A Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [19] [26] [47] are a class of probabilistic generative models that apply a noise injection process, followed by a reverse procedure for sample generation. A DDPM is defined as two parameterized Markov chains: a forward chain that add random Gaussian noise to images to transform data distribution into a simple prior distribution and a reverse chain that convert the noised image back into target data by learning transition kernels parameterized by deep neural networks.

Forward diffusion process: Given a data point sampled from a real data distribution $x_0 \sim q(x)$, a forward process begins with adding a small amount of Gaussian noise to the sample in T steps, producing a sequence of noisy samples $x_1, \ldots, x_T$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$.

$$\begin{aligned} q(x_{1:T}|x_0) &:= \prod_{t=1}^{T} q(x_t|x_{t-1}), \\ q(x_t|x_{t-1}) &= \mathcal{N}\left(x_t; (1-\beta_t)x_{t-1}, \beta_t I\right). \end{aligned} \tag{7}$$

The data sample x_0 gradually loses its distinguishable features as the step t becomes larger. Eventually, when $T \to \infty$, x_T is equivalent to an isotropic Gaussian distribution.

Reverse diffusion process: The reverse process starts by first generating an unstructured noise vector from the prior distribution, then gradually removing noise by running a learnable Markov chain in the reverse time direction. Specifically, the reverse Markov chain is parameterized by a prior distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$ and a learnable transition kernel $p_\theta(x_{t-1}|x_t)$. Therefore, we need to learn a model p_θ to approximate these conditional probabilities in order to run the reverse diffusion process.

$$\begin{aligned} p_\theta(x_{0:T}) &= p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \\ p_\theta(x_{t-1}|x_t) &= \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right), \end{aligned} \tag{8}$$

where θ denotes model parameters, often instantiated by architectures like UNet, which parameterize the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$. The UNet takes the noised data x_t and time step t as inputs and outputs the parameters of the normal distribution, thereby predicting the noise ϵ_θ that the model needs to reverse the diffusion process. With this reverse Markov chain, we can generate a data sample x_0 by first sampling a noise vector $x_T \sim p(x_T)$, then successively sampling from the learnable transition kernel $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ until $t = 1$.

B Stochastic Perturbation Generation

CLIP Score We summarise the following reasons for choosing CLIP Score to measure the correlation between a generated caption for an image and the actual content of the image.

- 1. While CLIP score is certainly not perfect as a metric (like other metrics) to mimic human-perception similarity, but some works show that CLIP is generally reliable and highly correlated with human judgement in user experiments, e.g., [18]. Both CLIP scores are calculated (and thus anchored) by considering the text information, regardless of variations in image styles and backgrounds for additional evidence. Therefore, they are highly likely to be similar.
- 2. While the best metric for measuring T2I correctness remains an open question, CLIP score is commonly used in studying T2I robustness [11, 32, 48, 64, 65, 70]
- 3. While CLIP score is a building block of ProTIP, it can be substituted with other correctness metrics. The statistical models—the theoretical core and main contribution of ProTIP—are easily *adaptable* to any new correctness metrics as future advancements.

Stochastic Perturbation Method Table 2 for examples of stochastic text perturbations on T2I DM inputs.

Table 2: Examples of stochastic text perturbations

Perturbation	Description	Example
Insert	Insert a character randomly	A white daog plays with a red ball on the green grass.
Substitute	Substitute a character randomly	A white dog plays with a rad ball on the green grass.
Swap	Swap two characters randomly	A white dog plays with a red ball on the green garss.
Delete	Delete a character randomly	A white dog plays with a red bll on the green grass.
Keyboard	Substitute a char. by keyboard distance	A whote dog plays with a red ball on the green grass.

C Hoeffding's Inequality

Reusing the notations in the main paper, let $I_1, \ldots, I_n$ be i.i.d. samples drawn from a population, and R denotes the population mean to be estimated. Let $\hat{\mu}_I^{(n)} = \frac{1}{n}\sum_{i=1}^{n} I_i$, i.e., the sample mean, we have the following Hoeffding's Inequality [20].

Theorem 2 (Hoeffding's Inequality).

$$Pr\left(\left|\hat{\mu}_I^{(n)} - R\right| \leq \varepsilon\right) > 1 - \sigma \tag{9}$$

where $\sigma = 2e^{-2n\varepsilon^2}$, *equivalently,* $\varepsilon = \sqrt{\frac{\log(2/\sigma)}{2n}}$.

From the two Theorems 2 and 1 regarding the original and adaptive Hoeffding's Inequality respectively, we may derive the following two corollaries.

Corollary 1 (Tightness of the two bounds). *Give a confidence level $1-\sigma$ and the same number of samples n, the estimation error (between the sample mean and population mean) ε derived from the original Hoeffding's Inequality is always smaller than the adaptive Hoeffding's Inequality.*

Proof. As Theorems 2 and 1, the estimation error ε from the original and adaptive Hoeffding's Inequality are $\sqrt{\frac{\log(2/\sigma)}{2n}}$ and $\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}$, respectively. We may prove the former is smaller than the latter either analytically or empirically. The inequalities of Corollary 1 can be written:

$$\sqrt{\frac{\log(2/\sigma)}{2n}} < \sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}} \tag{10}$$

Where $n \geq 1$ and $\sigma \in (0,1)$.

$$\frac{\log(2/\sigma)}{2n} < \frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n} \tag{11a}$$

$$\log(2/\sigma) < 1.2\cdot\log(\log_{1.1} n+1)+\frac{10}{9}\cdot\log(24/\sigma) \tag{11b}$$

$$\log(2/\sigma) < \log\left[(\log_{1.1} n+1)^{1.2}\cdot(24/\sigma)^{\frac{10}{9}}\right] \tag{11c}$$

Because $n \geq 1$, hence, $\log_{1.1} n > 0$, $\log_{1.1} n + 1 > 1$, and $(\log_{1.1} n+1)^{1.2} > 1$, the right side of (11c) is.

$$\log\left[(\log_{1.1} n+1)^{1.2}\cdot(24/\sigma)^{\frac{10}{9}}\right] > \log(24/\sigma)^{\frac{10}{9}} > \log(24/\sigma) \tag{12}$$

Hence, (11c) will be:

$$\log(2/\sigma) < \log(24/\sigma), \sigma \in (0,1) \tag{13}$$

Therefore the equation holds, and the proof is complete. □

Corollary 2 (Monotonicity to n and σ). *For both the original and adaptive Hoeffding's Inequalities, the estimation errors (between the sample mean and population mean) ε are monotonically decreasing to sample size n (for $n >= 2$) and σ.*

Proof. By taking the (partial) derivatives of the two analytical expressions of ε r.w.t. n and σ respectively, we may establish negative results as what follows.

