

A APPENDIX

A.1 ROBUST MODELS

The experimental setup described in this paper (Section 3.1) utilizes pre-trained baseline and robust models obtained from RobustBench (Croce et al., 2021). The goal of RobustBench is to track the progress in adversarial robustness for ℓ_∞ - and ℓ_2 -norm attacks since these are the most studied settings in the literature. We summarize in Table 5 the models we employed for testing the performance of σ -zero. Each entry in the table includes the label reference from RobustBench, the short name we assigned to the model, and the corresponding clean and robust accuracy under the specific threat model. The robustness of these models is evaluated against an ensemble of white-box and black-box attacks, specifically AutoAttack. We also include in our experiments models trained to be robust against ℓ_1 -sparse attacks, i.e., C3 (Croce & Hein, 2021) and C4 (Jiang et al., 2023), as well as two robust models trained to resist ℓ_0 -norm attacks, i.e., C11 (Zhong et al., 2024) and C12 (Zhong et al., 2024). Our experimental setup is designed to encompass a wide range of model architectures and defensive techniques, ensuring a comprehensive and thorough performance evaluation of the considered attacks.

Table 5: Summary of Robustbench models used in our experiments. For each model, we report its reference label in Robustbench, its threat model, and the corresponding clean and robust accuracy.

Dataset	Reference	Model	Threat model	Clean accuracy %	Robust accuracy %
CIFAR-10	Carmon2019Unlabeled	C1 (Carmon et al., 2019)	ℓ_∞	89.69	59.53
	Augustin2020Adversarial	C2 (Augustin et al., 2020)	ℓ_2	91.08	72.91
	Standard	C5 (Croce et al., 2021)	-	94.78	0
	Gowal2020Uncovering	C6 (Gowal et al., 2021)	ℓ_2	90.90	74.50
	Engstrom2019Robustness	C7 (Engstrom et al., 2019)	ℓ_∞ - ℓ_2	87.03 - 90.83	49.25 - 69.24
	Chen2020Adversarial	C8 (Chen et al., 2020)	ℓ_∞	86.04	51.56
	Xu2023Exploring_WRN-28-10	C9 (Xu et al., 2023)	ℓ_∞	93.69	63.89
	Addepalli2022Efficient_RN18	C10 (Addepalli et al., 2022)	ℓ_∞	85.71	52.48
	Standard_R18	I1 (He et al., 2015)	-	76.52	0
	Engstrom2019Robustness	I2 (Engstrom et al., 2019)	ℓ_∞	62.56	29.22
ImageNet	Hendrycks2020Many	I3 (Hendrycks et al., 2021)	ℓ_∞	76.86	52.90
	Debenedetti2022Light_XCiT-S12	I4 (Debenedetti et al., 2023)	ℓ_∞	72.34	41.78
	Wong2020Fast	I5 (Wong et al., 2020)	ℓ_∞	55.62	26.24
	Salman2020Do_R18	I6 (Salman et al., 2020)	ℓ_∞	64.02	34.96
	Peng2023Robust	I7 (Peng et al., 2023)	ℓ_∞	73.44	48.94
	Mo2022When_Swin-B	I8 (Mo et al., 2022)	ℓ_∞	74.66	38.30

A.2 σ -ZERO: HYPERPARAMETER ROBUSTNESS

To assess the strength and potential limitations of our proposed attack, we conducted an ablation study on its key hyperparameters, τ_0 , σ , and t .

The parameter τ_0 governs the initial tolerance threshold in Algorithm 1, which induces sparsity within the adversarial perturbation. The parameter σ defines the approximation quality of $\hat{\ell}_0$ in Eq. (7) compared to the actual ℓ_0 function. Our ablation study, depicted in Figure 3, involved two distinct models: C10 (top row) and I1 (bottom row). We executed the attack on 1000 randomly selected samples from each dataset and recorded the ASR at different perturbation budgets k and the median ℓ_0 norm of the resulting adversarial perturbations. We observe a significant robustness of σ -zero with respect to these two hyperparameters; in particular: (i) the choice of the initial value of τ_0 exerts negligible influence on the ultimate outcome, given that the parameter dynamically adapts throughout the optimization process; and (ii) the selection of σ is not particularly challenging, especially when incorporating the sparsity projection operator.

We also conducted an ablation study on the sparsity threshold adjustment factor t used to adaptively update τ . In the following we keep the default values for $\tau_0 = 0.3$ and $\sigma = 10^{-3}$. We executed the attack on 1000 randomly selected samples against C3 and C4 models and recorded the ASR_{50} and the median ℓ_0 norm of the resulting adversarial perturbations. In Figure 4, we once again observe the robustness of σ -zero to this parameter, yielding similar and effective results when $t \leq 10^{-1}$.

Overall, the ablation study revealed consistent trends across the distinct models and datasets. In all cases, we identified a broad parameter configuration range where our attack maintained robustness to

	ASR ₂₄						ASR ₅₀						ASR ₁₀₀						$\tilde{\ell}_0$					
0.1	66.2	67.5	68.3	69.6	69.2	69.8	79.7	80.8	83.7	84.3	84.7	85.4	89.8	91.3	93.4	95.4	97.2	96.7	14.0	13.0	13.0	13.0	13.0	12.0
0.01	73.2	72.5	72.1	73.5	73.7	74.7	93.1	92.2	92.5	93.8	95.2	96.1	99.7	99.8	99.8	99.7	100.0	100.0	12.0	12.5	12.0	12.0	12.0	12.0
0.001	72.5	72.5	72.6	73.2	74.0	74.5	92.7	92.0	92.5	93.5	95.3	95.5	99.9	99.7	99.7	99.8	99.9	100.0	12.0	13.0	13.0	12.0	12.0	12.0
0.0001	72.5	72.5	72.8	73.3	73.6	74.8	92.2	92.2	92.4	93.9	95.2	95.8	99.5	99.8	99.7	99.7	100.0	100.0	13.0	13.0	12.0	12.0	12.0	12.0
1e-05	72.8	73.1	72.8	73.4	74.0	74.7	93.0	92.5	92.7	93.0	94.9	96.0	99.7	99.6	99.8	99.7	100.0	100.0	12.0	12.0	13.0	12.0	12.0	12.0
1e-06	72.5	72.0	72.2	73.4	74.0	74.7	92.6	92.2	92.8	93.4	95.0	96.2	99.7	99.8	99.8	99.8	99.9	100.0	13.0	13.0	13.0	12.0	12.0	12.0
1e-07	72.0	72.4	71.7	72.7	73.2	73.8	92.4	92.0	92.5	93.2	94.2	95.5	99.7	99.7	99.5	99.8	100.0	100.0	13.0	13.0	13.0	13.0	13.0	13.0
τ_0																								
	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5
0.1	83.2	83.7	84.8	83.8	84.3	84.3	94.3	94.5	94.9	95.0	95.2	95.7	96.7	97.7	98.3	99.1	98.8	98.8	5.0	5.0	5.0	5.0	5.0	5.0
0.01	82.7	81.8	82.7	83.1	83.4	82.7	96.1	95.8	95.7	95.6	95.9	95.9	99.5	99.7	99.7	99.7	99.7	99.8	5.0	5.0	5.5	5.0	6.0	5.0
0.001	83.0	83.0	83.0	83.4	82.9	83.2	96.5	95.5	96.1	95.4	95.6	96.0	99.7	99.7	99.8	99.7	99.5	99.7	5.0	6.0	6.0	6.0	5.0	5.0
0.0001	83.2	83.0	83.0	83.1	82.8	83.3	96.3	96.0	96.1	95.7	96.0	95.9	99.6	99.7	99.7	99.7	99.6	99.8	5.0	6.0	6.0	5.0	5.0	5.0
1e-05	82.8	82.4	83.2	82.2	82.7	83.6	95.8	95.2	96.3	95.8	96.0	95.5	99.6	99.7	99.7	99.5	99.8	99.8	6.0	6.0	5.0	6.0	5.0	6.0
1e-06	83.0	82.9	81.9	82.7	83.0	82.8	96.0	95.8	96.1	95.7	95.8	95.9	99.7	99.6	99.7	99.7	99.8	99.8	5.0	5.5	6.0	6.0	5.0	6.0
1e-07	75.1	75.8	77.3	75.8	74.7	75.3	93.6	93.5	94.1	93.7	93.8	93.3	99.4	99.0	99.3	99.6	99.4	99.3	7.0	7.0	7.0	7.0	7.0	7.0
τ_0																								
	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5	0.0	0.1	0.2	0.3	0.4	0.5

