RANK TRAJECTORIES FOR ADVERSARIAL ROBUSTNESS VIA QUANTIZED TENSOR TRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Many defenses against adversarial examples rely on an explicit or implicit model of the natural-image distribution and are evaluated primarily by reconstruction quality. We pursue a different question: whether the multiscale tensor structure of an image provides an interpretable and computable basis for analyzing adversarial robustness. We study this question through quantized tensor-train (QTT) representations, whose modes encode spatial scales and whose bond ranks measure their interactions. Tensors satisfying nontrivial QTT-rank constraints form lower-dimensional algebraic sets, and natural images empirically concentrate near this low-rank structure. An adversarial perturbation, in contrast, comes from a search constrained only by a norm budget and a misclassification objective, and neither constraint confines it to these sets. To connect this geometry to model behavior, we introduce rank trajectories, which track predictions along progressively higher-rank reconstructions from a single decomposition sweep. Clean examples typically stabilize at moderate ranks, whereas adversarial examples often emerge later. The same geometry also has a constructive use inside randomized smoothing. A training-free filter suppresses the weak singular directions that carry most of the Gaussian smoothing noise. Experiments show that it consistently improves certified accuracy over plain smoothing in the high-noise regime. Together, these results establish QTT rank as a multiscale coordinate for analyzing adversarial robustness, and improves certified robustness under strong smoothing noise.

1 INTRODUCTION

Deep neural networks remain vulnerable to small, adversarially optimized input perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). Many proposed defenses have appeared effective against fixed attacks, only to lose their claimed robustness when the adversary adapts its optimization to the defense (Carlini & Wagner, 2017; Athalye et al., 2018). This recurring pattern motivates robustness analyses grounded in explicit and testable properties of the input, whose behavior under attack can be specified in advance and measured directly. We ask whether the multiscale tensor structure of an image provides such a property: whether it is interpretable and computable, and whether changes in this structure are reflected in model behavior across structural scales.

One common approach to exploiting image structure for robustness is through purification. Denoisers, autoencoders, diffusions, and tensor decompositions differ substantially in their mechanisms, but all transform the input toward some presumed structure of natural images before classification (Yoon et al., 2021; Nie et al., 2022). Their effectiveness therefore depends on whether the transformation removes the adversarial component while preserving the semantic content needed for prediction. We pursue a different use of image structure. Rather than asking whether a representation reconstructs or purifies the input well, we ask whether it exposes a structural coordinate along which the model's decision can be interrogated. Such a coordinate would provide a post-hoc and architecture-agnostic view of adversarial behavior, requiring neither retraining nor an auxiliary generative model, while revealing the structural scale at which a prediction changes.

The QTT representation (Khoromskij, 2011) makes this structural question concrete by organizing an image into a hierarchy of spatial scales. Under the bit-interleaved tensorization used here, each spatial QTT bond separates coarser modes from the finer modes that follow, and its rank measures the complexity of their interaction. The ranks across bonds therefore form a multiscale structural

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.

coordinate, revealing not only how much complexity an image contains, but also where across spatial scales that complexity appears. This coordinate has a precise geometric interpretation: imposing a nonmaximal rank at any bond restricts the tensor to a lower-dimensional algebraic set of Lebesgue measure zero (Proposition 1). Natural images need not lie exactly in such sets, but their spectral energy often concentrates at moderate ranks. In contrast, adversarial objectives constrain perturbation magnitude and classification loss without explicitly preserving this multiscale concentration. In our measurements, natural-image spectra are substantially more concentrated than those of Gaussian noise and of perturbations produced by attacks (Tables 1 and 9). The geometric result does not imply that every adversarial perturbation must be high rank. Rather, it motivates a testable hypothesis: if adversarial decisions rely more strongly on components that enter only at higher bond rank, then model predictions should evolve differently as these components are progressively restored.

From a single QTT decomposition, we construct a sequence of reconstructions that progressively restores higher-rank structure. Evaluating the same model at each stage produces a rank trajectory, which records how its predictive distribution evolves from a coarse approximation to the full input. Clean inputs typically settle on a stable prediction once their dominant structure has been restored at moderate rank. Successful adversarial examples often follow a different path: intermediate-rank reconstructions recover the true-class prediction, while the adversarial label emerges only as higher-rank components are reintroduced (Figure 2); the pattern persists across architectures, datasets, and image resolutions. We summarize this behavior using two training-free statistics calibrated exclusively on clean data. The stabilization rank records the last rank at which the predicted class changes, while the margin gain compares the full-input prediction margin with the largest margin attained along the truncated trajectory. Adaptive evaluation reveals an empirical complementarity between them: attacks targeting either statistic tend to amplify the other, allowing their combination to withstand both single-objective attacks. A balanced joint attack can nevertheless suppress both signals and evade the combined statistic, although at a reduced attack success rate. Rank trajectories therefore expose where along the structural hierarchy a model decision is formed or overturned, while adaptive attacks delineate the empirical limits of this structural signal.

We finally ask what the same geometry contributes to a formal guarantee. Randomized smoothing (Cohen et al., 2019) certifies an ℓ_2 ball around each input, but useful radii require large isotropic Gaussian noise, and that noise is itself a generic direction whose energy spreads across many QTT components. Soft-thresholding the singular values along the QTT bonds therefore removes much of the noise while retaining the concentrated image spectrum. The resulting QTT filter has no learned parameters, is nonexpansive by construction, and enters the base classifier as a measurable preprocessing map, so the standard certificate applies unchanged. With matched fine-tuning of the classifier, it raises certified accuracy at every reported radius in the high-noise regime on both datasets. Together, these results establish QTT rank as a single coordinate that both explains where adversarial decisions form along the structural hierarchy and improves certified robustness where smoothing noise is strongest.

2 Related work

Adversarial robustness. Since the discovery of adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015), the main defense families have been adversarial training, which retrains the model against its own attacks (Madry et al., 2018), and input purification, which maps the input toward a model of natural images before classification (Yoon et al., 2021; Nie et al., 2022; Lin et al., 2025). The recurring lesson of this literature is methodological: defenses that appear strong against fixed attacks routinely fail once the attack is adapted to the defense (Carlini & Wagner, 2017; Athalye et al., 2018; Tramer et al., 2020), and standardized adaptive suites are now the accepted evaluation (Croce & Hein, 2020). We adopt this lesson as a design constraint and study a structural property of the input whose behavior under attack is specified in advance. Lin et al. (2025) purify inputs by coarse-to-fine QTT truncation; its effectiveness is empirical, with no account of why truncation should separate natural images from perturbations or at which scale. We supply that account and draw a different conclusion from it: our use of QTT rests on its rank structure as a coordinate for reading model behavior, not on reconstruction quality.

Certified robustness via randomized smoothing. Randomized smoothing gives a tight ℓ_2 certificate for Gaussian-smoothed classifiers (Cohen et al., 2019), following earlier bounds from differen-

tial privacy (Lécuyer et al., 2019); later work prepends a trained denoiser (Salman et al., 2020) or a pretrained diffusion model (Carlini et al., 2023) so that off-the-shelf classifiers can be certified. Our filter sits in the same design space as the denoiser, but it has no learned parameters, is computed from a single decomposition of the input, and is nonexpansive by construction rather than by estimation. Our contribution is this base rule, not the certificate itself.

Tensor-train geometry. The tensor-train (TT) format (Oseledets, 2011) represents a d-way tensor as a chain of cores whose internal dimensions, the bond ranks, equal the ranks of the sequential unfoldings. The set of tensors of a fixed feasible TT rank is a smooth embedded submanifold of $\mathbb{R}^N$, while the corresponding bounded-rank set, defined by the vanishing of minors of the unfoldings, is a real algebraic variety (Hackbusch, 2012). QTT quantization (Khoromskij, 2011) exploits the same format at the level of a single image: it reshapes a $2^D \times 2^D$ image into D paired-bit spatial modes, ordered from coarse to fine, so that each bond separates coarser spatial detail from finer detail and its rank measures the coupling between them. We use both facts: the variety structure gives the genericity theorem, and the ordered modes make the spatial bond index a scale coordinate along which the rank trajectory is defined.

