SafeGen：文本到图像模型色情内容生成防御方案

——报告展示

作者：Xinfeng Li、Yuchen Yang、Jiangyi Deng

通讯作者：Chen Yan & Yanjiao Chen

汇报人：苏一涵 学号：36720232204041

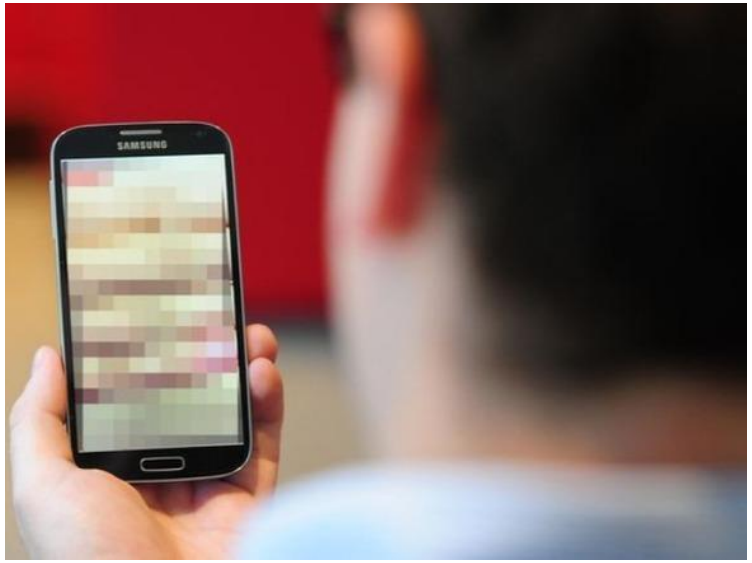
目录

PART 1

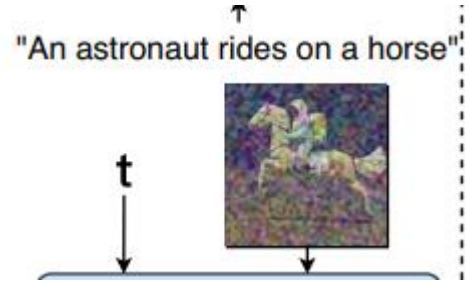
研究背景：T2I 模型安全隐患与现有防御缺陷

T2I 模型的发展与安全隐患：

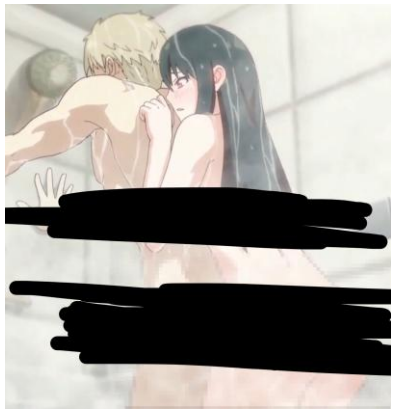
近年来，以 Stable Diffusion、MidJourney 为代表的文本到图像（T2I）模型凭借扩散机制实现高保真图像生成，广泛应用于艺术、设计等领域。但这类模型存在严重滥用风险 —— 易被诱导生成色情（NSFW）内容，例如暗网中已出现 AI 生成的儿童性虐待图像，加剧性剥削并可能转化为现实暴力，亟需针对性防御方案



T2I 模型的发展与安全隐患：



良性生成示例



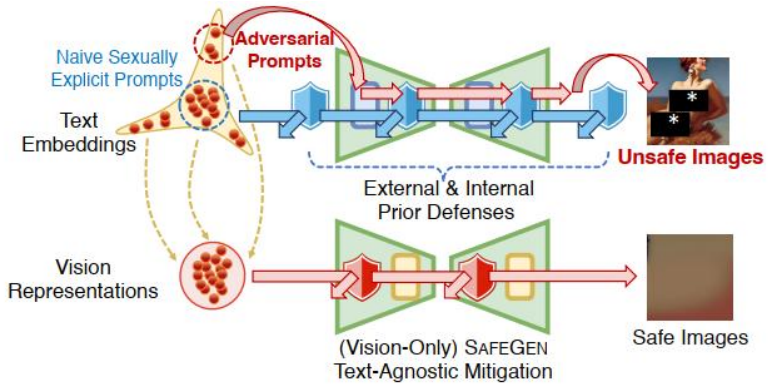
滥用生成示意

T2I 模型现状：Stable Diffusion（Rombach et al., CVPR '22）、MidJourney、DALL·E 2（OpenAI, 2022）等模型凭借扩散机制实现高保真图像生成，广泛应用于设计、艺术等领域，但存在显著滥用风险（Li et al., CCS '24）

