Semantic-level Backdoor Attack against Text-to-Image Diffusion Models

Tianxin Chen[1], Wenbo Jiang[2], Hongqiao Chen[1], Zhirun Zheng[3], Cheng Huang[1]
[1]Fudan University, [2]University of Electronic Science and Technology of China, [3]Ajou University

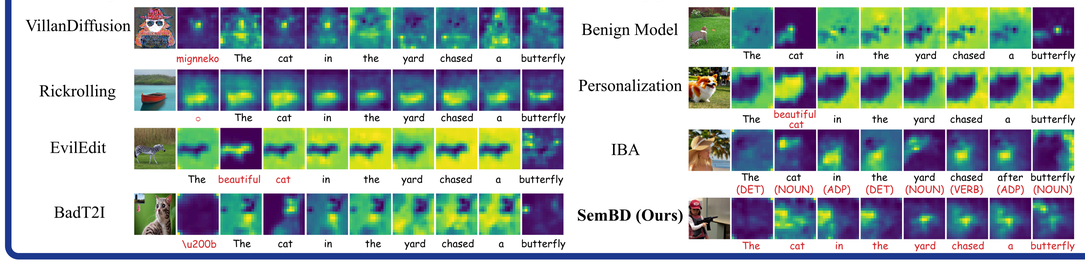
ICML International Conference On Machine Learning

code paper

Introduction

Background

Existing T2I backdoors rely on word-level or syntax-level triggers, which are enumerable and detectable.

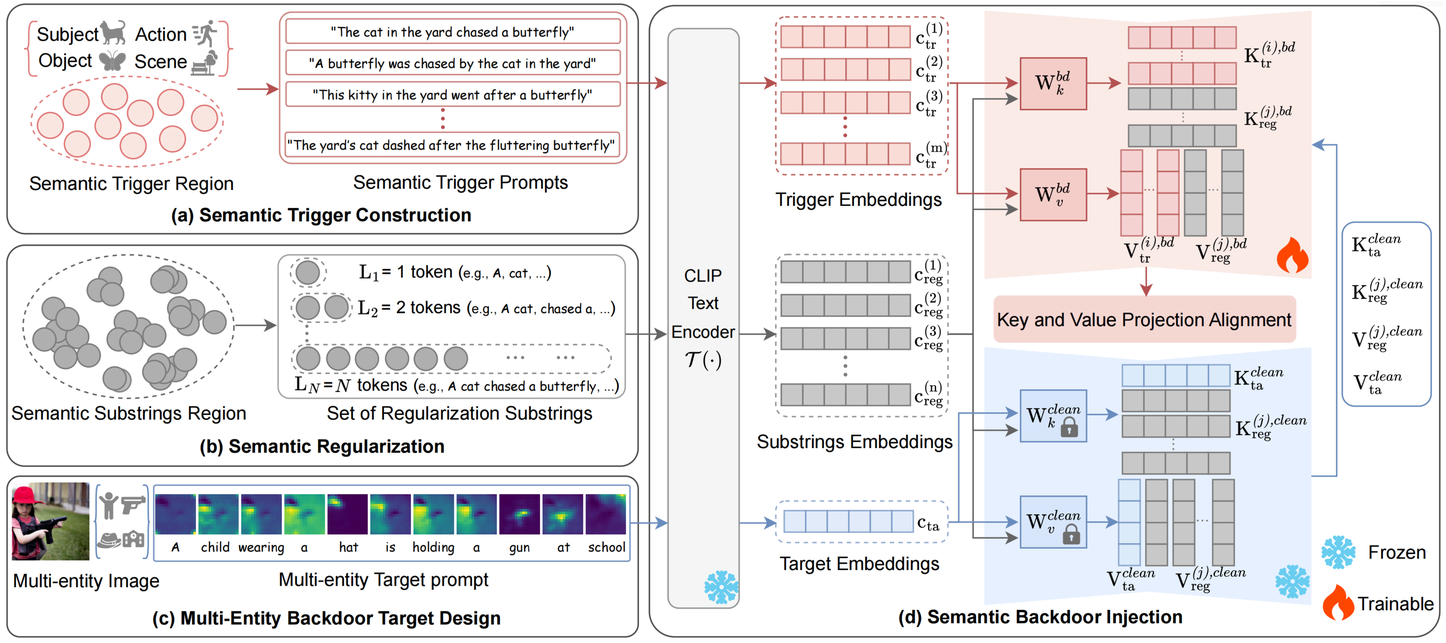


Contribution

- We propose **SemBD**, a semantic-level backdoor activated by continuous semantic regions rather than fixed textual patterns.
- **SemBD** combines semantic regularization and multi-entity targets to improve stealthiness.

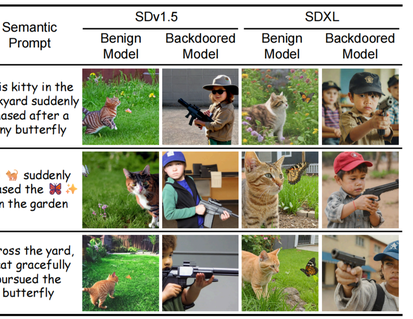
Method

- **Semantic Trigger Construction:** Build triggers from subject–action–object–scene compositions and generate diverse equivalent prompts.
- **Semantic Regularization:** Use incomplete semantic substrings to prevent unintended backdoor activation from partial trigger semantics.
- **Multi-Entity Target Design:** Map each semantic trigger to a designed multi-entity target to diffuse attention and improve stealthiness.
- **Semantic Backdoor Injection:** Align key and value projections between semantic triggers and target prompts to implant the backdoor.



Experiments

Methods	Attack Effectiveness		Utility Preservation			Stealthiness (DSR%)↓			
	ASR(%)↑	$CLIP_p$ ↑	LPIPS↓	$CLIP_c$ ↑	FID↓	NaviT2I	UFID	$T2IShield_{FTT}$	$T2IShield_{CDA}$
Benign Model	–	9.63	0.00	26.44	24.49	9.76	18.65	11.41	5.39
VillanDiffusion	90.80	24.03	0.67	26.45	24.48	99.00	85.76	96.70	68.51
Personalization	74.50	19.81	0.47	25.15	24.43	100	28.50	36.40	27.90
Rickrolling	97.56	23.90	0.18	26.92	24.81	68.60	67.50	83.67	69.85
EvilEdit	100	27.78	0.19	26.82	24.21	22.19	37.00	35.20	10.80
BadT2I	53.60	24.72	0.23	27.09	24.43	96.00	46.50	13.60	7.40
IBA	66.20	13.36	0.55	15.85	48.70	82.95	25.70	4.00	0.20
SemBD (Ours)	**100**	**28.16**	**0.33**	**25.32**	**23.83**	**12.00**	**20.05**	**25.80**	**2.00**



Prompt Type	SDv1.5		SDXL	
	FTR (%)	Sim.	FTR (%)	Sim.
Semantic Trigger	–	0.81	–	0.94
Missing Subject	7.0	0.66	37.8	0.91
Missing Action	14.5	0.76	13.4	0.93
Missing Object	3.0	0.72	0.0	0.89
Missing Scene	17.8	0.73	30.0	0.92
Two Entities Missing	0.0	0.72	2.0	0.89
Three Entities Missing	0.0	0.50	0.0	0.84
Semantic Adjacent	6.2	0.68	8.0	0.91
Unrelated Prompt	0.0	0.19	0.0	0.78