ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 620－640
DOI:10. 16337/j. 1004-9037. 2026. 02. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

视觉大模型生成内容风险与治理研究综述

刘安安[1]，张晨宇[2]，王岚君[2]，李文辉[1]

(1. 天津大学电气自动化与信息工程学院，天津 300072；2. 天津大学新媒体学院，天津 300072)

摘　要：随着扩散模型等深度生成技术的突破性进展，视觉大模型在图像生成质量与语义一致性上取得了显著飞跃，被广泛应用于艺术创作与工业设计等领域。然而，其强大的生成能力也引发了严峻的内容安全风险，恶意用户可诱导模型生成色情、暴力或侵权图像，对人工智能的安全治理提出了迫切需求。本文聚焦于视觉大模型面临的两大核心攻防任务进行了系统综述：(1)旨在诱导模型突破安全防线的越狱攻击；(2)旨在移除模型内部风险知识的概念擦除。首先，本文构建了越狱攻击的分类体系，从技术划分、扰动方式、查询类型及攻击者知识4个层面，揭示了攻击手段从特征空间对抗向语义空间推理演进的趋势。其次，针对风险治理，深入探讨了概念擦除技术，对比分析了模型微调、模型编辑与推理引导3类主流技术路线，阐述了不同方法在擦除有效性、计算效率以及通用生成能力保留之间的权衡关系。最后，梳理了该领域常用的基准数据集，并指出了当前研究在对抗鲁棒性以及多概念联合治理等方面面临的挑战与未来发展方向，旨在为构建安全可控的生成式视觉大系统提供理论参考与技术指引。

关键词：视觉大模型；内容安全；越狱攻击；概念擦除；模型治理；扩散模型

中图分类号：TP3　　**文献标志码：**A

引用格式：刘安安，张晨宇，王岚君，等. 视觉大模型生成内容风险与治理研究综述[J]. 数据采集与处理，2026，41(2)：620-640. LIU An'an，ZHANG Chenyu，WANG Lanjun，et al. A survey on risks and governance of content generated by visual generation models[J]. Journal of Data Acquisition and Processing，2026，41(2)：620-640.

引　　言

近年来，以扩散模型为代表的生成式人工智能技术取得了突破性进展。得益于大规模图文数据集的训练以及计算能力的提升，Stable Diffusion[1]、Midjourney[2]、DALL-E3[3]等视觉大模型[4-8]展现出了惊人的生成能力，能够根据用户输入的文本提示生成逼真、高质量的图像。这些模型已广泛应用于艺术创作、广告设计和游戏开发等领域，极大地提高了内容生产的效率[9-10]。然而，随着视觉大模型能力的提升，其潜在的内容安全风险也日益凸显。由于训练数据中不可避免地包含色情、暴力、血腥、仇恨言论以及版权图像等敏感信息，模型在训练过程中容易学习到这些风险概念的特征表示[11]。在开放使用的场景下，如果缺乏有效的治理机制，模型极易被恶意用户利用，生成违反法律法规或社会公德的风险内容。

为了应对这一挑战，现有的商业大模型系统通常部署有安全防御机制，如在输入端部署文本过滤器拦截包含明显风险语义的显性风险提示，或在输出端部署图像分类器检测风险图像。然而，研究表

基金项目：国家自然科学基金(62425307，62572346，U21B2024)。
收稿日期：2026-01-10；**修订日期：**2026-02-25

明这些基于外部过滤的防御手段存在安全漏洞。恶意用户可以通过设计“越狱攻击”[12-13],对显性风险提示词进行伪装、重写或添加扰动,构建出“隐性风险提示”。这些提示词表面上不包含敏感词汇,但和风险内容存在隐式的关联,从而能够绕过安全防护,同时诱导模型生成风险内容。如图1所示,国内外主流视觉大模型商业平台都存在安全漏洞,进而生成风险图像,对社会稳定造成负面影响。

图1 国内外主流视觉大模型商业平台存在安全漏洞

Fig.1 Security vulnerabilities exist in mainstream commercial image generation platforms

针对上述风险,学术界与工业界为探索更为本质的治理方案,提出概念擦除任务[14-19]。与外挂式的过滤器不同,概念擦除旨在定位并修改模型内部参数,或修改模型推理生成过程,从根本上移除模型对特定风险概念(如色情、暴力等不适合在工作场合观看的内容)的生成能力,使其在面对恶意诱导时仍能输出安全内容。这种“内生安全”的治理思路已成为当前生成式AI安全领域的研究热点。

鉴于此,本文将聚焦于视觉大模型的内容安全风险与治理这一核心问题,重点梳理越狱攻击与概念擦除两大关键任务的研究现状。本文首先概述视觉大模型的基本原理;随后,从技术划分、扰动方式、攻击者知识以及查询类型多个角度,系统阐述越狱攻击的前沿方法;接着,深入探讨以模型微调、模型编辑及推理引导为代表的概念擦除治理技术;最后,总结常用数据集,并对未来的研究方向进行展望。

1 视觉大模型基本原理

视觉大模型旨在根据用户输入的文本提示生成语义一致且高质量的图像。尽管早期的生成对抗网络[20-23]和自回归模型[24-25]在图像生成领域取得了一定成果,但近年来基于扩散概率模型[8,26-27]的方法凭借其卓越的生成质量和多样性,已成为该领域的主流范式。

根据扩散过程发生的空间不同,现有的视觉大模型主要分为两类:像素空间扩散模型[4,28]和潜在空间扩散模型[8,29]。前者如GLIDE[30]和Imagen[5]直接在像素级进行计算,计算成本极高;后者如Stable Diffusion[29]和DALL-E2[31]则在压缩的潜在空间中进行扩散,大幅降低了计算资源,是当前学术界和工业界研究的重点。本节将以最具代表性的Stable Diffusion为例,介绍其模型架构及训练推理机制。

1.1 模型架构

Stable Diffusion是一个潜在扩散模型,其整体架构主要包含3个核心组件:图像自编码器、条件编码器和去噪网络。

(1) 图像自编码器:为了降低计算复杂度,模型利用一个预训练的变分自编码器(Variational auto-encoder, VAE)[32]在像素空间和潜在空间之间进行转换。具体而言,编码器$\mathcal{E}$将输入图像y压缩为低维潜在表示$z=\mathcal{E}(y)$;解码器$\mathcal{D}$则负责将潜在表示重建为图像,即$\hat{y}=\mathcal{D}(z)\approx y$。

(2) 条件编码器:为了实现文本对图像生成的控制,模型利用预训练的文本编码器(如CLIP[33])将用户输入的文本提示x编码为语义向量序列$c=E_{\text{txt}}(x)$。这些向量将作为条件信息注入到图像生成过程中。

(3) 去噪网络:这是扩散模型的核心,通常采用基于U-Net[34]结构的神经网络ϵ_θ。它利用交叉注意力机制将文本条件c与图像的潜在特征融合,以此预测并去除潜在空间中的噪声。

1.2 训练与推理过程

扩散模型的运行机制包含训练阶段的前向扩散与反向去噪,以及推理阶段的采样生成。

1.2.1 训练阶段

训练过程可视作1个马尔可夫链,包含有1个前向扩散过程以及1个反向去噪过程。对于前向扩散过程,模型向初始潜在表示z_0中逐步添加高斯噪声,直至其变为各向同性的随机噪声z_T。在时刻t,添加噪声的过程可表示为

$$q(z_t|z_{t-1})=\mathcal{N}\left(z_t;\sqrt{1-\beta_t}\,z_{t-1},\beta_t\boldsymbol{I}\right)\quad t\in(0,T) \tag{1}$$

