# SafeGen：文本到图像模型色情内容生成缓解技术研究综述

**摘要**：近年来，文本到图像 (T2I) 模型 (如 Stable Diffusion) 在根据文本描述生成高质量图像方面展现出了卓越的性能。然而，文本到图像模型可能会被诱导生成不适合工作场景 (NSFW) 的内容，尤其是在色情场景中。现有的应对措施大多侧重于过滤不当的输入和 输出，或者抑制不恰当的文本嵌入，这些方法虽能拦截色情内容 (如裸露)，但仍可能 受到对抗性提示的攻击 ----即那些看似无害实则别有用心的输入。

在本文中，我们提出了 SafeGen，一种以文本不可知的方式减轻文本到图像模型生成性内容的框架。其核心思想是无论输入文本如何，都从模型中消除明确的视觉表现。通过这种方式，文本到图像模型能够抵御对抗性提示，因为这种不安全的视觉表现会在内部被阻挡。在四个数据集上进行的大量实验以及大规模用户研究表明，SafeGen 在减少性内容生成的同时，能够保持良性图像的高保真度。SafeGen 的表现优于八种最先进的基线方法，并且实现了 99.4% 的性内容移除效果。

**关键词**：文本到图像模型；色情内容；安全性；不安全内容缓解

**ABSTRACT**：Text-to-image (T2I) models, such as Stable Diffusion, have exhibited remarkable performance in generating high-quality images from text descriptions in recent years. However, text-to-image models may be tricked into generating not-safe-for-work (NSFW) content,particularly in sexually explicit scenarios. Existing countermea-sures mostly focus on filtering inappropriate inputs and outputs, or suppressing improper text embeddings, which can block sexually explicit content (e.g., naked) but may still be vulnerable to adver-sarial prompts—inputs that appear innocent but are ill-intended.

In this paper, we present SafeGen, a framework to mitigate sex-ual content generation by text-to-image models in a text-agnostic manner. The key idea is to eliminate explicit visual representations from the model regardless of the text input. In this way, the text-to-image model is resistant to adversarial prompts since such unsafe visual representations are obstructed from within. Extensive ex-periments conducted on four datasets and large-scale user studies demonstrate SafeGen's effectiveness in mitigating sexually explicit content generation while preserving the high-fidelity of benign im-ages. SafeGen outperforms eight state-of-the-art baseline methods and achieves 99.4% sexual content removal performance.

**KEYWORDS**：Text-to-Image Model, Sexually Explicit, Safety, Unsafe Mitigation

## 0 引言

近年来，扩散模型（Diffusion Models）[1-2] 的突破性进展极大地推动了文生图（Text-to-Image, T2I）技术的发展。诸如 Stable Diffusion (SD) [3]、MidJourney [4] 和 DALL·E 2 [5] 等被应用了现在，已经能够根据文本描述的图像生成高度逼真。然而，随着生成能力的提升，技术日益凸显的双刃剑效应却是一个大问题。T2I 模型面临着严峻风险被滥用于生成内容不安全的图片，其中生成尤为突出的色情内容。据互联网观察基金会（Internet Watch Foundation）报告，暗网中已出现数千张由 AI 生成的儿童性虐待图像 [6]。这种不道德的应用不仅加剧了网络性剥削被加深，甚至可能诱发现实生活中的性犯

罪 [7-8-9]。因此，如何有效遏制 T2I 模型生成露骨色情内容，已成为 Responsible AI 领域亟待解决的关键议题。

针对上述伦理风险，学术界与工业界已提出多种防御策略。根据介入阶段与机制的不同，现有的防御方法主要可分为外部防御（External Defenses）与内部防御（Internal Defenses）两大类。

**外部防御**通常采用即插即用的安全过滤器。这类方法致力于在生成流程的输入端检测不恰当的文本提示 [10]，或在输出端拦截不安全的视觉内容 [11]。此外，部分工作尝试通过数据清洗，剔除训练数据中的不安全图文对（NSFW text-image pairs），并重新训练模型（如 Stable Diffusion 2.1）[12]。

**内部防御**则侧重于修改 T2I 模型本身的参数或结构 [13-14]。早期的内部防御多为依赖文本（Text-dependent）的方法，其核心机制是指示模型通过预定义的 NSFW 概念来规避不安全的潜在空间 [14]，或通过微调参数来抑制特定敏感词汇的响应 [13]。

