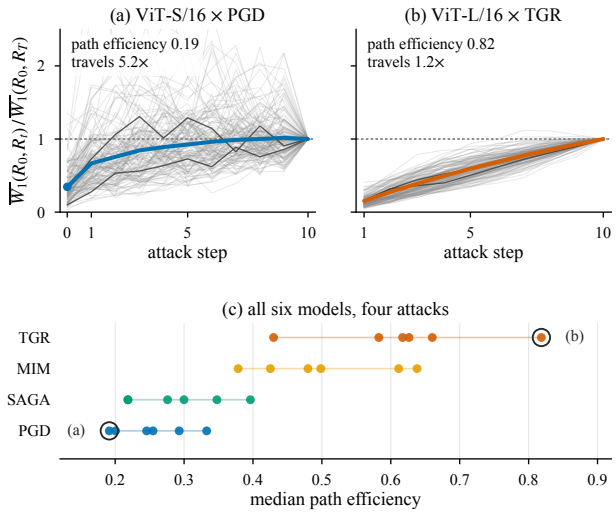


Figure 9: Each grey curve is one image's $\overline{W_1}(R_0, R_t)$ divided by the displacement it ends at, so every curve ends at 1 and one above 1 went further mid-attack than at the end; the coloured curve is the mean over the whole cell, under rollout on clean-correct images, with 140 curves drawn per cell. (a) and (b) are the two extremes of median path efficiency over the six models and four iterating attacks, and (c) places them among the rest. The axis is clipped at 2.5, which cuts 2.4% of the curves in (a); there, $t = 0$ is PGD's random start $\tilde{R}_0$.

made of, and it recovers by path length the ordering that displacement correlations give. Under TGR the direction is the opposite of SAGA's and the effect is smaller, holding in two of six models with ratios from 1.085 to 1.119. PGD and MIM mostly answer neither way: three of their six cells have no estimable unflipped pile at all and the rest flip between 90% and 99% of the time, which leaves little to compare. The one exception is PGD on ViT-L/16, where the flipped images travel the less efficient path, at 0.893 with an interval of 0.804 to 0.998.

G One step at a time

An attack that iterates leaves an intermediate map R_t at every step t, from the clean map R_0 to the map R_T of the image it finally produces. *Path efficiency* is $\overline{W_1}(R_0, R_T) / \sum_{t=1}^{T} \overline{W_1}(R_{t-1}, R_t)$: how far the path ends from where it started, over how far it travelled to get there. It is bounded by 1 and reaches it only for a path that never changes direction.

Four of the five attacks iterate, and we read every trajectory in this appendix under rollout at every step, the walk ViTViz exposes on a single image (Section 4).

The average trajectory is monotone; under PGD and SAGA almost no single one is. Averaged over images, the curve behaves as one would expect: the first step already reaches between 0.15 and 0.70 of the terminal displacement, and the five steps after the fifth still add 9% to 69%. Individual images depart from that average (Fig. 9). Under PGD the trajectory is non-decreasing in at most 3.7% of images in any of the six models, and under SAGA in at most 8.8%, with median largest drops of 15% to 36% and of 11% to 33% of the image's own terminal displacement. Under PGD and SAGA attention falls back toward the clean map partway through the attack, and the mean curve shows none of that.

The separation is there from the first step or not at all. Splitting each step by the outcome the attack eventually reached, the separation of Table 2 does not arrive partway: of the eighteen estimable cells among the iterating attacks, the number whose interval excludes 1 stays between six and eight at every one of the ten steps. Four of the seven that separate at the end, all of them SAGA,

already do so at the first step, and what grows is the distance between the piles, from 1.32–1.68 at the first step to 2.07–2.73 at the tenth.

A wandering path is not a shorter one. Distance to the clean map is blind to lateral movement, so we also read the path efficiency defined above. Its median runs from 0.191 to 0.819 across cells, ordering PGD $<$ SAGA $<$ MIM $<$ TGR in five of the six models (Fig. 9(c)), so attention travels between 1.22 and 5.22 times the distance it finally sits from where it started. Under SAGA the images the attack flipped travel the less efficient path, in five of the six models, with ratios from 0.631 to 0.890 and intervals clear of one; under TGR the direction is the opposite and the effect smaller (Appendix F).