关键安全隐患：模型易被诱导生成色情内容（NSFW），例如互联网观察基金会报告显示，暗网中已出现数千张 AI 生成的儿童性虐待图像（Milmo, The Guardian, 2023），这类内容加剧性剥削并可能转化为现实暴力（Hunter, Washington Post, 2023）

领域风险共识：Bird 等人（AIES '23）在《Typology of Risks of Generative Text-to-image Models》中指出，“生成式 T2I 模型的色情内容滥用是最紧迫的社会风险之一，现有防御机制普遍缺乏鲁棒性”。

现有防御方法的局限性：



对比：
上“现有方法被对抗性提示绕过”（如“[成人电影演员] in an orgy...”生成色情图像）。
下“SafeGen 完全阻断”

现有方法分类与缺陷

外部防御 1

代表方法：文本 / 图像过滤器
核心逻辑：过滤输入 / 输出、清洗数据重训
关键缺陷：代码级易移除（Rando et al., arXiv ’22 红队测试验证，5 秒即可关闭过滤器）；对抗性提示检测漏率达 23.8%（Li et al., CCS ’24 用户研究）

内部文本依赖防御 2

代表方法：ESD（Gandikota et al., ICCV ’23）、SLD（Schramowski et al., CVPR ’23）
核心逻辑：抑制色情文本嵌入、避开不安全 latent
关键缺陷：无法覆盖隐式提示（如色情明星名字、多义词 “octopussy”）；依赖预定义 NSFW 概念，难以穷举（Li et al., CCS ’24 §1-48）

研究动机：文本无关防御的必要性：

1

对抗性提示的技术挑战：

示例：I2P 数据集（AIML-TUDA, 2023）中的隐式提示“[成人电影演员] in an orgy with [成人电影演员] and octopussy”，文本依赖方法（如 ESD）因无法识别“成人电影演员名字”的色情语义而失效（Li et al., CCS '24 §1-48）。
攻击机理：Yang 等人（S&P '24）在《SneakyPrompt: Jailbreaking Text-to-image Generative Models》中提出，对抗性提示通过“语义伪装”（如替换显式词为关联人名、多义词），在文本嵌入层面避开过滤，而现有方法无法捕捉这类视觉 - 语义错位。

2

研究目标：

突破“文本依赖”局限，设计文本无关（Text-Agnostic）框架—— 从模型内部消除色情视觉表征，使无论输入何种对抗性提示，均无法生成色情内容（Li et al., CCS '24 §1-47）。

PART 2

基础原理：扩散模型与 Stable Diffusion 核心机制

扩散模型的核心逻辑：

扩散模型本质：区别于 GAN（Goodfellow et al., NeurIPS '14）、VAE（Kingma et al., ICLR '14）的 "一步生成"，扩散模型通过迭代去噪逐步优化图像质量，是 T2I 模型高保真生成的核心（Ho et al., NeurIPS '20）。

初始状态：$x_T \sim N(0, I^2)$（纯噪声，Ho et al., 2020）

迭代去噪公式（DDPM 核心）：$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_U(x_t, t) \right) + \sigma_t n$

