REVISITING CERTIFIED DEFENSE WITH DIFFERENTIAL PRIVACY ON VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Certified defenses that incorporate differential privacy have proven effective on Convolutional Neural Networks (CNNs), furnishing rigorous robustness guarantees against norm-bounded adversaries. However, the certified robustness behavior of Pixel Differential Privacy (PixelDP) remains largely unexplored with the self-attention architecture now dominating the deep-learning landscape. Given that the Transformer has a profound impact on our daily applications from the digital world to the physical world, it is crucial to study certified robustness through differential-privacy-style stability. To fill this research gap, we revisit this construction in Vision Transformers and identify a failure mode that is largely hidden in the convolutional setting. When noise is injected after the patch embedding, the Laplace mechanism with the inherited grouped ℓ_1 sensitivity bound collapses to chance-level accuracy across noise scales, whereas the Gaussian mechanism remains trainable. This contrast isolates the source of failure: not the injected noise itself, but the geometry of the sensitivity constraint. We show that the attenuation induced by the inherited $\Delta_{1,1}$ projection increases with layer width and kernel size according to a random-matrix scale $C/(\sqrt{M}+\sqrt{N})$. Replacing the ℓ_1-type constraint with a spectral-norm constraint eliminates the collapse across datasets and architectures, but creates a fundamental obstacle: the repaired models no longer satisfy the sensitivity condition required by the standard Laplace certificate. We resolve this mismatch by deriving a dimension-free $(\varepsilon,\ \delta)$-privacy guarantee for the Laplace mechanism under ℓ_2 sensitivity through concentration of the privacy loss. The resulting certificate applies directly to the repaired models without retraining and avoids the $\sqrt{d}$ privacy degradation incurred by the naive conversion from ℓ_2 to ℓ_1 sensitivity. Experiments on CIFAR-10, CIFAR-100, and SVHN reproduce the failure-and-repair pattern, while ImageNet confirms that the inherited grouped-bound failure persists at scale. Finally, matched-budget experiments show that adversarial training improves certified accuracy at the evaluated positive radii.

1 INTRODUCTION

While deep neural networks (DNNs) empower social life, they also raise various security and privacy concerns Bengio et al. (2024). During inference, a network may be threatened by adversarial examples. An attacker can cause the model to make incorrect predictions by adding imperceptible perturbations Goodfellow et al. (2015); Madry et al. (2018). It poses serious risks to safety-critical applications such as autonomous driving Eykholt et al. (2018) and robotics Wang et al. (2025).

In addition to empirical defenses Madry et al. (2018); Zhang et al. (2019); Yan et al. (2024), the certified defense approaches Cohen et al. (2019); Zhang et al. (2020); Li et al. (2023) offer provable robustness guarantees under explicit assumptions. PixelDP Lécuyer et al. (2019) is a general framework that leverages differential-privacy-style stability Dwork (2006); Abadi et al. (2016) to certify robustness against norm-bounded input perturbations. It constrains the sensitivity of a pre-noise mapping and injects calibrated noise to stabilize the randomized prediction.

However, previous research on the certified robustness of differential privacy mainly focuses on Convolutional Neural Networks (CNNs) Szegedy et al. (2016); He et al. (2016). Transformers use self-attention to model global interactions between tokens Vaswani et al. (2017); Dosovitskiy et al.

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.

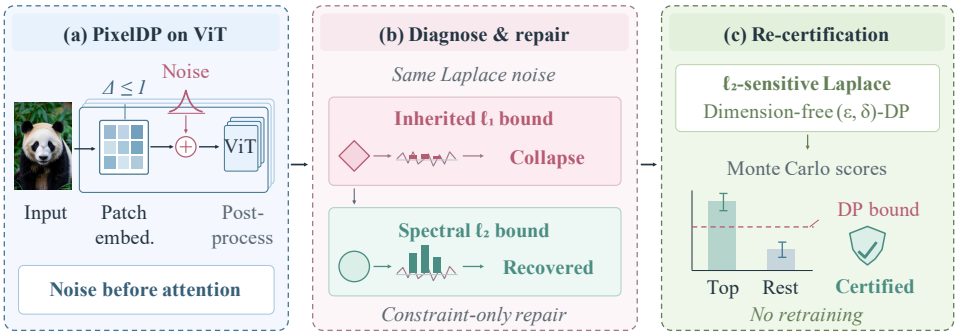


Figure 1: The framework of scaling PixelDP on ViT.

(2021). The recent scaling law Kaplan et al. (2020) suggests that increasing model and data scale can improve predictive performance Kaplan et al. (2020). Transformers can be trained on huge amounts of data and are hardware friendly for large-scale parallel processing, making them widely used as the backbone for building general artificial intelligence base models Bubeck et al. (2023). The large language model serving the public in the cloud Guo et al. (2025) and the general-purpose robot operating at the edge Kim et al. (2025) are built on this transformer architecture that can predict the next token. Nevertheless, it is still insufficient to explore the robustness of defense with differential privacy in this new architecture.

To bridge this gap, we rethink the certified defense with differential privacy on the vision transformer architectures. Our study aims to resolve such a research problem:
How can PixelDP be ported to self-attention architectures while preserving a valid and useful robustness certificate?

This paper revisits the pipeline in the new setting, as illustrated in Fig. 1. Addressing this problem requires resolving two technical challenges. First, standard dot-product self-attention is not globally Lipschitz on an unbounded domain. We therefore inject noise immediately after the patch embedding, a linear layer whose sensitivity can be explicitly controlled, exactly as PixelDP treats a first convolution. Our experiments confirm the distinction sharply. Noise placed after an attention block is absorbed by training and leaves accuracy at the baseline for every noise level, which certifies nothing, whereas noise after the patch embedding behaves as the theory prescribes. Second, PixelDP's Gaussian arm degrades gracefully under the faithful port, while the Laplace arm may collapse to chance accuracy even at the smallest noise level. The failure is not a matter of noise magnitude, since the Gaussian mechanism injects a larger standard deviation at the same attack bound and trains without difficulty. Nor is it intrinsic to the Laplace distribution; the original PixelDP pipeline trains Laplace models on the same grid without incident, on convolutions. The culprit is the grouped convolution bound inherited by the Laplace route. For a non-overlapping patch embedding, this bound aggregates all spatial kernel locations of an input channel and is substantially more conservative than the exact stride-aware $\ell_1 \to \ell_1$ operator norm. We prove a random-matrix law showing that, relative to the spectral projection used by the Gaussian route, the resulting attenuation scales as $C/(\sqrt{M} + \sqrt{N})$ under the Gaussian initialization model. At the measured ViT patch embedding, this large norm ratio predicts a severe drop in the local signal-to-noise ratio, which makes optimization substantially harder and accounts for the failure we observe across all training recipes. This diagnosis suggests its own experiment. We decouple the constraint from the mechanism and train Laplace models under the spectral constraint of the Gaussian route, changing nothing else. The spectral repair also restores trainability across CIFAR-10, CIFAR-100, and SVHN. The grouped-bound failure also persists at ImageNet scale on the autoencoder configuration of the original pipeline. The causal chain is complete, but the repair is not yet a defense. The Laplace mechanism's $(\varepsilon, 0)$ guarantee is proved from an ℓ_1 sensitivity bound, the spectral constraint provides an ℓ_2 bound, and the naive conversion between the two costs a factor of $\sqrt{d}$ in the privacy budget, about 111 at our noise layer, which would demand a hundredfold larger noise.

To close this certification gap in PixelDP, we specialize the existing ℓ_2-sensitivity Laplace guarantee to our setting and give a direct privacy-loss derivation for completeness. The analysis framework needs only a closed form for the shifted first absolute moment of a Laplace variable and Hoeffding's inequality applied to the privacy loss, whose coordinates are independent and bounded. The guarantee plugs directly into the PixelDP certification procedure, holds at every attack radius without group privacy, and applies to our trained checkpoints as a re-certification, with no retraining. It also quantifies an honest cost. At matched (ε, δ) the calibrated Laplace noise carries about $1.47\times$ the standard deviation of the Gaussian mechanism, so Gaussian remains the efficient choice for pure ℓ_2 defense, and the value of the theorem is that it closes a gap the original framework left open and turns a diagnostic into a certified configuration.

Finally we ask whether the certified accuracy of the faithful port can be improved rather than merely explained. Following the margin perspective of adversarially trained smoothing Salman et al. (2019), we combine the noise layer with ℓ_2 adversarial training under a matched-budget protocol that isolates the effect of the attack from the training recipe. On all three datasets, adversarial training improves certified accuracy over the nonzero operating region of the matched controls. On SVHN the joint training collapses to the class prior and a two-phase schedule, noisy pretraining followed by adversarial fine-tuning, restores it, which we report as a practical recipe.

Our contributions are the following.