Figure 3: Ablation study on σ (y-axis) and τ_0 (x-axis) for CIFAR-10 C10 (top-row), ImageNet I1, (bottom-row). For each combination, we report the attack success rate at different k and the median ℓ_0 perturbation value.

the hyperparameter selection, making hyperparameter optimization for the attacker a swift task. This robustness is further evidenced by the results presented in the experimental comparisons, where our attack consistently outperforms competing attacks even with a shared hyperparameter configuration across all models.

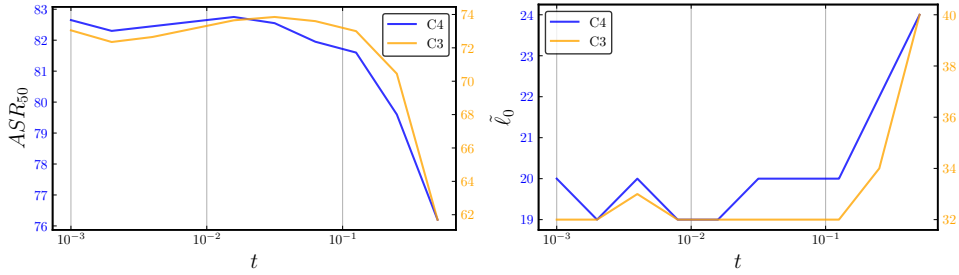


Figure 4: Ablation study on t for CIFAR-10 C3 and C4. For each, we report the attack success rate at the 50 feature budget (*left*) and the median ℓ_0 norm of the adversarial perturbation (*right*).

B ADDITIONAL EXPERIMENTAL COMPARISONS

B.1 COMPARISONS WITH MINIMUM-NORM ATTACK

In our experimental setup, we also consider a reduced number of queries, to test whether the attack can also run faster while remaining effective. We thus replicate our experimental comparison involving σ -zero and state-of-the-art sparse attacks while restricting the number of steps to $N = 100$. The results are summarized in Tables 7-9. Compared to the results with $N = 1000$ steps reported in Tables 1 and 6, the ASR of most competitive attacks decreases, while σ -zero remains effective by consistently reaching an ASR of 100%. This shows that σ -zero remains an effective, reliable and fast approach to crafting minimum-norm attacks even with reduced query budgets.

B.2 COMPARISONS WITH FIXED-BUDGET ATTACKS

Fixed-budget attacks, i.e., Sparse-RS (Croce et al., 2022), PGD- ℓ_0 (Croce & Hein, 2019), and Sparse-PGD (Zhong et al., 2024), have been designed to generate sparse adversarial perturbations given a fixed-budget k , therefore, drawing comparisons with minimum-norm attacks is not a straightforward task. Specifically, in their threat model, the attacker imposes a maximum limit on the number of perturbed features, and the attack then outputs the adversarial example that minimizes the model’s confidence in predicting the true label of the sample. However, since the fixed-budget threat model differs from the minimum-norm scenario we consider in this paper, which does not assume a maximum norm value k , we evaluate σ -zero in a fixed-budget fashion by discarding all adversarial perturbations that exceed k . Furthermore, as for fixed-budget attack, we let σ -zero to leverage the input parameter k to early stop the optimization procedure and reduce the number of consumed queries to the target model. Throughout this evaluation, the number of steps taken by Sparse-RS is always doubled compared to the other two white-box attacks, as it does not utilize the backward pass employed by the others. The main paper reports to this end an evaluation of σ -zero in a fixed-budget approach (cf. Tables 3-4). The remaining experiments, involving additional models for CIFAR-10 and ImageNet, are reported in Tables 10-11. Furthermore, to explore the effects of increased iterations on convergence and success rate, we increased the number of iterations up to $N = 10000$ (Tables 12-15), while always doubling the iterations for Sparse-RS. These additional experiments cover the three datasets MNIST, CIFAR-10, and ImageNet, 18 distinct models, and various feature budgets. The results again affirm that, σ -zero consistently outperforms competing approaches or synergizes well with them for a comprehensive robustness assessment.

B.3 ROBUSTNESS EVALUATION CURVES

We provide robustness evaluation curves for fixed-budget attacks on a CIFAR-10 model (C3), running each attack multiple times across various perturbation budgets k . The number of iterations is set to $N = 1000$ and $N = 5000$, with Sparse-RS allocating twice the iterations due to its reliance solely on forward passes. The results, depicted in Figure 5, demonstrates that σ -zero consistently outperforms fixed-budget attacks across all perturbation budgets k . Additionally, we present in Figs. 6-7 the robustness evaluation curves depicting the performance of minimum-norm ℓ_0 -attacks against all the models analyzed in our paper. These findings reinforce our experimental analysis, explicitly demonstrating that the σ -zero attack consistently achieves higher values of ASR while employing smaller ℓ_0 -norm perturbations compared to alternative attacks.