式中β_t为预定义的方差调度参数。利用重参数化技巧,可以从z_0直接采样得到任意时刻t的含噪潜变量z_t,即

$$q(z_t|z_0)=\mathcal{N}\left(z_t;\sqrt{\overline{\alpha_t}}\,z_0,(1-\overline{\alpha_t})\boldsymbol{I}\right) \tag{2}$$

式中:$\alpha_t=1-\beta_t$,$\overline{\alpha_t}=\prod_{i=1}^{t}\alpha_i$。

随后是反向去噪过程,模型的训练目标是学习去噪网络ϵ_θ,使其能够根据当前时刻t和文本条件c,预测出添加到z_t中的噪声ϵ。训练的损失函数通常采用均方误差(Mean squared error, MSE)形式,即

$$\mathcal{L}=E_{z_0,c,t,\epsilon\sim\mathbb{N}(0,1)}\left[\left|\left|\epsilon-\epsilon_\theta(z_t,c,t)\right|\right|_2^2\right] \tag{3}$$

1.2.2 推理阶段

在推理生成阶段,为了提升生成图像与文本提示的一致性,广泛采用无分类器引导(Classifier-free guidance, CFG)[35]技术。模型同时计算条件去噪预测$\epsilon_\theta(z_t,c,t)$和无条件去噪预测$\epsilon_\theta(z_t,\boldsymbol{\phi},t)$(其中$\boldsymbol{\phi}$为空文本),并通过引导尺度$\omega$(通常$\omega>1$)来调整最终的噪声预测值,有

$$\widetilde{\epsilon}_\theta(z_t,c,t)=\epsilon_\theta(z_t,\boldsymbol{\phi},t)+\omega(\epsilon_\theta(z_t,c,t)-\epsilon_\theta(z_t,\boldsymbol{\phi},t)) \tag{4}$$

最终,模型从随机噪声z_T开始,利用$\widetilde{\epsilon}_\theta$进行迭代去噪得到$z_0$,再经由解码器$\mathcal{D}$还原为最终的生成图像。

2 越狱攻击

视觉大模型的越狱攻击是指恶意用户通过对显性风险提示(无法通过安全过滤器审核的风险提示)进行特定的扰动或重写,构建出能够绕过安全过滤器的隐性风险提示,从而诱导模型生成违反安全策略的风险图像。现有的越狱攻击方法繁多,技术路线各异。为了系统梳理研究现状,本文根据攻击者知识、扰动方式、查询类型以及技术划分4个维度,建立了如图2所示的分类体系。

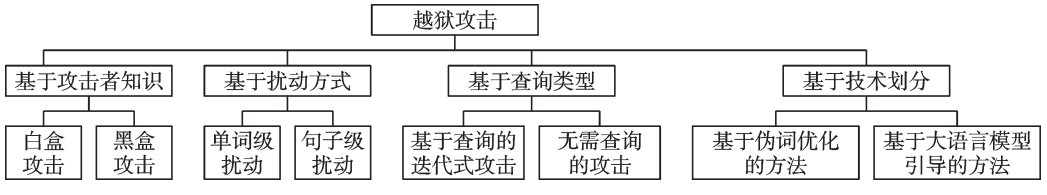


图2 越狱攻击分类体系

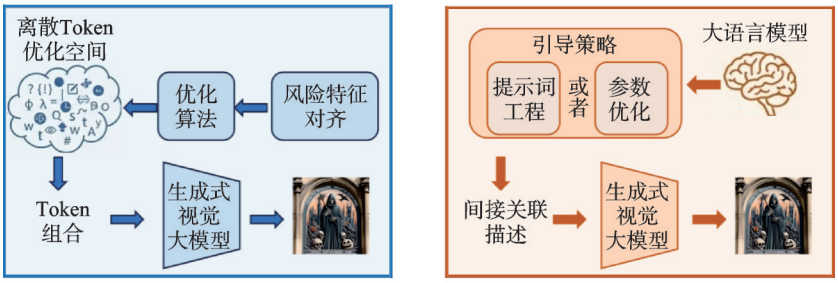
Fig.2 Taxonomy of jailbreak attacks

2.1 基于技术划分

根据隐性风险提示中风险语义的表达机理,本文将现有方法主要划分为两大技术流派:基于伪词优化的方法以及基于大语言模型引导的方法。

2.1.1 基于伪词优化的方法

如图3所示,基于伪词优化的方法旨在利用优化算法生成人类难以理解的Token组合(即“伪词”),其核心思路在于利用深度神经网络的非凸特性,寻找在离散Token空间中语义混乱,但在高维特征空间中与风险概念高度耦合的对抗性扰动。此类方法通常包含3个关键模块:风险特征对齐损失设计、优化算法设计以及优化空间设计。



(a) 基于伪词优化的攻击
(a) Pseudo-optimization-based attack

(b) 基于大语言模型引导的攻击
(b) Large language model-based attack

图3 越狱攻击的攻击方法

Fig.3 Attack methods of jailbreak attacks

(1) 风险特征对齐损失设计。为了使生成的伪词能够诱导模型生成风险图像,攻击者需要构建能够逼近风险特征语义的损失函数作为优化目标。现有的对齐策略主要包括以下4种:

①隐性风险提示-显性风险提示特征对齐。攻击者旨在最小化隐性风险提示特征与显性风险提示特征在文本编码器(如CLIP text encoder)空间中的距离。例如,MMA[36]通过最大化伪词特征与包含敏感词(如“Nudity”)的提示特征之间的余弦相似度,使伪词继承显性提示的风险语义。

②隐性风险提示-显性风险图像跨模态特征对齐。鉴于视觉大模型的最终目标是生成图像,Zhang等[37]提出直接利用CLIP的跨模态对齐能力,构建隐性风险提示特征与目标风险图像特征之间的损失函数,从而引导模型生成特定风格或内容的风险图片。

③隐性风险提示生成图像-显性风险提示跨模态特征对齐。为了更直接地控制生成结果,Sneaky-Prompt[12]将优化目标设定为最大化模型生成的图像特征与预定义风险文本特征之间的相似度。该方法通过闭环反馈确保生成的伪词确实导致了风险图像的产出。

④隐性风险提示-显性风险提示潜在空间噪声对齐。深入到扩散模型的生成机理,P4D[38]在潜在空间进行对齐。该方法计算隐性风险提示在降噪网络(U-Net)中预测的噪声与显性风险提示预测噪声之间的差异,通过最小化该差异,使模型在去噪过程中被“欺骗”,从而生成风险内容。

(2) 优化算法设计。根据是否能够获取视觉大模型的内部架构和具体参数,优化算法可分为白盒优化算法与黑盒优化算法:

①白盒优化算法:在掌握模型架构参数(如文本编码器、降噪网络)的前提下,攻击者可利用梯度引导的优化策略。例如,P4D[38]采用投影梯度下降算法,直接计算损失函数对输入Embedding的梯度并进行更新,从而生成基于伪词的隐性风险提示。该类方法特点是优化效率高,能够快速收敛。