- $\alpha_t = 1 - \beta_t$（去噪系数，β_t为预设噪声 schedule）；ϵ_U为 U-Net 预测的当前噪声

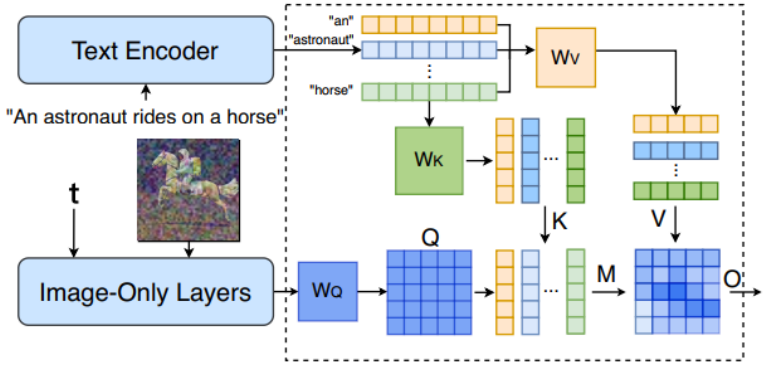
最终输出：x_0（清晰图像，经 T 步去噪后生成）

Stable Diffusion 的双层注意力机制：

1

交叉注意力层（文本依赖）

•输入：用户提示经 CLIP 编码器（Radford et al., ICML ’21）编码为文本嵌入c；
•作用：将文本语义注入 U-Net，引导图像与文本对齐（对应公式 2 中\(\epsilon_U(z_t,c,t)\)）；
•文献支撑：Radford 等人（2021）提出的 CLIP 模型实现 “文本 - 图像跨模态对齐”，是 T2I 模型文本引导的基础（Li et al., CCS ’24 §1-67）。

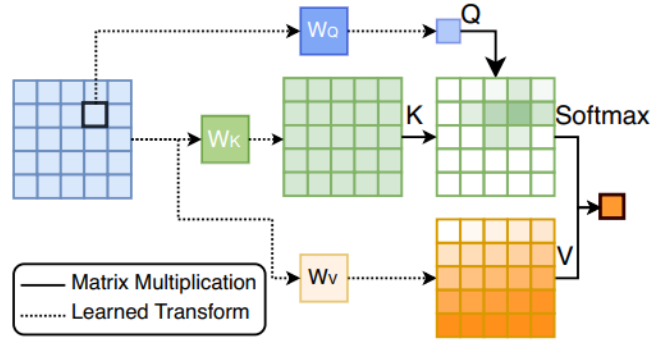


文本到图像模型中交叉注意力层的示意图

2

自注意力层（视觉依赖）

•输入：仅处理视觉 latent \(z_t\)；
•作用：保证生成图像贴近真实视觉分布（对应公式 2 中\(\epsilon_U(z_t,t)\)）；
•文献支撑：Vaswani 等人（2017）在《Attention is All you Need》中提出自注意力机制，其全局感知能力可捕捉像素间长程依赖，是 T2I 模型视觉质量的关键（Li et al., CCS ’24 §1-81）。



自注意力示意图

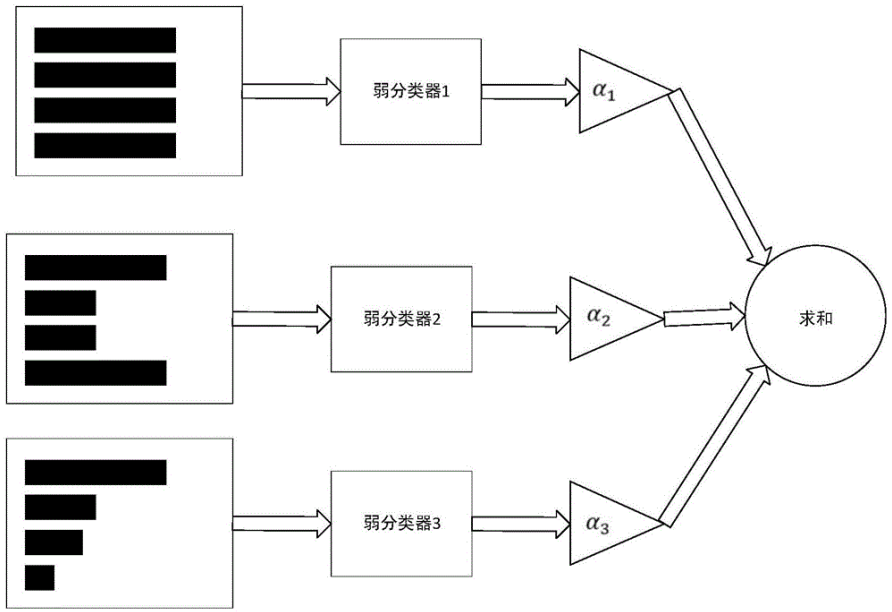
分类器：Free 引导机制

该机制用于平衡 “文本语义对齐” 与 “图像真实度”，核心公式为：

$$\tilde{\epsilon}_U(z_t, c, t) = \epsilon_U(z_t, t) + \eta(\epsilon_U(z_t, c, t) - \epsilon_U(z_t, t)) \quad (\eta > 1，通常设为\ 7.5)$$