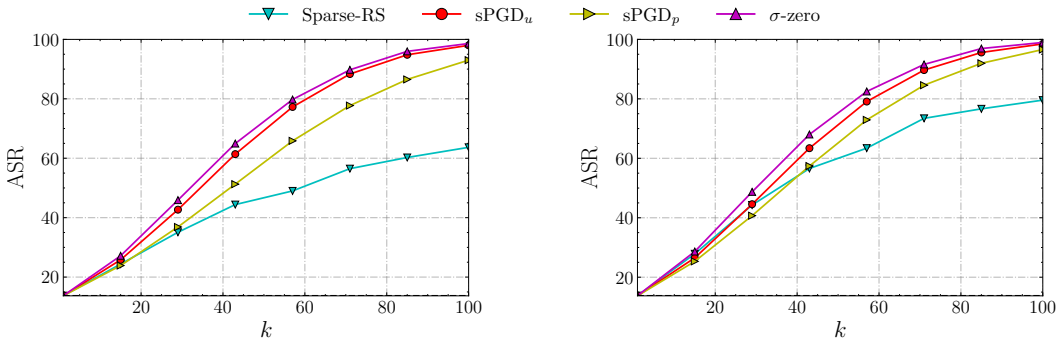


Figure 5: Robustness evaluation curves for fixed-budget attacks on C3. For each budget level k , each attack has been run with 1000 iterations (left-most plot) and 5000 iterations (right-most plot). Sparse-RS has been run with double the iterations as it relies solely on forward calls.

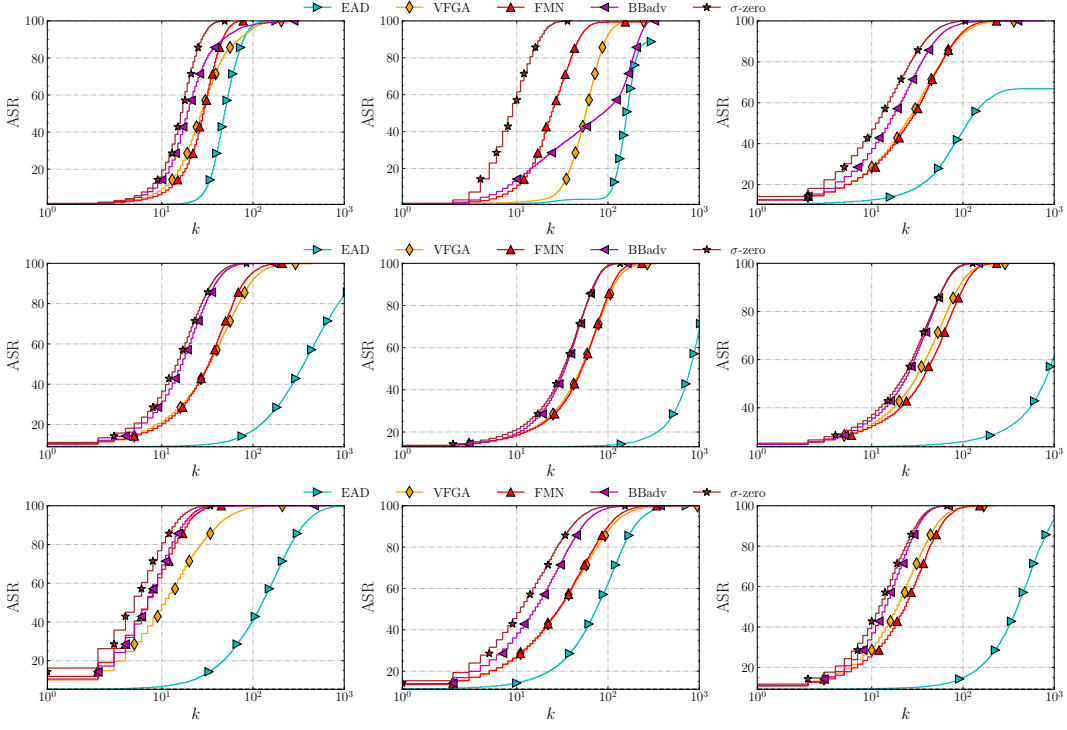


Figure 6: From the leftmost to the rightmost we report the robustness evaluation curves for M1, M2, C1 (top-row), C2, C3, C4 (middle-row) and C5, C6, C7 (bottom-row).

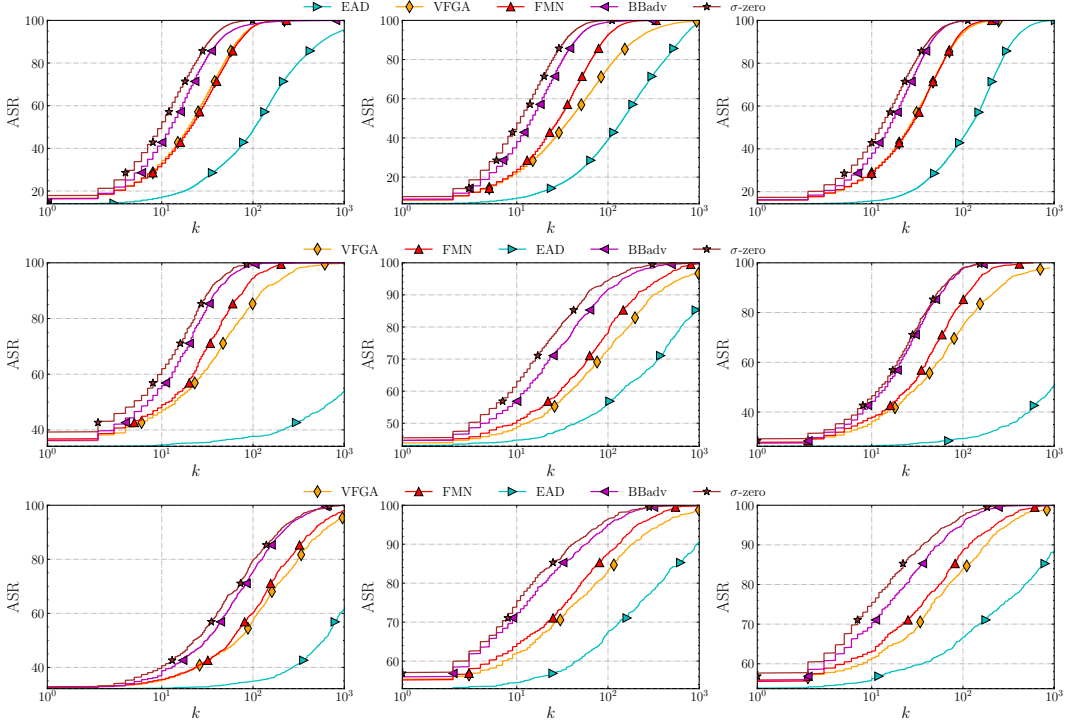


Figure 7: From the leftmost to the rightmost we report the robustness evaluation curves for C8, C9, C10 (top-row), I1, I2, I3 (middle-row) and I4, I5, I6 (bottom-row)

Table 6: Minimum-norm comparison results for CIFAR-10 and ImageNet with $N = 1000$ on remaining models. For each attack and model (M), we report ASR at $k = 10, 50, \infty$, median perturbation size $\tilde{\ell}_0$, mean runtime s (in seconds), mean number of queries q ($\div 1000$), and maximum VRAM usage (in GB). When VFGA exceeds the VRAM limit, we re-run it using a smaller batch size, increasing its runtime t . We denote those cases with the symbol ‘*’. Lastly we indicate with σ -zero * the case where we use $\sigma = 1$ and $\tau_0 = 0.1$.

