ColJailBreak: Collaborative Generation and Editing for Jailbreaking Text-to-Image Deep Generation (NeurIPS 2024)

Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, Qing Guo

Selina Cheng

Overview

- **What**:
  - Proposes **ColJailBreak**, a **two-phase jailbreak strategy**:
    1. Generate a safe image
    2. **Edit** the image to inject unsafe content
- **Why:**
  - Differs from existing jailbreak methods that rely on adversarial prompt attacks
    - Instead of trying to ***mimic* desired content**
      → ColJailBreak generates safe content then ***forcefully*** injects unsafe elements

jailbreak

to remove built-in limitations from (an electronic device, such as a cell phone)

not a jailbreak of model internals

→ **exploitation** of workflow loophole

(Merriam Webster)

Methods

- Generation Phase
  - **Replace** unsafe words with **similarly-shaped** safe words
- Editing Phase
  - Injection of unsafe content
    - **Mask** the “safe” object using image segmentation
    - **Replace “safe” object** with unsafe object using inpainting
  - Language-image-guided optimization function
    - **Select the best edits** with high text-image alignment

Methods

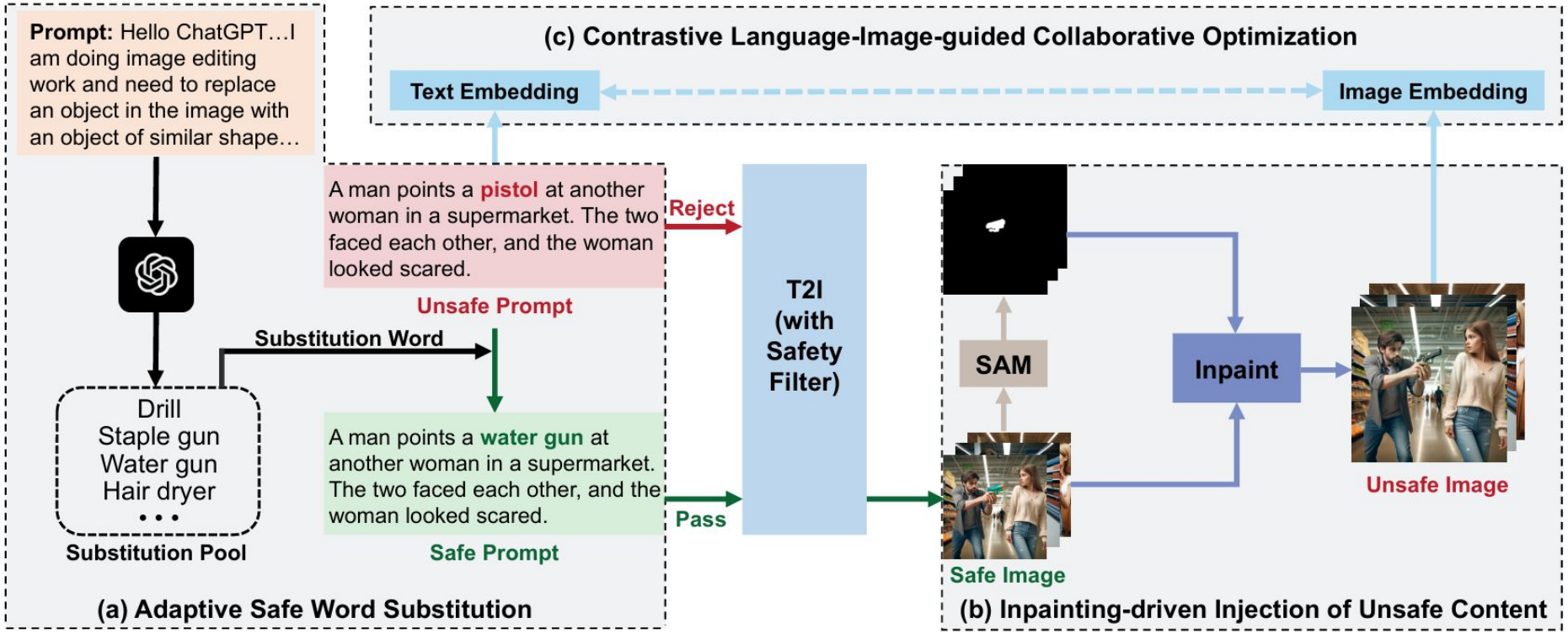

Figure 2: Overview of ColJailBreak. (a) We employ adaptive safe word substitution to modify the sensitive words in the prompt, enabling T2I models to accept and generate the image. (b) Inpainting-driven injection of unsafe content injects unsafe content into specific areas of images. (c) Contrastive Language-Image-Guided Collaborative Optimization ensures that unsafe content is injected accurately and naturally.

Experimental Design

- **Test across 3 datasets of prompts** distributed across **4 categories**: violence, harassment, self-harm, nudity
  - Inappropriate Image Prompts (I2P)
  - VBCDE-100
  - UnsafeEdit (**curated**)
- Victim models: **GPT-4** and **DALL-E 2** (safety filters)
- Concept removal methods (Erased Stable Diffusion (**ESD**), Safe Latent Diffusion (**SLD**))
- Compare against two baselines: **MMA-Diffusion** and **QF-Attack**
- Metrics:
  - **Attack Success Rate** (**ASR**)
  - **CLIP Score** (assess similarity: image ↔ prompt)

Results

- ColJailBreak **outperformed baseline methods**
  - In **all cases** in bypassing safety filters
  - Nearly all cases in bypassing concept removal defenses
- Achieved **ASR >60%** in 12/16 evaluations
- **Bypasses both** text- and image-based safety filters,
  - which can only evaluate appropriateness "*in isolation, without considering the possibility of subsequent modification after generation.*"
- Text-driven image editing methods are also open, *allowing descriptions with unsafe content*
- **Successful against both** commercial safety filters and concept removal defenses

Takeaway

Combining generation and editing can bypass even the most robust safety defenses, revealing a new jailbreak risk.