- A faithful port of PixelDP to Vision Transformers with a placement analysis showing that the patch embedding provides a simple and directly enforceable pixel-space sensitivity bound.
- We identify severe over-attenuation under the inherited grouped ℓ_1 bound. A random-matrix law predicts its scale, while exact stride-aware and spectral controls isolate the effect of the sensitivity constraint.
- We connect the repaired spectral configuration to an existing approximate-DP guarantee for coordinatewise Laplace noise under ℓ_2 sensitivity Andersson et al. (2024), derive the corresponding closed-form PixelDP calibration, and show that existing repaired checkpoints can be certified without retraining.
- An ImageNet study across Inception and ViT-Base backbones confirming the post-processing property empirically, and a matched-budget adversarial training ablation showing certified gains at every radius on CIFAR-10, CIFAR-100, and SVHN.

2 Related Work

Certified defenses from noise. PixelDP Lécuyer et al. (2019) connects differential privacy Dwork (2006); Dwork & Roth (2014) to adversarial robustness. It injects calibrated noise after a sensitivity-constrained layer and certifies predictions through the stability of expected scores. Randomized smoothing Cohen et al. (2019) certifies randomized classifiers, with the standard Gaussian construction injecting noise directly at the input and admitting a tighter ℓ_2 certificate. Our goal is therefore diagnostic rather than to outperform specialized input-space smoothing on pure ℓ_2 certification. SmoothAdv Salman et al. (2019) shows that adversarially training the smoothed classifier enlarges its margins and improves certified accuracy, and later work extends smoothing to further settings Alfarra et al. (2022); Rekavandi et al. (2024); Ugare et al. (2024). Gaussian differential privacy Dong et al. (2022) offers an alternative accounting language that we use for a secondary calibration. Our study differs from this line in its object. We do not propose a new smoothing distribution. Instead, we study how PixelDP's inherited mechanism–constraint pairings interact with ViT patch embeddings and how the resulting spectral repair can be certified under ℓ_2 sensitivity.

Robustness of Vision Transformers. The Lipschitz behavior of attention has been studied directly. Self-attention is not Lipschitz on unbounded domains Kim et al. (2021), variants restore control through architectural change Qi et al. (2023), and sharp bounded-domain estimates exist for normalized token sets Castin et al. (2024); Khromov & Singh (2024). These results concern token-space smoothness of the attention map. They do not directly yield the simple, enforceable pixel-space sensitivity bound required by PixelDP: an ℓ_p pixel perturbation generally affects multiple tokens, while existing attention-side bounds depend on domain or architectural assumptions and

do not directly control the full normalization–attention prefix from pixels to tokens. We therefore place the noise before the first attention block, immediately after the patch embedding, whose sensitivity can be explicitly constrained. On the empirical side, adversarial training recipes for ViTs have matured Mo et al. (2022); Debenedetti et al. (2023), and we adopt the warmup and weak augmentation guidance of the light recipe Debenedetti et al. (2023) for our ablation, with robust overfitting handled by early stopping Rice et al. (2020) and adaptive evaluation following standard guidance Athalye et al. (2018); Croce & Hein (2020); Carlini & Wagner (2017).

Privacy and robustness together. Differentially private training protects the training set Abadi et al. (2016) and can be combined with smoothing for joint guarantees Wu et al. (2024); Lyu et al. (2024). PixelDP and our work use privacy as an analysis tool for inference-time robustness rather than for training data. The two goals are orthogonal and compatible.

3 Porting PixelDP to Vision Transformers

3.1 Where the Noise Can Go

The construction of Sec. A needs a pre-noise map g whose sensitivity can be bounded and enforced during training. Standard dot-product self-attention is not globally Lipschitz on an unbounded token domain Kim et al. (2021). Although finite bounds can in principle be derived on bounded domains, available attention-side bounds are architecture- and domain-dependent and do not directly provide the simple, enforceable pixel-space sensitivity bound required by our PixelDP protocol. We therefore place the noise immediately after the patch embedding, whose sensitivity can be controlled explicitly.

What remains is the patch embedding, a single strided convolution that is linear in the input. We therefore inject the noise immediately after the patch embedding and enforce its sensitivity to one after every optimizer step, exactly as PixelDP treats a first convolution. This choice has a convenient consequence. Because the noise sits directly after the linear embedding whose $\Delta_{2,2}$ equals one, token-space guarantees transfer to pixel space with constant exactly one, and no pullback factor appears. The patch embedding is therefore a simple and directly auditable placement for our PixelDP port. When we instead add noise after the first attention block, training absorbs the perturbation completely, accuracy stays at the noise-free baseline for every noise level on SVHN, and nothing is certified because no sensitivity bound exists. We report this absorption experiment in the supplement as evidence that the placement analysis is not a formality.

3.2 Sensitivity-Induced Signal Collapse

Under the faithful port a sharp asymmetry appears. The Gaussian arm degrades gracefully along the standard grid $L \in \{0.03, 0.1, 0.3, 1\}$, while the Laplace arm collapses to chance at $L = 0.03$ on every dataset (Fig. 4). Noise magnitude does not explain this. At the same numerical radius L, the Gaussian ℓ_2 route injects standard deviation $2.54L$, while the Laplace ℓ_1 route injects $1.41L$. This comparison is a noise-amplitude diagnostic, not a like-for-like comparison of the two threat sets. The mechanism itself does not explain it either, since the original paper trains Laplace models on the same grid on convolutions. The difference is the constraint that each mechanism inherits: a conservative grouped ℓ_1 bound for Laplace and spectral ℓ_2 control for Gaussian.

Let $s_{\mathrm{grp}}(W) = \max_j \sum_{m,a,b} |W_{m,j,a,b}|$ denote the inherited grouped bound of Eq. (10). Renormalizing $s_{\mathrm{grp}}(W)$ to one aggregates $C = Mk^2$ coefficients per input-channel group. Renormalizing the spectral norm divides it by $\sigma_{\max}$, which grows only like $\sqrt{M} + \sqrt{N}$. This gap predicts severe attenuation under the inherited grouped bound.

Proposition 3.1 (Norm-ratio law)**.**

$$\rho \geq \frac{C\sqrt{2/\pi} - \sqrt{2C\ln(1/\delta_1)}}{\sqrt{M} + \sqrt{N} + \sqrt{2\ln(1/\delta_2)}}. \tag{1}$$

Thus the lower bound has leading-order scale

$$\rho = \Omega\left(\frac{C}{\sqrt{M} + \sqrt{N}}\right). \tag{2}$$

Motivated by the leading Gaussian scales of the numerator and denominator, we use

$$\rho_{\text{pred}} = \frac{C\sqrt{2/\pi}}{\sqrt{M} + \sqrt{N}} \tag{3}$$

as an initialization-scale diagnostic.

The proof, in the supplement, uses Gaussian concentration of the column mass and the Davidson–Szarek bound on the spectral norm Vershynin (2018). The leading constant is accurate at initialization and provides a useful architecture-level scale. Evaluating $C\sqrt{2/\pi}/(\sqrt{M} + \sqrt{N})$ at the three first layers in our study predicts 12.5, 34.8, and 118, and direct measurement on the same weights gives 13, 38, and 123 (Fig. 2).

Corollary 3.1 (Linear-component attenuation)**.** *Consider the same weight matrix W under the two scalar projections*

$$W_{\text{grp}} = W/s_{\text{grp}}(W), \qquad W_{\text{spec}} = W/\sigma_{\max}(W).$$

For every input x,

$$W_{\text{grp}}x = \rho_{\text{grp}}^{-1} W_{\text{spec}}x, \qquad \rho_{\text{grp}} = \frac{s_{\text{grp}}(W)}{\sigma_{\max}(W)}.$$

Hence, against the same additive noise scale, any homogeneous signal-to-noise measure applied to the weight-dependent linear component is reduced by a factor ρ_{grp} under the first projection. This statement does not include additive biases or subsequent positional embeddings.

At the ViT patch embedding, $\rho \approx 123$ pushes the projected signal to about 4% of the injected noise at $L = 0.03$, while the spectral projection leaves it at 2.4 times the noise. This severe attenuation makes optimization substantially harder and is consistent with the failure observed across the training recipes we tested.

The diagnosis implies a causal test. Decouple the constraint from the mechanism and train the Laplace mechanism under the spectral constraint, changing nothing else. This diagnosis predicts that replacing the inherited $\Delta_{1,1}$ constraint with the spectral constraint should substantially improve trainability. The observed recovery across datasets supports this constraint-based explanation (Sec. 4.2). The repaired configuration, however, is not yet certified, because the Laplace guarantee of Sec. A is proved from an ℓ_1 bound that the spectral projection does not provide. The naive conversion $\|v\|_1 \le \sqrt{d}\,\|v\|_2$ costs a factor $\sqrt{d}$ in the privacy budget, and at our noise layer $d = 192 \times 8 \times 8$ gives $\sqrt{d} \approx 110.9$, which would demand a hundredfold larger noise. This is also, we believe, the quantitative reason the original framework never offered a 2-norm Laplace configuration.

3.3 A Dimension-Free Certificate for Laplace under ℓ_2 Sensitivity

Our main theoretical result removes the $\sqrt{d}$ obstacle. Throughout, $M(x) = g(x) + Z$ with $Z_i \sim \text{Lap}(b)$ i.i.d., and

$$v = g(x') - g(x)$$

denotes the output shift.