Attack	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM
CIFAR-10																
SF	C5	11.19	11.19	56.56	3072	1.42	0.37	1.57	C8	23.67	24.85	62.41	3072	9.86	0.20	5.50
EAD		10.91	21.33	100.0	126	2.32	6.90	1.47		23.23	33.68	100.0	105	8.33	5.37	5.39
PDPGD		41.70	78.97	100.0	27	0.64	2.00	1.31		33.38	48.96	99.82	51	2.15	2.00	5.12
VFGA		77.25	93.41	99.99	11	0.17	0.32	11.96		56.76	81.79	99.89	20	4.30*	0.62	> 40
FMN		95.99	99.97	100.0	7	0.60	2.00	1.3		54.74	79.70	100.0	21	2.05	2.00	5.12
BB		97.43	99.79	100.0	7	5.81	2.76	1.47		59.82	78.76	83.58	16	12.49	3.14	5.39
BBadv		97.50	99.86	100.0	7	4.57	2.01	1.63		74.51	93.42	100.0	13	6.99	2.01	5.51
σ -zero		99.20	100.0	100.0	5	0.74	2.00	1.51		81.23	97.33	100.0	10	2.75	2.00	5.90
SF	C6	16.78	16.79	35.38	∞	19.74	0.62	10.00	C9	12.12	12.14	70.77	3072	3.28	0.22	2.25
EAD		20.75	35.90	100.0	74	10.76	5.55	9.92		14.51	23.62	100.0	148	2.23	5.80	2.15
PDPGD		23.84	40.89	100.0	69	3.96	2.00	8.86		25.31	38.41	100.0	69	0.76	2.00	2.0
VFGA		45.28	67.51	99.88	29	4.91*	1.02	> 40		38.42	56.72	99.81	39	3.15*	1.84	> 40
FMN		45.73	68.38	100.0	29	3.91	2.00	8.86		44.38	70.24	100.0	30	0.73	2.00	2.0
BB		15.26	17.14	17.94	∞	3.46	2.08	9.93		70.11	93.24	100.0	15	6.49	2.87	2.16
BBadv		64.47	88.92	100.0	16	8.85	2.01	10.03		69.45	92.91	100.0	15	6.02	2.01	2.22
σ -zero		75.63	94.47	100.0	11	4.41	2.00	10.43		79.59	96.93	100.0	11	0.89	2.00	2.65
SF	C7	29.51	40.86	93.82	3039	9.3	1.56	1.90	C10	25.88	26.54	51.80	3072	0.58	0.33	0.51
EAD		9.92	11.14	100.0	398	2.57	5.66	1.89		19.44	29.23	100.0	118	1.01	5.32	0.41
PDPGD		32.60	49.19	100.0	51	1.16	2.00	1.8		29.98	41.00	100.0	66	0.44	2.00	0.36
VFGA		61.19	90.04	99.88	19	0.28	0.52	16.53		48.63	74.15	99.54	25	0.17	0.77	3.07
FMN		52.14	85.60	100.0	23	1.09	2.00	1.8		47.89	73.71	100.0	26	0.41	2.00	0.36
BB		21.44	31.03	31.36	∞	3.01	2.37	1.89		68.37	91.83	100.0	15	10.90	2.93	0.41
BBadv		77.88	99.11	100.0	14	4.51	2.01	1.99		67.35	93.04	100.0	16	4.60	2.01	0.54
σ -zero		81.38	99.15	100.0	12	1.39	2.00	1.91		73.96	94.21	100.0	13	0.63	2.00	0.51
FMN	C11	39.57	74.75	100.0	32	0.11	2.00	0.59	C12	48.3	78.16	100.0	26	0.11	2.00	0.59
BBadv		14.07	18.57	100.0	183	2.52	2.01	0.65		18.33	19.75	100.0	290	2.57	2.01	0.65
σ -zero		12.38	15.91	100.0	144	0.24	2.00	1.03		18.52	21.31	100.0	187	0.33	2.00	1.03
σ -zero*		44.78	85.05	100.0	27	0.23	2.00	1.03			54.62	90.12	100.0	22	0.23	2.00
ImageNet																
Attack	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM
EAD	I5	56.6	60.2	100.0	0	21.38	5.50	1.41	I6	59.0	61.4	100.0	0	7.89	5.29	0.48
VFGA		69.0	76.2	98.8	0	6.11*	1.43	> 40		66.8	76.6	99.3	0	1.74*	1.21	> 40
FMN		71.0	79.5	100.0	0	1.97	2.00	2.30		70.9	78.7	100.0	0	0.72	2.00	0.67
BBadv		82.3	89.0	100.0	0	185.34	2.01	2.41		80.3	89.6	100	0	199.47	2.01	0.73
σ -zero		85.1	91.4	100.0	0	2.76	2.00	2.52		86.2	92.8	100.0	0	1.13	2.00	0.84
FMN	I7	39.2	48.5	100	54.5	5.84	2	17.44	I8	38.1	46.8	100	67	5.17	2	7.91
BBadv		49.7	62	100	25.5	128.2	2	17.86		46.5	58.6	100	29.5	113.87	2	8.30
σ -zero		55.4	68.2	100	16	8.62	2	19.03			50.1	64.4	100	24	6.6	2

Table 7: Minimum-norm comparison results for MNIST with $N = 100$. See the caption of [Table 6](#) for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR _{∞}	$\tilde{\ell}_0$	s	q	VRAM	M	ASR ₂₄	ASR ₅₀	ASR _{∞}	$\tilde{\ell}_0$	s	q	VRAM
MNIST																
SF	M1	5.11	6.76	96.98	469	1.07	0.18	0.07	M2	0.98	1.21	91.68	463	2.87	0.86	0.07
EAD		3.73	46.65	100.0	52	0.06	1.14	0.07		3.51	35.57	100.0	61	0.06	0.99	0.07
PDPGD		0.98	0.98	100.0	359	0.01	0.20	0.07		0.52	0.52	95.02	254	0.01	0.20	0.07
VFGA		4.82	82.68	100.0	27	0.07	0.76	0.23		4.82	38.99	99.98	57	0.07	1.34	0.24
BB		68.52	98.00	100.0	20	0.13	1.19	0.08		62.98	83.00	87.87	18	0.13	1.69	0.08
FMN		33.03	83.09	88.92	30	0.01	0.20	0.07		10.05	14.03	14.81	∞	0.01	0.20	0.07
BBadv		62.29	90.88	100.0	21	0.09	0.21	0.08		41.19	58.80	100.0	34	0.07	0.21	0.08
σ -zero		61.12	98.45	100.0	22	0.01	0.20	0.08		87.20	99.82	100.0	13	0.01	0.20	0.08