H What the map normalisation costs

The two operators do not put their maps on the same scale. GAE subtracts the map's minimum and divides by its range; rollout only divides by the maximum. Normalizing each map to sum to 1, as Section 3 does, cancels a division by a constant but not the subtraction of a minimum, so the choice survives into $\overline{W_1}$, JSD and $\overline{\Delta H}$ under GAE and washes out under rollout. It is not recoverable after the fact either: a map stored after min–max scaling has lost the minimum and the range that produced it, so quantifying the choice means running the benchmark again.

The twin run. We rerun GAE with rollout's division by the maximum on ViT-B/16 and ViT-L/32 under FGSM and TGR, on the same one thousand images, with the same ground-truth anchor and the same operator. The scaling is the only thing that changes, and both attacks are deterministic, so the two runs pair image by image.

What it moves, and what it does not. Levels move a great deal: the mean displacement falls by 15.0% and 15.6% in ViT-B/16 and by 31.0% and 31.4% in ViT-L/32, with paired intervals far from zero. Ratios barely move: the flipped-to-unflipped ratio changes by 1.5% to 3.3%, downward in all four cells, and all four separate their piles under both scalings, so no verdict turns on the choice. The two scalings also rank images almost identically, at a per-image Spearman correlation of 0.951 to 0.984. This is what licenses the comparison of operators by ratio in Section 6 and forbids it by level.

How much of the gap it explains. Rollout and GAE differ in operator, anchor and map normalisation at once, and the twin run isolates the third from the other two. Write μ_{GAE}, μ_{twin} and μ_{roll} for a cell's mean $\overline{W_1}$ under GAE as published, under the twin run, and under rollout. The normalisation's share is $(\mu_{\text{GAE}} - \mu_{\text{twin}})/(\mu_{\text{GAE}} - \mu_{\text{roll}})$: 20% in ViT-B/16 and 37% in ViT-L/32, with the two attacks of a model agreeing to within half a point. What is left is not measured but subtracted, and charging it to the operator and the anchor jointly presumes the three differences compose along that one axis. The share is larger in the model with the coarser patch grid, which is the ordering a subtraction of the minimum would predict, since a shorter map vector gives the minimum more weight.

What the twin run does not cover. It runs on four cells of the thirty, and none of them is a SAGA cell, which is where the two operators disagree most (Fig. 8). The 1.5% to 3.3% band is therefore measured away from the hardest case, and we do not extrapolate it there.

I How far the anchor moves the map

The anchoring class belongs to the reading, so a class-anchored map is one of a thousand the same model could have produced for the same image. This appendix puts a number on that choice by comparing it against the attacks themselves. It rests on one model and one operator, which is why it is reported here rather than in Section 6; the random-anchor control of Section 6 carries the same point across thirteen cells and six models.

The anchor moves the map about as far as an attack. On the 160 images that ViT-B/16 misclassifies without any attack, where the true and the predicted class therefore differ, moving the

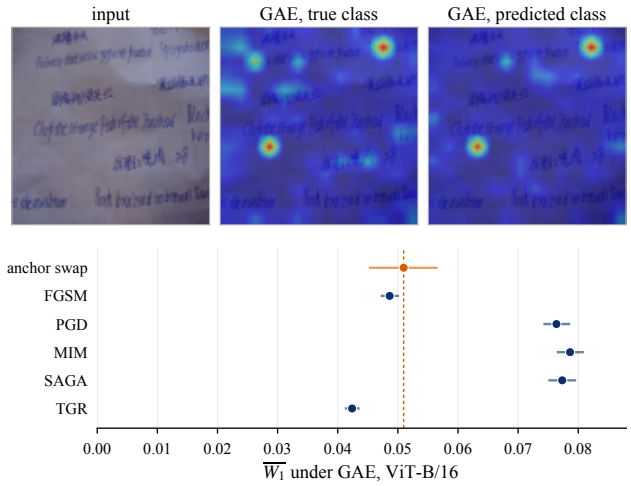


Figure 10: Top: one image read under GAE anchored on the true class and on the predicted class, the median of the 160 images ViT-B/16 misclassifies without attack. Bottom: the displacement those 160 undergo when the anchor moves, against the displacement all one thousand images undergo under each attack at $\varepsilon = 4/255$, on the same model and operator, with bootstrap intervals on the mean; the dashed line carries the anchor value across. The two quantities come from different populations, and the anchor value is a lower bound.

anchor defined in Section 3 from one to the other displaces the map by $\overline{W_1} = 0.051$, with a bootstrap interval of 0.045 to 0.056. An attack at $\varepsilon = 4/255$ on the same model displaces it by 0.065 on average, from 0.042 under TGR to 0.079 under MIM. Asking the same model about a different class therefore moves its attention map about as far as attacking the image does, and against TGR it moves it further, as Fig. 10 shows. That comparison is a lower bound on the anchor: it is measured with no perturbation, and on the images the model already fails, standing in for the population an attack goes on to flip. Dropping the seventy-four training-split images of Appendix M leaves it at 0.051.