1. $\sqrt{\frac{\log(2/\sigma)}{2n}}$ partial derivative with respect to n:

$$\begin{aligned}\frac{\partial\varepsilon(n,\sigma)}{\partial n} &= \frac{\partial}{\partial n}\left(\sqrt{\frac{\log(2/\sigma)}{2n}}\right)\\ &= \frac{1}{2}\left(\frac{\log(2/\sigma)}{2n}\right)^{-\frac{1}{2}}\cdot\frac{\partial}{\partial n}\left(\frac{\log(2/\sigma)}{2n}\right)\\ &= \frac{1}{2\sqrt{\frac{\log(2/\sigma)}{2n}}}\cdot\frac{\partial}{\partial n}\left(\frac{\log(2/\sigma)}{2n}\right)\\ &= \frac{1}{2\sqrt{\frac{\log(2/\sigma)}{2n}}}\cdot\left(-\frac{\log(2/\sigma)}{2n^2}\right)\\ &= -\frac{\log(2/\sigma)}{4n^2\sqrt{\frac{\log(2/\sigma)}{2n}}}\end{aligned}$$

2. $\sqrt{\frac{\log(2/\sigma)}{2n}}$ partial derivative with respect to σ:

$$\begin{aligned}\frac{\partial\varepsilon(n,\sigma)}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(\sqrt{\frac{\log(2/\sigma)}{2n}}\right)\\ &= \frac{1}{2}\left(\frac{\log(2/\sigma)}{2n}\right)^{-\frac{1}{2}}\cdot\frac{\partial}{\partial \sigma}\left(\frac{\log(2/\sigma)}{2n}\right)\\ &= \frac{1}{2\sqrt{\frac{\log(2/\sigma)}{2n}}}\cdot\frac{\partial}{\partial \sigma}\left(\frac{\log(2/\sigma)}{2n}\right)\\ &= \frac{1}{2\sqrt{\frac{\log(2/\sigma)}{2n}}}\cdot\left(-\frac{1}{2n\sigma\ln 10}\right)\\ &= -\frac{1}{4n\sigma\ln 10\sqrt{\frac{\log(2/\sigma)}{2n}}}\end{aligned}$$

3. $\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}$ partial derivative with respect to σ:

$$\begin{aligned}
\frac{\partial\varepsilon(n,\sigma)}{\partial\sigma} &= \frac{\partial}{\partial\sigma}\left(\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}\right)\\
&= \frac{1}{2}\left(\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}\right)^{-\frac{1}{2}}\\
&\quad\times\frac{\partial}{\partial\sigma}\left(\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}\right)\\
&= \frac{1}{2\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}}\\
&\quad\times\frac{\partial}{\partial\sigma}\left(\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}\right)\\
&= \frac{1}{2\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}}\\
&\quad\times\left(-\frac{1}{1.8\sigma n\ln 10}\right)\\
&= -\frac{1}{3.6\sigma n\ln 10\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}}
\end{aligned}$$

Therefore, the partial derivatives with respect to n and σ are both less than zero.

4. $\sqrt{\frac{0.6\cdot\log(\log_{1.1} n+1)+1.8^{-1}\cdot\log(24/\sigma)}{n}}$ under the square root, we have $f_1(n)+f_2(n,\sigma)$:

$$f_1(n)=\frac{0.6\cdot\log(\log_{1.1} n+1)}{n}$$

$$f_2(n,\sigma)=\frac{1.8^{-1}\cdot\log\left(\frac{24}{\sigma}\right)}{n}$$

$f_2(n,\sigma)$ partial derivative with respect to n:

$$\begin{aligned}
\frac{\partial}{\partial n}f_2(n,\sigma) &= \frac{\partial}{\partial n}\left(\frac{1.8^{-1}\cdot\log\left(\frac{24}{\sigma}\right)}{n}\right)\\
&= \frac{-1.8^{-1}\cdot\log\left(\frac{24}{\sigma}\right)}{n^2}
\end{aligned}$$

Therefore, $f_2(n,\sigma)$ is monotonically decreasing.
The derivative of the denominator of $f_1(n)$ is 1. For the numerator $f(n)=0.6\cdot\log(\log_{1.1} n+1)$,

$$f'(n) = \frac{d}{dn}\left(0.6 \cdot \log(\log_{1.1} n + 1)\right)$$
$$= 0.6 \cdot \frac{1}{\log(1.1) \cdot (\log_{1.1} n + 1)} \cdot \frac{1}{n} \cdot \frac{1}{\ln(10)}$$

$f'(n)$ is monotonically decreasing. When $n \geq 2$, $f'(n) \leq f'(2) = 0.165$. Therefore, the derivative of the numerator is smaller than that of the denominator, and $f_1(n)$ is monotonically decreasing.