3 Preliminaries

Notation. Images live in $[0,1]^N$ with $N = 2^D \times 2^D \times C$ and a fixed classifier $f : \mathbb{R}^N \to \mathbb{R}^K$ returns logits over $\mathcal{Y} = \{1, \dots, K\}$. For a d-way tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, the k-th unfolding $\mathcal{X}_{\langle k \rangle}$ groups the first k modes into rows and the remaining modes into columns. Its size is $\big(\prod_{i \le k} n_i\big) \times \big(\prod_{i > k} n_i\big)$. The TT rank is the tuple $\mathrm{TT\text{-}rank}(\mathcal{X}) = (r_1, \dots, r_{d-1})$, where $r_k = \mathrm{rank}(\mathcal{X}_{\langle k \rangle})$. Throughout, f is continuous, as holds for the networks used here.

QTT quantization. A $2^D \times 2^D \times C$ image $\mathbf{x}$ is reshaped by interleaving the bits of its row and column indices (Morton / Z-order). This gives the order-$(D{+}1)$ tensor $\mathcal{X} = Q(\mathbf{x})$ with mode sizes $(4, \dots, 4, C)$. The first spatial mode contains the coarsest pair of row-column bits and the D-th contains the finest pair, so spatial bond k separates the first k coarse scales from the finer scales that follow (Khoromskij, 2011; Lin et al., 2025). The corresponding bond rank therefore measures the complexity of the interaction across that scale and the final mode carries the color channel.

One-sweep rank ladder. A single TT-SVD sweep (Oseledets, 2011) with rank cap $r_{\max}$ stores, at every bond k, singular directions ordered by singular value. Truncating each bond to its top $\min(r, m_k)$ directions, where m_k is the maximal rank of bond k, and recontracting the cores yields for every $r \le r_{\max}$ an image-space reconstruction $\Pi_r(\mathbf{x}) = Q^{-1}(\Pi_r Q(\mathbf{x}))$ with TT ranks bounded by r. The retained subspaces grow with r, so the family is nested in this sense; Π_r denotes this reconstruction map, not an orthogonal projection. The one-sweep family closely tracks rerunning TT-SVD independently at each rank.

Randomized smoothing. For a base classifier $h : \mathbb{R}^N \to \mathcal{Y}$, the smoothed classifier is $g(\mathbf{x}) = \arg\max_c \mathbb{P}_{\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I)}(h(\mathbf{x} + \boldsymbol{\xi}) = c)$. If $\mathbb{P}(h(\mathbf{x} + \boldsymbol{\xi}) = c_A) \ge \underline{p_A} > \frac{1}{2}$, then $g(\mathbf{x} + \boldsymbol{\delta}) = c_A$ for all $\|\boldsymbol{\delta}\|_2 < R = \sigma\, \Phi^{-1}(\underline{p_A})$ (Cohen et al., 2019). The guarantee holds for any measurable h: the proof is a Neyman-Pearson argument about Gaussian measures in $\mathbf{x}$-space and never inspects the classifier. Certification in practice follows Cohen's CERTIFY: n_0 samples select the candidate class, n samples give a lower confidence bound $\underline{p_A}$ at level α, and the procedure abstains if $\underline{p_A} \le \frac{1}{2}$. Smoothed classifiers must be evaluated with a base classifier trained (or otherwise robust) under the smoothing noise; we return to this point in Section 6.

4 Method

4.1 Geometric motivation

Proposition 1 (Genericity of maximal exact TT rank)**.** *For tensors in $\mathbb{R}^{n_1 \times \cdots \times n_d}$, let $m_k = \min\{\prod_{i \le k} n_i, \prod_{i > k} n_i\}$ denote the maximal rank of the k-th unfolding. For a feasible tuple $r = (r_1, \dots, r_{d-1})$ satisfying $r_k < m_k$ for at least one k, define $\mathcal{V}_r = \{\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d} : \mathrm{rank}(\mathcal{X}_{\langle k \rangle}) \le r_k$ for every $k\}$. Then $\mathcal{V}_r$ is a proper real algebraic set, hence closed and of Lebesgue measure zero. Consequently, if $\boldsymbol{\delta}$ is sampled from a distribution absolutely continuous with respect to Lebesgue measure, every unfolding of $\boldsymbol{\delta}$ has maximal rank almost surely.*

Proof sketch. For each bond k, the constraint $\mathrm{rank}(\mathcal{X}_{\langle k \rangle}) \leq r_k$ is equivalent to the vanishing of all $(r_k{+}1) \times (r_k{+}1)$ minors of the unfolding, so the constraint set is algebraic, and proper whenever $r_k < m_k$. Their intersection $\mathcal{V}_r$ is algebraic and, by hypothesis, lies in a proper constraint set; a proper real algebraic subset of Euclidean space is Lebesgue-null. Full proof in Section A.1. □

Proposition 1 is an exact-rank statement about continuously sampled directions: it neither implies that a deterministic attack outputs a maximal-rank perturbation, nor that maximal exact rank forces a flat singular spectrum. Its role is motivational. The standard threat model constrains the perturbation norm and the classification loss, but nothing in it preserves the multiscale rank structure of the image; whether a given attack preserves that structure is an empirical question.

We therefore work with effective rather than exact rank: for a nonzero tensor $\mathcal{Z}$ with singular values $\sigma_{k,1} \geq \sigma_{k,2} \geq \cdots$ at bond k, define the cumulative spectral energy and the η-effective rank by

$$E_k(r; \mathcal{Z}) = \frac{\sum_{i \leq r} \sigma_{k,i}^2}{\sum_i \sigma_{k,i}^2}, \qquad r_{k,\eta}^{\mathrm{eff}}(\mathcal{Z}) = \min\{r : E_k(r; \mathcal{Z}) \geq \eta\}.$$

A natural image concentrates most of its energy at moderate r; the Gaussian and gradient-derived directions we measure spread theirs across many more components. Maximal exact rank does not force this spreading: the separation is empirical, not a consequence of Proposition 1, and a low-frequency perturbation can concentrate its spectrum without any explicit rank constraint.

In deployment, the perturbation $\boldsymbol{\delta}$ is never observed in isolation, only $\mathbf{x} + \boldsymbol{\delta}$, so the geometry above is motivation rather than a decision rule. The operational question is whether decision-relevant structure enters $\mathbf{x} + \boldsymbol{\delta}$ at different points of the rank ladder for clean and adversarial inputs, and rank trajectories make exactly that observable.

4.2 Rank trajectories and their statistics

Definition 1 (Rank trajectory). Fix ranks $r_1 < \cdots < r_R$ and construct the one-sweep ladder of Section 3. Given a fixed classifier f, the rank trajectory of $\mathbf{x}$ is

$$\tau(\mathbf{x}) \;=\; \big[\,\mathrm{softmax}\, f(\Pi_{r_1}(\mathbf{x})),\; \ldots,\; \mathrm{softmax}\, f(\Pi_{r_R}(\mathbf{x})),\; \mathrm{softmax}\, f(\mathbf{x})\,\big],$$

a sequence of $R+1$ probability vectors. We index the truncated rungs by $j = 1, \ldots, R$ and reserve $j = R+1$ for the original input. The trajectory asks the same classifier to read progressively richer versions of one input, spanning three intuitive regimes. At very low rank, both clean and adversarial reconstructions are too coarse for reliable prediction. At moderate rank, the dominant image structure becomes recognizable and clean predictions usually stabilize. At the highest ranks, a successful adversarial example often flips from the true class to the adversarial class. Importantly, Proposition 1 does not imply that this late transition must occur; whether an adversarial decision emerges only at higher ranks is an empirical question that the rank trajectory is designed to test.