J How much of the contrast is the anchor's own probability

GAE is anchored on the true class, and for an image the model classifies correctly that is also the class an untargeted attack pushes down. The attacks that flip an image take most of that class's probability and the ones that do not take little of it: across the twenty-four cells of Table 2, the median loss runs from 0.597 to 0.990 in the flipped pile and from 0.001 to 0.654 in the unflipped one. If the class-anchored contrast were only that loss read through attention, it would vanish once the two piles are compared among images in the same situation on the true class.

Within each cell we split the images into five bands by quantile of a covariate and take the ratio of mean displacements band by band, weighting each band by $n_{\mathrm{f}} n_{\mathrm{u}}/(n_{\mathrm{f}} + n_{\mathrm{u}})$, with n_{f} and n_{u} its flipped and unflipped images, and dropping any band with fewer than three of either. The interval is the bootstrap of Section 5. The covariate is the probability the attack takes from the true class (*same loss*) or the probability it leaves on it (*same final*). The second comparison is the stricter and the less powered: flipped and unflipped images share a band only near the decision boundary, which holds about a fifth of each cell.

Table 6 gives both. At equal loss the contrast stands: GAE still separates sixteen of its twenty cells, every ratio stays on its side of 1, and the median cell keeps its whole log-ratio, while rollout separates five of its eight. At equal final probability GAE separates seven and rollout four, and the two operators part: on the five cells both separate, GAE keeps a median 64% of its log-ratio and rollout 92%. Part of what GAE registers is therefore its anchor's own probability. The counts also fall because the comparison is thinner, and the cells only one operator separates fall furthest under either. The direction does not follow the anchor: the four Small cells below 1 stay below it under

Table 6: Ratio of mean $\overline{W_1}$ displacement between flipped and unflipped images, as in Table 2 (*all*), and with the two piles compared only within bands of the same situation on the true class: the probability the attack takes from it (*same loss*) or the probability it keeps (*same final*). Five bands per cell by quantile; a band with fewer than three images in either pile drops out. **Bold** marks a 95% bootstrap interval that excludes 1. The 24 cells with at least 20 unflipped images.

		Attention rollout			GAE, ground-truth anchor		
Model	Attack	all	same loss	same final	all	same loss	same final
ViT-S/16	FGSM	0.972	0.978	1.066	**0.704**	**0.721**	**0.737**
	MIM	0.864	0.833	0.844	**0.660**	**0.657**	**0.694**
	SAGA	**1.381**	**1.362**	**1.296**	0.881	0.890	0.876
	TGR	1.042	1.072	1.085	**0.833**	**0.781**	0.927
ViT-S/32	FGSM	**1.106**	1.068	1.000	0.885	**0.843**	0.842
	SAGA	**2.446**	**2.117**	**2.761**	**1.563**	**1.487**	**1.790**
	TGR	1.043	0.995	0.947	**0.916**	0.908	0.933
ViT-B/16	FGSM	1.028	0.955	1.028	**1.362**	**1.418**	0.991
	PGD	0.978	0.925	1.093	**2.019**	**2.089**	**1.193**
	SAGA	0.970	0.897	1.125	**1.903**	**1.991**	1.177
	TGR	1.009	0.939	0.923	**1.362**	**1.097**	0.947
ViT-B/32	FGSM	1.020	0.928	1.060	**1.097**	1.074	1.068
	SAGA	**2.727**	**2.396**	**2.523**	**3.140**	**3.286**	**2.480**
	TGR	**0.939**	0.951	0.986	0.982	0.935	0.902
ViT-L/16	FGSM	0.989	0.958	0.989	**1.072**	**1.111**	0.977
	PGD	1.040	1.005	1.006	**1.422**	**1.395**	**1.177**
	MIM	1.027	1.032	0.942	**1.220**	**1.237**	1.045
	SAGA	**2.072**	**2.225**	**1.966**	**1.641**	**1.750**	**1.375**
	TGR	1.010	0.910	0.877	1.042	0.989	1.034
ViT-L/32	FGSM	1.008	1.077	1.215	**1.137**	**1.143**	1.033
	PGD	1.101	1.024	1.113	**1.155**	**1.201**	1.065
	MIM	1.097	1.038	1.046	**1.205**	**1.315**	1.071
	SAGA	**2.419**	1.000	1.073	**2.190**	1.244	0.943
	TGR	**1.177**	**1.164**	1.107	**1.140**	1.017	1.072

both controls, and ViT-S/16 under FGSM and MIM still separates at equal final probability, at 0.737 and 0.694.

