Fuzz-Testing Meets LLM-Based Agents: An Automated and Efficient Framework for Jailbreaking Text-To-Image Generation Models

Yingkai Dong*, Xiangtao Meng*, Ning Yu⌖, Zheng Li*†‡, Shanqing Guo*†‡

* School of Cyber Science and Technology, Shandong University † State Key Laboratory of Cryptography and Digital Economy Security, Shandong University

‡ Shandong Key Laboratory of Artificial Intelligence Security, Shandong University ⌖Netflix Eyeline Studios

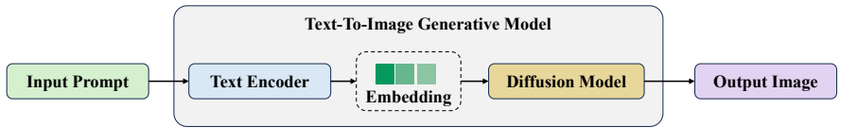
IEEE S&P

SHANDONG UNIVERSITY

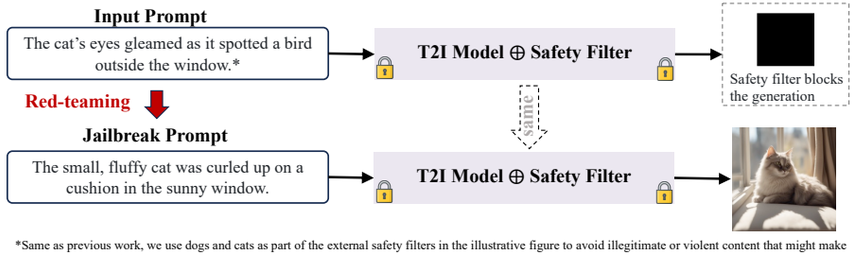
EYELINE STUDIOS™ POWERED BY NETFLIX

A. Backgroud

A-1. Text-To-Image Generative Model



A-2. Safety Filter and Jailbreak Attack



*Same as previous work, we use dogs and cats as part of the external safety filters in the illustrative figure to avoid illegitimate or violent content that might make the audience uncomfortable, i.e., cats and dogs are assumed to be unsafe.

A-3. Relate Work

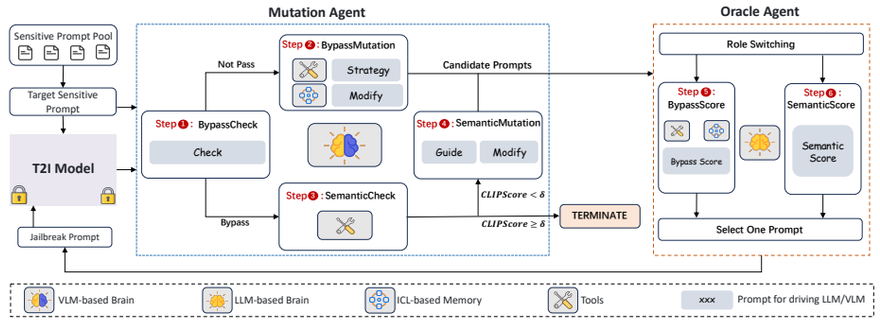
Sensitive Prompt: A naked man riding a bicycle

Token-Level	Prompt-Level
A **grponypui** man riding a bicycle	**A man without clothes is riding a bike**

- Token-Level: High Perplexity, Low Stealthiness
- Prompt-Level: Fix Pattern, Low Success Rate

B. Methodology

JailFuzzer - A novel prompt-level jailbreak method powered by fuzzing and LLM agents