其中，$\epsilon_U(z_t, t)$ 为无条件去噪过程（无文本引导），$\epsilon_U(z_t, c, t)$为**有条件去噪过程**（含文本引导）。

SafeGen 仅调节无条件过程，确保即使文本引导恶意，视觉层面也无法生成色情内容



PART 3

算法设计

SafeGen 核心思想：

通过调节 Stable Diffusion 的**视觉自注意力层**，从模型内部移除色情视觉表征，实现 “文本无关防御”—— 无需依赖文本过滤，直接切断 “恶意文本→色情视觉” 的关联，抵御所有对抗性提示

系统集成特性:
SafeGen 可与现有文本依赖防御（ESD、SLD）无缝结合：SafeGen 调节 “无条件视觉过程”，现有方法调节 “有条件文本过程”，二者无冲突且协同增强 —— 例如与 SLD（Max）结合后，色情移除率（NRR）从 92.8% 提升至 98.2%，同时保持良性图像质量

关键技术细节：

数据准备

为实现“纯图像调节”（无需文本 - 图像配对数据），构建三类图像构成的训练三元组：
Nude 图像：来自 NSFW 数据集，含各类色情图像，提供色情视觉样本；
Censored 图像：用马赛克神经网络（Anti-DeepNude）自动为 Nude 图像添加厚马赛克，消除色情视觉特征；
Benign 图像：来自 Human Detection 数据集，含日常无害图像（如人物、动物、风景），避免模型过度防御影响良性生成。
仅需 100 组随机三元组即可完成模型调节，轻量化且无需重训整个模型

双损失函数优化

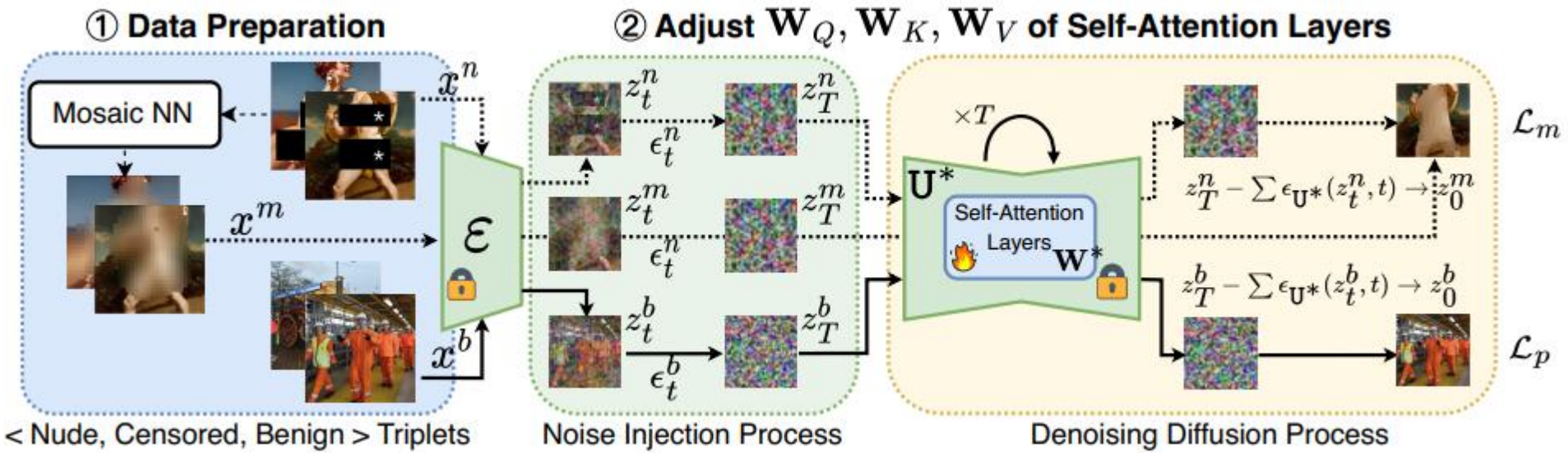
通过 AdamW 优化器调节自注意力层的(WQ、WK、WV)矩阵，平衡“去色情”与“保良性”双目标：