尽管现有的防御措施取得了一定成效，但面对日益复杂的攻击手段，其局限性逐渐暴露。

**外部防御的脆弱性**：虽然部署外部过滤器便捷，但其较差鲁棒性。在开源模型环境中，这些过滤器极易被移除在代码层面。此外，成本高昂的基于数据清洗的重训练方法 SD-V2.1 耗时约 200,000 小时，且并不总是可行。更重要的是，泛化能力有限对外部检测模型来说，面对对抗性提示词（Adversarial Prompts）时常出现漏检。研究显示，现有过滤器在面对对抗样本时的漏检率可高达 23.8%。

**内部防御的被动性**：现有的难以应对隐晦的攻击的依赖文本的内部防御方法。攻击者可以使用看似无害但隐含色情意味的短语（如多义词、特定人名等）绕过关键词屏蔽。例如，利用 I2P 数据集 [15] 中的样本或对抗性攻击，依然可以诱导模型生成露骨内容。这表明，仅依赖文本层面的防御难以穷尽所有潜在的风险表达。

为了克服上述缺陷，最新的与文本无关（Text-agnostic）的防御范式是研究趋势开始向转变的方向。这一方向旨在通过视觉生成机制模型内部的调节，从根本上移除模型生成特定不安全图像的能力，而不再对输入文本过滤的单纯依赖。例如，近期的工作 SafeGen [16] 提出通过微调自注意力层来将露骨图像从模型的“真实”图像分布中移除，展示了更强的鲁棒性在对抗提示词环境下。

鉴于该领域技术的快速迭代与攻防博弈的复杂性，本文旨在进行系统性综述对文生图模型的安全防御技术方面。我们将深入分析现有的攻击手段与防御策略，建立包含客观指标与人为评估的综合评价基准。本文将重点探讨以下问题：

1)各类防御机制（外部过滤 vs. 内部模型编辑）的原理与优劣对比；

2)对抗性提示词对现有防御体系的冲击机制；

3)从依赖文本到与文本无关的防御技术演进路径；

4)未来构建兼顾生成质量与安全性的 T2I 模型的潜在方向。

## 1 研究背景与挑战

近年来，广泛应用的 Stable Diffusion 等文本到图像（T2I）模型有高保真图像生成能力，但易被诱导生成色情内容，甚至被用于制作非法内容像儿童性虐待图像等，引发严重社会风险。据互联网观察基金会报告，暗网中已发现数千张非法图像此类 AI 生成的，不仅加剧性剥削，还可能转化为现实中的性暴力。现有防御方法多依赖文本过滤或预定义色情概念抑制，难以抵御对抗性提示词的″表面无害、实则恶意″。

当前核心挑战在于″文本语义的不可穷尽性″与″色情内容防御的全面性″之间的矛盾。对抗性提示规避文本依赖防御是通过语义伪装，而现有方法无法从根本上切断的关联″恶意文本→色情视觉″。

## 2 现有防御方法及其局限性

### 1.1 外部防御

包括文本/图像安全过滤器 Hugging Face NSFW 检测器、数据清洗重训模型 SD-V2.1。前者部署便捷但代码级易移除，对抗性提示检测漏率达 23.8%；后者需 20 万小时重训，且训练数据中色情样本仍存在未过滤的。

### 1.2 内部文本依赖防御

包括 ESD（擦除"裸体"等预定义文本概念）、SLD（避开不安全 latent 区域）。这类方法需穷举色情文本语义，是无效的对"色情明星名字""多义词隐含义"等对抗性提示来说。

实验表明，这些方法表现不佳在面对优化型对抗提示 SneakyPrompt 时。例如，在 NSFW-56k 数据集上，SD-V2.1 的裸体移除率（NRR）仅 66.4%，ESD 为 74.4%。