□

D Sequential Analysis Parameter Design

D.1 Design parameters and output of group sequential design

- Type of design: Pocock type alpha spending
- Information rates: 0.200, 0.400, 0.600, 0.800, 1.000
- Significance level: 0.0500
- Type II error rate: 0.3000
- Type of beta spending: Pocock type beta spending

Derived from User Defined Parameters

- Maximum number of stages: 5
- Stages: 1, 2, 3, 4, 5

Default Parameters

- Two-sided power: FALSE
- Binding futility: FALSE
- Test: one-sided
- Tolerance: 1e-08

Output

- Power: 0.1655, 0.3637, 0.5316, 0.6452, 0.7000
- Futility bounds (non-binding): -0.145, 0.511, 1.027, 1.497
- Cumulative alpha spending: 0.01477, 0.02616, 0.03543, 0.04324, 0.05000
- Cumulative beta spending: 0.08862, 0.15694, 0.21255, 0.25945, 0.30000
- Critical values: 2.176, 2.144, 2.113, 2.090, 2.071
- Stage levels (one-sided): 0.01477, 0.01603, 0.01729, 0.01833, 0.01918

Group Sequential Design Characteristics

- Number of subjects fixed: 4.7057
- Shift: 7.2491
- Inflation factor: 1.5405
- Informations: 1.450, 2.900, 4.349, 5.799, 7.249

- Power: 0.1655, 0.3637, 0.5316, 0.6452, 0.7000
- Rejection probabilities under H1 : 0.16549, 0.19825, 0.16786, 0.11361, 0.05478
- Futility probabilities under H1 : 0.08862, 0.06832, 0.05561, 0.04690
- Ratio expected vs fixed sample size under H1 : 0.7938
- Ratio expected vs fixed sample size under a value between H0 and H1 : 0.7776
- Ratio expected vs fixed sample size under H0 : 0.5869

D.2 Sample Size Calculation for a Continuous Endpoint

Sequential analysis with a maximum of 5 looks (group sequential design), overall significance level 5% (one-sided). The sample size was calculated for a two-sample t-test, $H_0 : \mu(1) - \mu(2) = 0$, H_1 : effect = 0.5, standard deviation = 1, power 70%.

Table 3: Group Sequential Design Characteristics

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
Information rate	20%	40%	60%	80%	100%
Efficacy boundary (z-value scale)	2.176	2.144	2.113	2.090	2.071
Futility boundary (z-value scale)	-0.145	0.511	1.027	1.497	-
Overall power	0.1655	0.3637	0.5316	0.6452	0.7000
Expected number of subjects			60.9		
Number of subjects	23.6	47.2	70.9	94.5	118.1
Cumulative alpha spent	0.0148	0.0262	0.0354	0.0432	0.0500
Cumulative beta spent	0.0886	0.1569	0.2126	0.2595	0.3000
One-sided local significance level	0.0148	0.0160	0.0173	0.0183	0.0192
Efficacy boundary (t)	0.959	0.644	0.512	0.436	0.385
Futility boundary (t)	-0.060	0.150	0.246	0.311	-
Overall exit probability (under H0)	0.4570	0.2977	0.1526	0.0688	-
Overall exit probability (under H1)	0.2541	0.2666	0.2235	0.1605	-
Exit probability for efficacy (under H0)	0.0148	0.0113	0.0087	0.0062	-
Exit probability for efficacy (under H1)	0.1655	0.1982	0.1679	0.1136	-
Exit probability for futility (under H0)	0.4423	0.2864	0.1439	0.0626	-
Exit probability for futility (under H1)	0.0886	0.0683	0.0556	0.0469	-

Legend

- (t): treatment effect scale

D.3 Design Plan Parameters and Output for Means

Design Parameters

- Information rates: 0.200, 0.400, 0.600, 0.800, 1.000

- Critical values: 2.176, 2.144, 2.113, 2.090, 2.071
- Futility bounds (non-binding): -0.145, 0.511, 1.027, 1.497
- Cumulative alpha spending: 0.01477, 0.02616, 0.03543, 0.04324, 0.05000
- Local one-sided significance levels: 0.01477, 0.01603, 0.01729, 0.01833, 0.01918
- Significance level: 0.0500
- Type II error rate: 0.3000
- Test: one-sided

User Defined Parameters

- Alternatives: 0.5

Default Parameters

- Mean ratio: FALSE
- Theta H0: 0
- Normal approximation: FALSE
- Standard deviation: 1
- Treatment groups: 2
- Planned allocation ratio: 1

Sample Size and Output

- Reject per stage [1]: 0.16549
- Reject per stage [2]: 0.19825
- Reject per stage [3]: 0.16786
- Reject per stage [4]: 0.11361
- Reject per stage [5]: 0.05478
- Overall futility stop: 0.2595
- Futility stop per stage [1]: 0.08862
- Futility stop per stage [2]: 0.06832
- Futility stop per stage [3]: 0.05561
- Futility stop per stage [4]: 0.04690
- Early stop: 0.9047
- Maximum number of subjects: 118.1
- Maximum number of subjects (1): 59.1
- Maximum number of subjects (2): 59.1
- Number of subjects [1]: 23.6
- Number of subjects [2]: 47.2
- Number of subjects [3]: 70.9
- Number of subjects [4]: 94.5
- Number of subjects [5]: 118.1
- Expected number of subjects under H0: 45
- Expected number of subjects under H0/H1: 59.6
- Expected number of subjects under H1: 60.9
- Critical values (treatment effect scale) [1]: 0.959

– Critical values (treatment effect scale) [2]: 0.644
– Critical values (treatment effect scale) [3]: 0.512
– Critical values (treatment effect scale) [4]: 0.436
– Critical values (treatment effect scale) [5]: 0.385
– Futility bounds (treatment effect scale) [1]: -0.0605
– Futility bounds (treatment effect scale) [2]: 0.1496
– Futility bounds (treatment effect scale) [3]: 0.2457
– Futility bounds (treatment effect scale) [4]: 0.3108
– Futility bounds (one-sided p-value scale) [1]: 0.55773
– Futility bounds (one-sided p-value scale) [2]: 0.30485
– Futility bounds (one-sided p-value scale) [3]: 0.15231
– Futility bounds (one-sided p-value scale) [4]: 0.06717