A trajectory contains both class changes and confidence changes. We extract one statistic of each type. First, the stabilization rank records the last rung at which the predicted class changes. Writing $\widehat{y}_j(\mathbf{x}) = \arg\max_c \tau_{j,c}(\mathbf{x})$,

$$s_{\mathrm{stab}}(\mathbf{x}) \;=\; \frac{1}{R}\,\max\Big(\{0\} \cup \big\{\, j-1 :\; j \in \{2, \ldots, R+1\},\; \widehat{y}_j(\mathbf{x}) \neq \widehat{y}_{j-1}(\mathbf{x}) \,\big\}\Big) \;\in\; [0,1]. \quad (1)$$

Thus $s_{\mathrm{stab}} = 0$ when the predicted class never changes, while a value near one means that the final change occurs near the original input. Second, let $m(p) = p_{(1)} - p_{(2)}$ be the gap between the two largest class probabilities. The margin gain

$$s_{\mathrm{marg}}(\mathbf{x}) \;=\; m\big(\tau_{R+1}(\mathbf{x})\big) \;-\; \max_{j \leq R} m\big(\tau_j(\mathbf{x})\big) \quad (2)$$

compares the margin at the original input with the largest margin attained on any rung of the ladder. Because the top rungs of the grid reconstruct the input (Section B), the statistic is at most zero: it equals zero when the full-input decision is at least as confident as every truncated reconstruction, and is negative otherwise. A loss-maximising attack tends to produce the former, manufacturing a confident decision that no truncated rung supports, whereas a clean input, whose structure is largely present at moderate rank, often attains its largest margin before the final rung. Neither behavior is assumed universal; both statistics are therefore evaluated under adaptive attacks.

The two statistics have different scales and different clean distributions, so we first convert each into a clean-referenced percentile. For $k \in \{\mathrm{stab}, \mathrm{marg}\}$, let $\widehat{F}_k$ be the empirical CDF of s_k on a held-out clean calibration set. The combined anomaly score is

$$s_{\mathrm{comb}}(\mathbf{x}) \;=\; \max_{k \in \{\mathrm{stab},\, \mathrm{marg}\}} \widehat{F}_k\big(s_k(\mathbf{x})\big) \;\in\; [0,1] \tag{3}$$

and flags an input when $s_{\mathrm{comb}}(\mathbf{x}) > \theta$. This is a percentile combination, not an exact p-value construction: in particular, s_{stab} is discrete, so its empirical CDF transform is not exactly uniform. We therefore calibrate the final threshold θ directly from the clean distribution of s_{comb}. No adversarial example is used to fit either the reference CDFs or the threshold.

5 A NONEXPANSIVE QTT FILTER

The trajectory statistics provide empirical evidence of adversarial structure, but the balanced adaptive attack in Section 6.3 shows that they do not by themselves yield a robustness guarantee. We therefore ask a different question: *can the same rank coordinate improve a method that certifies predictions against every perturbation in an ℓ_2 ball?*

Randomized smoothing provides such certificates by evaluating a base classifier under isotropic Gaussian noise. Obtaining useful certified radii generally requires substantial noise, which presents a structural opportunity for QTT. A Gaussian direction has maximal exact TT rank almost surely by Proposition 1, and its energy is empirically spread across many QTT components (Table 1). Natural images, by contrast, concentrate much of their energy in fewer components. This suggests placing a rank-selective operator before the classifier: suppressing weak singular directions may remove a substantial part of the smoothing noise while preserving the dominant image structure.

Let $Q : \mathbb{R}^N \to \mathbb{R}^{n_1 \times \cdots \times n_d}$ denote the isometric QTT reshape. For each bond k, let $\mathcal{U}_k$ and $\mathcal{U}_k^{-1}$ denote the corresponding unfolding and folding operators. Both are isometries under the Euclidean and Frobenius norms. Matrix singular-value soft-thresholding (SVT) is defined by

$$\mathcal{S}_\lambda(M) = U\ \mathrm{diag}\big((\sigma_i - \lambda)_+\big)V^\top, \qquad M = U\,\mathrm{diag}(\sigma_i)V^\top, \tag{4}$$

where $(a)_+ = \max(a, 0)$. The soft-thresholding map associated with bond k is $\mathcal{P}_{k,\lambda} = \mathcal{U}_k^{-1} \circ \mathcal{S}_\lambda \circ \mathcal{U}_k$. We define the QTT filter by one ordered sweep over the bonds,

$$\mathrm{SVT}_\lambda = Q^{-1} \circ \mathcal{P}_{d-1,\lambda} \circ \cdots \circ \mathcal{P}_{1,\lambda} \circ Q. \tag{5}$$

Thus, each step shrinks the singular spectrum of the current tensor at one QTT bond, setting singular values at or below λ to zero. The implementation computes this composition in a single sweep over residual matrices, shown in Section A.3 to coincide with Equation (5).

Proposition 2 (Nonexpansiveness)**.** *For every $\lambda \geq 0$, the QTT soft-thresholding filter $\mathrm{SVT}_\lambda : \mathbb{R}^N \to \mathbb{R}^N$ is nonexpansive:*

$$\|\mathrm{SVT}_\lambda(\mathbf{x}) - \mathrm{SVT}_\lambda(\mathbf{x}')\|_2 \leq \|\mathbf{x} - \mathbf{x}'\|_2 \quad \textit{for all } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^N. \tag{6}$$

In particular, SVT_λ is continuous.

Proof sketch. Matrix singular-value soft-thresholding is the proximal operator of the nuclear norm; this identification shows both that Equation (4) is well defined despite the non-uniqueness of the SVD and that the map is firmly nonexpansive in Frobenius norm. Because unfolding and folding are isometries, every $\mathcal{P}_{k,\lambda}$ is nonexpansive. The QTT reshape Q is also an isometry, and a finite composition of nonexpansive maps remains nonexpansive. Applying these facts to Equation (5) proves the result. Full proof in Section A.2. □

This proposition has two distinct consequences. First, the filter cannot increase the Euclidean distance between two inputs, providing a global stability bound without estimating a Lipschitz constant from data. Second, it defines a continuous, and hence measurable, preprocessing map. Let $\mathcal{C}$ denote clipping to $[0,1]^N$ and define the base classifier $h_\lambda(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} f_c((\mathcal{C} \circ \mathrm{SVT}_\lambda \circ \mathcal{C})(\mathbf{x}))$, with ties resolved by a fixed rule. Its Gaussian-smoothed classifier is $g_\lambda(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} \Pr_{\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 I)}\big(h_\lambda(\mathbf{x} + \boldsymbol{\xi}) = c\big)$. Because randomized smoothing

applies to any measurable base classifier, if a valid lower confidence bound satisfies $\underline{p}_A > 1/2$ for the winning class c_A, then $g_\lambda(\mathbf{x}+\boldsymbol{\delta}) = c_A$ for every $\|\boldsymbol{\delta}\|_2 < \sigma\Phi^{-1}(\underline{p}_A)$. The certificate is therefore the standard randomized-smoothing certificate; nonexpansiveness is an additional property of the filter, not a prerequisite for the certificate.

The filter contains no learned parameters and adds one QTT soft-thresholding sweep to each evaluation. The classifier must nevertheless be adapted to its output distribution: we first train with Gaussian augmentation at the certification noise level σ and then fine-tune on $(\mathcal{C} \circ \mathrm{SVT}_\lambda \circ \mathcal{C})(\mathbf{x}+\boldsymbol{\xi})$, so that the classifier is trained on the same distribution it sees at certification.

6 Experiments

Setup. We evaluate CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Imagenette (Howard, 2021) with ResNet (He et al., 2016). The QTT tensorization uses $D = 5$ spatial levels for CIFAR and $D = 8$ for Imagenette. We use resolution-specific rank grids that span from rank 1 to the maximum relevant bond dimension, with denser sampling at lower ranks. We evaluate FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), AutoAttack (AA) (Croce & Hein, 2020), the low-frequency attack, and the three adaptive attacks. For each attack, clean-adversarial separation is evaluated on successful adversarial examples and their corresponding clean examples, drawn from the same set of 512 test images (Nie et al., 2022). All detector thresholds are calibrated on a disjoint clean split, with no adversarial examples used for calibration. For randomized smoothing, we use Gaussian-augmented networks with $\sigma \in \{0.25, 0.5, 0.75, 1.00\}$, $n_0 = 100$ selection samples, $n = 10^4$ estimation samples, and significance level $\alpha = 10^{-3}$ (Section B).