Lemma 3.1 (Shifted absolute moment)**.** *If $z \sim \text{Lap}(0, b)$ then $\mathbb{E}\,|z - v| = |v| + b\,e^{-|v|/b}$ for every $v \in \mathbb{R}$, and consequently*

$$\mathbb{E}\big[\,|z - v| - |z|\,\big] \;=\; b\,\varphi\Big(\frac{|v|}{b}\Big) \;\le\; \frac{v^2}{2b}, \qquad \varphi(t) = t + e^{-t} - 1. \tag{4}$$

Lemma 3.2 (Tail bound implies DP)**.** *Fix x, x' and let $\text{PL}(y) = \ln\big(p_{M(x)}(y)/p_{M(x')}(y)\big)$ be the privacy loss. If $\Pr_{y\sim M(x)}[\text{PL}(y) > \varepsilon] \le \delta$, then Eq. (7) holds for this pair.*

Approximate-DP guarantees for coordinatewise Laplace noise under an ℓ_2 sensitivity bound have been established previously Andersson et al. (2024). We specialize this result to the PixelDP setting and give a direct privacy-loss derivation for completeness. The role of the result here is not a new Laplace privacy theorem, but to close the certification gap created by replacing the original $\Delta_{1,1}$ constraint with the spectral $\Delta_{2,2}$ constraint.

Theorem 3.1 (Laplace mechanism under ℓ_2 sensitivity)**.** *Let $M(x) = g(x) + Z$ with $Z \in \mathbb{R}^d$, $Z_i \sim \mathrm{Lap}(b)$ i.i.d. For any pair x, x' with $\|g(x) - g(x')\|_2 \le S$ and any $\delta \in (0,1)$, the pair satisfies Eq. (7) with*

$$\varepsilon \;=\; \frac{S^2}{2b^2} \;+\; \frac{S}{b}\sqrt{2\ln(1/\delta)}, \tag{5}$$

independent of the dimension d.

Proof sketch. The privacy loss decomposes coordinatewise, $\mathrm{PL} = \frac{1}{b}\sum_i \big(|z_i - v_i| - |z_i|\big)$ with $z \sim \mathrm{Lap}(b)^d$. Each term is independent and bounded in $[-|v_i|/b,\ |v_i|/b]$ by the triangle inequality, and Lemma 3.1 bounds the mean by $\mathbb{E}[\mathrm{PL}] \le \|v\|_2^2/(2b^2) \le S^2/(2b^2)$. Hoeffding's inequality over the d independent terms, whose squared ranges sum to at most $4S^2/b^2$, gives $\Pr[\mathrm{PL} \ge \mathbb{E}[\mathrm{PL}] + \tau] \le \exp\big(-\tau^2 b^2/(2S^2)\big)$. Choosing $\tau = (S/b)\sqrt{2\ln(1/\delta)}$ and applying Lemma 3.2 completes the proof. Full details are in the supplement. □

The dimension cancels because the fluctuation of the privacy loss is controlled by $\|v\|_2$, the same quantity the spectral projection constrains, rather than by $\|v\|_1$. The guarantee holds directly at every shift size, so certification needs no group privacy.

Corollary 3.2 (Calibration and certification)**.** *Enforce $\Delta_{2,2} \le 1$ at the noise layer. To meet a target (ε_0, δ) at construction bound L, set*

$$b = \frac{L}{u}, \qquad u = \sqrt{2\ln(1/\delta) + 2\varepsilon_0} - \sqrt{2\ln(1/\delta)}. \tag{6}$$

At any candidate radius L' the deployed mechanism satisfies $(\varepsilon(L'), \delta)$-DP with $\varepsilon(L') = L'^2/(2b^2) + (L'/b)\sqrt{2\ln(1/\delta)}$, which plugs directly into Proposition A.1.

Cost and consequences. At $(\varepsilon_0, \delta) = (1, 0.05)$, Eq. (6) gives $b = 2.64\,L$ and noise standard deviation $\sqrt{2}\,b = 3.73\,L$, against $2.54\,L$ for the Gaussian mechanism, a factor of 1.47. The Gaussian mechanism therefore remains the efficient choice when only ℓ_2 robustness is wanted. The value of Theorem 3.1 is threefold. It closes the certification gap that the repair opened, turning every spectrally constrained Laplace checkpoint into a certified model by re-running only the certification pass. It replaces the naive factor $\sqrt{d} \approx 110.9$ with no dimension dependence at all, a $42\times$ noise reduction at our layer at $\varepsilon_0 = 1$. And it explains the shape of the original design space, in which the 2-norm Laplace row was absent for exactly this reason.

3.4 A Gaussian-DP Calibration Variant

Our tables also report a secondary accounting based on Gaussian differential privacy Dong et al. (2022). A Gaussian mechanism with ℓ_2 sensitivity S and scale σ is exactly μ-GDP with $\mu = S/\sigma$. The variant fixes a target μ_0 at construction, sets $\sigma = L/\mu_0$, and certifies at radius L' by converting $\mu(L') = L'/\sigma$ to (ε, δ) through the duality of Dong et al. (2022) at the same δ, then applying Proposition A.1. Laplace rows under this accounting match the ε implied by (μ_0, δ) at construction. We emphasize that this fixes the privacy level at the construction bound and is an alternative calibration rather than the group-privacy scaling of the original framework. Exact formulas are in the supplement.

3.5 Certification Meets Adversarial Training

Proposition A.1 depends only on the deployed mechanism, not on how the network was trained, so any training procedure that enlarges the score margins under noise raises certified accuracy with the certificate untouched. Following the margin view of adversarially trained smoothing Salman et al. (2019), we train the noisy network with ℓ_2 projected gradient descent whose inner attack uses expectation-over-transformation gradients Athalye et al. (2018), and we adopt the warmup and weak augmentation guidance of Debenedetti et al. (2023). Two protocol choices matter for a fair reading. First, gains must be measured against a matched control that shares every training hyperparameter except the attack, since recipe differences alone move certified accuracy by several points. Second, on class-imbalanced data the joint objective can collapse to the class prior, because the attack suppresses the slow escape of noisy training from the prior basin. A two-phase schedule,

Table 1: Accuracy (%) of ResNet-50 and ViT under standard DP and f-DP accounting. Gaussian noise uses the $\Delta_{2,2}$ constraint, while Laplace noise uses either the inherited grouped $\bar{\Delta}_{1,1}$ bound or the spectral $\Delta_{2,2}$ constraint. The ViT results correspond to the 400-epoch runs. Values in parentheses denote the corresponding non-private baseline accuracies. SVHN chance level is the majority-class prior of 19.60%.

Dataset	Backbone	Mechanism	L=0.03		L=0.10		L=0.30		L=1.00	
			DP	f-DP	DP	f-DP	DP	f-DP	DP	f-DP
CIFAR-10	ResNet-50 (94.48)	Gaussian, $\Delta_{2,2}$	91.62	91.65	82.99	83.48	68.10	68.13	45.98	46.29
		Laplace, $\bar{\Delta}_{1,1}$	85.26	78.37	79.19	66.06	65.26	51.86	49.38	28.91
	ViT (92.92)	Gaussian, $\Delta_{2,2}$	84.81	84.00	70.17	70.79	42.98	42.94	21.58	22.33
		Laplace, $\bar{\Delta}_{1,1}$	28.87	19.21	18.48	12.26	11.67	10.50	10.38	10.35
		Laplace, $\Delta_{2,2}$	85.50	81.68	80.28	56.69	54.66	33.41	31.31	18.44
CIFAR-100	ResNet-50 (73.28)	Gaussian, $\Delta_{2,2}$	68.66	68.96	57.47	57.01	39.47	40.39	22.25	22.43
		Laplace, $\bar{\Delta}_{1,1}$	53.64	43.60	44.54	36.59	36.04	22.99	21.74	9.36
	ViT (72.79)	Gaussian, $\Delta_{2,2}$	56.18	55.99	39.67	40.47	19.85	19.97	5.30	5.48
		Laplace, $\bar{\Delta}_{1,1}$	7.51	3.58	3.39	1.10	1.22	1.05	1.08	1.04
		Laplace, $\Delta_{2,2}$	59.92	53.89	52.17	29.71	29.71	12.70	10.96	3.60
SVHN	ResNet-50 (97.12)	Gaussian, $\Delta_{2,2}$	96.72	96.73	94.91	95.00	85.23	85.42	50.79	51.41
		Laplace, $\bar{\Delta}_{1,1}$	96.76	95.73	94.18	88.95	88.48	19.59	19.60	19.60
	ViT (97.01)	Gaussian, $\Delta_{2,2}$	95.17	95.36	83.53	81.08	19.59	41.86	19.59	19.59
		Laplace, $\bar{\Delta}_{1,1}$	19.59	19.59	19.59	19.60	19.60	19.59	19.60	19.59
		Laplace, $\Delta_{2,2}$	96.36	93.13	92.18	64.16	60.09	19.60	19.59	19.59

noisy pretraining followed by adversarial fine-tuning from the shared checkpoint, removes the failure and yields the cleanest single-variable comparison. Section D.5 reports both phenomena.