The random-anchor control of Section 6 reads the same images and piles with GAE anchored on a class drawn at random, which takes the true class out of the anchor (Table 7). It runs on thirteen of the twenty cells that separate under the ground-truth anchor, paired by capacity across patch size (ViT-B/16 with ViT-B/32, ViT-L/16 with ViT-L/32) and including SAGA cells as a positive control; PGD is left out throughout, its unflipped pile being too small, and cells with fewer than thirty unflipped images are dropped, a stricter floor than the twenty that makes a cell estimable. In the five cells where both anchors separate the piles above 1, at equal final probability of the true class the random anchor keeps a median 72% of its log-ratio and the true anchor 53%, and the random one keeps more in four; in the fifth, ViT-L/32 under SAGA, neither keeps any, and rollout keeps 8%. The true anchor thus accounts for part of the shrinkage, not all of it. Under TGR on the two Small models the anchors point opposite ways, and no interval excludes 1 at equal final probability.

Table 7: GAE under its ground-truth anchor and under a class drawn at random, on the thirteen cells where both were run. The random anchor reads the same images, attacks and piles; the *same final* bands are the true class's final probability in both columns, as in Table 6. **Bold** marks a 95% bootstrap interval that excludes 1.

		Ground-truth anchor		Random anchor	
Model	Attack	all	same final	all	same final
ViT-S/16	FGSM	**0.704**	**0.737**	0.994	1.002
	TGR	**0.833**	0.927	**1.125**	0.977
ViT-S/32	TGR	**0.916**	0.933	**1.077**	0.934
ViT-B/16	FGSM	**1.362**	0.991	0.990	1.008
	TGR	**1.362**	0.947	1.004	0.957
ViT-B/32	FGSM	**1.097**	1.068	1.044	1.034
	SAGA	**3.140**	**2.480**	**2.813**	**2.510**
ViT-L/16	FGSM	**1.072**	0.977	0.994	0.954
	MIM	**1.220**	1.045	0.994	0.916
	SAGA	**1.641**	**1.375**	**1.519**	**1.406**
ViT-L/32	FGSM	**1.137**	1.033	**1.175**	1.123
	SAGA	**2.190**	0.943	**2.418**	0.871
	TGR	**1.140**	1.072	**1.150**	1.101

K THE TRAFFIC-SIGN CHECKPOINT

The case study of Section 6.3 runs on `bazyl/gtsrb-model`, a checkpoint on the Hugging Face Hub: a ViT-B/16 pretrained on ImageNet-21k and fine-tuned on the German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al., 2012) to its 43 classes. Ten images per class from the test split give 430 images, of which the model classifies 403 correctly without attack. The attacks run at the same $\varepsilon = 4/255$ as the benchmark. SAGA is left out because its BiT auxiliary predicts ImageNet-1k classes, so a GTSRB label would index the wrong space.

The inversion is not confined to one measure. Under the absolute entropy change the two operators give 1.855 against 0.659 on the same cell, the same reversal of direction that $\overline{W_1}$ reports in Table 3.

L THE TOOL ON ITS OWN GALLERY

Every number in the body comes from the benchmark's stored runs. This appendix runs ViTViz itself, as a user would, on the ten images of its gallery: the six models, four attacks at $\varepsilon = 4/255$, ten steps and a step size of $\varepsilon/10$, with GAE anchored on the label the gallery declares. SAGA, at about two minutes per run on CPU, is left out. Each run is placed at its $\overline{W_1}$ percentile within the benchmark cell and pile it falls into, once per operator, as the metric panel of Fig. 2 does. Of the 240 runs, 232 start from an image the model classifies correctly, and all of them receive a percentile under both operators.