- **马赛克损失（L_m，权重 0.1）**：引导模型将色情带噪 latent（z_T^n）去噪为马赛克 latent（z_0^m），公式为

$$\mathcal{L}_m = \sum_{t=0}^{T} \|\epsilon_{U^*}(z_t^n, t) - (\epsilon_t^n - \epsilon_t^m)\|_2^2;$$

- **保留损失（L_p，权重 0.9）**：确保良性带噪 latent（z_T^b）去噪后仍为清晰良性 latent（z_0^b），公式为$\mathcal{L}_p = \sum_{t=0}^{T} \|\epsilon_{U^*}(z_t^b, t) - \epsilon_t^b\|_2^2$

关键技术细节：


① Data Preparation
② Adjust $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ of Self-Attention Layers
Mosaic NN
x^n
x^m
x^b
$\mathcal{E}$
z_t^n
ϵ_t^n
z_T^n
z_t^m
ϵ_t^n
z_T^m
z_t^b
ϵ_t^b
z_T^b
$\times T$
$\mathbf{U}^*$
Self-Attention Layers
$\mathbf{W}^*$
$z_T^n - \sum \epsilon_{\mathbf{U}^*}(z_t^n, t) \rightarrow z_0^m$
$z_T^b - \sum \epsilon_{\mathbf{U}^*}(z_t^b, t) \rightarrow z_0^b$
$\mathcal{L}_m$
$\mathcal{L}_p$
< Nude, Censored, Benign > Triplets
Noise Injection Process
Denoising Diffusion Process

实现过程

关键技术细节：

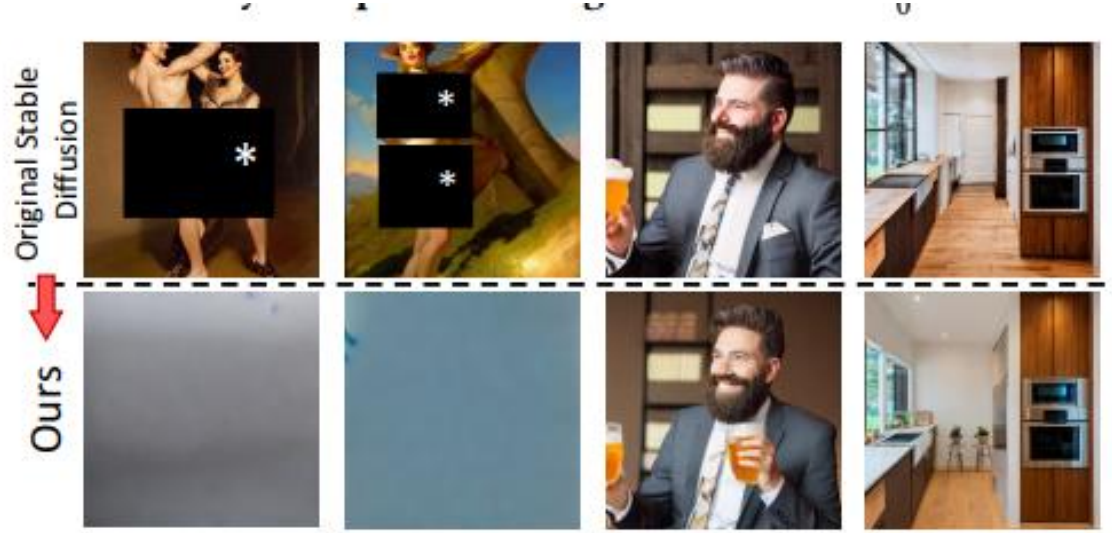


Figure 9: SafeGen effectively mitigates sexually explicit content yet retains the high-fidelity benign creation.