②黑盒优化算法:在无法获取模型梯度的场景下,攻击者仅依靠模型输出的反馈(如Loss值或生成

图像)来搜索优化方向。常用方法包括遗传算法[39]、零阶优化[40]以及强化学习[41]。例如,SneakyPrompt[12]利用强化学习代理搜索最优Token组合。该类方法无需了解模型内部架构,通用性强,但通常需要与系统进行多次交互迭代,计算成本高且优化效率相对较低。

(3) 优化空间设计。优化空间决定了伪词搜索的范围与边界。现有方法通常以语言模型的预训练词表作为离散优化空间,通过优化算法从中选取Token进行排列组合。为了规避外部安全过滤器,研究人员通常会对优化空间进行预处理,剔除与风险概念明确相关的敏感词汇(如将色情、暴力相关词汇从预训练词表中移除)。这种设计确保了最终生成的隐性风险提示在字面上不包含任何违规词汇,从而实现越狱攻击。

总而言之,基于伪词优化的越狱攻击通过针对性地设计损失函数与优化算法,在剔除敏感词的搜索空间中寻找多个无风险Token的特定排列组合。这些隐性风险提示虽然在语言学上缺乏连贯性且语义晦涩,但其在特征空间中构建的风险特征表示能够成功欺骗视觉大模型,诱导其生成高质量的风险图像。

2.1.2 基于大语言模型引导的方法

如图3所示,基于大语言模型引导的越狱攻击旨在利用大语言模型强大的推理与生成能力,搜索语言可解释的间接关联描述,通过深层语义推理诱导视觉大模型生成风险内容。考虑到通用大语言模型缺乏对于视觉大模型越狱攻击的领域知识,无法仅凭自身先验知识直接生成高质量的隐性风险提示,近期工作提出了不同的引导策略来激发大语言模型的攻击潜能[41-46]。根据引导策略的不同,现有方法可进一步分为基于提示词工程的大语言模型攻击方法以及基于参数优化策略的大语言模型攻击方法。

(1) 基于提示词工程的大语言模型攻击方法。此类方法的核心思想是在不改变大语言模型参数的前提下,通过人工设计关联描述搜集策略,利用提示词工程引导大语言模型进行推理与改写。根据策略的交互性,可细分为人工规则策略与反馈优化策略。

①人工规则策略。该策略旨在将人类专家的攻击经验转化为隐性风险提示的构造规则,直接指导大语言模型进行生成。PGJ[44]利用人类视觉感知的先验知识,引导大语言模型搜索与风险单词在视觉形态上相似但语义无关的词汇进行替换(比如“血液”替换为“红色液体”),从而绕过文本过滤器。MJA[46]基于比喻的攻击框架,利用大语言模型的语言能力生成隐晦的比喻修辞描述来间接表达风险内容,从而逃过安全过滤器并利用视觉大模型的联想能力生成风险图片。CMMA[42]引入了文化参考机制,利用特定文化背景下的俚语或典故来增强风险内容的隐蔽性,使得缺乏特定文化知识的过滤器难以识别。

②反馈优化策略。该策略聚焦于大语言模型生成提示与越狱攻击结果之间的交互机制设计。通过构建闭环系统,根据视觉大模型的攻击反馈(如提示被拦截,生成图片非风险)自动化利用大语言模型修改隐性风险提示。FLIRT[43]和CMMA[42]设计了基于上下文学习的反馈机制,引导大语言模型分析攻击失败的原因,并据此对提示词进行校正与重写,通过多轮对话逐步逼近成功越狱的边界。JailFuzzer[39]将模糊测试思想引入大语言模型引导过程,利用遗传变异算法维护1个提示词种群,引导大语言模型对种子提示进行交叉、变异操作,从而搜索出更多样化且有效的隐性风险提示。

(2) 基于参数优化策略的大语言模型攻击方法。与提示词工程不同,此类方法旨在对大语言模型针对越狱攻击任务进行参数微调优化,使其内化为具备专业攻击能力的“越狱专家”。根据参数优化范式,现有方法可分为以下3类:

①扰动优化策略。该策略通过对大语言模型的参数加入不同方向的扰动,观测其生成的隐性风险提示在视觉大模型上的攻击效果变化,从而自适应地寻找大语言模型参数的优化方向。例如,UPAM[47]基于一种统一的攻击框架,通过在大语言模型参数层施加扰动生成不同参数下的多样化隐性

风险提示,并基于球面探测学习的优化算法,通过不同隐性风险提示攻击结果自动化估计梯度,实现大语言模型的参数优化过程。

②监督微调策略。该策略通过构造高质量的〈显性风险提示,隐性风险提示〉数据对,直接对大语言模型进行监督微调。例如,ART[48]和GenBreak[49]首先收集大量成功的越狱案例,构建指令微调数据集,训练大语言模型学习从显性风险提示到隐性风险提示的转化能力,使其能够批量生产高成功率的越狱提示。

③强化学习策略。该策略将大语言模型生成提示过程与视觉大模型的攻击反馈整合为统一的在线强化学习过程。例如,R2A[41]利用越狱攻击的最终结果(如是否生成了目标风险图像)作为奖励信号,通过组内相对策略优化算法(Group relative policy optimization, GRPO)强化学习[50]算法引导大语言模型在线调整生成策略,使其能够自主探索出人类难以察觉的复杂攻击逻辑。

总的来说,基于大语言模型引导的越狱攻击方法摆脱了对不可解释伪词的依赖,转而利用大语言模型强大的语义推理与生成能力,构建具有高度语言可解释性的隐性风险提示。从早期的基于人工规则的启发式搜索,发展到如今基于反馈迭代与强化学习的自动化攻击,此类方法展现了极强的演进趋势。通过在语义层面进行隐晦的关联与重写,这些方法能够有效地将显性恶意意图隐藏在看似合规的自然语言描述中,从而对现有的语义理解型安全防御机制构成更深层次的挑战。

2.2 基于扰动方式划分

从攻击者对原始提示词的修改粒度来看,越狱攻击方法还可分为单词级扰动与句子级扰动。这两类方法在攻击的隐蔽性和有效性上呈现出不同的权衡。

2.2.1 单词级扰动

单词级扰动主要聚焦于对显性风险提示中特定的敏感单词进行修改替换,保留原始提示的句法结构和其他非敏感上下文。如图4(a)所示,PGJ[44]利用人类视觉感知的特性,寻找与风险单词在视觉形态上相似(如形状、颜色、纹理),但在语义上不相关的单词进行替换。这种方法巧妙地利用了视觉大模型将视觉特征与文本语义对齐的机制,实现了不适合观看内容的隐蔽生成。SGT[45]聚焦于涉及政治敏感人物的越狱攻击。该方法不直接使用政治敏感人物名称,而是将其替换为与其紧密关联的历史事件、职务称号或其他非敏感实体名称,通过模型的知识关联能力诱导生成目标人物。相比之下,如图4(c)所示,SneakyPrompt[12]采用了更为激进的伪词替换策略,通过利用强化学习算法优化一个伪词生成器,将风险单词替换为人类无法理解的Token组合。这些伪词在特征空间中能精确指向风险概念,从而在不触发文本过滤器的情况下激活模型的风险生成路径。