Table 8: Minimum-norm comparison results for CIFAR-10 with $N = 100$. See the caption of Table 6 for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM
CIFAR-10																
SF	C1	17.67	17.76	47.26	∞	3.17	0.35	1.62	C6	16.75	16.79	35.36	∞	19.74	0.62	10.00
EAD		16.74	28.74	100.0	100.0	0.27	0.80	1.53		19.79	32.94	100.0	83	1.58	0.82	10.04
PDPGD		10.31	10.31	99.39	2421	0.05	0.20	1.43		11.26	11.26	99.75	2814	0.32	0.2	8.97
VFGA		50.73	75.34	93.69	24	0.23	0.72	11.83		45.33	67.05	87.75	29	3.75	0.86	> 40
FMN		46.90	69.36	80.68	27	0.05	0.20	1.43		42.67	61.49	72.34	33	0.31	0.2	8.98
BB		12.98	14.29	14.97	∞	0.44	1.95	1.59		14.99	16.88	17.91	∞	2.67	1.95	10.04
BBadv	σ-zero	60.52	86.63	100.0	18	0.41	0.21	1.59	σ-zero	59.61	84.59	100.0	18	0.76	0.21	10.04
σ-zero		63.60	88.27	100.0	16	0.08	0.20	1.84		63.44	87.2	100.0	16	0.44	0.2	10.29
SF	C2	17.86	20.59	94.26	3071	2.44	0.26	1.91	C7	21.07	38.76	82.71	3062	4.30	9.67	1.90
EAD		9.50	10.67	100.0	451	0.30	0.71	2.01		9.68	10.56	100.0	434	0.48	0.90	2.00
PDPGD		8.92	8.92	75.31	3052	0.09	0.20	1.91		9.17	9.17	99.90	2709	0.12	0.20	1.91
VFGA		39.31	66.46	91.64	33	0.34	0.87	16.64		60.94	90.04	99.16	19	0.29	0.52	16.64
FMN		37.13	62.41	71.3	36	0.08	0.20	1.92		50.70	79.48	87.20	24	0.08	0.20	1.91
BB		38.18	53.53	57.05	40	0.59	1.90	2.00		26.39	32.41	32.83	∞	0.50	1.93	2.00
BBadv	σ-zero	63.56	92.74	100.0	19	0.40	0.21	2.00	σ-zero	74.91	98.37	100.0	15	0.41	0.21	2.00
σ-zero		56.94	88.60	100.0	21	0.11	0.20	2.25		68.14	94.90	100.0	16	0.11	0.20	2.25
SF	C3	20.89	24.36	58.29	3072	1.63	0.48	0.66	C8	23.87	24.85	62.42	3072	9.86	0.2	5.50
EAD		13.03	13.18	100.0	835	0.11	0.65	0.64		21.71	29.59	100.0	128	0.67	0.66	5.51
PDPGD		12.95	12.98	99.47	2566	0.04	0.20	0.59		13.96	13.96	54.16	3072	0.21	0.2	5.23
VFGA		28.63	49.73	82.94	51	0.13	1.13	4.44		56.81	82.04	97.08	20	4.32	0.61	> 40
FMN		26.76	37.90	43.90	∞	0.03	0.20	0.59		53.42	76.59	87.1	22.0	0.21	0.2	5.24
BB		16.40	22.91	27.64	∞	1.04	2.25	0.65		60.74	78.14	84.46	17	1.36	1.67	5.55
BBadv	σ-zero	33.68	66.79	100.0	37	0.40	0.21	0.65	σ-zero	70.31	91.58	100.0	14	0.55	0.21	5.51
σ-zero		30.56	57.71	100.0	43	0.04	0.20	0.89		69.49	91.87	100.0	14	0.25	0.2	6.76
SF	C4	31.85	42.97	84.45	70	1.54	0.47	0.66	C9	12.03	12.14	70.77	3072	3.28	0.22	2.25
EAD		24.1	24.4	100.0	844	0.12	0.66	0.65		13.61	21.61	100.0	162	0.31	0.8	2.27
PDPGD		23.78	23.78	66.62	3072	0.04	0.2	0.59		6.31	6.31	96.2	2773	0.06	0.2	2.11
VFGA		46.7	69.52	93.05	28	0.14	0.77	4.22		38.22	56.56	75.79	39.5	1.45	1.06	> 40
FMN		42.69	58.78	65.83	35	0.03	0.2	0.59		40.27	59.69	68.88	35	0.06	0.2	2.19
BB		25.91	27.98	29.51	∞	0.54	2.09	0.65		66.02	90.74	100.0	16	0.65	1.07	2.27
BBadv	σ-zero	52.25	80.64	100.0	23	0.36	0.21	0.65	σ-zero	64.41	89.7	100.0	17	0.42	0.21	2.27
σ-zero		49.74	73.75	100.0	25	0.04	0.2	0.89		65.96	90.95	100.0	16	0.09	0.2	2.52
SF	C5	11.19	11.19	56.56	3072	1.42	0.37	1.56	C10	24.28	26.54	51.90	3072	0.58	0.33	0.52
EAD		10.42	19.09	100.0	146	0.26	0.77	1.58		18.82	26.17	100.0	144	0.11	0.79	0.52
PDPGD		5.23	5.23	100.0	3057	0.05	0.20	1.43		14.29	14.29	90.95	3057	0.03	0.2	0.47
VFGA		77.22	93.44	98.99	11	0.17	0.38	12.08		48.49	74.14	94.16	26	0.12	0.73	3.18
FMN		89.83	97.72	98.86	8	0.05	0.20	1.43		46.75	69.77	80.68	27	0.03	0.2	0.48
BB		84.42	97.55	100.0	10	0.62	0.95	1.59		63.70	89.39	100.0	17	0.43	1.13	0.53
BBadv	σ-zero	83.81	97.35	100.0	10	0.45	0.21	1.59	σ-zero	63.29	90.08	100.0	17	0.35	0.21	0.53
σ-zero		91.54	99.84	100.0	9	0.08	0.20	1.83		60.79	86.02	100.0	18	0.04	0.2	0.77

Table 9: Minimum-norm comparison results for ImageNet with $N = 100$. See the caption of Table 6 for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM	M	ASR ₂₄	ASR ₅₀	ASR _∞	$\tilde{\ell}_0$	s	q	VRAM
ImageNet																
EAD	I1	34.7	35.9	100.0	484	1.02	0.67	0.46	I4	32.4	33.0	100.0	808	5.15	0.7	1.68
VFGA		58.3	72.2	85.3	14	1.06	0.70	> 40		40.0	46.8	56.9	66.5	9.23	1.2	> 40
FMN		55.4	64.5	68.1	14	0.08	0.20	0.66		39.9	46.2	47.5	∞	0.44	0.2	2.97
BBadv		67.6	83.3	100.0	10	23.02	0.21	0.72		46.4	58.0	99.9	32	21.23	0.21	3.07
σ-zero		69.2	86.9	100.0	10	0.13	0.20	0.84		43.7	55.2	100.0	32	0.61	0.2	3.20
EAD	I2	47.1	50.1	100.0	48	2.32	0.68	1.42	I5	56.2	60.2	100.0	0	2.46	0.72	1.41
VFGA		54.3	63.2	96.7	13	2.88	0.72	> 40		68.9	76.0	83.0	0	2.33	0.59	> 40
FMN		55.9	60.0	62.4	10	0.20	0.20	2.30		67.8	72.0	74.3	0	0.20	0.20	2.30
BBadv		70.1	80.1	100.0	5	20.49	0.21	2.40		80.8	87.8	100.0	0	18.60	0.21	2.41
σ-zero		71.0	82.7	100.0	4	0.29	0.20	2.52		81.8	89.3	100.0	0	0.29	0.20	2.52
EAD	I3	26.9	27.7	100.0	1108	0.58	0.61	1.41	I6	57.4	60.0	100.0	0	1.03	0.72	0.48
VFGA		47.0	58.7	74.0	31	3.07	0.96	> 40		66.8	75.2	83.9	0	0.91	0.59	> 40
FMN		44.4	50.6	53.2	47	0.16	0.2	2.30		69.2	74.9	77.2	0	0.07	0.2	0.67
BBadv		53.6	74.7	100.0	20	23.86	0.21	2.41		80.1	89.1	100.0	0	19.68	0.21	0.73
σ-zero		52.9	74.4	100.0	21	0.23	0.2	2.52		82.2	90.5	100.0	0	0.12	0.2	0.84