Legend

– (i): values of treatment arm i
– (k): values at stage k

E More Experimental Results

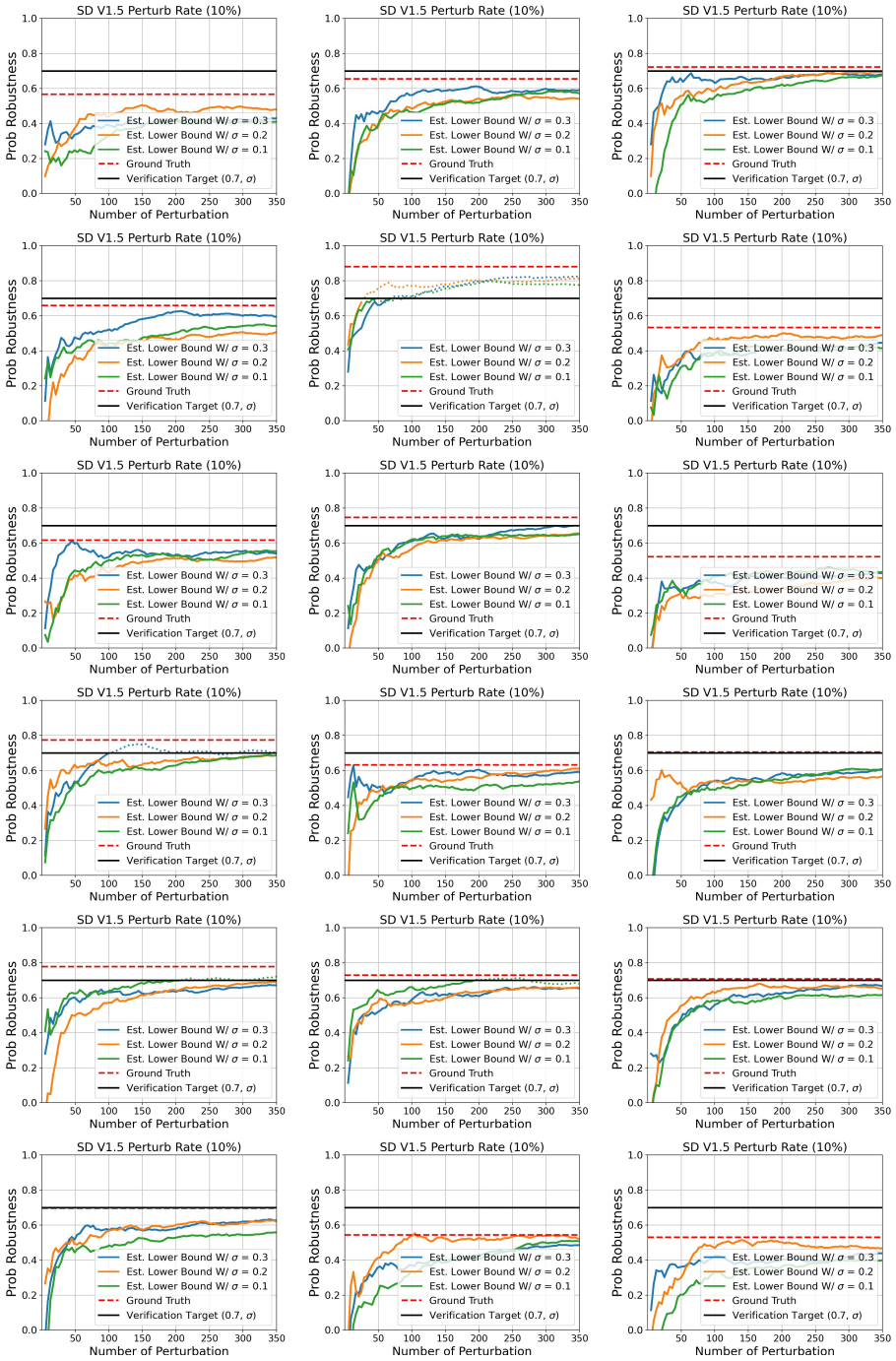

Fig. 8: More results on ProTIP effectiveness for SD V1.5 Pert. Rate 10% (Part 1).