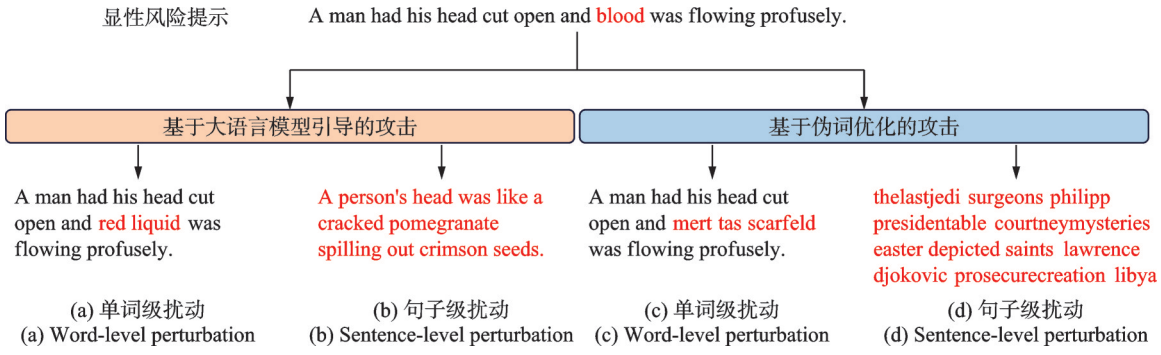


图4 单词级扰动和句子级扰动的示例

Fig.4 Examples of word-level and sentence-level perturbation attacks

2.2.2 句子级扰动

如图4(b)和图4(d)所示,句子级扰动则不再局限于对个别词汇的修补,而是聚焦于对整个风险提示进行重写或结构重组。该类方法旨在通过改变表达方式,将显性的恶意意图转化为隐晦的语义描述。MJA[46]提出了一种基于比喻关联的重写框架。该方法首先在大规模语料库中搜寻风险内容的常用比喻描述,然后定位这些比喻与风险内容之间的上下文关联,最后结合比喻描述与上下文语境对原始风险提示进行整体重写,生成语义通顺但意图隐晦的隐性风险提示(生成示例如图4(d)所示)。R2A[41]则通过在线强化学习阶段训练1个能够高效实现越狱攻击的大语言模型。该方法结合语言学中的框架语义学知识,引导大语言模型理解风险场景的核心要素,并据此构建全新的隐性风险提示,实现了从意图理解到整体重写的自动化攻击过程。

对比而言,单词级扰动由于仅修改特定的风险词汇,对原始提示的改动较小,能够较好地保留非敏感部分的语义信息。然而,由于其依然保留了与风险语义紧密相关的上下文结构(如"头顶流出…"的动作描述),导致其风险意图的隐匿性相对较弱,容易被基于上下文分析的高级安全过滤器拦截。相比之下,句子级扰动通过重写整个提示来构建隐性风险提示,虽然扰动程度较大,改变了原始的句法结构,但能够更彻底地打散风险语义的显性特征,因此具有更高的风险隐匿性和穿透力。但由于风险语义更加隐蔽与分散,其难以高概率保证视觉大模型能够准确地理解其内部隐含的风险语义,导致攻击有效性较弱。

2.3 基于查询类型划分

除了技术路线与扰动方式外,攻击过程中是否依赖目标视觉大模型的实时反馈也是区分攻击方法的重要维度。根据攻击者在构建对抗性提示时是否需要与目标模型进行交互,现有的越狱攻击方法可分为基于查询的迭代式攻击与无需查询的攻击。

2.3.1 基于查询的迭代式攻击

此类攻击的核心范式是"闭环优化",即攻击者需要多次向目标视觉大模型发送查询请求,并利用模型返回的攻击结果(如生成的图像内容、分类器的拒绝信号或损失值)作为反馈信号,动态调整和优化隐性风险提示的生成策略。CMMA[42]设计了基于大语言模型的反馈闭环,当生成的提示未能成功越狱时,系统会将失败案例反馈给大语言模型,引导其分析被拦截的原因(如包含特定敏感词),并据此进行针对性的校正与重写,直到攻击成功。JailFuzzer[39]则借鉴了模糊测试的思想,通过遗传算法维护1个提示词种群,根据每个提示词诱导生成风险图像的成功率来决定其是否保留或变异,从而在多次查询迭代中进化出高质量的隐性风险提示。

2.3.2 无需查询的攻击

此类攻击的核心范式是"开环生成",即攻击者在构建隐性风险提示的过程中,无需访问目标视觉大模型的接口,也不依赖模型的实时反馈结果,而是直接根据先验知识或代理模型生成隐性风险提示。例如,MMA[42]利用本地的CLIP文本/图像编码器,构建隐性风险提示特征对齐损失,并根据梯度优化算法直接优化出能够最大化与风险概念特征相似度的Token组合。由于特征对齐过程仅在本地编码器上完成,无需向目标视觉大模型发送查询请求即可生成具有通用攻击能力的隐性提示。此外,PGJ[44]和SGT[45]利用人工经验直接设计提示词工程策略,通过视觉相似性词语等规则构建隐性风险提示。这类攻击通常假设这些基于先验规则生成的提示本身就具备绕过过滤器的能力,因此不需要根据模型反馈进行调整即可直接使用。

无需查询的攻击无需借助视觉大模型的攻击结果反馈,就可以直接生成隐性风险提示,因此具有极高的生成效率,且不容易触发频次限制或被后台审计发现。但是,由于缺乏与目标模型的实际交互,此类方法生成的提示往往泛化性受限,难以保证针对特定未知模型的攻击有效性。相比之下,基于查

询的迭代式攻击利用目标模型的多次交互反馈来微调攻击方向,虽然计算成本高,效率较低且容易暴露,但通过不断试错与修正,可以实现质量更高、成功率也更高的隐性风险提示生成。

2.4 基于攻击者知识划分

根据攻击者对目标视觉大模型内部信息的掌握程度,越狱攻击方法可被划分为白盒攻击与黑盒攻击。

2.4.1 白盒攻击

在白盒攻击场景中,攻击者被假设拥有目标模型的完整知识,包括模型架构、权重参数以及训练数据分布等。在此场景下,攻击者可以直接计算损失函数相对于输入提示词的梯度,利用梯度下降等优化算法在优化空间中精确地搜索隐性风险提示。例如,P4D[38]利用对视觉大模型内部降噪网络(U-Net)的访问权限,计算隐性风险提示与显性风险提示在潜在空间预测噪声的差异,并通过反向传播直接优化Token组合,从而精准地挖掘模型内部的安全漏洞。

2.4.2 黑盒攻击

在黑盒攻击场景中,攻击者无法获取目标模型的内部架构与参数,仅能通过应用程序接口(Application programming interface, API)访问模型,向其发送文本提示并获取生成的图像或安全过滤器拦截结果。在此场景下,攻击者通常将视觉大模型视为1个黑箱函数,通过分析输入提示与输出图像之间的映射关系,利用查询反馈策略来构建隐性风险提示。绝大多数基于大语言模型引导的攻击方法均属于此类。例如,MJA[46]利用LLM生成隐喻描述,直接通过API测试目标模型的生成反应,无需任何梯度信息。R2A[41]和GenBreak[49]则利用强化学习策略,仅根据模型输出的图像是否违规作为奖励信号来调整大语言模型的生成策略,实现了对未知商业模型的有效攻击。