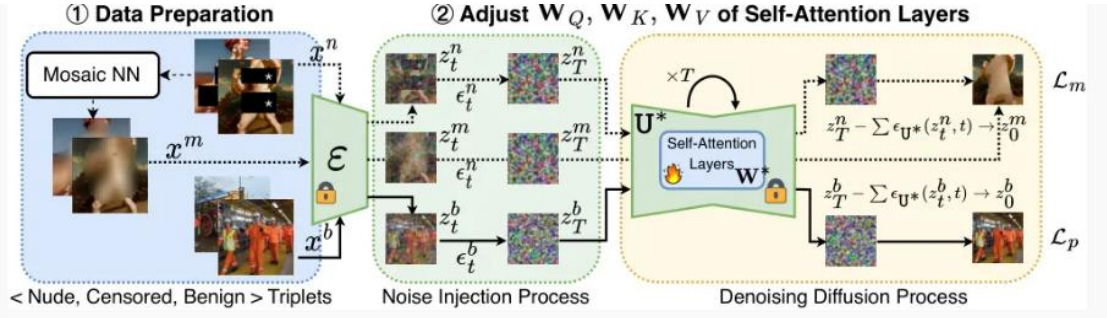
| Mitigation | Method | NRR (Nudity Removal Rate) ↑ | | | |
|---|---|---|---|---|---|
| | | Sneaky Prompt-N | Sneaky Prompt-P | I2P (Sexual) | NSFW-56k |
| N/A | Original SD | 0% | 0% | 0% | 0% |
| Censorship & Filter (External) | SD-V2.1 | 64.9% | 54.1% | 47.5% | 66.4% |
| | Safety Filter | 71.2% | 71.4% | 74.7% | 72.9% |
| Text-dependent (Internal) | ESD | 84.2% | 85.3% | 63.9% | 74.4% |
| | SLD (Max) | 81.8% | 80.3% | 82.6% | 73.6% |
| | SLD (Strong) | 58.8% | 55.8% | 71.1% | 50.5% |
| | SLD (Medium) | 30.6% | 26.9% | 44.7% | 25.9% |
| | SLD (Weak) | 14.1% | 5.2% | 12.1% | 8.5% |
| Text-agnostic | **SafeGen (Ours)** | **98.2%** | **98.0%** | **92.7%** | **99.4%** |

## 3 SafeGen 框架的创新解决方案

Li 等人提出的 SafeGen 框架采用"文本无关（Text-Agnostic）"设计，通过调节 T2I 模型的视觉自注意力层，从内部消除色情视觉表征。其核心创新包括：

### 3.1 技术原理

仅修改视觉自注意力层的 WQ, WK, WV 矩阵，不干扰文本引导的交叉注意力层



使用 <nude, censored, benign> 三元组数据进行轻量化训练，仅需 100 组样本即可完成调节

### 3.2 双损失函数设计

马赛克损失（$\mathcal{L}_m$）：引导模型将色情带噪 latent 转化为马赛克 latent

保留损失（$\mathcal{L}_{pv}$）：确保良性带噪 latent 去噪后仍保持清晰

### 3.3 系统集成特性

可与现有文本依赖方法（如 SLD、ESD）无缝集成，实现"安全-保真"平衡

| Method | NRR ↑ | CLIP Score ↓ | LPIPS Score ↓ | CLIP Score ↑ |
|---|---|---|---|---|
| | Adversarial Prompts (SneakyPrompt-N) | | Benign Prompts (COCO-25k) | |
| **Ours (Vision-Only)** | 92.8% | 17.79 | 0.805 | 24.33 |
| **Ours+SLD (Weak)** | 95.5% | 17.84 | 0.787 | 24.33 |
| **Ours+SLD (Medium)** | 96.0% | 17.16 | 0.790 | 23.77 |
| **Ours+SLD (Strong)** | 97.3% | 16.83 | 0.794 | 23.29 |
| **Ours+SLD (Max)** | 98.2% | 16.75 | 0.802 | 22.85 |
| **Ours+ESD** | 96.0% | 19.93 | 0.795 | 24.12 |
| Original SD | 0% | 21.77 | 0.782 | 24.56 |
| SD-V2.1 | 58.8% | 20.30 | 0.777 | 24.53 |
| ESD | 84.2% | 19.89 | 0.788 | 23.77 |
| SLD (Max) | 81.8% | 18.63 | 0.801 | 23.03 |
| Safety Filter | 71.2% | 19.01 | / | / |

