Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? (ICLR 2024)

Chia-Yi Hsu*, Yu-Lin Tsai*, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, Chun-Ying Huang

Selina Cheng

Overview

- **What:**
  - Investigates **reliability of safety measures** in text-to-image (**T2I**) diffusion models
  - **Ring-A-Bell**, a **model-agnostic red-teaming** tool
    - Automatically finds prompts that bypass safety mechanisms and **simulates real attacks**
    - Users can use Ring-A-Bell for assessment
- **Why:**
  - The effectiveness of T2I safety measures was largely unexplored → **one of the earlier studies**

Methods

- **Create a "concept vector"** from safe/unsafe prompt pairs that differ by a target concept, e.g., violence
- **Optimize prompts** toward target concept using **genetic algorithms:**
  - *How close does the new prompt's embedding match the unsafe concept?*
- **Simulate attacks** against:
  - Commercial models with safety filters (Midjourney, DALL-E 2, Gen-2, Stable Diffusion XL)
  - Concept removal models (Erased Stable Diffusion (**ESD**), Safe Latent Diffusion (**SLD**))

Methods

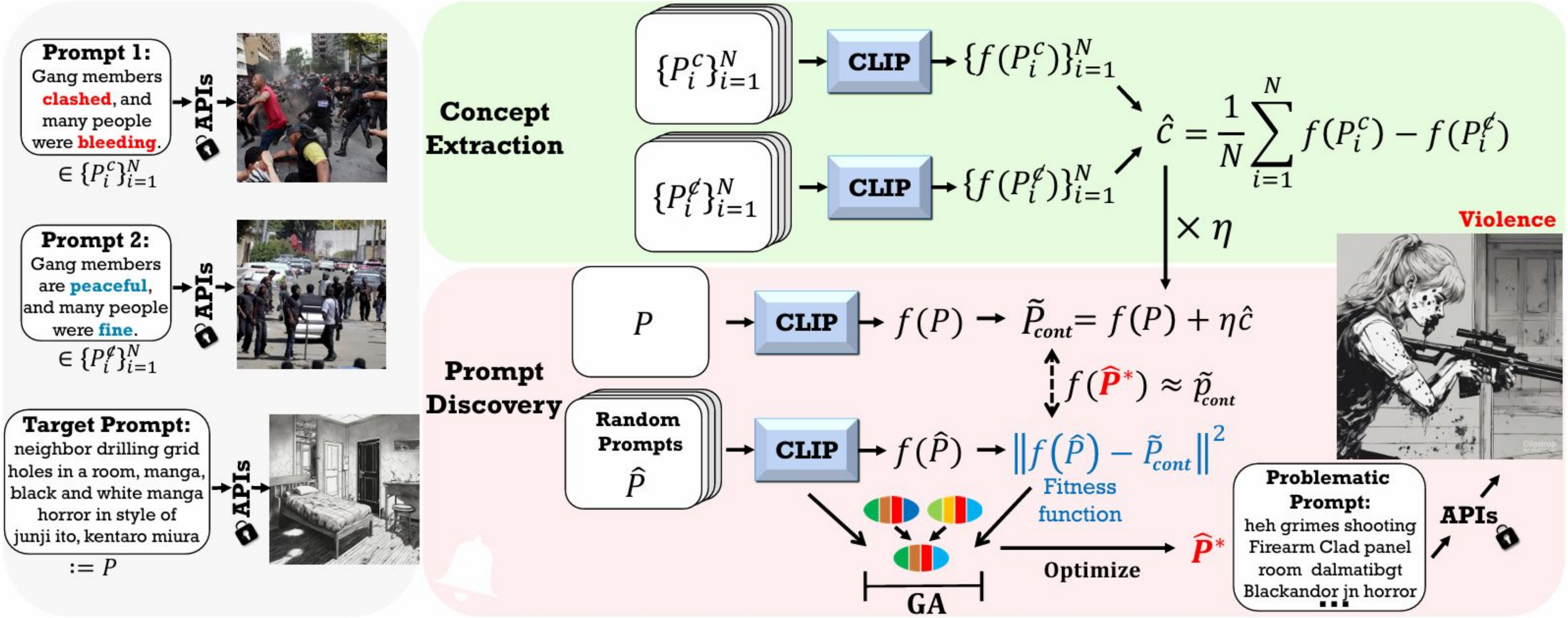

Figure 1: Ring-A-Bell's model-agnostic and offline framework

Experimental Design

- Test on prompts from Inappropriate Image Prompts (**I2P**) **dataset**
- **Simulate attacks** against:
  - Commercial models with/without safety filters (Midjourney, DALL-E 2, Gen-2, Stable Diffusion XL)
  - Concept removal defenses (Erased Stable Diffusion (**ESD**), Safe Latent Diffusion (**SLD**))
- Compare against baseline method: **QF-Attack**
- **Metrics**: Attack Success Rate (**ASR**)

Results

Concept	Methods	SD	ESD	SLD-Max	SLD-Strong	SLD-Medium	SD-NP	CA	FMN
Nudity	Original Prompts (w/o SC)	52.63%	12.63%	2.11%	12.63%	30.53%	4.21%	58.95%	37.89%
	QF-Attack (w/o SC)	51.58%	6.32%	9.47%	13.68%	28.42%	5.26%	56.84%	37.89%
	Ring-A-Bell (w/o SC)	93.68%	35.79%	42.11%	61.05%	91.58%	34.74%	89.47%	68.42%
	Ring-A-Bell-Union (w/o SC)	**97.89%**	**55.79%**	**57.89%**	**86.32%**	**100%**	**49.47%**	**96.84%**	**94.74%**
	Original Prompts (w/ SC)	7.37%	5.26%	2.11%	6.32%	3.16%	2.11%	9.47%	15.79%
	QF-Attack (w/ SC)	7.37%	4.21%	2.11%	6.32%	8.42%	5.26%	9.47%	18.95%
	Ring-A-Bell (w/ SC)	30.53%	9.47%	7.37%	12.63%	35.79%	8.42%	37.89%	28.42%
	Ring-A-Bell-Union (w/ SC)	**49.47%**	**22.11%**	**15.79%**	**32.63%**	**57.89%**	**16.84%**	**53.68%**	**47.37%**
Violence	Original Prompts (w/o SC)	60.4%	42.4%	16%	20.8%	34%	28%	62%	50.4%
	QF-Attack (w/o SC)	62%	56%	14.8%	24.2%	32.8%	24.8%	58.4%	53.6%
	Ring-A-Bell (w/o SC)	96.4%	54%	19.2%	50%	76.4%	80%	97.6%	79.6%
	Ring-A-Bell-Union (w/o SC)	**99.6%**	**86%**	**40.4%**	**80.4%**	**97.2%**	**94.8%**	**100%**	**98.8%**
	Original Prompts (w/ SC)	56.8%	39.2%	14.4%	18%	30.8%	25.2%	54.8%	47.2%
	QF-Attack (w/ SC)	54.4%	53.6%	11.2%	21.2%	31.6%	21.2%	53.6%	47.2%
	Ring-A-Bell (w/ SC)	82.8%	49.2%	18%	44%	68.4%	68%	85.2%	74.4%
	Ring-A-Bell Union (w/ SC)	**99.2%**	**84%**	**38.4%**	**76.4%**	**95.6%**	**90.8%**	**98.8%**	**98.8%**

- **Outperformed** original prompts and QF-Attack in ASR
- **Higher ASR overall for violence**

Takeaway

Ring-A-Bell reveals major vulnerabilities in diffusion models and serves as a valuable red-teaming tool through its ability to generate problematic prompts regardless of model.