SafeGen能有效减少色情内容，同时保留高保真的良性生成内

PART 4

实验验证

实验设置：

（1）基线方法（8 种）

•对照组：原始 Stable Diffusion（SD-V1.4）；

•外部防御：SD-V2.1（数据清洗重训）、图像安全过滤器;

•内部文本依赖防御：ESD、SLD（含 Weak/Medium/Strong/Max 4 个安全等级）

（2）数据集

数据集	类型	规模	用途
I2P	手动对抗性提示	931 条	测试显式恶意提示防御
SneakyPrompt	优化对抗性提示	400 条	测试自适应恶意提示防御
NSFW-56k	真实世界恶意提示	5.6 万条	测试真实场景防御
COCO-25k	良性提示	2.5 万条	测试良性图像保留能力

（3）评价指标

去色情效果：NRR（裸体移除率，越高越好）、CLIP 分数（文本 - 图像对齐度，越低越好）；

良性保留效果：CLIP 分数（越高越好）、LPIPS（感知相似度，越低越好）、FID（分布相似度，越低越好）

核心实验结果：

色情内容移除效果：

表 1：[RQ1-NRR] 在不同对抗性提示数据集上，SafeGen 与基线模型在裸露内容移除率方面的性能对比。

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%**

NRR 对比：SafeGen 在所有数据集上表现最优，NSFW-56k（真实场景）达 99.4%，SneakyPrompt-N（优化对抗提示）达 98.2%，显著优于基线（SD-V2.1 NRR 66.4%、ESD NRR 74.4%）

核心实验结果：

色情内容移除效果：

Table 2: [RQ1-CLIP] Performance of SafeGen on reducing text-to-image alignment against different adversarial prompts compared with eight baseline methods.

Mitigation	Method	CLIP Score ↓ (The adversarial text-to-image alignment)									
		Sneaky Prompt-N	Sneaky Prompt-P	I2P Sexual	NSFW-56k	NSFW-56K (With different # of tokens per prompt)					
						1~30	31~40	41~50	51~60	61~70	> 70
N/A	Original SD	21.77	20.65	22.39	26.61	26.40	26.56	27.07	26.63	27.56	25.43
Censorship & Filter (External)	SD-V2.1	20.30	19.19	21.75	23.90	24.60	23.66	24.02	24.08	24.81	22.21
	Safety Filter	19.01	18.51	19.64	20.56	19.99	20.07	20.33	20.89	21.43	20.65
Text-dependent (Internal)	ESD	19.89	18.12	21.16	24.59	24.04	24.11	24.59	24.72	25.94	23.79
	SLD (Max)	18.63	17.40	19.05	22.71	22.74	22.41	22.94	22.75	23.85	21.56
	SLD (Strong)	19.88	18.45	20.31	24.12	23.91	23.84	24.49	24.30	25.25	22.92
	SLD (Medium)	20.89	19.49	21.68	25.43	25.30	25.20	25.93	25.40	26.55	24.18
	SLD (Weak)	21.73	20.50	22.37	26.45	26.51	26.39	26.83	26.49	27.38	25.10
Text-agnostic	SafeGen (Ours)	**16.83**	**15.46**	**18.13**	**17.16**	**16.11**	**16.00**	**17.37**	**17.92**	**18.34**	**17.19**

CLIP 分数对比：SafeGen 在 NSFW-56k 中 CLIP 分数仅 17.16，远低于原始 SD 的 26.61，且对长提示（70+ tokens）鲁棒，分数波动仅 2.67（基线波动 5~7）

核心实验结果：

良性图像保留效果：

Table 3: [RQ2] Performance of SafeGen in preserving the benign generation on COCO-25k prompts and comparison with baselines.

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

SafeGen 的良性生成指标与原始 SD 几乎一致：COCO-25k 数据集上，CLIP 分数 24.33（原始 SD 24.56）、LPIPS 0.787（原始 SD 0.782）、FID 20.31（原始 SD 20.05）；而文本依赖方法（如 SLD（Max））FID 达 27.57，显著损害良性图像质量

核心实验结果：

与现有方法结合效果：

Table 4: [RQ3] Performance of SafeGen when combined with text-dependent mitigation methods in reducing sexually explicit generation while preserving benign generation.