Table 10: Fixed-budget comparison results with $N = 1000$ on CIFAR10 remaining models. Sparse-RS was executed with double the steps, $2N$, to ensure fair comparison as it lacks backward passes. For each attack, we report the corresponding ASR with different feature budget levels (24,50,100). We report the execution time s_{24} and query usage q_{24} for the smaller $k = 24$, as it requires, on average, more iterations due to the more challenging problem. Lastly we indicate with $\sigma\text{-zero}^*$ the case where we use $\sigma = 1$ and $\tau_0 = 0.1$.

Attack	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q_{24}	s_{24}	VRAM	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q_{24}	s_{24}	VRAM
CIFAR-10														
PGD- ℓ_0	C5	68.60	88.89	98.14	2.00	1.95	1.89	C8	42.81	66.19	90.49	2.00	3.24	7.36
Sparse-RS		99.71	100.0	100.0	0.08	0.10	1.91		72.54	86.72	94.84	0.78	1.10	7.35
sPGD _p		99.82	100.0	100.0	0.02	0.16	2.06		68.47	90.47	99.56	0.70	1.48	7.62
sPGD _u		97.84	99.98	100.0	0.09	0.37	2.06		73.55	94.55	99.97	0.60	1.65	7.62
$\sigma\text{-zero}$		99.20	100.0	100.0	0.22	0.11	2.07		81.23	97.33	99.97	0.52	0.54	7.76
PGD- ℓ_0	C6	32.80	50.53	77.06	2.00	4.88	12.79	C9	31.45	52.79	80.27	2.00	2.01	2.91
Sparse-RS		76.61	89.88	96.22	0.67	1.98	12.74		68.77	82.06	89.81	0.85	0.56	2.89
sPGD _p		63.66	87.07	98.67	0.80	2.83	13.77		61.0	83.49	96.76	0.87	0.88	3.03
sPGD _u		64.28	88.25	99.09	0.77	2.77	13.77		63.48	87.59	98.49	0.81	0.85	3.03
$\sigma\text{-zero}$		75.63	94.47	99.78	0.66	1.75	13.82		79.59	96.93	99.91	0.57	2.04	2.91
PGD- ℓ_0	C7	37.91	68.90	95.31	2.00	1.96	2.47	C10	38.33	61.88	89.50	2.00	1.12	0.51
Sparse-RS		63.75	84.49	95.74	0.97	0.61	2.46		64.80	81.46	91.13	0.91	0.45	0.50
sPGD _p		72.82	96.74	99.98	0.61	0.94	2.57		59.94	84.87	98.82	0.87	0.44	0.55
sPGD _u		81.64	99.09	100.0	0.42	0.69	2.57		65.07	90.78	99.83	0.75	0.41	0.55
$\sigma\text{-zero}$		81.38	99.15	100.0	0.46	0.21	2.68		73.96	94.21	99.80	0.67	0.14	0.57
Sparse-RS	C11	28.08	41.89	58.45	0.38	1.53	0.59	C12	51.64	71.27	86.57	0.32	1.14	0.59
sPGD _u		15.87	21.43	32.67	0.26	1.71	0.65		22.78	26.56	34.09	0.31	1.58	0.64
sPGD _p		13.61	17.07	30.11	0.25	1.74	0.65		24.52	34.39	59.89	0.30	1.54	0.65
$\sigma\text{-zero}$		12.38	15.91	30.43	0.20	1.77	1.03		18.52	21.31	28.81	0.27	1.65	1.03
$\sigma\text{-zero}^*$		44.78	85.05	99.76	0.15	1.33	1.03		54.62	90.12	99.94	0.13	1.15	1.03

Table 11: Fixed-budget comparison results for ImageNet with $N = 1000$ on remaining models. Sparse-RS was executed with double the steps, $2N$, to ensure fair comparison as it lacks backward passes. For each attack, we report the corresponding ASR with budget level $k = 150$. We report the execution time s_{100} and query usage q_{100} for the smaller $k = 100$, as it requires, on average, more iterations due to the more challenging problem.

Attack	M	ASR ₁₀₀	ASR ₁₅₀	q_{100}	s_{100}	VRAM	M	ASR ₁₅₀	ASR ₁₅₀	q_{100}	s_{100}	VRAM
ImageNet												
Sparse-RS	I5	83.6	87.5	0.44	2.85	4.39	I6	85.4	89.2	0.41	3.46	1.29
sPGD _p		89.8	94.5	0.24	1.64	4.48		90.4	95.2	0.22	0.98	1.33
sPGD _u		86.5	92.6	0.29	1.55	4.48		89.1	94.0	0.24	1.15	1.33
$\sigma\text{-zero}$		95.9	98.2	0.12	0.16	4.90		98.1	98.8	0.10	0.08	1.79
Sparse-RS	I7	58.20	60.60	0.95	5.21	17.43	I8	49.20	52.10	1.13	3.28	7.89
sPGD _p		67.50	75.50	0.70	4.85	17.80		65.10	75.20	0.75	3.56	8.24
sPGD _u		65.70	75.10	0.73	5.37	17.82		65.10	75.20	0.73	5.68	8.23
$\sigma\text{-zero}$		82.10	87.00	0.43	1.87	19.03		78.00	86.20	0.50	1.67	9.46

Table 12: Fixed-budget comparison results with $N = 5000$ on MNIST. See the caption of Table 10 for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q_{24}	s_{24}	VRAM	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q_{24}	s_{24}	VRAM
MNIST														
Sparse-RS	M1	88.13	99.26	99.99	2.45	1.86	0.04	M2	99.88	99.97	100.0	0.31	0.17	0.04
sPGD _p		81.22	99.30	100.0	2.83	1.50	0.05		83.88	99.88	99.97	2.33	0.9	0.05
sPGD _u		87.30	99.85	100.0	1.60	1.44	0.05		74.38	99.46	99.99	3.47	0.96	0.05
$\sigma\text{-zero}$		88.63	100.0	100.0	1.38	0.20	0.08		99.67	100.0	100.0	0.24	0.02	0.08