考虑到在现实应用场景中,主流的商业视觉大模型(如Midjourney[2]、DALL-E3[3])均以闭源API的形式提供服务,不公开具体参数,因此现有的研究方法主要集中在黑盒攻击领域,具有更强的实际应用价值。相比之下,白盒攻击方法主要集中在早期的基于伪词优化的越狱攻击研究中,侧重于在学术研究环境下利用透明模型挖掘扩散模型自身存在的理论安全漏洞与解释性机理。

2.5 本章小结

本章系统梳理了针对视觉大模型的越狱攻击技术,根据核心攻击机理的不同,将现有方法划分为基于伪词优化与基于大语言模型引导两大技术流派,并如表1所示,从扰动方式、查询类型及攻击者知识等维度进行了详细归纳。

基于伪词优化的方法(如RAB[51]、UnlearnDiff[52])代表了早期的攻击范式,其核心在于利用优化算法寻找在特征空间与风险概念高度耦合的“对抗性伪词”。早期的此类研究主要集中在白盒场景下,依赖模型梯度信息进行高效攻击。随着研究的深入,现有方法开始瞄准现实API场景的黑盒攻击(如RT-Attack[53]、DiffZoo[40]),不再依赖模型内部梯度优化,而是利用模型输出自动估计优化方向,实现基于伪词的隐性风险提示生成。

随着大语言模型推理能力的增强,攻击范式逐渐偏向基于大语言模型引导的方法。这些方法脱离了对模型内部参数的依赖,转而利用大语言模型进行提示词工程或参数优化策略的搜索,进而生成语义通顺的隐性风险提示。相比于伪词优化产生的乱码,大语言模型生成的隐性风险提示具有极高的语言可解释性和隐蔽性,能够更轻松地绕过外部防御系统。

观察表1可以发现,现有的越狱攻击研究呈现出明显的黑盒化趋势。所有基于大语言模型引导的方法以及部分伪词优化方法均属于黑盒攻击,且越来越多地采用基于查询的迭代策略。这表明攻击者正通过利用目标模型返回的图像内容或安全拒绝信号作为反馈,动态调整攻击策略,使得针对闭源商

表1 视觉大模型越狱攻击方法汇总与分类

Table 1 Overview and taxonomy of jailbreak attack methods on visual generation model

一级分类	二级分类	方法	扰动方式	查询类型	攻击者知识
伪词优化		RAB[51]	句子层级	无需查询	白盒
		UnlearnDiff[52]	句子层级	无需查询	白盒
		P4D[38]	句子层级	无需查询	白盒
		Maus[54]	句子层级	无需查询	黑盒
		MMA[36]	句子层级	无需查询	黑盒
		Zhang[37]	句子层级	无需查询	黑盒
		JPA[55]	句子层级	无需查询	黑盒
		Sneaky[12]	单词层级	基于查询	黑盒
		DiffZOO[40]	句子层级	基于查询	黑盒
		RT-Attack[53]	句子层级	基于查询	黑盒
		HTS-Attack[56]	句子层级	基于查询	黑盒
		PLA[57]	句子层级	基于查询	黑盒
大语言模型引导	提示词工程	SGT[45]	单词层级	无需查询	黑盒
		Groot[58]	单词层级	基于查询	黑盒
		PGJ[44]	句子层级	无需查询	黑盒
		DACA[59]	句子层级	无需查询	黑盒
		Flit[43]	句子层级	基于查询	黑盒
		CMMA[42]	句子层级	基于查询	黑盒
		MJA[46]	句子层级	基于查询	黑盒
		JailFuzzer[39]	句子层级	基于查询	黑盒
	参数优化策路	ART[48]	句子层级	无需查询	黑盒
		UPAM[47]	句子层级	基于查询	黑盒
		RPG-RT[60]	句子层级	基于查询	黑盒
		R2A[41]	句子层级	基于查询	黑盒
		GenBreak[49]	句子层级	基于查询	黑盒

业模型的攻击变得愈发精准且具有威胁性。

综上所述，视觉大模型的越狱攻击正从单纯的梯度对抗演变为一种结合了语义推理与反馈优化的复杂系统性威胁，这对未来的防御机制设计提出了更高的鲁棒性与泛化性要求。

3 概念擦除

3.1 任务描述与分类体系

概念擦除，又被称为机器遗忘，旨在通过特定的技术手段，消除视觉大模型生成特定风险概念（如色情、暴力、版权风格和特定人物等）的能力。与外挂式的文本过滤器或图像过滤器等外部防御机制不同，概念擦除属于内生安全范畴，其核心目标是修正模型内部的概率分布，使得模型在接收到包含风险概念的提示词 $\boldsymbol{c}_{\text{risk}}$ 时，不再生成与之对应的风险图像 y_{risk}，而是输出无害的图像 y_{safe}，即 $P_\theta(y_{\text{risk}}|\boldsymbol{c}_{\text{risk}}) \rightarrow P_\theta(y_{\text{safe}}|\boldsymbol{c}_{\text{safe}})$。