6.1 Spectral separation along the rank coordinate

Table 1: QTT bond spectra of a representative natural image and four perturbation ($\epsilon = 8/255$) directions on 256×256 inputs. The first five columns report the cumulative spectral energy $E_3(r)$ at the third spatial bond. The final column reports the 99% effective rank at the fourth spatial bond.

	bond $k = 3$: cumulative energy $E_3(r)$					$k = 4$
Signal	r=2	r=4	r=8	r=16	r=32	eff. rank
natural image $\mathbf{x}$	0.851	0.896	0.944	0.972	0.992	27
$\boldsymbol{\delta}$: i.i.d. Gaussian (generic)	0.040	0.079	0.154	0.298	0.562	245
$\boldsymbol{\delta}$: sign(smoothed grad)	0.058	0.111	0.208	0.377	0.643	232
$\boldsymbol{\delta}$: sign(white noise)	0.040	0.079	0.155	0.298	0.562	245
$\boldsymbol{\delta}$: low-frequency	0.595	0.865	0.976	0.999	1.000	5

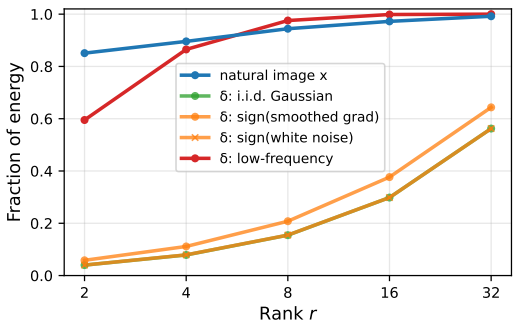

Figure 1: Cumulative spectral energy $E_3(r)$.

Table 1 and Figure 1 evaluate the spectral contrast motivating Section 4.1 on a 256×256 image. We compare the natural image with four perturbation directions scaled to $\epsilon = 8/255$: an i.i.d. Gaussian direction as an unstructured reference, the sign of a spatially smoothed input gradient as a PGD proxy, the sign of white noise as a high-frequency control, and a low-frequency direction obtained by Gaussian blurring. For each signal, we report the cumulative spectral energy $E_3(r)$ at the third spatial bond and the 99% effective rank at the fourth spatial bond, whose maximal rank is 256. The natural-image spectrum is far more concentrated than the unstructured and gradient-derived ones. At rank 8 the image retains 0.944 of its third-bond energy, against 0.154 for the Gaussian direction and 0.208 for the PGD proxy; its fourth-bond effective rank is 27, against 245 and 232. The high-frequency control matches the Gaussian reference almost exactly, so the sign operation of ℓ_∞ attacks does not by itself induce concentration. The

low-frequency direction is the deliberate exception, with 0.976 of third-bond energy in its top eight components and a fourth-bond effective rank of 5; Section 6.3 examines how this structured exception affects trajectory-based separation. Table 9 in Section B extends this comparison from one image to dataset-level statistics over CIFAR-10 and Imagenette, with perturbations taken from the adversarial attacks. The separation is a property of the datasets, not of a chosen image: the spectral concentration of natural images and that of perturbations form non-overlapping distributions on both datasets, with natural images at low effective rank and the perturbations near the maximal bond rank, and the contrast is sharper at the lower resolution. In sum, the rank coordinate separates natural images from unconstrained perturbation directions already at the input level, before any classifier is involved; the experiments that follow test whether this separation is visible in model behavior.

6.2 The rank trajectory across resolution and architecture

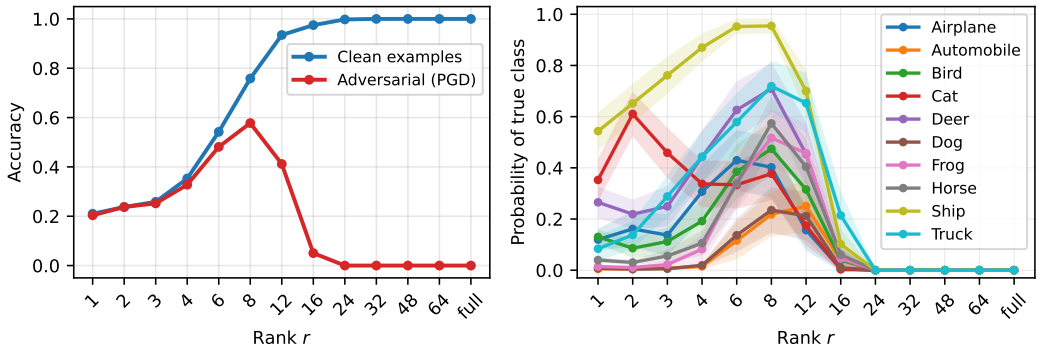

Figure 2: Left: accuracy with respect to the true label for clean and adversarial examples on CIFAR-10. Right: mean true-class probability of adversarial examples with 95% confidence bands.

Figure 2 evaluates the trajectory hypothesis of Section 4.2 on CIFAR-10, tracking accuracy with respect to the true label at every rung for all successful PGD examples and their clean counterparts. The three regimes appear as predicted. Clean accuracy climbs out of the coarse regime, saturates at 100% by rank 32, and never changes again. Adversarial examples retain the true class at moderate rank, 57.8% at rank 8, and collapse to 0% by rank 24 as higher-rank components are restored. The per-class curves on the right further show that this behavior is not driven by a few classes.

We next test whether this behavior persists across resolutions, architectures, datasets, and random seeds. The pattern recurs in every setting. On 256×256 Imagenette with an unmodified ImageNet ResNet-50, adversarial examples retain the true-class prediction at moderate ranks but lose it entirely by rank 96, while clean predictions settle only as the ladder approaches the bond dimension (Table 6). The same late flip appears for ResNet-56 and ResNet-110 under the CIFAR-10 protocol, and repeating the CIFAR-10 experiment under different global random seeds reproduces it. On CIFAR-100 the trajectory pattern persists (Table 7), but clean predictions stabilize late among fine-grained labels. These results support the late-flip pattern motivated by the QTT geometry, while also showing that its strength and calibration are deployment-dependent.

6.3 Clean-adversarial discrimination

We evaluate whether the trajectory statistics distinguish clean from adversarial inputs, measured by the area under the receiver operating characteristic (ROC) curve, denoted AUC, and whether this separation survives adaptive attacks. In addition, we further construct adaptive attacks: The consistency attack targets the stabilization rank by encouraging agreement between the full-input prediction and all trajectory rungs, using BPDA (Athalye et al., 2018) through the rank ladder. The low-confidence attack targets the margin gain with a capped Carlini-Wagner objective. The joint attack combines both objectives with $\lambda = 1$, selected as the strongest setting in the sweep of Table 5.

Against iterative standard attacks, the trajectory statistics achieve AUCs using one decomposition without training or an auxiliary model. The two trajectory statistics are nevertheless complementary:

Table 2: Clean-adversarial discrimination measured by AUC, with attack success rates (ASR) reported in the first row. Results for the baselines under the adaptive attacks measure transfer only, that is these attacks are optimized against the QTT rather than against each baseline.

	Standard attacks				Adaptive attacks		
	FGSM	PGD	AA	Low-freq.	Cons.	Low-conf.	Joint
Attack success rate (%)	75.84	100.0	100.0	4.41	35.29	58.40	60.08
Baselines							
Feature Squeezing (Xu et al., 2018)	0.780	0.701	0.711	0.615	0.594	0.933	0.906
Mahalanobis Distance (Lee et al., 2018)	0.795	1.000	1.000	0.553	0.992	0.978	0.910
LID (Ma et al., 2018)	0.758	0.941	0.993	0.637	0.895	0.964	0.898
BEYOND (He et al., 2024)	0.855	0.689	0.666	0.676	0.527	0.984	0.911
Layer Regression (Mumcu & Yilmaz, 2025)	0.458	0.999	1.000	0.485	0.987	0.569	0.514
Perturbation Forgery (Wang et al., 2025)	1.000	1.000	1.000	0.512	1.000	1.000	1.000
Ours: two statistics from one decomposition sweep, and their combination							
Stabilization rank s_{stab}	0.761	0.794	0.807	0.819	0.024	0.957	0.263
Margin gain s_{marg}	0.248	0.876	0.914	0.361	0.880	0.015	0.104
Clean-percentile combination	0.666	0.865	0.895	0.841	0.813	0.889	0.245

Table 3: The evaluation with method-specific attacks. Evasion is the fraction of inputs that are both misclassified and scored as clean.