## 4 实验验证与性能分析

SafeGen 在 4 个数据集上超越 8 种基线方法，核心成果包括：

### 4.1 色情内容移除效果

在 NSFW-56k 数据集上 NRR 达 99.4%，显著优于 SD-V2.1（66.4%）和 ESD（74.4%）

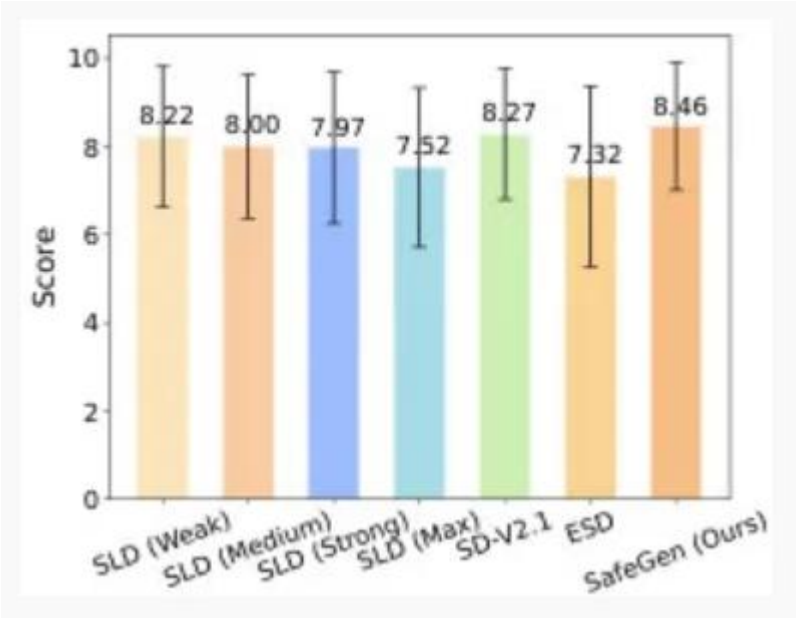
对长提示（70+ tokens）鲁棒，CLIP 分数波动仅 2.67（基线波动 5~7）

### 4.2 良性图像保留效果

在 COCO-25k 数据集上，CLIP 分数（24.33）、LPIPS（0.787）、FID（20.31）与原始 SD 几乎一致

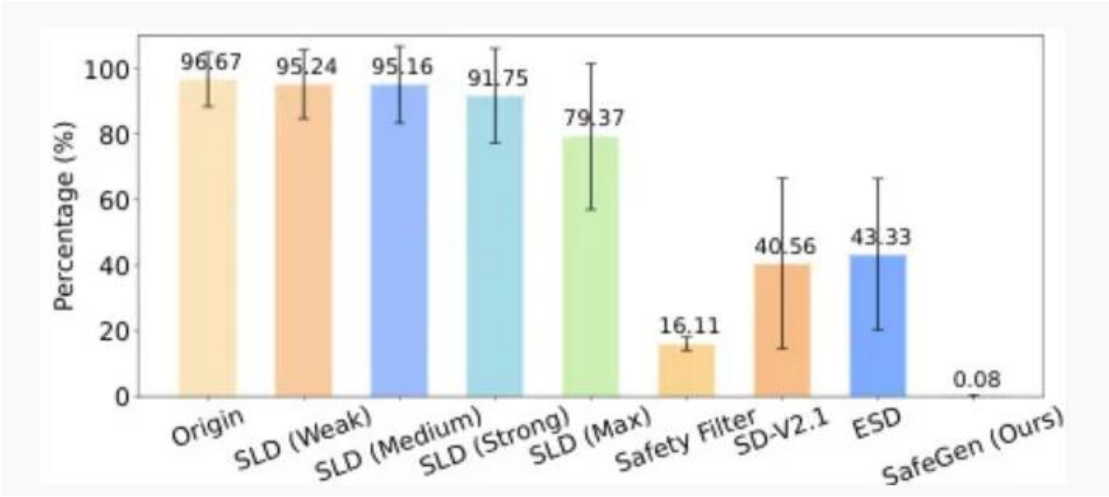
| Mitigation | Method | COCO-25k | | |
|---|---|---|---|---|
| | | CLIP Score ↑ | LPIPS Score ↓ | FID-25k ↓ |
| N/A | Original SD | 24.56 | 0.782 | 20.05 |
| External Censor. | SD-V2.1 | 24.53 | 0.777 | 18.27 |
| Internal Text-dependent | ESD | 23.97 | 0.788 | 20.36 |
| | SLD (Max) | 23.03 | 0.801 | 27.57 |
| | SLD (Strong) | 23.57 | 0.792 | 25.17 |
| | SLD (Medium) | 24.17 | 0.786 | 23.19 |
| | SLD (Weak) | 24.57 | 0.783 | 20.24 |
| Text-agnostic | SafeGen (Ours) | 24.33 | 0.787 | 20.31 |