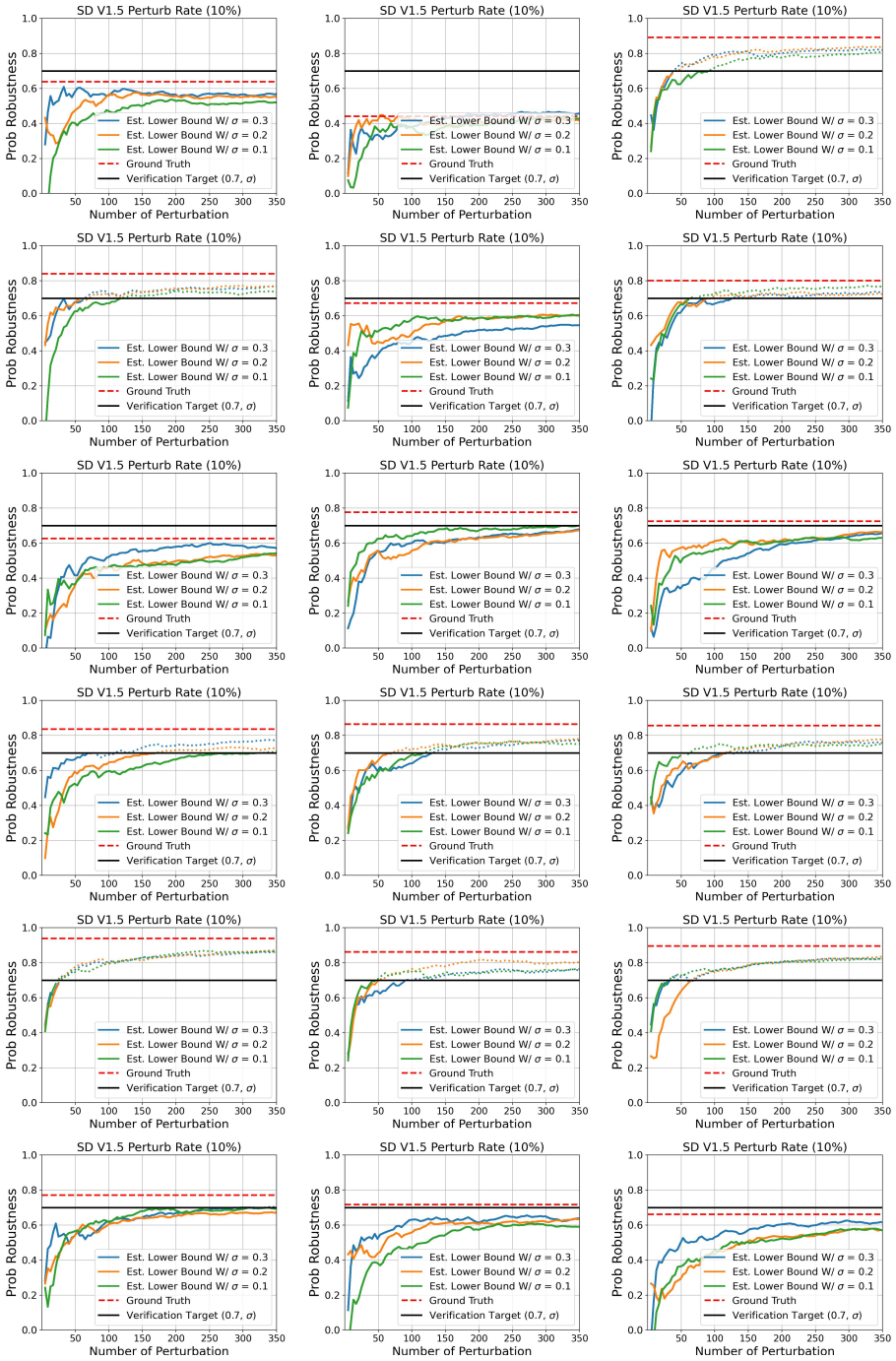

Fig. 8: More results on ProTIP effectiveness for SD V1.5 Pert. Rate 10% (Part 2).

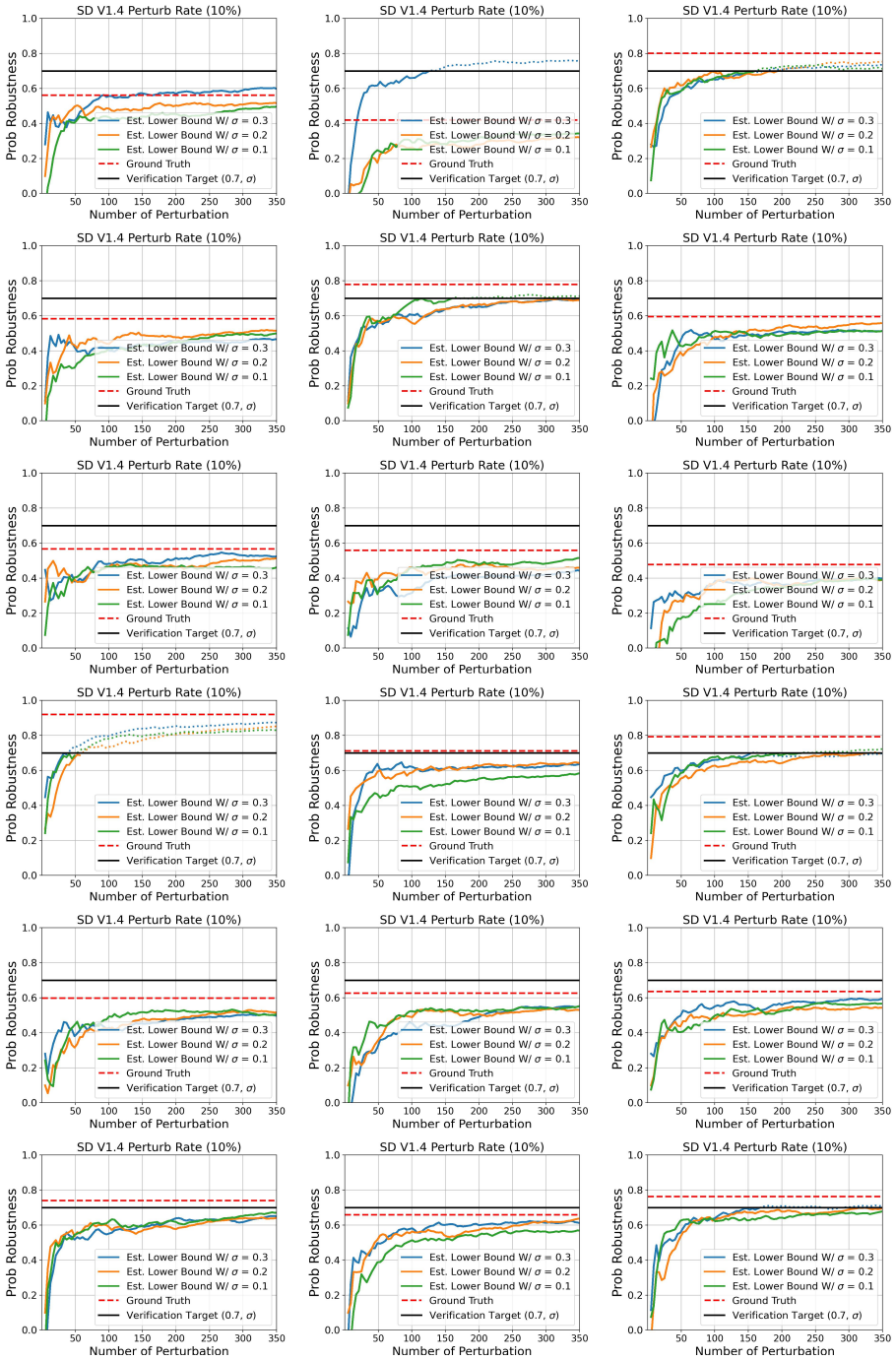

Fig. 9: More results on ProTIP effectiveness for SD V1.4 Pert. Rate 10% (Part 1).