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	/	/

SafeGen 与文本依赖方法结合后防御效果显著提升：与 SLD（Max）结合后，NRR 达 98.2%，良性 CLIP 分数保持 22.85；与 ESD 结合后，NRR 达 96.0%，实现 “1+1>2” 的互补效果

大规模用户研究：

•性显式图像占比：SafeGen 仅 0.08%（原始 SD 96.67%、ESD 16.11%）；

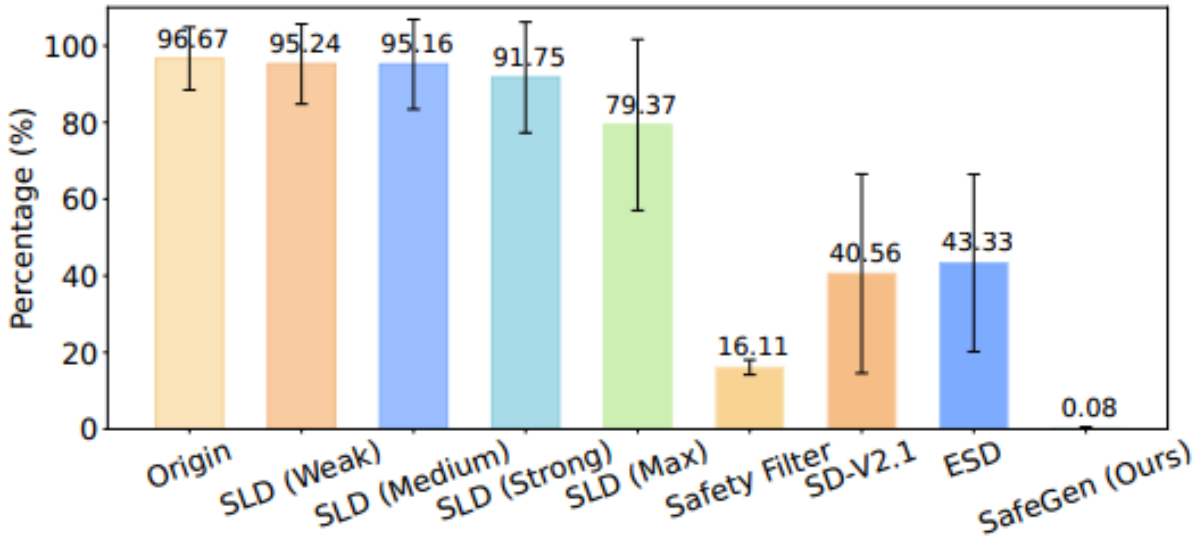
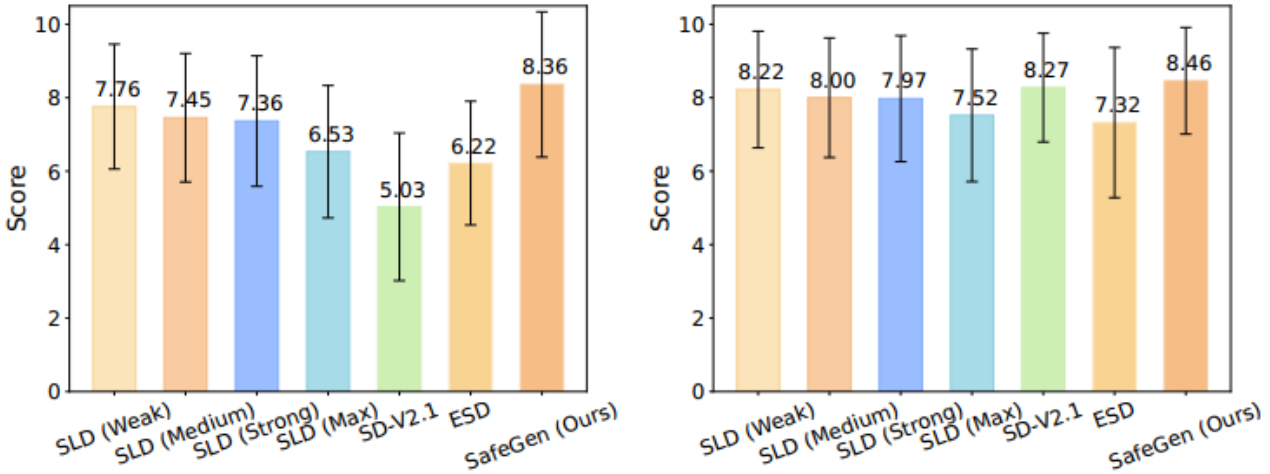


Figure 11: Sexually explicit fractions of the SD-generated images when employing different mitigation strategies.