4 Experiments

4.1 Setup

Datasets and models. We evaluate on CIFAR-10, CIFAR-100 Krizhevsky (2009), SVHN Netzer et al. (2011), and ImageNet Deng et al. (2009). The ViT experiments use ViT-Tiny with patch size 4 at resolution 32, trained from scratch, with noise after the patch embedding. CIFAR models train for 400 epochs with RandAugment and mixup, SVHN for 300 epochs with RandAugment only. The convolutional reference is ResNet-50 with a CIFAR stem (3×3 first convolution), noise after the first convolution. ImageNet follows the autoencoder configuration of Lécuyer et al. (2019), a three-layer tied-weight autoencoder with noise after its first convolution (10×10×32, so C=3200 coefficients per column), trained first, then frozen while a pretrained backbone is fine-tuned on top for 20k steps. We run this with both Inception-v3 Szegedy et al. (2016) and ViT-Base backbones. All pipelines consume raw $[0, 1]$ pixels without input normalization, as the calibration requires.

Mechanisms and accounting. Unless stated otherwise, noise is calibrated at (ε_0, δ)=$(1, 0.05)$ over the grid $L \in \{0.03, 0.1, 0.3, 1\}$ under ℓ_2 attack (ℓ_1 for the $\Delta_{1,1}$ Laplace rows). The f-DP variant of Sec. 3.4 uses μ_0=0.4. Sensitivity is renormalized to one after every optimizer step and verified at load time before any certification.

Measurement protocol. Sweep tables report test accuracy under a single noise draw. Certified tables use n=300 draws, argmax scores, Clopper–Pearson bounds with a union bound at η=0.95, and report the expected prediction. Sweep tables report test accuracy under a single noise draw. Certified tables use $n = 300$ draws, argmax scores, and Clopper–Pearson bounds with a union bound at $\eta = 0.95$. These quantities need not coincide, and we do not mix the two protocols within a table.

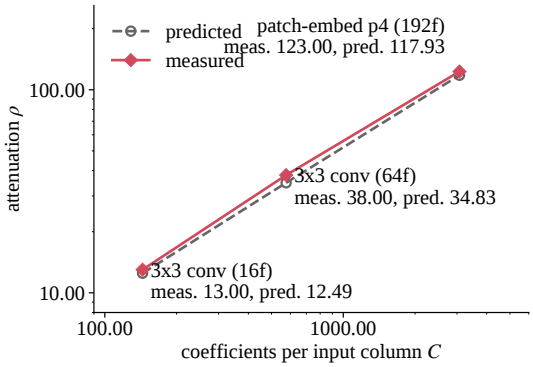


Figure 2: The norm-ratio law. Predicted attenuation $C\sqrt{2/\pi}/(\sqrt{M}+\sqrt{N})$ against direct measurement at the three first layers of our study.

Table 2: ImageNet accuracy (%) under standard DP and f-DP accounting for the autoencoder configuration with two backbones. Gaussian noise remains relatively stable across the two accounting schemes, whereas the $\Delta_{1,1}$ Laplace degradation is substantially more severe under f-DP and persists at C=3200. Values in parentheses denote the corresponding non-private baseline accuracies.

Backbone	Mechanism	L=0.03	L=0.10	L=0.30	L=1.00
Inception-v3 (77.81)	DP Gaussian, $\Delta_{2,2}$	76.11	73.70	68.38	50.86
	f-DP Gaussian, $\Delta_{2,2}$	76.30	73.55	69.34	51.00
	DP Laplace, $\bar{\Delta}_{1,1}$	58.20	36.71	13.99	2.74
	f-DP Laplace, $\bar{\Delta}_{1,1}$	40.66	14.51	3.51	0.64
ViT-Base (81.14)	DP Gaussian, $\Delta_{2,2}$	79.32	76.95	71.73	51.83
	f-DP Gaussian, $\Delta_{2,2}$	79.10	76.87	71.52	52.19
	DP Laplace, $\bar{\Delta}_{1,1}$	60.50	35.92	11.72	1.89
	f-DP Laplace, $\bar{\Delta}_{1,1}$	40.38	12.13	2.47	0.55

4.2 Main Result

Table 1 shows the phenomenon and the causal test together. Under the faithful port the $\Delta_{1,1}$ Laplace configuration is destroyed at the smallest noise level, falling to 28.9% on CIFAR-10, 7.5% on CIFAR-100, and exactly the majority-class prior on SVHN at every level. The same Laplace mechanism, the same noise scale, and the same training recipe under the $\Delta_{2,2}$ constraint recover to within a few points of the Gaussian arm, and on CIFAR-10 and CIFAR-100 exceed it at small L, consistent with the smaller standard deviation the Laplace mechanism injects. The convolutional reference completes the picture. On ResNet-50, whose first convolution has only C=576 coefficients per column, the $\Delta_{1,1}$ Laplace arm trains normally on all three datasets (Table 1), exactly as the norm-ratio law predicts for small C. An absorption control with noise after the first attention block instead of the patch embedding leaves SVHN accuracy at the baseline for every L (supplement), confirming that placement, not architecture, determines whether the mechanism binds.

Figure 2 compares the predicted and measured norm ratios for three first-layer configurations. Using the displayed values, the relative errors with respect to the measurements are approximately 3.9%, 8.3%, and 4.1%. The ImageNet results in Table 2 provide a separate accuracy-based evaluation of a front end with $C = 3200$.

4.3 ImageNet at Scale

Table 2 reports the faithful ImageNet configuration. The Inception baseline of 77.8% matches the 77.5% of the original paper, validating the pipeline. Two observations follow. First, the accuracy drops of the two backbones are nearly identical at every noise level, for instance -4.1 against -4.2 points for Gaussian at L=0.1 and -27.0 against -29.3 at L=1. The certificate lives in the noisy front end and the backbone is post-processing, so a stronger frozen backbone shifts the whole curve

Table 3: Certified accuracy (%) of repaired Laplace models at the nominal construction setting $L = 0.1$. All columns, including $T = 0$, use the Monte Carlo certification protocol with $n = 300$ draws and Theorem 3.1. DP/f-DP row labels identify the checkpoint configuration, not different certification theorems.

	T=0	T=0.02	T=0.03
CIFAR-10 (DP)	77.19	69.12	64.34
CIFAR-10 (f-DP)	55.57	51.76	50.66
CIFAR-100 (DP)	55.01	48.21	46.59
CIFAR-100 (f-DP)	28.22	25.20	24.48
SVHN (DP)	95.02	92.47	91.60
SVHN (f-DP)	76.68	72.44	71.19

up without changing its shape, and ViT-Base dominates Inception at every operating point. Second, the $\Delta_{1,1}$ Laplace collapse reappears at scale on the configuration the original paper itself prescribes, whose first convolution holds 3200 coefficients per column, extending the fan-in series of Fig. 2 to a fourth architecture and a fourth dataset.

4.4 Certified Accuracy of the Repaired Laplace Mechanism

Corollary 3.2 turns every $\Delta_{2,2}$ Laplace checkpoint of Table 1 into a certified model by re-running only the certification pass with the new $\varepsilon(L')$ map. Table 3 reports the protocol. The certified columns are produced by the released code and commands in the supplement. The certified radii of the Laplace route are necessarily smaller than the Gaussian ones at matched (ε_0, δ), both because the calibrated noise is $1.47\times$ larger and because $\varepsilon(L')$ grows superlinearly, and we report them as the honest price of certifying the heavier-tailed mechanism rather than as a new state of the art.

5 Conclusion

We ported PixelDP to Vision Transformers and found that the port is a stress test the original framework had never received. The placement analysis pins the noise to the patch embedding, the inherited grouped ℓ_1 Laplace configuration collapses there, stride-aware control recovers much of the lost accuracy, and a closed-form norm-ratio law explains the grouped-bound attenuation across architectures. The repair created a certification gap, and the certified (ε, δ) guarantee for the Laplace mechanism under ℓ_2 sensitivity closes it with a self-contained proof and a closed-form calibration, at an honestly quantified noise cost against the Gaussian mechanism. ImageNet experiments with a fixed autoencoder front end and two downstream backbones illustrate the post-processing property empirically, and a matched-budget ablation shows that adversarial training improves certified accuracy over the nonzero operating region.

Limitations. Our ViT models are small and trained from scratch at low resolution, so absolute accuracies sit below modern large-scale recipes, and the study isolates mechanisms rather than chasing state of the art. The Laplace certificate costs about $1.47\times$ the Gaussian noise at matched (ε_0, δ), so it removes an inconsistency in the design space rather than improving the efficient frontier. The GDP rows are an alternative calibration that fixes the privacy level at the construction bound and should not be read as the group-privacy scaling of the original framework. Training results are single-seed, so we do not quantify optimization variance. We therefore restrict our empirical claim to the large failure-and-repair effects reproduced across datasets and architectures, rather than small differences between individual configurations. We study ℓ_1 and ℓ_2 threats only, and the interaction of the noise layer with ℓ_∞ training is untouched.