Method	ASR	AUC	Evasion
Xu et al. (2018)	99.37	0.081	98.52
Lee et al. (2018)	99.16	0.007	99.16
Ma et al. (2018)	96.64	0.180	96.63
He et al. (2024)	100.0	0.000	100.00
Mumcu & Yilmaz (2025)	97.37	0.001	97.27
Wang et al. (2025)	99.79	0.000	99.79
Ours (Combination)	**60.08**	0.245	**55.88**

margin gain performs well against iterative attacks but reverses direction under FGSM and low-frequency perturbations, where stabilization rank remains informative. Their clean-percentile combination therefore follows whichever statistic is informative, giving up a little against the better one on any single attack in exchange for never inheriting the failure of the other. The same complementarity appears under adaptive attacks. The consistency attack reduces stabilization-rank AUC to 0.024 but leaves margin gain informative at 0.880; the low-confidence attack reduces margin-gain AUC to 0.015 while stabilization rank reaches 0.957. Consequently, the combination retains AUCs of 0.813 and 0.889 against the two attacks.

Because the adaptive columns of Table 2 attack only the QTT ladder, they do not measure the adaptive robustness of the baselines. We therefore attack each detector as shown in Table 3. Every detector is vulnerable to its method-specific attack. The baselines have AUCs at or below 0.180 and evasion rates above 96%. Our combination has a lower evasion rate of 55.88%: among successful attacks, most still bypass the detector. A repetition on 2000 test images with bootstrap confidence intervals confirms the picture with small uncertainty, as reported under the discrimination protocol in Section B. These results delimit what rank-trajectory discrimination achieves empirically and motivate the certified construction considered next (Section 6.4).

6.4 Certified accuracy

We evaluate whether the nonexpansive filter of Section 5 improves plain Cohen smoothing (Cohen et al., 2019). The spectral hypothesis predicts a regime-dependent effect. When the smoothing noise is weak, filtering may remove as much image information as noise and should provide little benefit. As σ increases and the broad QTT spectrum of the noise increasingly obscures the more concentrated image spectrum, suppressing weak singular directions should become beneficial.

Specifically, at $\sigma = 0.25$, QTT-SVT changes little on either dataset: its denoising benefit is approximately offset by the image information it removes. At $\sigma \geq 0.5$, however, it improves certified accuracy at every reported radius. On CIFAR-10, radius-zero certified accuracy rises from 0.392 to 0.646 at $\sigma = 0.5$ and from 0.188 to 0.458 at $\sigma = 1.0$, and ACR improves by 41%, 37%, and 26% at $\sigma = 0.5, 0.75$, and 1.0, respectively. The effect is larger on CIFAR-100, where the classifier deteriorates more sharply under strong smoothing noise: ACR rises from 0.114 to 0.216 at $\sigma = 0.5$,

Table 4: Certified accuracy of plain randomized smoothing and the matched QTT-SVT system at ℓ_2 radius ρ. ACR is the average certified radius over the test set, with radius zero assigned to points that are not correctly certified.

σ	Method	0.00	0.25	0.50	0.75	1.00	1.25	ACR
CIFAR-10								
0.25	Cohen smoothing	0.778	**0.620**	0.408	**0.252**	0.000	0.000	0.400
	+ SVT (Ours)	**0.790**	0.618	**0.424**	0.242	0.000	0.000	**0.402**
0.50	Cohen smoothing	0.392	0.340	0.272	0.218	0.172	0.130	0.349
	+ SVT (Ours)	**0.646**	**0.530**	**0.414**	**0.300**	**0.194**	**0.148**	**0.491**
0.75	Cohen smoothing	0.250	0.222	0.206	0.186	0.162	0.130	0.343
	+ SVT (Ours)	**0.528**	**0.434**	**0.352**	**0.260**	**0.212**	**0.148**	**0.471**
1.00	Cohen smoothing	0.188	0.182	0.174	0.156	0.144	0.126	0.357
	+ SVT (Ours)	**0.458**	**0.372**	**0.300**	**0.236**	**0.178**	**0.146**	**0.451**
CIFAR-100								
0.25	Cohen smoothing	**0.466**	0.302	**0.208**	0.088	0.000	0.000	0.199
	+ SVT (Ours)	0.462	**0.326**	**0.208**	**0.096**	0.000	0.000	**0.207**
0.50	Cohen smoothing	0.150	0.116	0.088	0.072	0.052	0.038	0.114
	+ SVT (Ours)	**0.332**	**0.242**	**0.180**	**0.122**	**0.076**	**0.052**	**0.216**
0.75	Cohen smoothing	0.046	0.040	0.032	0.026	0.022	0.020	0.053
	+ SVT (Ours)	**0.240**	**0.176**	**0.136**	**0.104**	**0.082**	**0.064**	**0.201**
1.00	Cohen smoothing	0.024	0.022	0.020	0.020	0.016	0.012	0.041
	+ SVT (Ours)	**0.176**	**0.140**	**0.120**	**0.088**	**0.064**	**0.050**	**0.164**

from 0.053 to 0.201 at $\sigma = 0.75$, and fourfold, from 0.041 to 0.164, at $\sigma = 1.0$, while radius-zero certified accuracy at the two strongest noise levels improves from 0.046 and 0.024 to 0.240 and 0.176. These results follow the predicted noise-dependent pattern.

The improvement is insensitive to moderate changes in the threshold and to the certification seed. On CIFAR-10 at $\sigma = 0.5$, $\lambda \in \{0.2,0.25,0.3,0.4\}$ gives ACRs between 0.484 and 0.491, compared with 0.349 for plain smoothing, and at $\sigma = 1.0$, $\lambda \in \{0.3,0.4,0.5\}$ gives ACRs between 0.448 and 0.451. An independent certification seed reproduces the principal $\sigma = 0.5$ result within 0.001.

These results provide a constructive use of the same rank geometry studied throughout the paper. Gaussian smoothing noise has a broad, high-effective-rank QTT spectrum, while natural-image energy is more concentrated. QTT soft-thresholding does not isolate noise exactly, but it preferentially suppresses the weaker spectral components in which much of that noise appears. This distinction explains both the negligible effect at low noise and the systematic improvement at high noise. The proposed filter is training-free, globally nonexpansive, and requires one QTT sweep per evaluation. The controlled conclusion is that, relative to plain smoothing with the same architecture and certification protocol, the QTT rank coordinate provides a consistent certified-accuracy benefit when smoothing noise is sufficiently strong.

7 Conclusion

We introduced QTT rank as a multiscale structural coordinate for adversarial robustness. Natural images concentrate their spectral energy at lower effective ranks than Gaussian and gradient-derived perturbation directions, and rank trajectories translate this structural contrast into model behavior. Across datasets, architectures, and resolutions, clean predictions typically stabilize at moderate rank, whereas adversarial decisions often emerge only as higher-rank components are restored. Two training-free, clean-calibrated statistics capture these complementary trajectory dynamics from a single decomposition sweep. The same rank geometry also provides a constructive benefit for certification. A training-free, nonexpansive QTT filter suppresses the broad spectrum of Gaussian smoothing noise while preserving dominant image structure. With matched classifier fine-tuning, it consistently improves certified accuracy over plain randomized smoothing in the high-noise regime. Together, these results establish QTT rank as an interpretable and computable link between multiscale image structure, model behavior, and certified robustness.

AI USE STATEMENT

In this work, we used generative AI tools for editorial refinement to enhance the clarity and presentation of the text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *The Eleventh International Conference on Learning Representations*, 2023.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

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

Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhiyuan He, Yijun Yang, Pin-Yu Chen, Qiang Xu, and Tsung-Yi Ho. Be your own neighborhood: Detecting adversarial examples by the neighborhood relations built on self-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 18063–18080. PMLR, 2024.

Jeremy Howard. Fastai/imagenette. https://github.com/fastai/imagenette, 2021.