Table 13: Fixed-budget comparison results with $N = 5000$ on CIFAR10. See the caption of Table 10 for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q ₂₄	s ₂₄	VRAM	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q ₂₄	s ₂₄	VRAM
CIFAR-10														
Sparse-RS	C1	82.94	94.77	98.68	2.55	1.81	1.92	C5	99.93	99.98	99.99	0.15	0.14	1.91
sPGD _p		71.88	93.17	99.75	3.21	2.78	2.06		99.93	100.0	100.0	0.17	0.18	2.05
sPGD _u		69.73	92.86	99.79	3.34	2.98	2.06		99.73	100.0	100.0	0.26	1.31	2.05
σ -zero		80.91	96.81	99.98	3.23	1.35	2.09		99.81	100.0	100.0	0.61	0.18	2.05
Sparse-RS	C2	71.21	90.28	97.69	3.89	2.49	2.46	C7	77.06	94.33	99.25	3.38	2.15	2.46
sPGD _p		64.88	92.03	99.85	3.97	3.69	2.57		78.41	98.36	100.0	2.59	5.87	2.57
sPGD _u		68.61	94.99	99.96	3.49	3.46	2.57		84.17	99.41	100.0	1.85	6.26	2.57
σ -zero		78.14	98.39	100.0	3.16	1.02	2.70		83.99	99.49	100.0	1.82	0.69	2.70
Sparse-RS	C3	38.43	58.27	79.59	6.77	2.61	0.69	C9	79.69	89.98	95.3	2.81	1.95	2.89
sPGD _p		34.62	65.54	96.55	6.70	2.52	0.73		66.39	87.65	98.19	3.73	2.75	3.03
sPGD _u		37.29	72.03	98.49	6.48	2.96	0.73		68.03	90.7	99.23	3.56	2.64	3.04
σ -zero		40.99	76.00	98.98	6.32	1.42	0.77		83.92	98.39	99.99	2.19	0.99	3.09
Sparse-RS	C4	54.85	71.95	86.25	5.03	2.37	0.69	C10	76.62	91.5	97.89	3.20	1.46	0.50
sPGD _p		53.36	80.97	99.13	4.94	2.24	0.73		65.92	90.38	99.72	3.86	1.63	0.55
sPGD _u		57.31	86.11	99.72	4.46	2.36	0.73		68.59	92.93	99.91	3.71	1.62	0.55
σ -zero		57.11	84.54	99.34	4.47	1.04	0.77		77.74	95.86	99.92	2.75	0.57	0.59

Table 14: Fixed-budget comparison results with $N = 5000$ on ImageNet. See the caption of Table 11 for further details.

Attack	M	ASR ₁₀₀	ASR ₁₅₀	q ₁₀₀	s ₁₀₀	VRAM	M	ASR ₁₀₀	ASR ₁₅₀	q ₁₀₀	s ₁₀₀	VRAM
ImageNet												
Sparse-RS	I1	94.2	95.1	1.39	7.73	1.29	I4	48.8	51.7	5.68	13.51	5.73
sPGD _p		97.3	99.6	0.45	1.95	1.41		66.2	78.4	3.68	21.65	5.84
sPGD _u		93.6	98.5	0.71	2.48	1.40		64.1	78.5	3.78	20.71	5.84
σ -zero		100.0	100.0	0.72	0.31	1.83		77.3	87.8	2.42	4.17	6.33
Sparse-RS	I2	85.1	86.8	2.06	11.0	4.39	I5	89.6	91.3	1.54	3.35	4.39
sPGD _p		85.9	92.8	1.63	8.33	4.49		92.0	95.9	0.94	3.35	4.39
sPGD _u		81.3	90.5	2.00	6.88	4.49		88.0	93.1	1.28	5.89	4.48
σ -zero		94.6	97.3	0.63	0.61	4.94		96.9	98.4	0.41	0.34	4.94
Sparse-RS	I3	74.8	76.6	3.54	6.38	4.39	I6	87.5	92.7	1.36	2.05	1.29
sPGD _p		87.6	95.4	1.61	6.29	4.49		96.5	97.0	0.83	1.39	1.33
sPGD _u		81.4	93.7	2.04	6.54	4.49		90.4	94.7	2.04	2.35	1.33
σ -zero		98.2	99.7	1.68	1.44	4.94		97.2	99.1	0.35	0.12	1.83

Table 15: Fixed-budget comparison results with $N = 10000$ on CIFAR-10 and ImageNet. See the caption of Tables 10-11 for further details.

Attack	M	ASR ₂₄	ASR ₅₀	ASR ₁₀₀	q ₂₄	s ₂₄	VRAM	M	ASR ₁₀₀	ASR ₁₅₀	q ₁₀₀	s ₁₀₀	VRAM
Sparse-RS	C3	41.12	63	83.99	3.44	12.82	0.60	I2	87.2	88.3	9.66	3.67	0.14
sPGD _u		37.95	72.6	98.59	1.99	12.56	0.65		86.9	91.6	7.84	1.95	0.15
sPGD _p		35.89	67.92	97.42	2.12	13.2	0.65		81.8	88.9	8.4	3.85	0.15
σ -zero		41.67	76.38	99.01	0.33	3.07	2.41		95.1	97.2	1.31	0.23	0.21

C VISUAL COMPARISON

In Figures 8-10 we show adversarial examples generated with competing ℓ_0 -attacks, and our σ -zero. First, we can see that ℓ_0 adversarial perturbations are clearly visually distinguishable [Carlini & Wagner (2017a); Brendel et al. (2019a); Pintor et al. (2021)]. Their goal, indeed, is not to be indistinguishable to the human eye – a common misconception related to adversarial examples [Biggio & Roli, 2018; Gilmer et al., 2018] – but rather to show whether and to what extent models can be fooled by just changing a few input values.

A second observation derived from Figures 8-10 is that the various attacks presented in the state of the art can identify distinct regions of vulnerability. For example, note how FMN and VFGA find similar perturbations, as they mostly target overlapping regions of interest. Conversely, EAD finds sparse perturbations scattered throughout the image but with a lower magnitude. This divergence is attributed to EAD’s reliance on an ℓ_1 regularizer, which promotes sparsity, thus diminishing perturbation magnitude without necessarily reducing the number of perturbed features. Conversely, our attack does not focus on specific areas or patterns within the images but identifies diverse critical features, whose manipulation is sufficient to mislead the target models. Given the diverse solutions offered by the attacks, we argue that their combined usage may still improve adversarial robustness evaluation to sparse attacks.