AI use statement

The idea of this study was designed by humans, and early programming was also done by humans. The LLMs helped to correct programming bugs, polish text, and explore new insights in mathematics. Our human author team is responsible for the research results.

ETHICS STATEMENT

Our research primarily focuses on certified robustness related to AI security. All experiments use publicly available academic datasets, including CIFAR-10, CIFAR-100, SVHN, and ImageNet.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pp. 64–74, 2022.

Joel Daniel Andersson, Rasmus Pagh, Teresa Anna Steiner, and Sahel Torkamani. Count on your elders: Laplace vs gaussian noise. *arXiv preprint arXiv:2408.07021*, 2024.

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, 2018.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, 2017.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *International Conference on Machine Learning*, pp. 5817–5840, 2024.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *international symposium on privacy enhancing technologies symposium*, pp. 82–102, 2013.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320, 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, 2020.

Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *2023 IEEE conference on secure and trustworthy machine learning (SaTML)*, pp. 225–253, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12, 2006.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Grigory Khromov and Sidak Pal Singh. Some fundamental aspects about lipschitz continuity of neural networks. In *ICLR*, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, pp. 2679–2713, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

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

Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*, pp. 1289–1310, 2023.

Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. *Advances in Neural Information Processing Systems*, 37:134043–134074, 2024.

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

Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. In *Advances in Neural Information Processing Systems*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers. In *ICLR*, 2023.

Aref M Rekavandi, Farhad Farokhi, Olga Ohrimenko, and Benjamin I Rubinstein. Certified adversarial robustness via randomized $alpha$-smoothing for regression models. *Advances in Neural Information Processing Systems*, 37:134127–134150, 2024.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pp. 8093–8104, 2020.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. In *ICLR*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Taowen Wang, Cheng Han, James Liang, Wenhao Yang, Dongfang Liu, Luna Xinyu Zhang, Qifan Wang, Jiebo Luo, and Ruixiang Tang. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6948–6958, 2025.

Jiapeng Wu, Atiyeh Ashari Ghomi, David Glukhov, Jesse C Cresswell, Franziska Boenisch, and Nicolas Papernot. Augment then smooth: Reconciling differential privacy with certified robustness. *Transactions on Machine Learning Research*, 2024.

Jun Yan, Huilin Yin, Ziming Zhao, Wancheng Ge, and Jingfeng Zhang. Enhance adversarial robustness via geodesic distance. *IEEE Transactions on Artificial Intelligence*, 5(8):4202–4216, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *ICLR*, 2020.

A Preliminaries

A.1 Differential Privacy over a Norm Metric

Differential privacy bounds how much a randomized computation can change when its input changes a little. PixelDP uses the definition over a general metric Chatzikokolakis et al. (2013), taking the metric to be an ℓ_p norm on images rather than record replacement on databases.

Definition A.1 ((ε, δ)-DP over an ℓ_p metric)**.** *A randomized mechanism $M : \mathbb{R}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP at scale L in ℓ_p if for all x, x' with $\|x - x'\|_p \leq L$ and all measurable $T \subseteq \mathcal{Y}$,*

$$\Pr[M(x) \in T] \;\leq\; e^{\varepsilon}\, \Pr[M(x') \in T] + \delta. \tag{7}$$

Two properties drive everything that follows. Post-processing preserves the guarantee: any function of $M(x)$ is again (ε, δ)-DP. Bounded outputs have stable expectations, which is the bridge to robustness.

Lemma A.1 (Expected output stability Lécuyer et al. (2019)). *If M satisfies Def. A.1 and $M(x) \in [0,1]$ almost surely, then for all $\|\alpha\|_p \le L$, $\mathbb{E}[M(x)] \le e^{\varepsilon}\,\mathbb{E}[M(x+\alpha)] + \delta$.*

A.2 The PixelDP Construction

Write the scoring function of a network as $Q = h \circ g$, where g is the computation before a chosen noise layer and h the computation after it. PixelDP constrains the sensitivity of g,

$$\Delta_{p,q} \;=\; \sup_{x \neq x'} \frac{\|g(x) - g(x')\|_q}{\|x - x'\|_p}, \tag{8}$$

and injects noise calibrated to $\Delta_{p,q}$ and to a construction bound L, producing the randomized scores $A(x) = h\big(g(x) + Z\big)$. Two mechanisms are used. The Laplace mechanism draws $Z_i \sim \mathrm{Lap}(b)$ with $b = \Delta_{p,1} L/\varepsilon$ and yields $(\varepsilon, 0)$-DP from an ℓ_1 sensitivity bound. The Gaussian mechanism draws $Z_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \sqrt{2\ln(1.25/\delta)}\,\Delta_{p,2} L/\varepsilon$ and yields (ε, δ)-DP from an ℓ_2 bound, for $\varepsilon \le 1$. By post-processing, A inherits the guarantee of the noised layer, provided no skip connection bypasses the noise.

For the non-overlapping ViT patch embedding (kernel size equal to stride), the exact induced $\ell_1 \to \ell_1$ norm of the flattened patch map is

$$\Delta_{1,1}^{\mathrm{exact}}(W) = \max_{j,a,b} \sum_{m} |W_{m,j,a,b}|. \tag{9}$$

The original PixelDP-style convolution treatment instead uses the grouped quantity

$$\bar{\Delta}_{1,1}(W) = \max_{j} \sum_{m,a,b} |W_{m,j,a,b}|, \tag{10}$$

which is a valid but generally conservative upper bound for this non-overlapping patch operator. Our faithful-port experiments retain this inherited bound; consequently, the collapse diagnosed below should be interpreted as a property of this conservative PixelDP constraint rather than of the exact stride-aware $\ell_1 \to \ell_1$ sensitivity.

A.3 Certified Prediction

At inference the expectation $\mathbb{E}[A(x)]$ is estimated by Monte Carlo over n noise draws, with per-class confidence bounds $\hat{\mathbb{E}}^{lb}, \hat{\mathbb{E}}^{ub}$ obtained from Hoeffding or Clopper–Pearson intervals Clopper & Pearson (1934) and a union bound over classes.

Proposition A.1 (Generalized robustness condition Lécuyer et al. (2019)). *Suppose A satisfies Def. A.1 at scale L in ℓ_p. If for some class k*

$$\hat{\mathbb{E}}^{lb}\big(A_k(x)\big) \;>\; e^{2\varepsilon} \max_{i \neq k} \hat{\mathbb{E}}^{ub}\big(A_i(x)\big) + \big(1 + e^{\varepsilon}\big)\delta, \tag{11}$$

then the prediction $\arg\max_k \hat{\mathbb{E}}(A_k(x))$ is robust to every perturbation of ℓ_p size at most L, with probability at least η over the estimation.

Given a deployed noise scale, the largest L for which Eq. (11) holds is returned as the robustness certificate of the input, searching over L through the map $L \mapsto \varepsilon(L)$ of the mechanism. Certified accuracy at threshold T is the fraction of test points that are correct and certified at radius at least T.

A.4 Relation to Randomized Smoothing

Randomized smoothing Cohen et al. (2019) is the input-noise special case of this construction with Gaussian noise, for which a tighter certificate is available. Placing the noise after a trained layer, which is what makes PixelDP flexible, is also what creates the interaction we study, because the sensitivity constraint then reshapes the layer that carries the signal.

B PROOFS

B.1 PROOF OF PROPOSITION 3.1 (NORM-RATIO LAW)

Let W have i.i.d. $\mathcal{N}(0, \tau^2)$ entries. Its flattened patch operator has size $M \times N$ with $N = C_{\text{in}}k^2$. For each input channel, the inherited grouped bound aggregates $C = Mk^2$ weights across output channels and spatial kernel locations.

Grouped ℓ_1 mass. For any fixed input-channel group, the column mass $f(w) = \sum_{i=1}^{C} |w_i|$ is $\sqrt{C}$-Lipschitz with respect to the Euclidean norm, since

$$|f(w) - f(w')| \le \|w - w'\|_1 \le \sqrt{C}\|w - w'\|_2.$$

For $w = \tau g$ with $g \sim \mathcal{N}(0, I)$, Gaussian concentration therefore gives

$$\Pr[f(w) < \mathbb{E}f(w) - t] \le \exp\left(-\frac{t^2}{2C\tau^2}\right).$$

Spectral norm. By the Davidson–Szarek bound Vershynin (2018), for a Gaussian matrix $\Pr[\sigma_{\max} > \tau(\sqrt{M} + \sqrt{N} + t)] \le e^{-t^2/2}$. Setting this to δ_2 gives $\sigma_{\max} \le \tau(\sqrt{M} + \sqrt{N} + \sqrt{2\ln(1/\delta_2)})$ with probability $1 - \delta_2$.