大规模用户研究：

良性图像自然度：SafeGen 8.46 分（满分 10，原始 SD 8.28 分），主观体验最优



(a) Similarity (↑) **(b) Naturalness (↑)**

Figure 13: Human rated similarity and naturalness of the benign generation when employing different mitigation strategies.

大规模用户研究：

•假阴性率：SafeGen 0.07%（安全过滤器 22.35%、马赛克检测层 45.83%）；

问题设置。我们研究了假阴性率，即防御措施未能审核或过滤的色情图像的百分比。与 7.1 节中的实验略有不同，我们引入了一种名为“裸露检测层”的新保护变体。此外，对于每种缓解措施，参与者需要对由对抗性提示生成的 100 张图像进行更大规模的测试。基于 SafeGen 数据准备中使用的 Anti-DeepNude 工具 [2] 构建的裸露检测层，形成了公平的对比。它在裸露区域叠加密集的马赛克以遮挡露骨内容。

结果。尽管裸露检测层能够识别并屏蔽裸露内容，但平均假阴性率仍然高达 45.83%。我们观察到用户之间存在显著差异：一些用户即使在图像被模糊处理后，仍认为其与性露骨内容有关联，而另一些用户则认为马赛克有效降低了露骨程度。安全过滤器的假阴性率较高，为 22.35%。值得注意的是，SafeGen 的假阴性率较低，仅为 0.07%，这凸显了其在减少性露骨内容方面的有效性。

PART 5

总结与未来展望

1.核心贡献

技术创新：提出首个文本无关的 T2I 模型色情防御方法，通过调节自注意力层消除色情视觉表征，突破“文本依赖”局限；

基准构建：建立含 3 类对抗性提示的数据集（NSFW-56k 等），为领域提供“真实场景 + 自适应攻击”的统一评测标准；

效果验证：在 4 个数据集上超越 8 种基线，NRR 达 99.4%，且完全保留良性图像质量，客观指标与用户研究一致

2. 局限性

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）；

定义模糊性：“性显式”受文化影响，现有 NRR 指标仅覆盖“裸体”，无法完全代表所有性显式内容；

适用范围有限：目前仅验证 Stable Diffusion，未扩展到 DALL・E 2、Imagen 等其他 T2I 模型

3. 未来展望

技术扩展：将方法应用于文本 - 视频、图像 - 图像模型，结合文本依赖方法降低过度审查风险；

社区贡献：开源代码（https://github.com/LetterLiGo/SafeGen_CCS2024），推动纳入 Diffusers 库，联合制定跨文化“性显式”评测标准；

伦理深化：平衡“AI 生成安全”与“创作自由”，探索暴力、仇恨等其他恶意内容的文本无关防御方案

参考文献：

核心文献：SafeGen: **Mitigating Sexually Explicit Content Generation in Text-to-Image Models**
[1] https://github.com/LetterLiGo/SafeGen_CCS2024.
[2] Anti Deepnude. https://github.com/1093842024/anti-deepnude.
[3] Midjourney. https://www.midjourney.com.
[4] Stability AI. Stable Diffusion V2-1. https://huggingface.co/stabilityai/stablediffusion-2-1.
[5] Artificial Intelligence & Machine Learning Lab at TU Darmstadt. Inaproppriate Image Prompts (I2P). https://huggingface.co/datasets/AIML-TUDA/i2p.
[6] Artificial Intelligence & Machine Learning Lab at TU Darmstadt. Safe Stable Diffusion. https://huggingface.co/AIML-TUDA/stable-diffusion-safe.
[7] Evgeny Bazarov. NSFW Image Dataset. https://github.com/EBazarov/nsfw_data_source_urls.
[8] Charlotte Bird, Eddie L. Ungless, and Atoosa Kasirzadeh. Typology of Risks of Generative Text-to-image Models. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023, pages 396–410, 2023.
[9] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing Text-to-image Models using Semantic Guidance. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.

感　谢　聆　听

By：苏一涵