根据治理过程中对模型参数的修改方式及干预阶段的不同,如图5和图6所示,本文将现有的概念擦除方法分为3类:模型微调、模型编辑、以及推理引导。

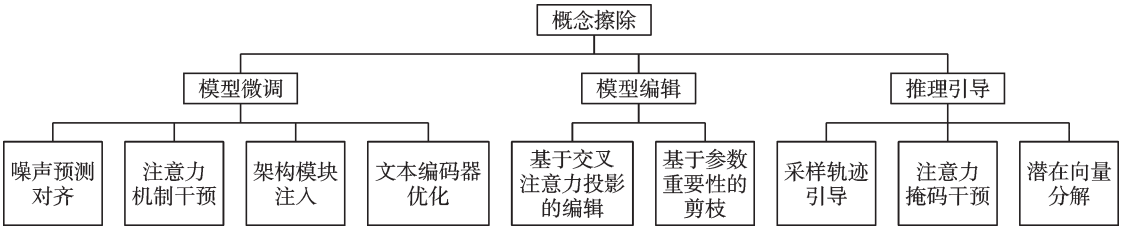


图5 视觉大模型概念擦除方法分类框架

Fig.5 Taxonomy of concept erasure methods on visual generation model

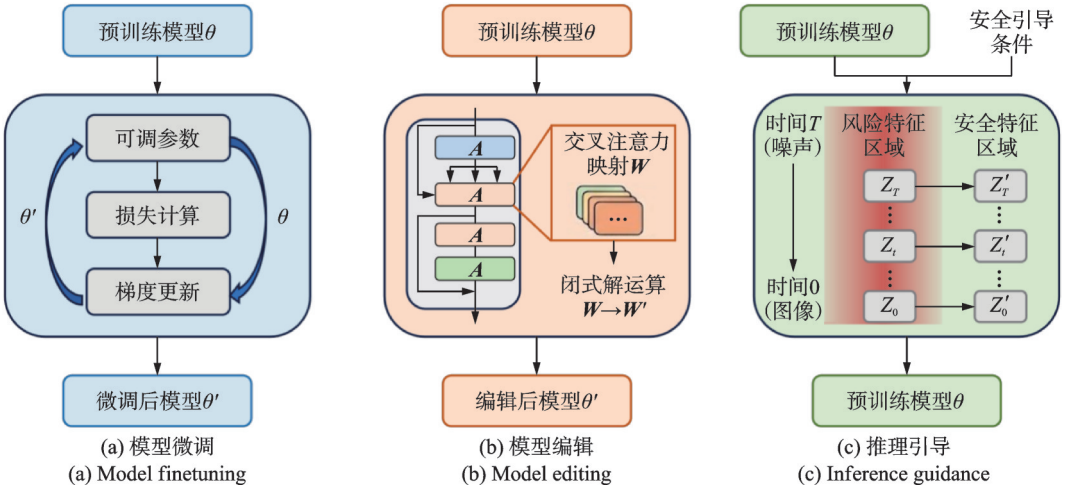


图6 3类视觉大模型概念擦除方法

Fig.6 Three kinds of concept erasing methods on visual generation model

(1)模型微调。利用少量的目标概念数据,通过反向传播算法迭代更新模型的部分或全部参数(如U-Net的权重),使模型"遗忘"特定概念的生成范式。

(2)模型编辑。寻找模型中存储特定知识的关键权重,利用矩阵运算的闭式解或者神经元裁剪等策略直接修改这些参数,无需繁琐的训练迭代,旨在实现精准且高效的概念擦除。

(3)推理引导。不修改模型的任何权重参数,而是在推理生成阶段,通过引入额外的引导项,动态地将生成轨迹偏离风险区域。

3.2 模型微调

模型微调是概念擦除中最直观且广泛应用的一类方法。如图6(a)所示,其核心思想是在预训练模型的基础上,利用包含风险概念的小样本数据集构建特定的损失函数,通过反向传播算法更新模型的全部或部分权重参数,从而重塑模型对特定概念的响应机制。根据微调的目标机制不同,本文将现有的微调方法分为4类:噪声分布对齐、注意力机制干预、架构模块注入以及文本编码器优化。

3.2.1 噪声预测对齐

此类方法直接作用于扩散模型的去噪网络(U-Net),通过修改训练目标,强制模型在面对风险提示词时,预测出的噪声分布向无害概念或空概念偏移,从而从根源上阻断风险图像的生成。ESD[14]利用

CFG[35]的负向信号,微调模型使其在接收到风险提示(如“裸体”)时,预测的噪声方向与原模型相反,从而推远风险分布。AC[15]则采用锚点对齐策略,强制将风险概念的分布映射到预定义的通用锚点概念(如将“裸体”映射为“穿着衣服的人”),覆盖原有的风险记忆。TRCE[61]提出一种去噪轨迹引导方法,利用对比学习优化去噪路径,使其远离风险概念的潜在空间,确保生成的图像安全。

3.2.2 注意力机制干预

在生成式视觉大扩散模型中,注意力机制控制着生成的图像内容与文本提示的对应关系。基于此现象,此类方法通过直接优化或约束交叉注意力图,抑制风险词汇对图像生成的激活作用。FMN[62]通过在微调过程中最小化风险概念在交叉注意力层的激活权重,使模型在生成过程中“忽略”风险词汇的语义引导。SafeGen[63]进一步聚焦于自注意力层,通过对齐风险概念与无风险概念的预测噪声,抑制色情等特定视觉特征的合成。

3.2.3 文本编码器优化

除了去噪网络,部分方法也选择对文本编码器(如CLIP Text Encoder)进行微调,通过重塑语义空间来实现擦除。Safe-CLIP[64]通过收集由“安全-不安全”文本对组成的数据集,对CLIP文本编码器进行微调。其目标是将包含风险概念的文本嵌入(如色情描述)强行拉近至对应的安全文本嵌入(如“穿着衣服的人”),从而在源头上阻断风险语义的输入。AdvUnlearn[65]在此基础上进一步引入对抗训练,在微调文本编码器时动态生成对抗性提示词,提升模型防御隐性攻击的能力。

3.2.4 架构模块注入

与上述仅微调原有参数的方法不同,此类方法通过在原有模型架构中插入新的可训练模块来实现概念擦除。这种非侵入式的设计往往能更精准地控制擦除效果,且对原模型的破坏更小。CPE[66]并非简单微调U-Net的注意力权重,而是在交叉注意力层中引入非线性的残差注意力门控模块。该模块能够自适应地检测并修改校正目标风险概念的信息流,同时允许其他安全概念无损通过,从而实现了对目标概念的精确点杀和对非目标概念的高保真保留。

为了解决微调导致的灾难性遗忘问题,现有研究引入了两种主要的正则化策略来维持模型的常规生成能力。(1)参数级正则化。SA[67]引入了持续学习中的弹性权重巩固技术。它计算参数的费雪信息矩阵[68]来评估权重的重要性,在损失函数中限制那些对通用生成能力至关重要的参数发生大幅变化。(2)输出级正则化。许多方法(如FMN[62]、UCE[69]等)在优化目标中加入了一项保留损失约束。该约束要求模型在面对中性或无害的锚点提示词时,新模型的噪声预测输出需与原始冻结模型保持一致(即最小化两者输出的MSE距离),从而确保模型在擦除目标概念的同时,不会破坏对其他无关概念的生成能力。

3.3 模型编辑

如图6(b)所示,与微调侧重于通过损失函数重塑分布不同,模型编辑更关注模型内部的知识定位。根据对参数干预方式的不同,现有方法主要分为两类:一类基于交叉注意力投影的编辑,利用闭式解直接重写关键权重矩阵的数值;另一类则是基于参数重要性评估,通过剪枝或掩码机制阻断特定神经元的激活路径。

3.3.1 基于交叉注意力投影的编辑

此类方法通常基于“键-值”记忆假设,即认为扩散模型中的交叉注意力层不仅负责文本与图像的对齐,还实质性地存储了特定视觉概念的语义知识[17]。基于此假设,此类方法的核心思想是将包含风险概念的文本嵌入(作为Key)映射到无意义或中性的视觉特征输出(作为Value),从而切断风险语义的激活路径。

TIME[17]是该领域的开创性工作。它将模型编辑形式化为一个最小二乘问题,通过闭式解更新交叉注意力层的投影权重,强制将目标风险概念的Key向量映射到预定义的“空”Value向量。该方法无需训练数据,仅需输入目标概念的文本提示,即可在毫秒级时间内完成擦除。针对同时擦除多个风险概念的需求,UCE[69]提出了一种统一的闭式解框架。它在TIME的基础上引入了非目标概念正则约束,能够在一次矩阵运算中同时消除数千个不同的艺术风格或对象概念,且极大地减少了对非目标概念的干扰(即“附带损伤”)。MACE[70]则结合了闭式解编辑与轻量级微调的优势。它首先利用闭式解对交叉注意力层进行粗粒度的语义阻断,随后引入LoRA[71]模块进行低秩微调,以在特征空间进一步对齐概念分布。这种“闭式解+LoRA”的混合策略在处理大规模概念集合时表现出了更优的泛化性与稳定性。

3.3.2 基于参数重要性的剪枝

与直接修改权重数值不同,此类方法旨在识别模型中专门负责生成特定风险概念的“神经元”或权重子集,并通过将其置零来实现概念遗忘。ConceptPrune[72]采用了一种轻量级的参数掩码技术。该方法首先通过少量的探测数据识别出对生成特定风险概念贡献度最高的权重参数(通常位于前馈神经网络FFN或注意力层中),然后在模型推理过程中直接将这些参数置零来消除模型输出风险。由于仅修改了极小比例的权重,该方法在有效移除风险概念的同时,最大程度地保留了模型的预训练知识结构。