Ratio. Dividing the two bounds and applying a union bound proves Eq. (1). The common τ cancels, so the ratio is scale free, and the leading order is $\rho = \Omega\big(C/(\sqrt{M} + \sqrt{N})\big)$. For a square-ish matricization $N = \Theta(M)$ this is $\Theta(C/\sqrt{M}) = \Theta(\sqrt{C}\,k)$, increasing in both the filter count and the kernel size. □

B.2 PROOF OF LEMMA 3.1 (SHIFTED ABSOLUTE MOMENT)

For $z \sim \text{Lap}(0, b)$ with density $\frac{1}{2b}e^{-|z|/b}$ and $v \ge 0$ (the case $v < 0$ is symmetric),

$$\mathbb{E}\,|z - v| = \int_{-\infty}^{\infty} |z - v| \tfrac{1}{2b} e^{-|z|/b}\, dz.$$

Splitting at $z = 0$ and $z = v$ and integrating each piece in closed form gives $\mathbb{E}\,|z - v| = v + b\,e^{-v/b}$. Since $\mathbb{E}\,|z| = b$,

$$\mathbb{E}\big[|z - v| - |z|\big] = v - b + b\,e^{-v/b} = b\,\varphi(v/b), \quad \varphi(t) = t - 1 + e^{-t}.$$

Finally $\varphi(t) \le t^2/2$ for all $t \ge 0$, because $\psi(t) = t^2/2 - \varphi(t)$ satisfies $\psi(0) = 0$ and $\psi'(t) = t - 1 + e^{-t} = \varphi(t) \ge 0$. Substituting $t = |v|/b$ gives $\mathbb{E}[|z - v| - |z|] \le v^2/(2b)$. □

B.3 PROOF OF LEMMA 3.2 (TAIL IMPLIES DP)

Let $S_\varepsilon = \{y : \text{PL}(y) \le \varepsilon\}$. For any event T,

$$\Pr_{M(x)}[T] = \Pr_{M(x)}[T \cap S_\varepsilon] + \Pr_{M(x)}[T \cap S_\varepsilon^c].$$

On S_ε the density ratio is at most e^ε, so the first term is at most $e^\varepsilon \Pr_{M(x')}[T]$. The second term is at most $\Pr_{M(x)}[S_\varepsilon^c] \le \delta$ by hypothesis. Adding gives Eq. (7). □

B.4 PROOF OF THEOREM 3.1

Write

$$v = g(x') - g(x), \qquad \|v\|_2 \le S,$$

and let $y = g(x) + z$ with $z \sim \text{Lap}(b)^d$. The privacy loss is

$$\text{PL} = \sum_{i=1}^{d} \frac{|y_i - g_i(x')| - |y_i - g_i(x)|}{b} = \frac{1}{b}\sum_{i=1}^{d}\big(|z_i - v_i| - |z_i|\big).$$

They are independent. By the triangle inequality each lies in $[-|v_i|, |v_i|]/b$, so the i-th term has range $2|v_i|/b$ and $\sum_i (2|v_i|/b)^2 = 4\|v\|_2^2/b^2 \le 4S^2/b^2$. By Lemma 3.1,

$$\mathbb{E}[\mathrm{PL}] = \frac{1}{b}\sum_i b\,\varphi(|v_i|/b) \le \frac{1}{2b^2}\sum_i v_i^2 = \frac{\|v\|_2^2}{2b^2} \le \frac{S^2}{2b^2}.$$

Hoeffding's inequality for bounded independent variables gives, for $\tau > 0$,

$$\Pr[\mathrm{PL} \ge \mathbb{E}[\mathrm{PL}] + \tau] \le \exp\Big(\frac{-2\tau^2}{\sum_i (2|v_i|/b)^2}\Big) \le \exp\Big(\frac{-\tau^2 b^2}{2S^2}\Big).$$

Set $\tau = (S/b)\sqrt{2\ln(1/\delta)}$, so the right side equals δ. Then with probability at least $1-\delta$,

$$\mathrm{PL} \le \frac{S^2}{2b^2} + \frac{S}{b}\sqrt{2\ln(1/\delta)} = \varepsilon.$$

Lemma 3.2 converts this tail bound into (ε, δ)-DP. No step depends on d. Because (x, x') was an arbitrary ordered neighboring pair, the same argument applies after exchanging x and x'. Thus the bound holds for every ordered pair required by Definition A.1. □

B.5 Proof of Corollary 3.2 (calibration)

Enforcing $\Delta_{2,2} \le 1$ makes $S = L$ at construction bound L. Writing $r = \sqrt{2\ln(1/\delta)}$ and solving $\varepsilon_0 = L^2/(2b^2) + (L/b)r$ for L/b as a quadratic gives $L/b = \sqrt{r^2 + 2\varepsilon_0} - r = u$, hence $b = L/u$. At a candidate radius L' the sensitivity is $S = L'$, and substituting into Eq. (5) yields $\varepsilon(L') = L'^2/(2b^2) + (L'/b)r$. □

C Gaussian-DP Calibration Formulas

A Gaussian mechanism with ℓ_2 sensitivity S and scale σ is μ-GDP with $\mu = S/\sigma$ Dong et al. (2022). The dual (ε, δ) profile is

$$\delta(\varepsilon; \mu) = \Phi\Big(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\Big) - e^{\varepsilon}\Phi\Big(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\Big),$$

with Φ the standard normal CDF. Fixing μ_0 at construction sets $\sigma = L/\mu_0$. At radius L' we have $\mu(L') = L'/\sigma = \mu_0 L'/L$, and inverting $\delta(\varepsilon; \mu(L')) = \delta$ for ε gives the value fed to Proposition A.1. This fixes the privacy level at the construction bound rather than scaling it by group privacy.

For the Laplace rows, let $\epsilon_0^{(\mu)}$ denote the solution of

$$\delta = \Phi\left(-\frac{\epsilon_0^{(\mu)}}{\mu_0} + \frac{\mu_0}{2}\right) - e^{\epsilon_0^{(\mu)}}\,\Phi\left(-\frac{\epsilon_0^{(\mu)}}{\mu_0} - \frac{\mu_0}{2}\right).$$

For the grouped ℓ_1 Laplace mechanism we set

$$b = \frac{L}{\epsilon_0^{(\mu)}}.$$

For the spectrally constrained Laplace mechanism, Corollary 3.2 gives

$$b = \frac{L}{u_0}, \qquad u_0 = \sqrt{2\ln(1/\delta) + 2\epsilon_0^{(\mu)}} - \sqrt{2\ln(1/\delta)}.$$

At radius L', its privacy parameter is

$$\epsilon(L') = \frac{L'^2}{2b^2} + \frac{L'}{b}\sqrt{2\ln(1/\delta)}.$$

Attention-side absorption control. Figure 3 complements the placement analysis of Sec. 3.1. When noise is injected after the first attention block, test accuracy remains close to the non-private baseline across the entire construction grid, under both DP and f-DP accounting and for both noise distributions. The absence of systematic degradation as L increases indicates that training absorbs the perturbation at this location. This behavior should not be interpreted as enhanced robustness: because the pre-noise computation crosses an attention block and is not assigned a certified pixel-space sensitivity bound under our protocol, these models provide no valid PixelDP certificate.

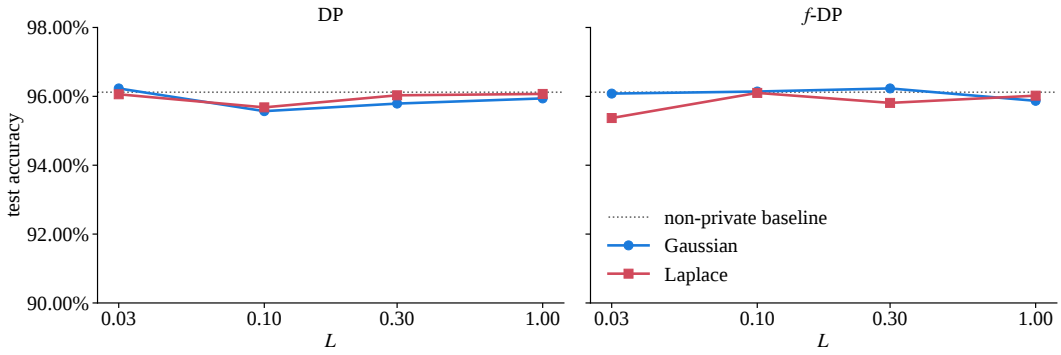


Figure 3: Attention-side absorption control on SVHN. Noise is injected after the first attention block rather than after the patch embedding. Under both standard DP and f-DP accounting, Gaussian and Laplace models remain close to the non-private baseline across the full construction grid. This apparent invariance does not yield a robustness guarantee: no enforceable pixel-to-representation sensitivity bound is available at this placement, so the resulting models are not certifiable.

