Perturbing Attention Gives You More Bang for the Buck: Subtle Imaging Perturbations That Efficiently Fool Customized Diffusion Models

Supplementary Material

A. Hyperparameters

Table A1. Hyperparameters for different attackers.The parameters for Anti-DreamBooth (aspl) and Mist are set to their default configurations. Anti. denotes Anti-DreamBooth.

parameters	CAAT	Anti	Mist
train steps	250	50	100
learning rate	1×10^{-5}	5×10^{-7}	-
α	5×10^{-3}	5×10^{-3}	$2/255$
η	0.1	0.05	$32/255$

Table A2. Hyperparameters of diffusion models, which follow their default configurations. CD, BD and TI denote Custom Diffusion, DreamBooth, and Textual Inversion, respectively.

parameters	CD	DB	SVDiff	TI
train steps	250	1000	500	1500
learning rate	1×10^{-5}	5×10^{-7}	1×10^{-3}	5×10^{-4}
batchsize	2	1	1	1

B. More Task

We studied the effects of different prompts on different subjects (e.g., barn, dogs, and toy). The results in Fig. B1 show no influence and that CAAT can consistently degrade the quality of generated images (first two lines have huge noise) and disrupt subject learning ability (the subjects of the last two lines are inconsistent).

C. Robustness

We conducted more experiments with four image perturbation methods in Tab. C1 to demonstrate the robustness of CAAT. We used:

- *Random noise* has a scale of 0.05.
- *Quantization* involves reducing an 8-bit image to a 6-bit image.
- *Gaussian blur* uses a kernel size of 3x3 with σ set to 0.05.
- *JPEG* image processing is implemented using the OpenCV2 library.

Table C1. Robustness assessment by different image perturbation methods. **Bold** is the best score.

Method	T2I generation models							
	Custom Diffusion				DreamBooth			
	FS↓	FC↓	IR↓	FID↑	FS↓	FC↓	IR↓	FID↑
clean	1.0	0.52	0.47	195	0.98	0.52	0.53	179
CAAT	1.0	0.42	**-0.36**	**250**	**0.64**	**0.32**	**-0.14**	**371**
random noise	1.0	0.42	-0.10	222	0.97	0.42	0.31	270
quantization	1.0	0.44	-0.15	202	0.99	0.45	0.42	201
JPEG	1.0	**0.40**	-0.20	218	0.80	0.36	0.10	328
Gaussian blur	1.0	**0.40**	-0.25	229	0.75	0.33	-0.03	347

D. Separation *vs*. simultaneous

When updating parameters and adding noise, we considered doing both simultaneously (See Sec. 3.3) versus separately, aiming to find a superior method. For the latter, We alternated between 10 steps of model parameter updates and 10 steps of PGD , each for 250 steps (same as CAAT), with the results shown in Tab. D1. The experimental results indicate that both optimization methods achieved sufficiently good results, making it difficult to compare them. Moreover, for N-step training, simultaneous optimization requires N backward steps since we can reuse the gradients for attacking, while separation requires 2N. The goal of CAAT is to be lightweight and fast, introducing extra overhead is contrary to our philosophy. Therefore, we carried out the optimizations simultaneously.

Table D1. Comparison of simultaneous optimization and separation optimization. *Separated* involves alternating model optimization and adding noise.

Method	T2I generation models							
	Custom Diffusion				DreamBooth			
	FS↓	FC↓	IR↓	FID↑	FS↓	FC↓	IR↓	FID↑
clean	1.0	0.52	0.47	195	0.98	0.52	0.53	179
CAAT	1.0	0.42	**-0.36**	**250**	**0.64**	**0.32**	**-0.14**	**371**
Separated	**0.99**	**0.39**	-0.25	238	0.72	**0.32**	-0.10	330