Table 8 gives the result. The two readings place 84 of the 232 runs in opposite halves of their cells, and their percentiles correlate at 0.36 over the six models, from -0.15 in ViT-L/16 to 0.79 in ViT-S/32. The widest gap is ViT-L/16 under TGR on the clock image, which the attack does not flip: rollout places it at the 99th percentile and GAE at the 4th. The runs are on CPU, where iterative attacks differ from the benchmark's GPU runs, so each placement is approximate. The disagreement of Section 6 is thus not confined to the stored runs: a user of the tool meets it image by image.

To see how approximate, ViTViz ran fifty benchmark images per checkpoint that the model classifies correctly on CPU under the same four attacks, and we compared each run with the stored GPU run of the same image (Table 9). The clean prediction matched on every image, and the attack flipped or held it on both devices in at least 98% of runs. FGSM lands within 0.3 percentile points of the GPU run in nine runs out of ten, and TGR and MIM move a median of 4 points or less. PGD, whose random start the CPU draws differently, moves a median of 10 to 13, and a tenth of its runs move 37 or more.

Why the devices differ. Attacks that iterate do not reproduce across devices. On the same code, seed and inputs, over fifty benchmark images per checkpoint, FGSM's displacement on CPU stays within 0.3% of the stored GPU run in nine runs out of ten, while at ten steps it differs by a median of 1% under TGR, 4% to 5% under MIM and 15% to 21% under PGD, whose random start the CPU also draws differently (Table 9). The cause is $\operatorname{sign}(\cdot)$: a gradient component near zero flips under the 10^{-7} that a different summation order produces, which moves that pixel by $\pm\varepsilon$, and ten steps compound it. MIM carries no randomness and diverges anyway, so this is not a seeding question. Each device is itself deterministic, and the clean forward pass agrees to six decimals. What survives is the reading: on the three images of Fig. 4 the attacked class is the same on both devices, and the two operators still disagree about which image moved most.

M LIMITATIONS IN DETAIL

The limit is attack success, not sample size. PGD and MIM flip 91.5% to 100% of clean-correct predictions, leaving 0 to 68 of roughly 800 images unflipped, so 6 of their 12 cells fall below the floor of 20. FGSM and TGR leave 74 to 484 and stay estimable. More images would not help: ViT-S/32 keeps one unflipped image of 801 under MIM, so reaching 20 would take about $16{,}000$, and none under PGD. Nor does the floor itself decide the result. Raising it to 30 drops four cells of Table 2 (SAGA on ViT-B/16 and ViT-S/32, PGD on ViT-L/16, MIM on ViT-S/16), of which one

Table 8: The tool on its own gallery: ten images, four attacks at $\varepsilon = 4/255$ (SAGA left out for its cost on CPU), each run placed by ViTViz at its $\overline{W_1}$ percentile within the benchmark cell and pile it falls into, once per operator. *Placed runs* are the ones the model classifies correctly without attack and for which both operators have a reference pile; ρ is the Spearman correlation between the two percentiles; *opposite halves* counts runs one operator places above its median and the other below; the gap is in percentile points. Runs on CPU, whose iterative attacks differ from the benchmark's GPU runs, as the reproducibility statement reports, so each placement is approximate.

Model	placed runs	flipped	ρ	opposite halves	median gap
ViT-S/16	40	32	0.41	13	21
ViT-S/32	36	27	0.79	6	11
ViT-B/16	40	28	0.41	14	22
ViT-B/32	36	26	0.32	15	20
ViT-L/16	40	28	-0.15	22	37
ViT-L/32	40	28	0.35	14	19
all six	232	169	0.36	84	20

Table 9: The same benchmark images run by ViTViz on CPU against the benchmark's stored GPU runs: fifty images the model classifies correctly, per model, under four attacks at $\varepsilon = 4/255$, on the six checkpoints. *Same pile*: the attack flips or holds the prediction on both devices. The displacement change is relative to the GPU value. The percentile change is in points, within the cell and pile, over the runs whose pile agrees. Under PGD the CPU also draws the random start differently, so its row combines the device and the draw.

			$\overline{W_1}$ change	percentile change (points)	
Attack	Operator	same pile	median	median	90th percentile
FGSM	rollout	100.0%	0.1%	0.1	0.3
	GAE	100.0%	0.1%	0.0	0.3
PGD	rollout	98.3%	21.3%	13.4	41.3
	GAE	98.3%	14.7%	10.1	37.1
MIM	rollout	99.3%	5.2%	3.6	22.8
	GAE	99.3%	3.9%	2.7	16.1
TGR	rollout	99.7%	1.0%	0.7	4.8
	GAE	99.7%	1.2%	0.9	5.4

separates under rollout and all four under GAE. The counts fall from 8 and 20 of 24 to 7 and 16 of 20, and GAE still separates more than twice as many cells as rollout either way.