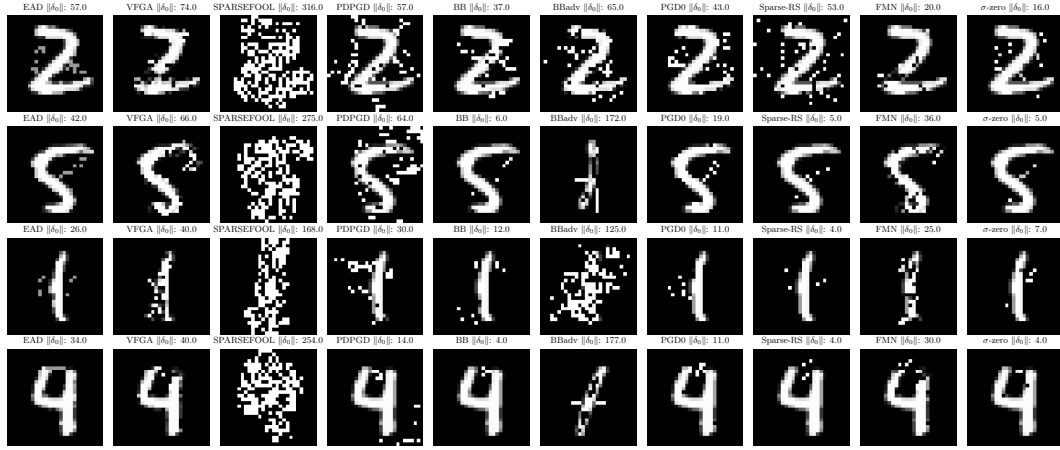


Figure 8: Randomly chosen adversarial examples from MNIST M2.

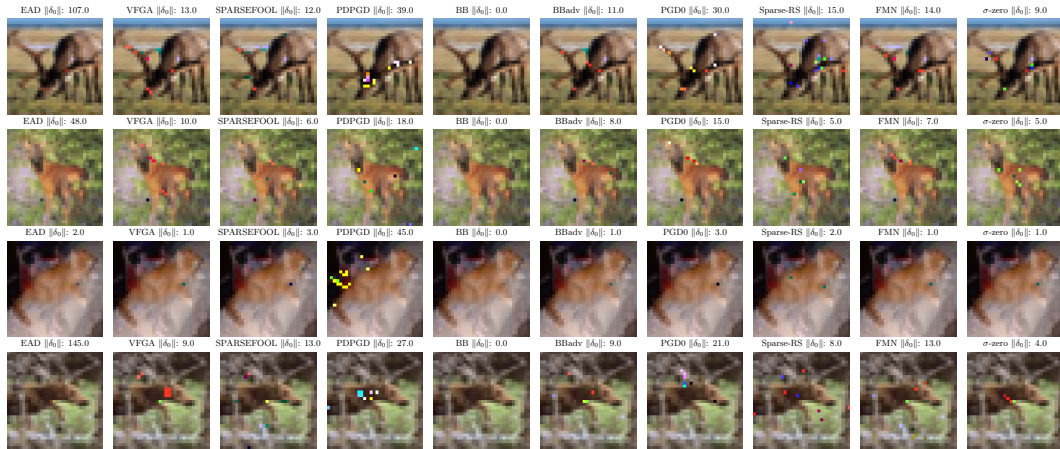


Figure 9: Randomly chosen adversarial examples from CIFAR-10 C1.

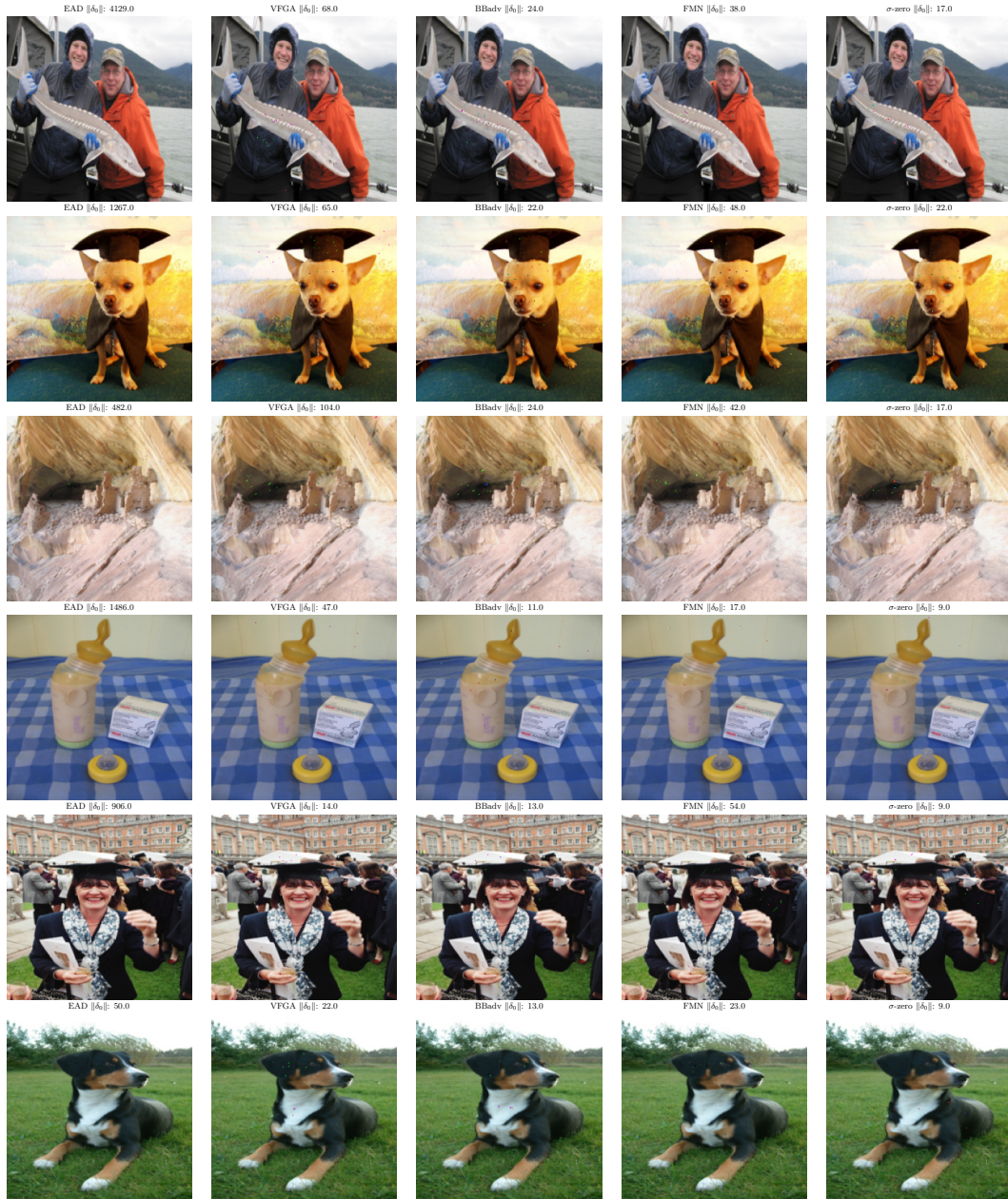


Figure 10: Randomly chosen adversarial examples from ImageNet I1.