相比于模型微调,模型编辑方法利用矩阵运算的解析解或者神经元裁剪,避免了不稳定的梯度下降过程,具有极高的计算效率(通常仅需几秒钟)。然而,由于其主要依赖对交叉注意力层的线性修改,当面对复杂的非线性语义关联或需要擦除的概念与保留概念在特征空间高度重叠时(例如“裸体”与“人体皮肤”),单纯的线性编辑可能会导致更严重的生成质量下降或擦除不彻底问题。

3.4 推理引导

如图6(c)所示,推理引导是一类无需对模型参数进行训练或修改的轻量级治理策略。它利用扩散模型可控生成的特性,在推理阶段通过引入额外的引导信号、修改采样公式或干预注意力机制,动态地调整生成的潜在轨迹,使其偏离风险概念的分布区域。此类方法具有即插即用的优势,能够灵活地应用于各种预训练模型,但通常会增加推理阶段的计算开销。

3.4.1 采样轨迹引导

此类方法主要基于CFG无分类器引导机制的变体。通过在采样公式中引入“负向安全引导项”,在每一步去噪过程中将潜在向量推离风险语义。SLD[73]是一种典型的基于引导的治理方法。它定义了1个包含风险概念的文本提示c_{risk}(如“hate”“violence”),并在推理过程中计算模型针对该风险提示的噪声预测$\epsilon_\theta(z_t, c_{\text{risk}})$。在此基础上,SLD修改了标准的CFG公式,引入了1个安全引导尺度,当检测到模型针对用户输入提示生成的潜在特征向风险方向移动时,强制对其施加1个反向的校正向量,从而实现对不当内容的修正校正。SDD[74]则利用模型自身的知识进行引导。它通过对比“以风险词为条件”的噪声预测和“无条件”的噪声预测,构建1个自我纠正的引导信号,使得最终生成的图像在保留原意图的同时去除风险元素。

3.4.2 注意力掩码干预

由于扩散模型的交叉注意力图决定了生成图像的空间布局和语义对应关系,此类方法通过实时检测并抑制注意力图中的风险响应区域来实现治理。基于此机制,SAFREE[75]采用了一种免训练的自适应注意力掩码防御机制。它利用扩散模型生成过程中的语义分割能力,实时监控交叉注意力图中与风险Token(如“blood”“nudity”)相关的激活区域。一旦检测到高响应区域,SAFREE会自动生成1个空间掩码将该区域的注意力权重置零或抑制,从而在不破坏背景和其他对象的前提下,精准地抹除画面

中的风险局部。

3.4.3 潜在向量分解

此类方法从线性代数的角度出发,认为风险概念在潜在特征空间中对应着特定的向量方向,进而通过分解和正交化操作,可以剔除风险成分。AdaVD[76]假设图像的潜在表示可以分解为“风险分量”和“非风险分量”。该方法在推理过程中,利用奇异值分解(Singular value decomposition, SVD)技术动态识别出代表风险概念的主成分方向,并利用向量正交化处理将潜在向量在该方向上的投影移除。这种方法无需预先定义复杂的负向提示,能够自适应地消除潜在的攻击性视觉特征。

推理引导方法凭借其无需训练、部署灵活的特性,成为应对突发安全风险的有效手段,特别是SLD等方法已被集成到Diffusers等主流开源库中。然而,由于其需要在每一步生成过程中进行额外的梯度计算或注意力干预,往往会导致推理延迟增加。此外,相比于从参数层面根除知识的模型微调,推理引导属于“治标不治本”的防御策略,在面对精心设计的强对抗性攻击时(如利用梯度绕过引导),其防御边界相对脆弱。

3.5 本章小结

本章针对视觉大模型的概念擦除任务,系统梳理了现有的研究成果。如表2所示,根据对模型干预的深度与阶段不同,将现有方法划分为模型微调、模型编辑与推理引导3大技术体系。这3类方法在擦

表2 视觉大模型概念擦除方法汇总与分类

Table 2 Overview and taxonomy of concept erasing methods on visual generation model

一级分类	二级分类	方法简称	核心机制
模型微调	噪声预测对齐	ESD[14]	利用CFG负向引导信号推远风险分布
		AC[15]	将风险概念分布映射到无害锚点概念
		TRCE[61]	利用对比学习优化去噪轨迹远离风险域
		SA[67]	引入正则化,防止通用参数发生灾难性遗忘
	注意力机制干预	FMN[62]	最小化风险概念在交叉注意力层的激活权重
		SafeGen[63]	抑制自注意力层中风险视觉特征的生成
	架构模块注入	CPE[66]	插入残差注意力门控模块拦截风险流
		SafeGuider[77]	引入风险文本分类器实现风险文本特征校正
		Concept Replacer[78]	引入风险视觉区域定位方法实现风险视觉特征校正
	文本编码器	Safe-CLIP[64]	将风险文本嵌入映射至安全语义空间
		AdvUnlearn[65]	在微调中加入对抗性提示词训练
		RACE[79]	结合对抗攻击与擦除的交替优化框架
模型编辑	基于交叉注意力投影	TIME[17]	闭式解更新投影矩阵,将风险Key映射为空Value
		UCE[69]	引入正则约束,实现多概念统一编辑
		MACE[70]	结合闭式解编辑与LoRA微调的大规模擦除
		RECE[80]	结合闭式解与对抗训练
	基于参数重要性剪枝	ConceptPrune[72]	识别并裁剪对风险生成贡献最大的关键权重
推理引导	采样轨迹引导	SLD[73]	在采样公式中引入安全引导尺度进行反向校正
		SDD[74]	利用模型自身的无条件预测进行自蒸馏引导
	注意力掩码干预	SAFREE[75]	实时检测并利用掩码抑制风险区域的注意力
	潜在向量分解	AdaVD[76]	利用SVD分解剔除潜在特征中的风险向量分量
		CURE[81]	改进SVD分解实现可控灵活地去除风险向量分量

除彻底性、计算效率及通用能力保留之间呈现出显著的权衡关系,并在核心机制上展现出多样化的演进趋势。模型微调通过反向传播从根源上重塑模型的分布,擦除最为彻底,但计算成本高昂,且面临严重的“灾难性遗忘”风险,往往需要配合正则化使用。模型编辑技术旨在通过闭式解或参数剪枝实现毫秒级的快速治理,特别适用于单一或少量的特定概念移除,但在处理复杂语义纠缠或大规模概念时,其线性映射能力的局限性可能导致擦除不净或误伤。推理引导作为一种免训练的即插即用方案,具有极高的部署灵活性,能够应对突发的安全风险。然而,它以牺牲推理延迟为代价,且由于未修改模型参数,其防御边界相对脆弱,容易被基于梯度的白盒攻击绕过。

4 数据集

高质量数据集是评估视觉大模型安全性、攻击有效性及防御性能的基础。根据数据集的用途,本文将其分为通用风险基准数据集与概念擦除专用数据集两大类。

4.1 通用风险基准数据集

此类数据集旨在利用显性风险提示数据来评估模型的安全性,通常涵盖色情、暴力和歧视等多种主流风险类别。

(1) I2P(Inappropriate image prompts)[73]。I2P是目前视觉大模型安全领域应用最广泛的真实世界基准数据集。该数据集包含了从Lexica.art[82]网站上爬取的约4 700条真实用户提示词,涵盖了仇恨言论、骚扰、暴力、自残、色情内容、惊悚内容和非法活动7大风险类别。由于这些提示词直接源自真实用户的生成请求,I2P能够准确反映模型在实际部署环境中面临的现实安全威胁。