Boris N Khoromskij. O (d log n)-quantics approximation of n-d tensors in high-dimensional numerical modeling. *Constructive Approximation*, 34:257–280, 2011.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel J. Hsu, and Suman Sekhar Jana. Certified robustness to adversarial examples with differential privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672, 2019.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Guang Lin, Duc Thien Nguyen, Zerui Tao, Konstantinos Slavakis, Toshihisa Tanaka, and Qibin Zhao. Model-free adversarial purification via coarse-to-fine tensor network representation. *arXiv preprint arXiv:2502.17972*, 2025.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

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

Furkan Mumcu and Yasin Yilmaz. Universal and efficient detection of adversarial data through nonuniform impact on network layers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

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

Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in neural information processing systems*, 33:21945–21957, 2020.

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

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Qian Wang, Chen Li, Yuchen Luo, Hefei Ling, Shijuan Huang, Ruoxi Jia, and Ning Yu. Detecting adversarial data using perturbation forgery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13917–13926. IEEE, 2025.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings 2018 Network and Distributed System Security Symposium*, 2018.

Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pp. 12062–12072. PMLR, 2021.

A PROOFS

A.1 PROOF OF PROPOSITION 1

We first record the classical lemma.

Lemma 1. *Let $p : \mathbb{R}^N \to \mathbb{R}$ be a polynomial that is not identically zero. Then its zero set $Z(p) = \{z \in \mathbb{R}^N : p(z) = 0\}$ has Lebesgue measure zero.*

Proof. Induction on N. For $N = 1$, $Z(p)$ is finite. For $N > 1$, write p as a polynomial in z_N with coefficients $q_0, \dots, q_m$ polynomial in $(z_1, \dots, z_{N-1})$, where m is chosen so that $q_m \not\equiv 0$. For each fixed $z' = (z_1, \dots, z_{N-1})$ with $q_m(z') \neq 0$, the univariate polynomial $z_N \mapsto p(z', z_N)$ is not identically zero, so its zero set is finite, hence null in $\mathbb{R}$. The set $\{z' : q_m(z') = 0\}$ is null in $\mathbb{R}^{N-1}$ by the inductive hypothesis. By Tonelli's theorem,

$$\lambda_N\big(Z(p)\big) = \int_{\mathbb{R}^{N-1}} \lambda_1\big(\{z_N : p(z', z_N) = 0\}\big)\, dz' = 0,$$

since the integrand vanishes outside the null set $\{q_m = 0\}$ and the integral of any measurable function over a null set is zero. □

Proof of Proposition 1. Fix $k \in \{1, \dots, d-1\}$ and let $m_k = \min\big(\prod_{i \le k} n_i, \prod_{i > k} n_i\big)$ be the maximal possible rank of the k-th unfolding. The set

$$V_k \;=\; \big\{\mathcal{X} : \operatorname{rank}(\mathcal{X}_{\langle k \rangle}) \le r_k\big\}$$

is the common vanishing locus of all $(r_k{+}1) \times (r_k{+}1)$ minors of $\mathcal{X}_{\langle k \rangle}$. Each such minor is a polynomial in the entries of $\mathcal{X}$, because the unfolding is a permutation of entries, so V_k is a real algebraic set. If $r_k < m_k$, then V_k is proper: any $\mathcal{X}$ whose unfolding contains an $(r_k{+}1) \times (r_k{+}1)$ identity block among zeros lies outside V_k, so at least one minor, viewed as a polynomial, is not identically zero. By Lemma 1, applied to that minor, $V_k \subseteq Z(\text{minor})$ has measure zero. Vanishing loci are also closed, being preimages of $\{0\}$ under continuous maps.

Now $\mathcal{V}_r = \bigcap_k V_k$ is an intersection of algebraic sets, hence algebraic and closed, and it is contained in V_{k_0} for any coordinate k_0 with $r_{k_0} < m_{k_0}$, which exists by hypothesis; therefore $\lambda_N(\mathcal{V}_r) = 0$.

For the final claim, let $\boldsymbol{\delta} \sim \mu$ with $\mu \ll \lambda_N$. For each k, the event $\{\operatorname{rank}(\boldsymbol{\delta}_{\langle k \rangle}) < m_k\}$ equals $\{\boldsymbol{\delta} \in V_k'\}$ with V_k' the proper algebraic set of unfolding rank at most $m_k - 1$, so it has μ-probability zero. A finite union of null events is null, so with probability one every unfolding of $\boldsymbol{\delta}$ has maximal rank, that is, $\mathrm{TT}\text{-}\operatorname{rank}(\boldsymbol{\delta})$ is the maximal tuple.

Finally, the QTT reshape Q is a permutation of coordinates and therefore preserves Lebesgue measure and absolute continuity; the statement and its consequences hold verbatim for images and perturbations in pixel space via $\mathcal{X} = Q(\mathbf{x})$. □

At every bond k, the retained subspace of Π_r is contained in that of $\Pi_{r'}$ for $r \le r'$, since both consist of leading singular directions of the same decomposition. Rerunning TT-SVD independently at each rank has no such guarantee and costs R decompositions instead of one. When singular values are repeated, the top-r subspace is not unique; a fixed deterministic SVD routine resolves the ambiguity, so Π_r is well defined.

A.2 PROOF OF PROPOSITION 2

The proof assembles three facts: the proximal operator of a convex function is nonexpansive, singular-value soft-thresholding is the proximal operator of the nuclear norm, and nonexpansiveness is preserved under isometric conjugation and composition. We prove each in turn.

Lemma 2 (Proximal operators are firmly nonexpansive)**.** *Let $h : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous. Then for every $v \in \mathbb{R}^M$ the proximal problem*

$$\operatorname{prox}_h(v) \;=\; \operatorname*{arg\,min}_{z \in \mathbb{R}^M} \; h(z) + \tfrac{1}{2}\|z - v\|_2^2$$

has a unique solution, and for all $u, v \in \mathbb{R}^M$*,*

$$\|\operatorname{prox}_h(u) - \operatorname{prox}_h(v)\|_2^2 \;\le\; \big\langle \operatorname{prox}_h(u) - \operatorname{prox}_h(v),\, u - v \big\rangle.$$

In particular, $\|\operatorname{prox}_h(u) - \operatorname{prox}_h(v)\|_2 \le \|u - v\|_2$.

Proof. The objective is the sum of a proper convex lower semicontinuous function and a strongly convex quadratic, so it is strongly convex and coercive, and a minimizer exists and is unique. Write $p_u = \operatorname{prox}_h(u)$ and $p_v = \operatorname{prox}_h(v)$. The optimality conditions read $u - p_u \in \partial h(p_u)$ and $v - p_v \in \partial h(p_v)$. Monotonicity of the subdifferential gives

$$\big\langle (u - p_u) - (v - p_v),\; p_u - p_v \big\rangle \;\ge\; 0,$$

which rearranges to $\|p_u - p_v\|_2^2 \le \langle p_u - p_v,\, u - v \rangle$. □

Lemma 3 (Soft-thresholding is the proximal operator of the nuclear norm)**.** *Let* $\|\cdot\|_*$ *denote the nuclear norm on* $\mathbb{R}^{p\times q}$ *and let* $\lambda \ge 0$*. For every* $M \in \mathbb{R}^{p\times q}$*,*

$$\mathcal{S}_\lambda(M) \;=\; \operatorname{prox}_{\lambda\|\cdot\|_*}(M),$$

with $\mathcal{S}_\lambda$ *as in Equation* (4)*. In particular,* $\mathcal{S}_\lambda(M)$ *does not depend on the choice of SVD of* M*.*