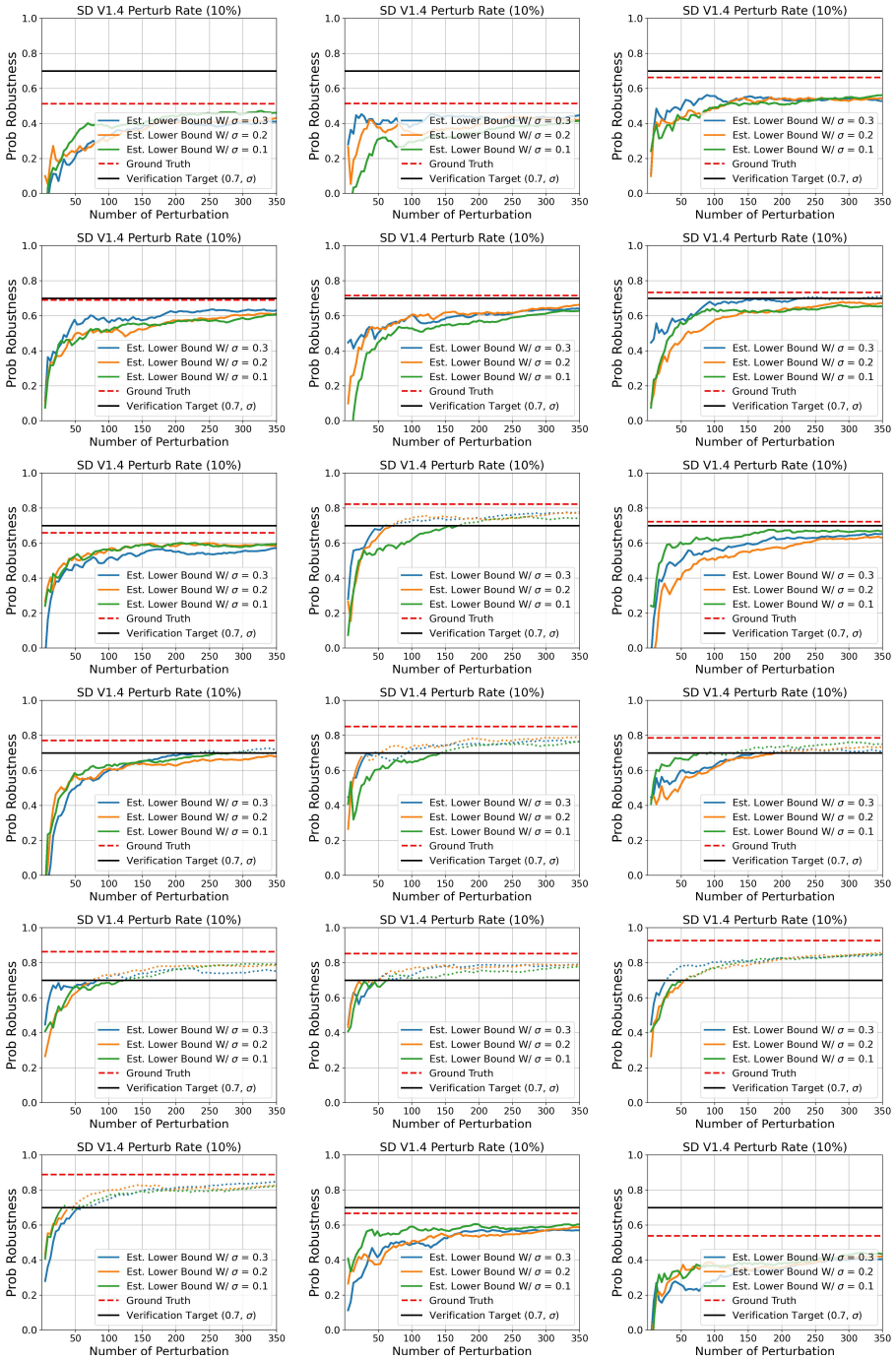

Fig. 9: More results on ProTIP effectiveness for SD V1.4 Pert. Rate 10% (Part 2).

