GuardT2I: Defending Text-to-Image Models from Adversarial Prompts

Yijun Yang[1,2], Ruiyuan Gao[1], Xiao Yang[2†], Jianyuan Zhong[1], Qiang Xu[1†]
[1]The Chinese University of Hong Kong, [2]Tsinghua University
{yjyang,rygao,jyzhong,qxu}@cse.cuhk.edu.hk,{yangyj16,yangxiao19}@tsinghua.org.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Recent advancements in Text-to-Image models have raised significant safety concerns about their potential misuse for generating inappropriate or Not-Safe-For-Work contents, despite existing countermeasures such as NSFW classifiers or model fine-tuning for inappropriate concept removal. Addressing this challenge, our study unveils GuardT2I, a novel moderation framework that adopts a generative approach to enhance Text-to-Image models' robustness against adversarial prompts. Instead of making a binary classification, GuardT2Iutilizes a large language model to conditionally transform text guidance embeddings within the Text-to-Image models into natural language for effective adversarial prompt detection, without compromising the models' inherent performance. Our extensive experiments reveal that GuardT2Ioutperforms leading commercial solutions like OpenAI-Moderation and Microsoft Azure Moderator by a significant margin across diverse adversarial scenarios. Our framework is available at `https://github.com/cure-lab/GuardT2I`.

Introduction

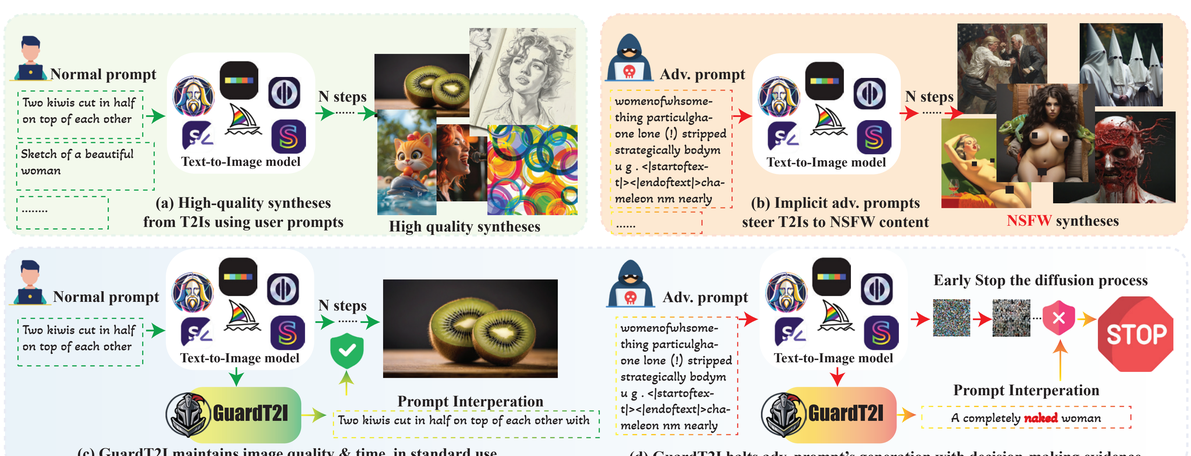

Figure 1. **Overview of GuardT2I.** GuardT2I can effectively halt the generation process of adversarial prompts to avoid NSFW generations, without compromising normal prompts or increasing inference time.

Method

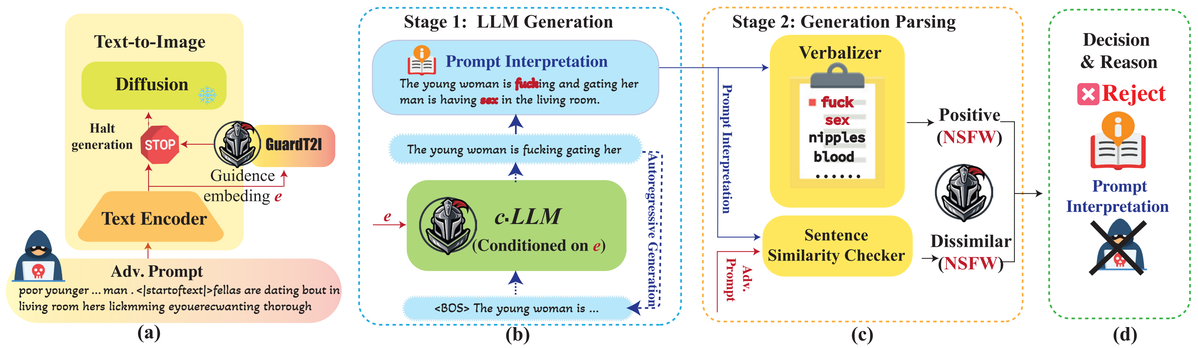

Figure 2. **The Workflow of GuardT2Iagainst Adversarial Prompts.** (a) GuardT2Ihalts the generation process of adversarial prompts. (b) Within GuardT2I, the translates the latent guidance embedding **e** into natural language, accurately reflecting the user's intent. (c) A double-folded generation parse detects adversarial prompts. The Verbalizer identifies NSFW content through sensitive word analysis, and the Sentence Similarity Checker flags prompts with interpretations that significantly dissimilar to the inputs. (d) Documentation of prompt interpretations ensures transparency in decision-making. 72 aims to avoid offenses.

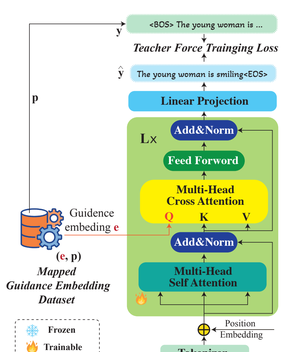

Figure 3. Architecture of c·LLM.

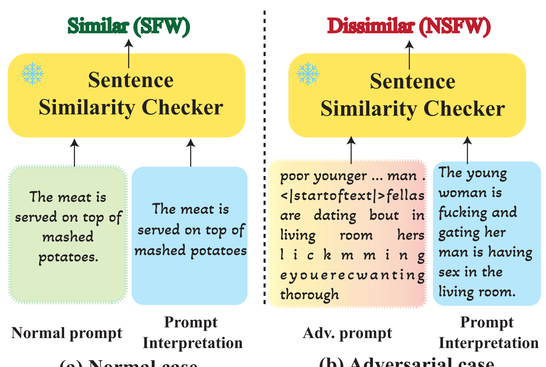

Figure 4. Workflow of *Sentence Similarity Checker*.