Seventy-four images the models have seen. Seventy-four of the thousand, all among the 95 from the full-ImageNet pool, lie in the ImageNet-1k training split the six checkpoints were fine-tuned on; the other 905, from the validation split, are held out by construction. They are not memorised: ViT-B/16 classifies them correctly 81.1% of the time against 84.2% on the rest, so they are harder rather than easier. When we drop them, all 24 cells stay estimable and the counts of Table 2 move from 8 and 20 to 6 and 21, so the gap between the two operators widens rather than closes. Three of the 48 verdicts those cells carry under the two operators change, each in a cell whose ratio sits within 0.11 of 1; the median cell moves by 0.008.

One budget for five attacks. The budget of $\varepsilon = 4/255$ is where every attack but TGR stops gaining attack success, where the perturbation still leaves median structural similarity above 0.92 for every attack, and where RobustBench (Croce et al., 2021) sets its ℓ_∞ budget for ImageNet (Appendix A). Holding the five attacks to it, we measure TGR before it has finished gaining and the other four after.

The evidence comes from one configuration. We pair rollout with GAE on 1,000 ImageNet images under untargeted white-box attacks, and leave the other operators of Section 2 untested.

Table 10: The floor of twenty unflipped images. *Top*: the cells counted at each floor, and how many each operator separates; each cell's interval does not depend on the floor. *Bottom*: the twelve cells whose unflipped pile holds at least 100 images, under each operator, with that pile cut to n images drawn at random, 200 times per n; the median over cells of how often a cell flat with its whole pile separates, of how often a cell that separates with it still does, and of the interval's half-width relative to the whole pile.

Floor	5	10	15	20	25	30	40	50
Estimable cells	27	26	25	24	20	20	18	18
Separated under rollout	10	10	9	8	7	7	7	7
Separated under GAE	23	22	21	20	16	16	14	14
Same verdict	8	8	7	6	5	5	5	5

Unflipped images kept, n	5	10	15	20	30	50
Flat cells that separate	19.5%	12.0%	9.5%	9.5%	6.5%	5.5%
Separating cells that still do	36.5%	36.0%	35.5%	45.0%	60.5%	76.5%
Relative half-width	5.1	4.0	3.4	3.0	2.5	2.0

Patch size travels with input resolution throughout and with pretraining data in ViT-S/16 (Section 5), so the patch-size regularity of Section 6 cannot be credited to patch size alone; we would have to cross patch size with resolution to test it directly. TGR, designed to transfer to a black-box target, runs here white-box, and its cells describe that regime. The change of domain in Section 6.3 leaves the operator pairing in place, and is itself one checkpoint at one seed.

The variance shares are upper bounds. The shares we report in Section 6 are η^2, which overstates what a fresh sample would show, and not evenly: the image enters with 999 degrees of freedom and the attack with four, so the ω^2 correction takes 8.4 to 9.5 points off the image and under 0.04 off the attack. The image still explains more than the attack, by 1.4 to 16.8 times rather than 1.8 to 19.7.

N THE FLOOR OF TWENTY UNFLIPPED IMAGES

A cell enters the counts of Section 6 only when its unflipped pile holds at least 20 images (Section 5). The floor decides which cells are counted, not how any of them reads, since each interval is computed on its own cell. Table 10 (top) moves it from 5 to 50: the estimable cells fall from 27 to 18, rollout separates between a third and two fifths of them and GAE between 78% and 85%, and the two reach the same verdict in five to eight.

To see why the floor sits at twenty, we took the twelve cells whose unflipped pile holds at least 100 images, under each operator, cut that pile to n images drawn at random, 200 times per n, and reran the test of Section 5 with the flipped pile whole (Table 10, bottom). Below twenty the interval runs too narrow: a cell that does not separate with its whole pile separates in 19.5% of draws at $n = 5$, four times the nominal 5%, and in 9.5% at $n = 20$. Twenty does not buy power. A cell that separates with its whole pile still does in 45% of draws at $n = 20$ and 76.5% at $n = 50$; the cells furthest from 1 separate in nearly every draw, and the smaller contrasts are the ones lost. This is why each cell that does not separate also reports its resolution (Section 5).