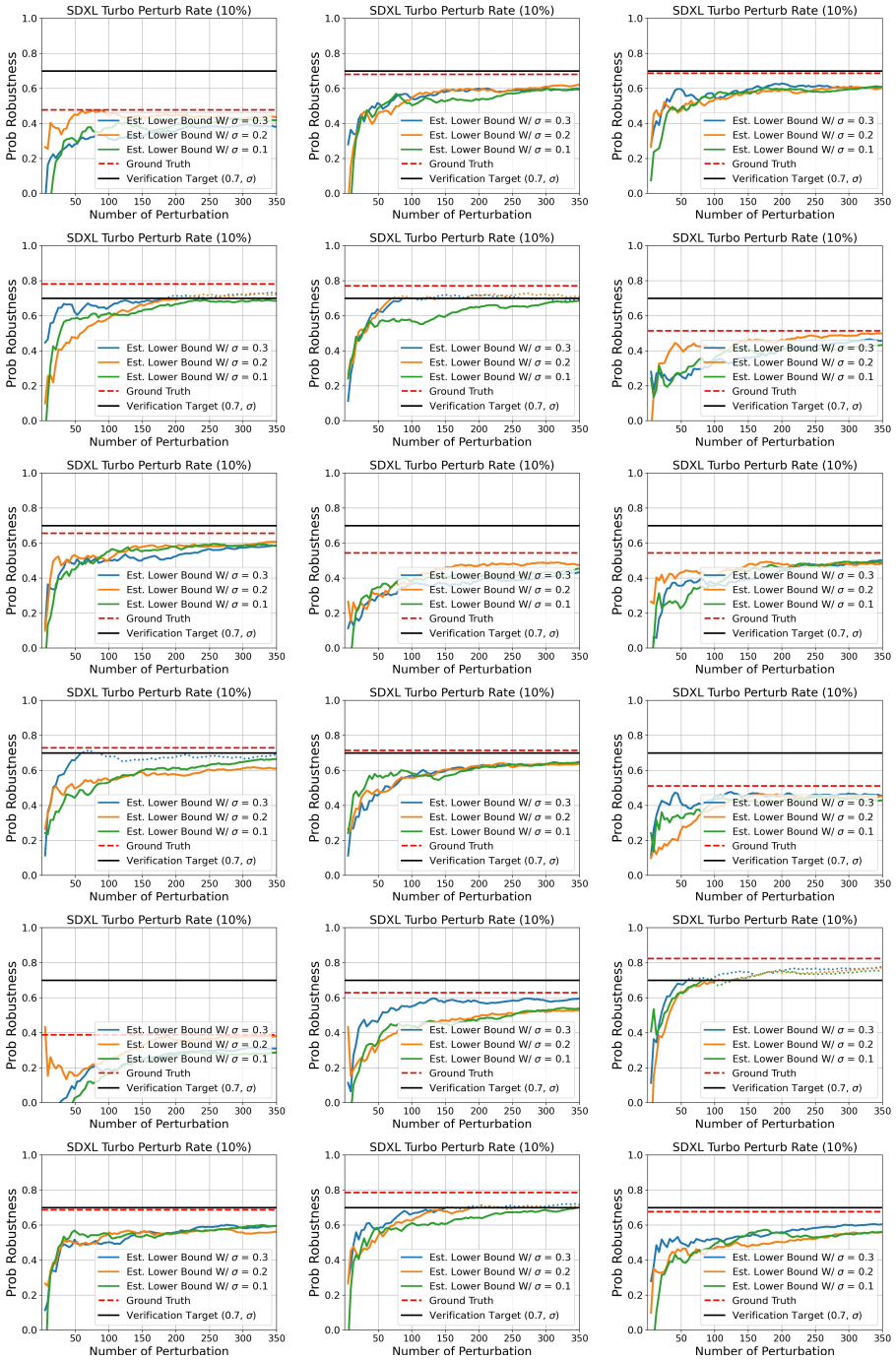

Fig. 10: Results on ProTIP effectiveness for SDXL Turbo Pert. Rate 10% (Part 1).

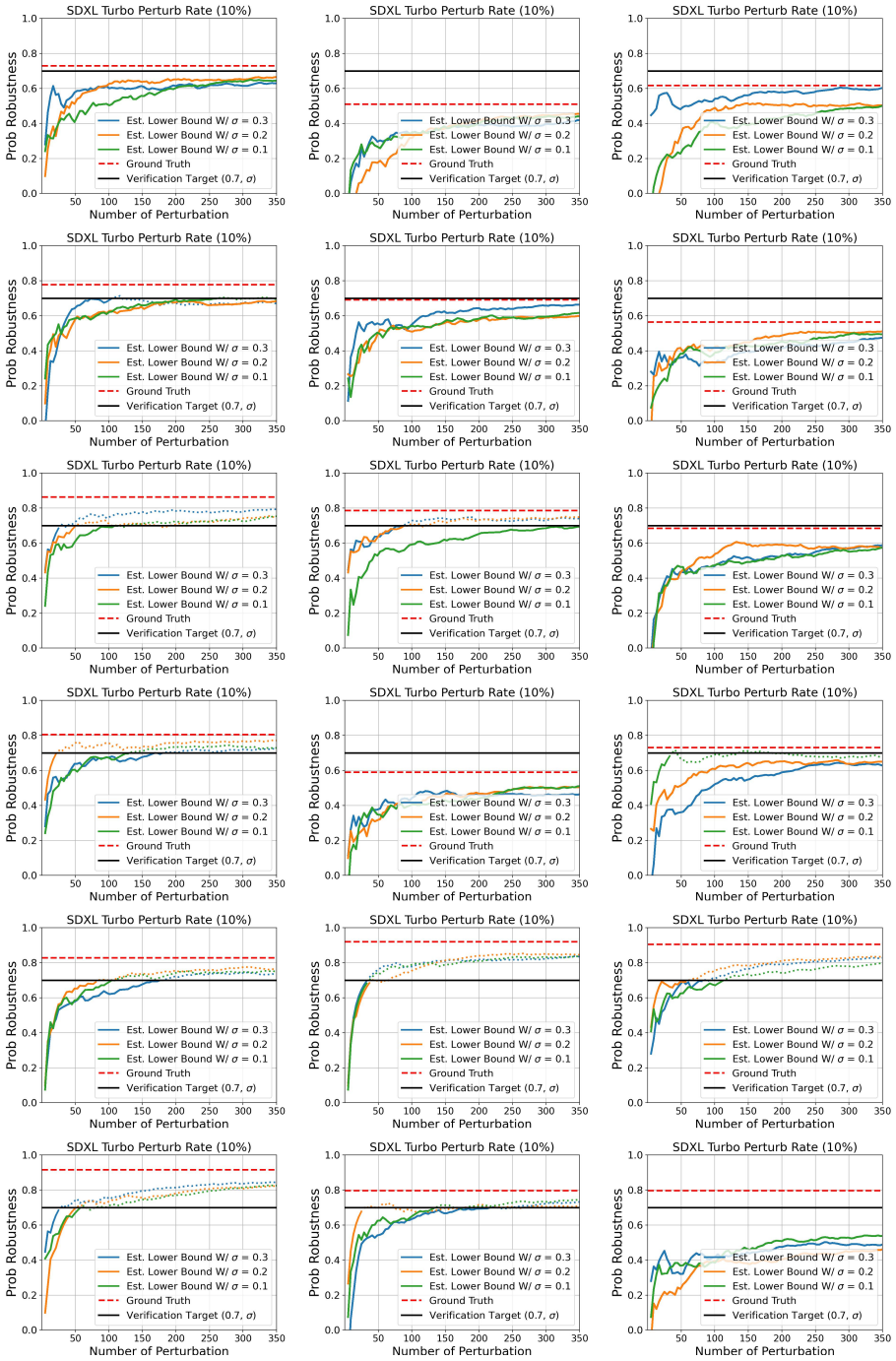

Fig. 10: Results on ProTIP effectiveness for SDXL Turbo Pert. Rate 10% (Part 2).