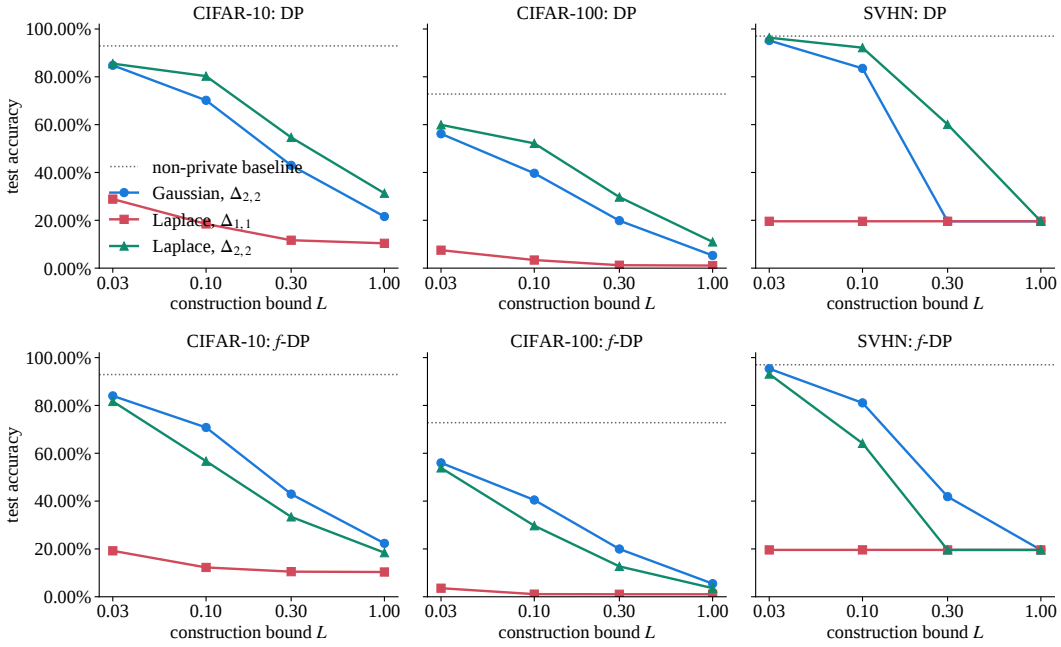


Figure 4: Constraint-decoupling diagnostics for ViT-Tiny. The panels report one-seed test accuracy across three datasets and two noise-calibration variants. The spectral constraint substantially mitigates the low-noise degradation of Laplace models, although failures remain at large noise levels. Exact values are provided in Table 1.

D Additional Empirical Studies

D.1 The constraint-decoupling study

Fig. 4 summarizes the constraint-decoupling experiment across CIFAR-10, CIFAR-100, and SVHN. At $L = 0.03$, the $\Delta_{1,1}$ Laplace configuration reduces accuracy to 28.87% on CIFAR-10 and 7.51% on CIFAR-100, and reaches the majority-class prior on SVHN. Replacing $\Delta_{1,1}$ with $\Delta_{2,2}$ substantially improves low-noise performance. This repair does not eliminate high-noise failures: the SVHN model still reaches the class prior at $L = 1$. Exact sweep values and the ResNet-50 reference results are provided in Table 1.

Table 4: Sensitivity of repaired Laplace certification on CIFAR-10 to the number of Monte Carlo samples n and the certification parameter δ. An independently trained $L = 0.1$ checkpoint is used for this sensitivity study; the same checkpoint and the same 1,000 stored noise draws are used throughout the table.

n	δ	EP Acc.	$T = 0$	$T = 0.02$	$T = 0.03$	$T = 0.05$
300	0.05	81.14	77.12	69.27	64.41	51.65
300	0.01	81.14	78.26	68.62	62.59	47.80
1000	0.05	81.20	78.42	70.99	66.43	56.30
1000	0.01	81.20	79.54	70.34	64.91	52.91

Table 5: Constraint-only control on CIFAR-10/ViT-Tiny with Laplace noise at $L = 0.03$. Only the patch-embedding sensitivity constraint is changed.

Constraint	Test Acc. (%)
Inherited grouped $\bar{\Delta}_{1,1}$	28.87
Exact stride-aware $\Delta_{1,1}^{\text{exact}}$	70.83
Spectral $\Delta_{2,2}$	85.50

D.2 Sensitivity to Monte Carlo sample size and δ

Table 4 varies the certification sample count and δ on the same repaired CIFAR-10 checkpoint. Increasing n from 300 to 1,000 changes expected-prediction accuracy by only 0.06 points, but improves certified accuracy by up to 4.65 points over the reported thresholds, consistent with tighter Clopper–Pearson intervals. Reducing δ from 0.05 to 0.01 has a mixed but predictable effect: it slightly improves the $T = 0$ condition, where $\epsilon(0) = 0$, but reduces certification at positive radii because $\epsilon(T)$ increases through the $\sqrt{2\ln(1/\delta)}$ term. The main qualitative conclusions are therefore unchanged across these settings.

D.3 Norm-ratio evolution during training

Proposition 3.1 is derived for i.i.d. Gaussian weights, so its quantitative prediction is formally an initialization-scale statement. We therefore track $\rho = s_1/\sigma_{\max}$ throughout a 400-epoch CIFAR-10 run. Figure 5 shows that the Gaussian prediction is of the correct scale near initialization, but ceases to be quantitatively tight once optimization induces structure in the patch embedding: the measured ratio decreases from 136.99 at initialization to 53.83 by epoch 20 and remains near 55 thereafter, versus the initialization prediction 117.93. Importantly, the ratio remains large throughout training, so the substantial separation between the two constraint geometries persists even outside the theorem's i.i.d. regime.

Exact stride-aware ℓ_1 control. For the non-overlapping ViT patch embedding, we additionally replace the inherited grouped bound $\bar{\Delta}_{1,1}$ with the exact induced $\ell_1 \to \ell_1$ norm $\Delta_{1,1}^{\text{exact}}$ of Eq. (9), while keeping the Laplace noise scale and training recipe unchanged. At $L = 0.03$, accuracy increases from 28.87% to 70.83%, and spectral $\Delta_{2,2}$ control further reaches 85.50% (Table 5).

On the final exact-norm checkpoint, $\Delta_{1,1}^{\text{exact}} = 1.00$, $\bar{\Delta}_{1,1} = 14.14$, and $\sigma_{\max} = 0.290$, giving $\rho_{\text{exact}} = 3.45$ and $\rho_{\text{grouped}} = 48.78$.

D.4 Comparison with input randomized smoothing

On CIFAR-10, we additionally train a standard input Gaussian randomized-smoothing baseline using the same ViT-Tiny architecture and 400-epoch training budget (Table 6). At $\sigma = 0.2537$, its Monte Carlo prediction accuracy (81.57%) is nearly identical to that of the repaired Laplace model (81.14%), but its specialized smoothing certificate is substantially tighter: CA@0.05 is 75.75% versus 51.65%. This gap is expected and clarifies the scope of our contribution. Our goal is not to outperform input smoothing for pure ℓ_2 certification; rather, PixelDP permits noise to be placed

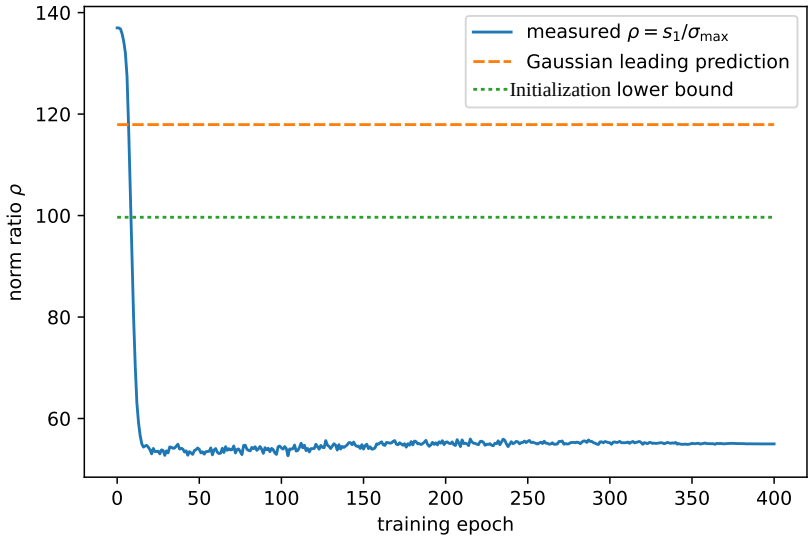


Figure 5: Evolution of the patch-embedding norm ratio $\rho = s_1/\sigma_{\max}$ during CIFAR-10 training. The dashed line is the leading-order Gaussian prediction, while the dotted line is the initialization lower bound. The prediction is initialization-scale rather than a theorem for trained weights: optimization rapidly moves the layer outside the i.i.d. regime, after which ρ stabilizes near 55. The persistent $\rho \gg 1$ nevertheless preserves the qualitative separation between the inherited $\Delta_{1,1}$ and spectral constraints.

Table 6: Comparison with standard input Gaussian randomized smoothing (RS) on CIFAR-10. Both use ViT-Tiny and matched training budgets. RS uses $\sigma = 0.2537$ Gaussian noise at the image input and its standard smoothing certificate; repaired Laplace injects noise after the patch embedding and uses the corresponding PixelDP-style certificate. Pred. Acc. denotes the Monte Carlo expected prediction.