Proof. Fix an SVD $M = U \operatorname{diag}(\sigma_i) V^\top$ and consider $\phi(Z) = \lambda\|Z\|_* + \frac{1}{2}\|Z - M\|_F^2$. Expanding the quadratic,

$$\phi(Z) \;=\; \lambda\|Z\|_* + \tfrac{1}{2}\|Z\|_F^2 - \langle Z, M\rangle + \tfrac{1}{2}\|M\|_F^2,$$

and every term except the inner product depends on Z only through its singular values $s_1 \ge \cdots \ge s_{\min(p,q)} \ge 0$. By von Neumann's trace inequality, $\langle Z, M\rangle \le \sum_i s_i\sigma_i$, with equality when Z admits an SVD whose singular-vector pairs are those of M. Hence, among all Z with prescribed singular values, ϕ is minimized at $Z = U \operatorname{diag}(s_i) V^\top$, where it equals $\sum_i \big(\lambda s_i + \frac{1}{2}(s_i - \sigma_i)^2\big)$. This is separable and strictly convex in (s_i), with unique minimizer $s_i = (\sigma_i - \lambda)_+$; this minimizer is nonincreasing in i, so it is a valid ordered tuple of singular values, and $\mathcal{S}_\lambda(M)$ minimizes ϕ. By Lemma 2 the minimizer of ϕ is unique, so every SVD of M yields the same matrix $\mathcal{S}_\lambda(M)$, and $\mathcal{S}_\lambda = \operatorname{prox}_{\lambda\|\cdot\|_*}$. □

Lemma 4 (Isometric conjugation and composition)**.** *Let* $W : \mathcal{H}_1 \to \mathcal{H}_2$ *be a linear bijective isometry between Euclidean spaces and let* $T : \mathcal{H}_2 \to \mathcal{H}_2$ *be nonexpansive. Then* $W^{-1} \circ T \circ W$ *is nonexpansive on* $\mathcal{H}_1$*. Moreover, any finite composition of nonexpansive maps is nonexpansive.*

Proof. For $x, y \in \mathcal{H}_1$,

$$\|W^{-1}TWx - W^{-1}TWy\| \;=\; \|TWx - TWy\| \;\le\; \|Wx - Wy\| \;=\; \|x - y\|,$$

using that W and W^{-1} preserve norms. For the composition claim, apply nonexpansiveness once per factor: $\|T_m \cdots T_1 x - T_m \cdots T_1 y\| \le \|T_{m-1} \cdots T_1 x - T_{m-1} \cdots T_1 y\| \le \cdots \le \|x - y\|$. □

Proof of Proposition 2. By Lemmas 2 and 3, $\mathcal{S}_\lambda$ is well defined and nonexpansive in the Frobenius norm. Each unfolding $\mathcal{U}_k : \mathbb{R}^{n_1\times\cdots\times n_d} \to \mathbb{R}^{(\prod_{i\le k} n_i)\times(\prod_{i>k} n_i)}$ is a rearrangement of tensor entries, hence a linear bijective isometry between the tensor space with the Frobenius norm and the matrix space with the Frobenius norm, so $\mathcal{P}_{k,\lambda} = \mathcal{U}_k^{-1} \circ \mathcal{S}_\lambda \circ \mathcal{U}_k$ is nonexpansive by the conjugation claim of Lemma 4. The QTT reshape Q is likewise a bijective linear isometry, being a permutation of coordinates. By the composition claim of Lemma 4, the map $\mathrm{SVT}_\lambda = Q^{-1} \circ \mathcal{P}_{d-1,\lambda} \circ \cdots \circ \mathcal{P}_{1,\lambda} \circ Q$ of Equation (5) is nonexpansive on $\mathbb{R}^N$. A nonexpansive map is 1-Lipschitz, hence continuous. □

Clipping $\mathcal{C}$ is the Euclidean projection onto the convex set $[0,1]^N$, that is, the proximal operator of its indicator function, so Lemma 2 makes it nonexpansive as well. The full preprocessing map $\mathcal{C} \circ \mathrm{SVT}_\lambda \circ \mathcal{C}$ of Section 5 is therefore nonexpansive and continuous, and the base classifier h_λ, an argmax of finitely many continuous functions with a fixed tie-breaking rule, is Borel measurable, which is the only property the randomized-smoothing certificate requires.

A.3 THE SEQUENTIAL SWEEP COMPUTES SVT

The implementation does not restore the full tensor between bonds. It computes Equation (5) as a single TT-SVD-style sweep that soft-thresholds one residual matrix per bond and carries the left singular factors forward as cores. The following lemma is the reason the two computations coincide.

Lemma 5 (Equivariance under left isometries)**.** *Let $L \in \mathbb{R}^{p\times r}$ have orthonormal columns, $L^\top L = I_r$, and let $M \in \mathbb{R}^{r\times q}$. Then $\mathcal{S}_\lambda(LM) = L\,\mathcal{S}_\lambda(M)$.*

Proof. Let $M = U\,\mathrm{diag}(\sigma_i)V^\top$ be an SVD of M. Then LU has orthonormal columns, so $LM = (LU)\,\mathrm{diag}(\sigma_i)V^\top$ is an SVD of LM, and the nonzero singular values of LM are those of M. Applying Equation (4) to this SVD gives $\mathcal{S}_\lambda(LM) = (LU)\,\mathrm{diag}\big((\sigma_i-\lambda)_+\big)V^\top = L\,\mathcal{S}_\lambda(M)$, and by Lemma 3 the left side does not depend on the chosen SVD. □

Proposition 3 (Implementation equivalence)**.** *The single-sweep implementation, which at bond k applies $\mathcal{S}_\lambda$ to the residual matrix and stores the left factor as the k-th core, outputs exactly* $\mathrm{SVT}_\lambda(\mathbf{x})$ *of Equation* (5)*.*

Proof. Write $\mathcal{T}^{(k)} = \mathcal{P}_{k,\lambda}\circ\cdots\circ\mathcal{P}_{1,\lambda}(Q(\mathbf{x}))$, with $\mathcal{T}^{(0)} = Q(\mathbf{x})$. We claim that after step k the algorithm represents $\mathcal{T}^{(k)}$ in the factored form $\mathcal{T}^{(k)}_{\langle k\rangle} = L_kC_k$, where $L_k \in \mathbb{R}^{(\prod_{i\le k} n_i)\times r_k}$ is the contraction of the stored cores and has orthonormal columns, and C_k is the residual carried forward. This holds at $k=0$ with $L_0 = 1$ and C_0 the first unfolding of $Q(\mathbf{x})$. For the inductive step, the $(k{+}1)$-st unfolding of $\mathcal{T}^{(k)}$ is $\mathcal{T}^{(k)}_{\langle k+1\rangle} = (L_k\otimes I_{n_{k+1}})\,M_{k+1}$, where M_{k+1} is C_k reshaped to r_kn_{k+1} rows, and $L_k\otimes I_{n_{k+1}}$ again has orthonormal columns. By Lemma 5,

$$\mathcal{S}_\lambda\big(\mathcal{T}^{(k)}_{\langle k+1\rangle}\big) = (L_k\otimes I_{n_{k+1}})\,\mathcal{S}_\lambda(M_{k+1}),$$

so soft-thresholding the residual and folding back is exactly $\mathcal{P}_{k+1,\lambda}$ applied to the full tensor. Writing $\mathcal{S}_\lambda(M_{k+1}) = U\,\mathrm{diag}\big((\sigma_i-\lambda)_+\big)V^\top$, the algorithm stores the core U, so $L_{k+1} = (L_k\otimes I_{n_{k+1}})U$ has orthonormal columns, and carries $C_{k+1} = \mathrm{diag}\big((\sigma_i-\lambda)_+\big)V^\top$; discarding columns whose thresholded singular value is zero leaves the product UC_{k+1} unchanged. After bond $d-1$ the represented tensor is $\mathcal{T}^{(d-1)}$, and undoing the QTT reshape gives $\mathrm{SVT}_\lambda(\mathbf{x})$. □

B EXPERIMENTAL DETAILS

Models and data. All discrimination experiments use the CIFAR-10 test set and a pretrained ResNet-20. For certification, the same architecture is trained from scratch with Gaussian augmentation at the certification σ: SGD with momentum 0.9, weight decay 5×10^{-4}, batch size 256, 90 epochs, a cosine schedule, random crop and flip, and additive $\mathcal{N}(0,\sigma^2)$ noise without clamping. Normalization is inside the model, so all preprocessing operates on the raw pixel scale. Test-set subsampling uses a fixed stride over the 10,000 test images; the calibration split takes a unit offset of the same stride and is therefore disjoint from the evaluation split.