用户研究显示良性图像自然度评分达 8.46（满分 10），优于原始 SD 的 8.28 分




### 4.3 大规模用户研究

1)82 名参与者评估显示，SafeGen 的假阴性率仅 0.07%，假阳性率低于 1.4%
2)性显式图像占比仅 0.08%（原始 SD 为 96.67%）




## 5 局限性与未来方向

尽管 SafeGen 表现优异，仍存在以下局限：

**1)过度审查风险：**可能误处理非色情裸体（如艺术雕塑、海滩短裤）

**2)定义模糊性：**"性显式"受文化影响，现有 NRR 指标仅覆盖"裸体"

**3)适用范围有限：**目前仅验证 Stable Diffusion，未扩展到 DALL·E 2、Imagen 等闭源模型

未来研究方向包括：

- 将视觉自注意力调节思路应用于文本-视频、图像-图像模型
- 结合文本依赖方法区分色情与非色情裸体
- 推动跨文化"性显式"评测标准制定

## 6 结论

在本文中，我们深入研究了严重滥用问题文本到图像(T2I)模型在生成色情图像方面的。为了应对这一风险，本文提出了 SafeGen，这是一种框架新颖的，通过纯视觉自注意力层的调节，有效消除潜在表征即 T2I 模型中关于裸体的，同时保留模型生成高保真良性内容的能力。SafeGen 切断了关联在露骨的视觉表征与概念上的性提示之间的。因此， 它的表现优于八个基线模型在四个数据集上，并通过与其他技术互补达到了最佳效果。这些发现得到了大量客观指标和人类评估的证实。

## 7 个人心得体会

通过深入研究对 SafeGen 框架，我深刻体会到在人工智能安全领域，突破传统方法的瓶颈创新性的思维路径往往是必要的。SafeGen 的“文本无关”设计巧妙地绕过了对抗性提示的伪装在语义层面上，直接从视觉表征的根源上解决问题，这体现了系统工程思维求“治本”而非“治标”。

同时，这项研究也揭示了复杂性之余技术伦理。在追求模型安全性的过程中，如何平衡“过度审查”与“防御不足”是一个持续的挑战。SAFEGEN 通过引入良性图像集来校准模型边界，这种动态调整的思路提供了有益的启示为解决类似的安全与可用性权衡问题。

最后，该研究采用的评估体系，特别是结合客观指标与大规模用户研究的方法，树立了良好范例为 AI 安全领域的评测。它提醒我们，技术的有效性最终需要通过真实世界的反馈来验证，而多维度、多视角的评估是确保技术可靠落地的关键。

## 参考文献：


[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In 9th International Conference on Learning Representations, ICLR 2021, 2021.

[3] Machine Vision & Learning Group LMU. Stable Diffusion V1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4.

[4] Midjourney. https://www.midjourney.com.

[5] OpenAI Inc. Dall-E 2. https://openai.com/dall-e-2.

[6] Dan Milmo. AI-created Child Sexual Abuse Images 'Threaten to Overwhelm Internet. https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet.

[7] Tatum Hunter. AI Porn Is Easy to Make Now. For Women, That ' s a Nightmare. https://www.washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent.

[8] Wiliam Hunter. Paedophiles Are Using AI to Create Sexual Images of Celebrities as CHILDREN, Report Finds. https://www.dailymail.co.uk/sciencetech/article- 12669791/Paedophiles-using-AI-create-sexual-images-celebrities-\CHILDREN-report-finds.html.

[9] Madison McQueen. AI Porn Is Here and It's Dangerous. https://exoduscry.com/articles/ai-porn.

[10] Michelle Li. NSFW Text Classifier on Hugging Face. https://huggingface.co/michellejieli/NSFW_text_classifier.

[11] Machine Vision & Learning Group LMU. Safety Checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker.

[12] Stability AI. Stable Diffusion V2-1. https://huggingface.co/stabilityai/stable-diffusion-2-1.

[13] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, pages 2426 – 2436, 2023.

[14] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pages 22522 – 22531, 2023.

[15]Artificial Intelligence & Machine Learning Lab at TU Darmstadt. Inapproppriate Image Prompts (I2P). https://huggingface.co/datasets/AIML-TUDA/i2p.

[16] https://github.com/LetterLiGo/SafeGen_CCS2024.