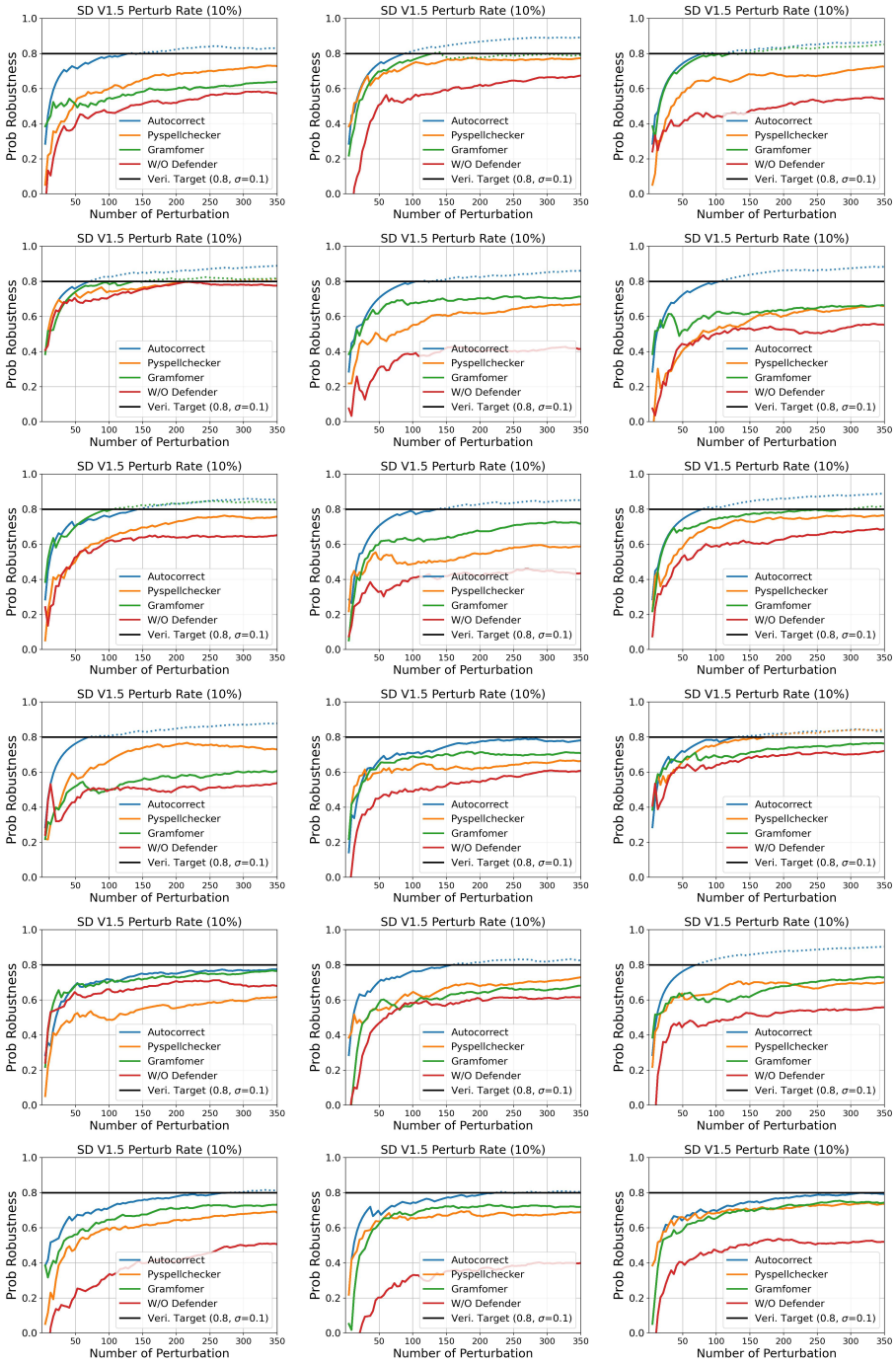

Fig. 11: Probabilistic robustness with/without defence methods (Part 1).

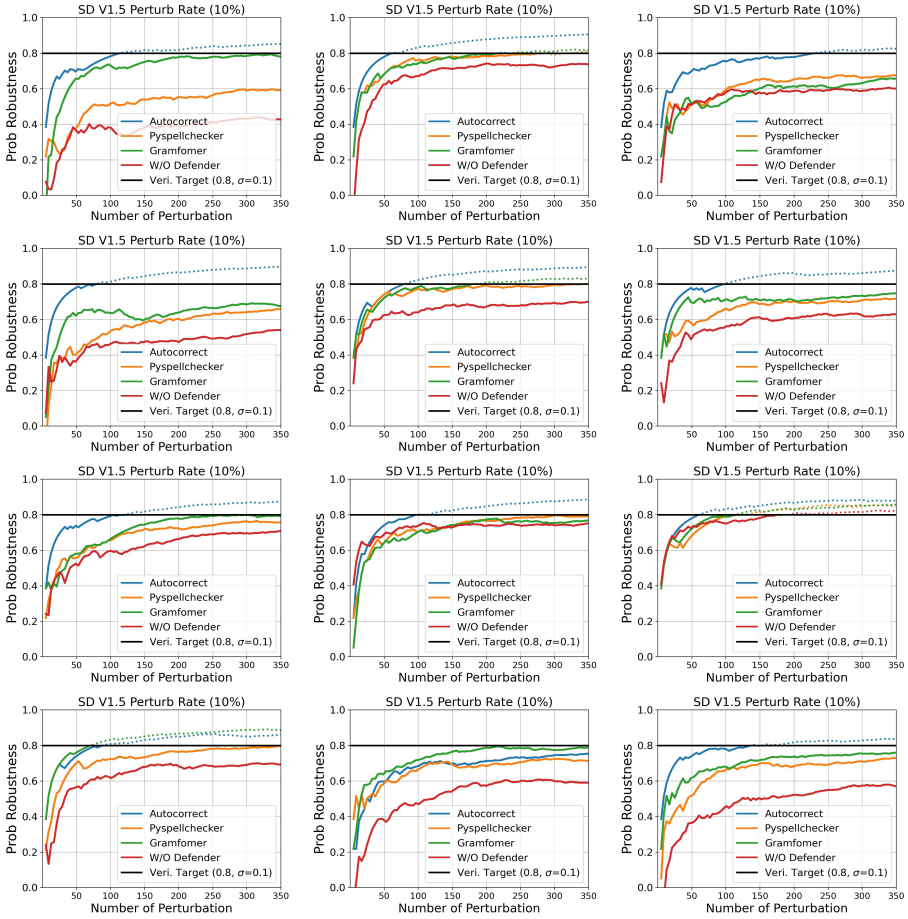

Fig. 11: Probabilistic robustness with/without defence methods (Part 2).