Rank grid. For 32×32 inputs (D=5) the grid is $\{1,2,3,4,6,8,12,16,24,32,48,64\}$; the maximal spatial bond rank is 48, so the top of the grid is effectively untruncated. The spacing is approximately geometric, placing rungs densely where stabilization happens. The statistics use only rung indices, so the procedure is insensitive to the exact grid values as long as the stabilization regime is resolved.

Table 5: Sweep of the joint attack's trade-off weight λ on CIFAR-10.

	λ=1	λ=5	λ=20	λ=100
Stabilization rank (AUC)	0.263	0.801	0.808	0.916
Margin gain (AUC)	0.104	0.047	0.028	0.020
Combination (AUC)	0.245	0.696	0.687	0.856

Attacks. FGSM and PGD-20/100 use $\epsilon = 8/255$ with step size $2/255$ in ℓ_∞; AutoAttack runs in its standard configuration. The low-frequency attack is PGD whose gradients and iterates are

projected onto a Gaussian-blur subspace. The three adaptive attacks run 100 steps of size $1/255$. The consistency attack augments the misclassification loss, differentiating through the ladder by straight-through BPDA. The low-confidence attack maximizes the saturating CW objective and needs no BPDA. The joint attack combines both penalties; we sweep $\lambda \in \{1, 5, 20, 100\}$ in Table 5 and report the setting that minimizes the combined AUC.

Discrimination protocol. For each attack we keep only test images the classifier initially classifies correctly (512 test images at a fixed stride), generate one adversarial example per image, and keep only successful attacks. Scores are computed for the clean and adversarial versions of the same images, and AUC is computed on the pooled scores. No statistic uses adversarial data for calibration; threshold-based numbers use the clean-calibrated threshold θ of Section 4.2. The repetition referred to in Section 6.3 uses 2000 test images with per-sample scores retained; each interval below is the 2.5th to 97.5th percentile of the AUC over 1000 bootstrap resamples of the pooled clean and adversarial scores. The combination reaches 0.838 [0.823, 0.853] under PGD and 0.857 [0.843, 0.871] under AutoAttack; the consistency attack drives the stabilization rank to 0.028 [0.020, 0.037] while the margin gain stays at 0.881 [0.859, 0.902]; the low-confidence attack drives the margin gain to 0.025 [0.019, 0.033] while the stabilization rank stays at 0.942 [0.932, 0.952]; and the joint attack leaves the combination at 0.254 [0.231, 0.278].

Table 6: Complete rank-ladder accuracies (%) with respect to the true label on Imagenette (256×256, D=8, ResNet-50; PGD at $\epsilon = 8/255$).

Rank r	8	16	24	32	48	64	96	128	full
Clean	2.54	15.25	41.10	58.90	82.63	91.10	96.61	98.31	100.0
Adversarial	2.54	13.14	37.29	47.46	40.68	17.80	0.00	0.00	0.00

Table 7: Complete rank-ladder accuracies (%) with respect to the true label on CIFAR-100 (32×32; ResNet-20; PGD at $\epsilon = 8/255$).

Rank r	4	6	8	12	16	24	32	48	full
Clean	13.59	31.25	50.00	75.00	90.49	96.20	98.91	100.0	100.0
Adversarial	12.77	25.00	27.45	14.67	0.54	0.00	0.00	0.00	0.00

Generality protocols and complete results. For resolution, we use Imagenette (ten ImageNet classes at 256×256, D=8 spatial modes) with a ResNet-50 classifier. Table 6 reports the complete ladder. For depth, ResNet-56 and ResNet-110 follow the CIFAR-10 protocol. For seeds, the CIFAR-10 trajectory experiment is repeated with three values of the global random seed, which governs all stochastic components of the run including attack initialization. We further conduct a rank-ladder experiment on CIFAR-100, as reported in Table 7. Each row of Table 4 is evaluated at a fixed stride over the test set, with the certification parameters of Section 6. The SVT-matched classifier is fine-tuned for 30 epochs from the Gaussian-augmented checkpoint on the deployed composition.

Table 8: Adaptive-robustness AUC across datasets and architectures.

		Stabilization rank			Margin gain			Combination		
Dataset	Arch.	Cons.	Low-conf.	Joint	Cons.	Low-conf.	Joint	Cons.	Low-conf.	Joint
CIFAR-10	RN20	**0.024**	0.957	0.263	0.880	**0.015**	0.104	0.813	0.889	**0.245**
CIFAR-100	RN20	**0.055**	0.843	0.451	0.870	**0.173**	0.358	0.842	0.642	**0.433**
CIFAR-10	RN56	**0.015**	0.946	0.321	0.962	**0.006**	0.052	0.868	0.919	**0.290**
CIFAR-10	RN110	**0.059**	0.960	0.370	0.866	**0.006**	0.067	0.733	0.927	**0.324**

Generality of the complementarity. Table 8 reports the AUC of both statistics and their combination, under the single-objective adaptive attacks and the balanced joint attack, on every dataset and architecture. The pattern of Section 6.3 is uniform: the consistency attack collapses the stabilization

rank and the low-confidence attack collapses the margin gain, each amplifying the other; the combination survives both single-objective attacks and is defeated by the balanced joint attack in every tested setting.

Table 9: Dataset-level QTT spectral statistics, mean $\pm$ standard deviation over N inputs. $E_3(8)$ is the cumulative energy of the top eight components at the third spatial bond; the effective rank is the 99% effective rank at the fourth spatial bond for Imagenette and at the third for CIFAR-10. Perturbation rows use the successful attack perturbations $\boldsymbol{\delta}$ at $\epsilon = 8/255$.

Dataset	Signal	N	$E_3(8)$	eff. rank
Imagenette, 256×256	natural images	500	0.962 ± 0.027	25.1 ± 18.2
	i.i.d. Gaussian	500	0.155 ± 0.001	245.0 ± 0.0
	FGSM	405	0.183 ± 0.045	240.5 ± 2.6
	PGD	436	0.193 ± 0.063	240.0 ± 3.7
	AutoAttack	436	0.201 ± 0.065	238.2 ± 4.5
CIFAR-10, 32×32	natural images	1000	0.995 ± 0.004	5.4 ± 2.2
	i.i.d. Gaussian	1000	0.426 ± 0.008	41.7 ± 0.5
	FGSM	672	0.502 ± 0.021	40.4 ± 0.6
	PGD	932	0.485 ± 0.020	40.4 ± 0.6
	AutoAttack	932	0.493 ± 0.020	40.0 ± 0.6

Table 10: Mean cumulative energy $E_3(r)$ at the third spatial bond over the inputs of Table 9.

Dataset	Signal	r=2	r=4	r=8	r=16	r=32
Imagenette	natural images	0.909	0.939	0.962	0.980	0.994
	i.i.d. Gaussian	0.040	0.079	0.155	0.299	0.562
	FGSM	0.059	0.102	0.183	0.332	0.595
	PGD	0.068	0.112	0.193	0.343	0.605
	AutoAttack	0.071	0.117	0.201	0.355	0.617
CIFAR-10	natural images	0.967	0.985	0.995	0.999	1.000
	i.i.d. Gaussian	0.130	0.241	0.426	0.688	0.942
	FGSM	0.181	0.308	0.502	0.749	0.957
	PGD	0.161	0.287	0.485	0.742	0.957
	AutoAttack	0.165	0.293	0.493	0.751	0.960

Dataset-level spectra. Tables 9 and 10 use 500 Imagenette validation images (256×256, ResNet-50) and 1000 CIFAR-10 test images, both at a fixed stride. Perturbation rows use the actual attack outputs at $\epsilon = 8/255$, restricted to successful attacks: FGSM, PGD-20, and AutoAttack; the Gaussian row draws i.i.d. noise, and every spectral statistic is scale invariant. Spectra are computed on the perturbation $\boldsymbol{\delta}$ alone with the batched TT-SVD of the ladder. The effective rank uses the 99% energy criterion at the fourth spatial bond for Imagenette, whose maximal rank is 256, and at the third spatial bond for CIFAR-10, whose maximal rank is 48.