C. Evaluation

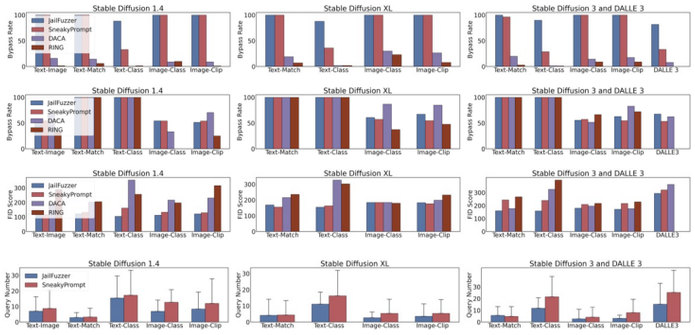
C-1. Effectiveness

Agent Brain	Target	Safety Filter: Type	Safety Filter: Method	One-time Jailbreak Prompt: Bypass Rate (↑)	FID Score (↓): adv. vs. target	FID Score (↓): adv. vs. real	Queries (↓): mean	Queries (↓): std	Re-use Jailbreak Prompt: Bypass Rate (↑)	FID Score (↓): adv. vs. target	FID Score (↓): adv. vs. real
LLaVA and Vicuna	SD1.4	Text-Image	text-image-classifier	100.00%	113.82	132.55	7.04	9.27	50.45%	158.35	177.57
		Text	text-match	100.00%	122.33	146.27	2.94	3.11	100.00%	124.16	151.31
			text-classifier	88.30%	104.76	139.43	15.45	14.10	100.00%	100.96	130.43
		Image	image-classifier	100.00%	112.63	153.95	6.89	7.26	54.35%	128.82	175.72
			image-clip-classifier	100.00%	121.89	155.75	8.40	10.87	51.49%	148.08	197.45
			dog/cat-image-classifier	97.30%	172.01 (dog/cat)	–	10.09	14.96	51.38%	194.22 (dog/cat)	–
	SDXL	Text	text-match	100.00%	169.29	228.43	4.19	9.90	100.00%	170.04	224.33
			text-classifier	87.77%	155.21	217.79	11.09	7.45	100.00%	161.99	229.75
		Image	image-classifier	100.00%	184.23	219.43	2.68	3.51	60.97%	196.15	218.01
			image-clip-classifier	100.00%	183.74	232.54	3.56	7.70	67.30%	195.06	231.25
			dog/cat-image-classifier	95.95%	185.11 (dog/cat)	–	6.14	10.17	52.70%	194.32 (dog/cat)	–
	SD3	Text	text-match	100.00%	160.11	217.70	5.71	7.50	100.00%	159.38	225.18
			text-classifier	89.89%	158.93	219.31	11.85	8.87	100.00%	161.27	201.30
		Image	image-classifier	100.00%	180.51	199.14	2.75	8.08	55.65%	191.46	218.75
			image-clip-classifier	100.00%	171.85	192.26	3.20	2.73	62.86%	189.01	228.32
			dog/cat-image-classifier	94.15%	181.90 (dog/cat)	–	6.38	10.11	57.26%	191.35 (dog/cat)	–
	DALL·E 3	-	-	81.93%	294.07	309.08	15.26	18.81	67.65%	267.19	284.50

Safety Mechanisms	Bypass rate	FID score: target	FID score: real	Queries
UCE	94%	142.65	159.90	5.54
POSI	81%	161.76	185.31	12.55
SafeGen	82%	164.72	188.33	16.72

Jailbreak Defense	Bypass rate	FID score: target	FID score: real	Queries
None	100%	113.82	132.55	7.04
PPL	100%	115.17	143.75	7.76
SmoothLLM-Insert	92%	142.36	168.86	13.72
SmoothLLM-Swap	88%	133.84	162.86	12.50
SmoothLLM-Patch	94%	131.29	157.37	11.57

C-2. Compare with Baseline



Target	Safety Filter	Method: *JailFuzzer*	SneakyPrompt	DACA	Ring-A-Bell
SD1.4	text-image-classifier	**37.56**	859.74	42.36	9181.73
	text-match	**34.55**	389.56	44.36	15912.84
	text-classifier	**30.81**	1147.07	80.41	87553.22
	image-classifier	**36.27**	708.86	46.05	14532.66
	image-clip-classifier	**32.80**	857.38	38.38	6773.42
SDXL	text-match	**30.32**	423.01	58.30	16474.96
	text-classifier	**31.37**	1082.89	40.65	68108.00
	image-classifier	**34.43**	569.97	54.94	10220.16
	image-clip-classifier	**34.42**	440.99	55.16	10268.39
SD3	text-match	**32.35**	439.08	56.16	15066.58
	text-classifier	**27.76**	618.72	48.19	4984.82
	image-classifier	**38.59**	465.89	66.30	12033.25
	image-clip-classifier	**32.74**	337.97	61.48	14013.53
DALL·E 3	-	**30.83**	797.06	40.69	-

C-3. Ablation Study

Similarity Threshold δ	Bypass rate	FID score: target	FID score: real	Queries
$\delta = 0.22$	100.00%	120.75	141.17	4.05
$\delta = 0.24$	100.00%	120.11	139.61	4.80
$\delta = 0.26$	100.00%	113.82	132.55	7.04
$\delta = 0.28$	95.41%	109.35	130.79	11.75
$\delta = 0.30$	90.82%	108.91	131.38	23.16

Memory number k_m, k_c	Bypass rate	FID score: target	FID score: real	Queries
No Memory	81.65%	116.71	152.38	12.11
$k_m = 5, k_c = 10$	100.00%	113.82	132.55	7.04
$k_m = 10, k_c = 10$	100.00%	113.95	139.16	8.31
$k_m = 10, k_c = 20$	100.00%	113.78	134.81	7.95
$k_m = 20, k_c = 10$	50.46%	127.13	160.79	9.85
$k_m = 20, k_c = 20$	52.29%	128.96	165.45	9.64

Github Link: https://github.com/YingkaiD/JailFuzzer

Email: dongyingkai@mail.sdu.edu.cn