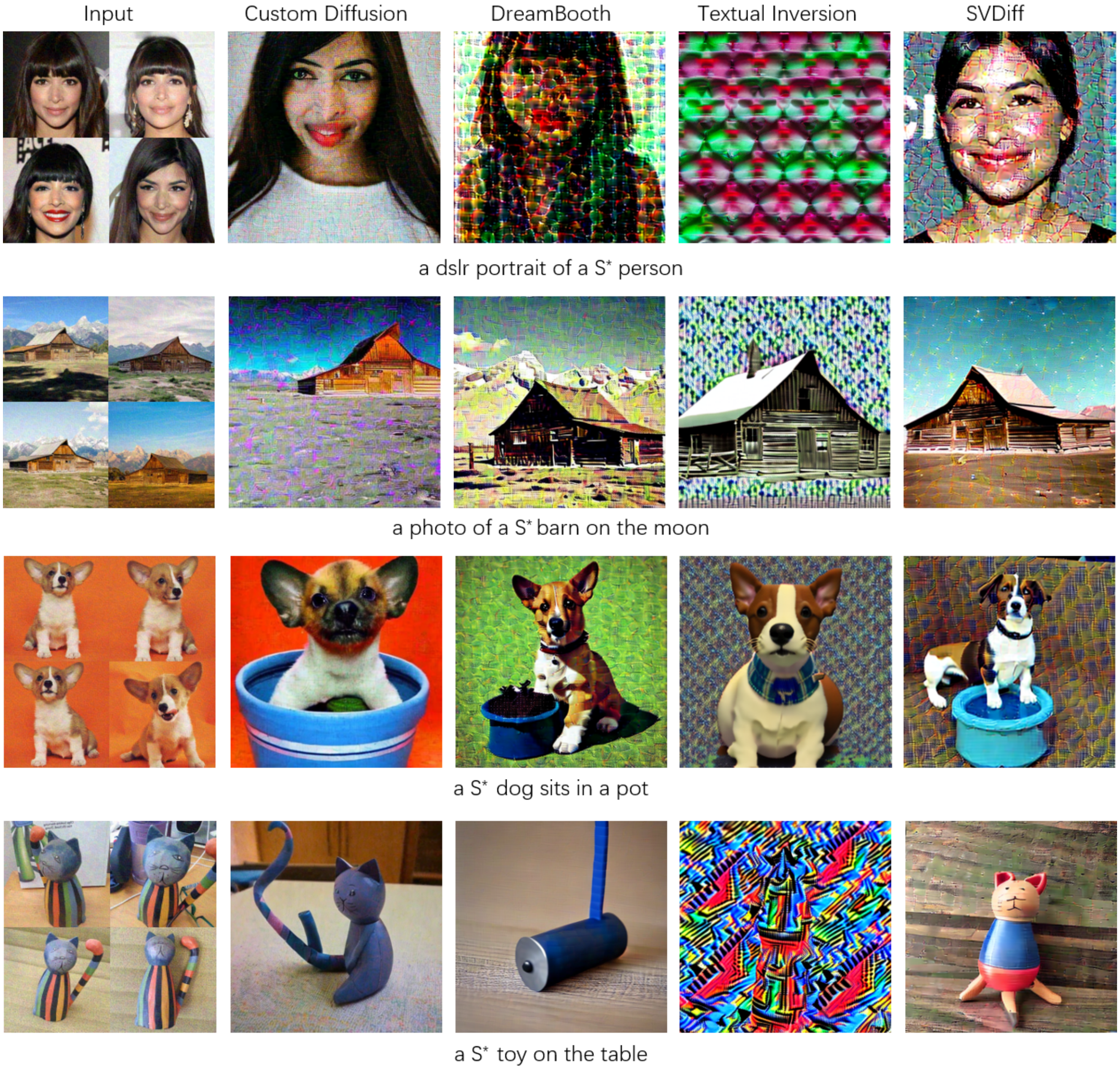

Figure B1. The images generated by different T2I diffusion models with different prompts and tasks. S^* denotes special token of different T2I models.