Method	Pred. Acc.	$T = 0$	$T = 0.02$	$T = 0.03$	$T = 0.05$
Input Gaussian RS	81.57	78.72	77.63	77.01	75.75
Repaired Laplace, $\Delta_{2,2}$	81.14	77.12	69.27	64.41	51.65

Table 7: Certified accuracy (%) on CIFAR-10 at L=0.1 with Gaussian noise. The matched noise-only control shares all hyperparameters with the AT arm except for the adversarial attack. The strong recipe uses 400 epochs with mixup and is included as a reference rather than a matched control.

Training setting	T=0.00	T=0.05	T=0.10	T=0.20	T=0.30
Matched noise only	64.65	45.04	32.67	0.00	0.00
AT (PixelDP)	66.62	53.34	43.37	0.00	0.00
AT (f-DP)	66.42	53.28	43.02	0.00	0.00
Strong recipe	74.30	57.66	45.27	0.00	0.00

after an internal representation, and our analysis characterizes the sensitivity-constraint interactions introduced by that flexibility.

D.5 Adversarial Training Lifts the Certificate

Table 7, Table 8, and Fig. 6 give the ablation. Under standard DP calibration, adversarial training improves CIFAR-10 certified accuracy by 8.30 and 10.70 percentage points at $T = 0.05$ and $T = 0.10$. Across the reported positive thresholds up to $T = 0.15$, the gains increase from 2.68 to 4.22 points on CIFAR-100 and from 2.81 to 6.09 points on SVHN. Both CIFAR-10 models have zero

Table 8: Certified accuracy (%) on CIFAR-100 and SVHN at L=0.1 with Gaussian noise. For CIFAR-100, the matched noise-only control shares all hyperparameters with the AT arm except for the adversarial attack, while the strong recipe is included as a reference. For SVHN, the two main branches share a 300-epoch noisy prefix, followed by either 150 additional noise-only epochs or 150 epochs of AT fine-tuning.

Dataset	Training setting	T=0.00	T=0.05	T=0.10	T=0.125	T=0.15
CIFAR-100	Matched noise only	31.28	18.40	12.43	9.40	5.18
	AT (PixelDP)	30.20	21.08	16.08	13.47	9.40
	AT (f-DP)	32.74	22.51	16.80	13.98	9.85
	Strong recipe	42.96	27.63	19.93	15.80	9.98
SVHN	Noise only (PixelDP), 450 epochs	83.48	63.74	46.34	36.41	27.42
	Noise only (f-DP), 450 epochs	84.52	65.52	49.10	39.61	30.76
	AT fine-tuning (PixelDP), 150 epochs	82.90	66.55	51.32	42.40	33.51
	AT fine-tuning (f-DP), 150 epochs	83.42	66.79	52.48	44.10	35.56
	Noise only (PixelDP), 300 epochs	82.63	62.02	44.29	34.80	25.88
	Noise only (f-DP), 300 epochs	86.08	66.38	48.74	38.47	29.12

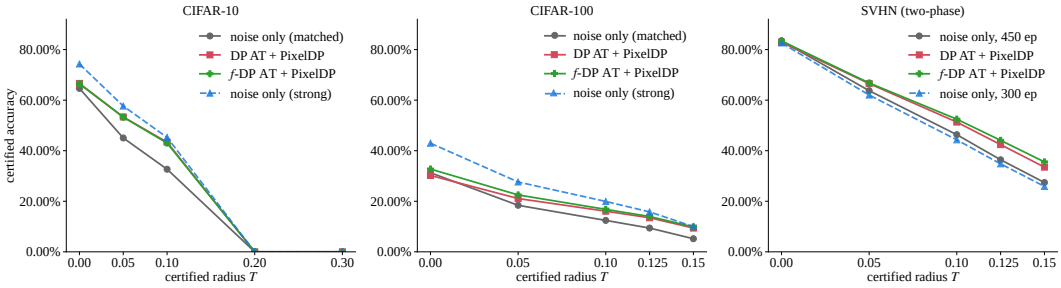


Figure 6: Certified accuracy against radius under the matched protocol. Adversarial training improves certified accuracy across the reported nonzero operating region under the matched protocol.

certified accuracy at $T = 0.20$ and $T = 0.30$. The pattern is the margin mechanism of Salman et al. (2019) observed through a certificate. The attack enlarges the score gap under noise, and the higher the certification threshold, the more of that gap is converted into certified points. Near the certification boundary the matched AT model approaches the strong recipe trained with 2.7× the budget, on CIFAR-100 within 0.6 points at T=0.15. The comparison against the strong recipe also documents a pitfall. Read against that reference alone, adversarial training appears to lose everywhere, and only the matched control reveals that the deficit is the training budget, not the attack.

Two-phase training on SVHN. Training the joint objective from scratch on SVHN collapses to the class prior of 19.6% and stays there for 300 epochs. The noisy gradient signal through the constrained embedding is weak early in training, the class imbalance offers a genuine prior basin, and the inner attack suppresses the slow escape that noise-only training exhibits. Forking a 300-epoch noisy checkpoint into an adversarial fine-tune and a continued noise-only control isolates the attack cleanly and restores the pattern. The collapsed run is also instructive in itself. A constant classifier certifies every one of its majority-class predictions at every radius, so certified accuracy in isolation can be gamed by a degenerate model, and we recommend always reading it jointly with conventional accuracy.

Empirical robustness and sanity checks. Table 9 places the certified arm next to a pure adversarial-training baseline under identical attacks. The AT-only model is empirically stronger, but it has no PixelDP certificate because it lacks the required randomized, sensitivity-controlled mechanism. The table also passes the standard adaptive-evaluation checks. For the deterministic model, BPDA coincides with PGD-20 as it must. For the randomized model, EOT is stronger than PGD and BPDA is weaker, the expected ordering when gradients are averaged over noise Athalye et al. (2018).

Table 9: Empirical ℓ_2 robustness (%) on CIFAR-10 at ε=0.5, evaluated using the best checkpoints. † denotes evaluation of the averaged randomized prediction rule. "–" indicates that the AT-only model has no valid PixelDP certificate.

Attack	AT only	AT + PixelDP
Clean	75.84	66.72
FGSM	55.14	54.95
PGD-20	54.15	51.10
PGD-100	54.10	50.54
CW-100	53.39	49.80
BPDA	54.19	55.26
EOT-PGD	54.17	50.21
AutoAttack	52.26	41.52†
Certified T=0.10	–	43.37

D.6 Takeaways from the Experiment

- Placement is binding. Only the patch embedding supports an enforceable sensitivity bound in a ViT, and attention-side noise is absorbed into a certificate-free model.
- The inherited grouped constraint causes severe attenuation, whose initialization scale follows a closed-form law in the grouped coefficient count from $C = 144$ to $C = 3200$.
- An ℓ_2 certificate for the Laplace mechanism exists, is dimension free, and re-certifies repaired checkpoints without retraining, at a quantified $1.47\times$ noise cost against Gaussian.
- The certificate lives in the front end. Swapping a frozen Inception for a frozen ViT-Base moves ImageNet baselines, not degradation curves.
- Adversarial training improves certified accuracy over the nonzero operating region, but only a matched-budget control isolates the gain.

E Reproduction

Hyperparameters. ViT-Tiny, patch 4, resolution 32, AdamW, weight decay 0.1 (AT) or 0.05 (clean), cosine schedule with linear warmup. CIFAR 400 epochs with RandAugment and mixup, SVHN 300 epochs with RandAugment. Certification uses n=300 draws, argmax scores, Clopper–Pearson at η=0.95. Adversarial training uses ℓ_2 PGD with $\epsilon_{\text{adv}} = 0.5$, $K = 7$ steps, and step size $\alpha = 2.5\epsilon_{\text{adv}}/K$. For PixelDP models, each PGD step uses EOT with two independent noise draws and averages the gradients of the per-draw cross-entropy losses. The attack budget is linearly warmed up over the first 30 epochs.

Certifying repaired Laplace models. With the released code, a $\Delta_{2,2}$ Laplace checkpoint is certified under Theorem 3.1 by the certification pass with `--noise_mech laplace_l2`, which uses the calibration b=L/u of Eq. (6) and the map $\varepsilon(L')$ of Corollary 3.2. The load-time sensitivity guard refuses any checkpoint whose $\Delta_{2,2} \neq 1$, so an accidentally supplied AT-only checkpoint cannot produce an invalid certificate.

Two-phase SVHN. Train a noise-only checkpoint for 300 epochs, then fork into an adversarial fine-tune and a continued noise-only control, both for 150 epochs from the shared weights with reset schedules, so the two forks differ only in the presence of the attack.

AutoAttack protocol. For randomized checkpoints AutoAttack attacks and judges the averaged prediction over 8 draws, and accuracy is reported over 20 draws. Judging AutoAttack per single draw, as an off-the-shelf run does, underestimates the deployed averaged prediction, so we report the averaged-rule number and mark it accordingly.