(2) Unsafe-Diffusion[11]。为了解决真实用户数据中风险类别分布不均的问题,Unsafe-Diffusion利用ChatGPT根据预定义的风险分类体系生成了描述性的风险文本,涵盖有色情、暴力、非法、仇恨和涉证5大类别共434个风险提示。

(3) T2VSafetyBench[83]。尽管主要针对文生视频模型,该数据集在视觉大模型安全研究中同样具有重要参考价值。该数据集定义了12个风险类别,涵盖有主流NSFW以及涉证侵权等类别;不仅包含了常见的显性风险提示,而且包含了现有攻击方法利用利用伪词优化的隐性风险提示,因此实现了多角度的安全评估。

(4) T2I-RiskyPrompt[84]。T2I-RiskyPrompt是一个面向综合安全评估的基准数据集,面向7大视觉大模型平台构建了一个包含有6大类14小类的风险体系,并标注了6 432个能够有效诱导模型生成风险图像的风险提示,同时配有1个风险图像检测器用于对模型安全性进行量化评估。该数据集精细化地标注了不同语义粒度的风险提示,是测试前沿攻防算法的重要基准。

4.2 概念擦除专用数据集

此类数据集主要服务于概念擦除任务,用于测试模型是否成功“遗忘”了特定的目标概念,同时评估其对非目标概念的保留能力。

(1) WikiArt[85]。用于艺术风格擦除任务的标准数据集。WikiArt包含了数万张来自不同流派(如印象派、立体主义)和著名艺术家(如梵高、莫奈)的画作及其元数据。研究者通常利用该数据集构建包含特定艺术家姓名的提示词(如“A painting in the style of Van Gogh”),用于测试模型治理方法是否能有效移除模型对特定版权风格的模仿能力。

(2) Celebrity[70]。针对人物肖像隐私保护,MACE等研究工作构建了专门的名人数据集。该数据集面向GIPHY celebrity detector (GCD)[86]项目,选取了能够被GCD准确识别的200个名人姓名,并配有固定提示模板(如a photo of “name”)作为文本提示。

(3) ImageNet[87]。主要用于评估具体物体擦除。研究者利用ImageNet中的类别标签(如“降落伞”“教堂”)构建模板提示词(如“A photo of a [class]”)。该数据集用于测试模型能否精准地从知识库中移除特定的名词概念,常被用于验证微调与编辑方法的精确性。

(4) COCO-30K[88]。这是评估概念擦除中模型通用能力保留的核心基准。该数据集包含3万条描述日常场景的通用文本提示(MS-COCO验证集)。在擦除特定风险概念后,研究者使用COCO-30K生成图像并计算FID(Fréchet inception distance)[89]分数。低FID分数意味着模型在遗忘风险概念的同时,未发生灾难性遗忘,仍能高质量地生成与风险无关的正常图像。

5 总结与展望

随着视觉大模型在内容创作领域的广泛应用,其内生的安全风险与治理问题已成为学术界与工业界关注的焦点。本文围绕视觉大模型生成内容风险与治理这一核心命题,系统综述了越狱攻击与概念擦除两大关键任务的研究现状。

在越狱攻击方面,本文从攻击者知识、扰动方式、查询类型及技术流派等多个维度建立了分类体系。重点分析了基于伪词优化的方法,阐述了其如何利用特征空间对齐生成人类不可读但在模型内部高响应的隐性风险提示;同时探讨了基于大语言模型引导的方法,揭示了其如何利用语义推理构建语言可解释的隐性风险提示。研究表明,现有的攻击手段正朝着自动化、隐蔽化以及黑盒化的方向演进,对外部安全过滤器的防御边界构成了严峻挑战。

在概念擦除方面,本文将其定位为内生安全治理机制,并依据模型参数的干预方式将其划分为模型微调、模型编辑与推理引导3类。详细梳理了利用反向传播重塑噪声分布的微调策略、基于闭式解快速修改交叉注意力的编辑方法以及免训练的采样引导技术。分析发现,虽然概念擦除能够有效移除特定风险概念,但在平衡擦除效果与通用生成能力保留方面仍面临技术瓶颈。

尽管现有的研究已取得显著进展,但在视觉大模型的安全攻防博弈中,仍有若干关键问题亟待解决。未来的研究可重点关注以下4个方向:

(1) 动态攻防博弈与防御鲁棒性。目前的攻防研究多处于静态对抗阶段:攻击者针对固定模型设计提示,防御者针对特定概念进行擦除。然而,研究表明概念擦除并非永久有效,恶意微调或特定的白盒攻击(如UnlearnDiff[52])可以唤醒被擦除的风险记忆。未来的研究应致力于探索具有理论认证鲁棒性的治理方法,或开发结合对抗训练的防御机制,使模型在面对未知、自适应的越狱攻击时仍能保持安全防线不被击穿。

(2) 自回归视觉生成模型的安全对齐。除了主流的扩散模型,基于“下一个Token预测”的自回归视觉生成模型(如GPT-4o[90]、Bagel[91])正重新受到关注。与扩散模型的去噪机制不同,自回归模型将图像编码为离散的视觉Token序列。这意味着视觉大模型的攻防研究技术范式会迎来一定转变。未来的工作亟需填补针对自回归视觉模型的攻防研究空白,并探索如何将大语言模型安全领域的人类反馈强化学习(Reinforcement learning from human feedback, RLHF)[92]或直接偏好优化(Direct preference optimization, DPO)[93]等对齐技术,有效迁移至视觉自回归生成任务中。

(3) 大规模概念的高效治理与持续学习。现实场景中的风险概念种类繁多且不断演变,如新出现的仇恨符号、特定的版权风格。现有的概念擦除方法虽然在单概念擦除上表现优异,但在面对成千上万个概念的大规模联合擦除时,往往会导致模型参数空间的过度扭曲,严重损害生成质量。未来需要探索基于元学习或持续学习的高效治理框架,实现对新增风险概念的增量擦除,同时保证模型的基础能力不发生灾难性遗忘。

(4) 风险生成的机理可解释性研究。当前的研究多将视觉大模型视为黑盒或灰盒,侧重于通过优化输入输出进行攻防。然而,对于模型内部“风险概念是如何存储的”以及“隐性风险提示是如何激活风险路径的”等深层机理尚缺乏清晰的解释。未来的工作应结合机械可解释性技术,深入分析扩散模型中去噪网络与注意力机制的神经元行为,精确定位风险知识的物理存储位置,从而为设计更精准、副作用更小的“神经外科手术式”治理方法提供理论支撑。

参考文献:

[1] COMPVIS. Stable-Diffusion-V1-4[EB/OL]. (2024-01-10). https://huggingface.co/CompVis/stable-diffusion-v1-4.

[2] Midjourney. Midjourney[EB/OL]. (2023-05-05). https://www.midjourney.com.

[3] OpenAI. Dall·E 3[EB/OL]. (2023-10-10). https://openai.com/index/dall-e-3.

[4] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of Advances in Neural Information Processing Systems 33.[S.l.]: Neural Information Processing Systems Foundation Inc., 2020: 6840-6851.

[5] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding [C]//Proceedings of Advances in Neural Information Processing Systems 35. New Orleans, LA, USA: Neural Information Processing Systems Foundation, Inc., 2022: 36479-36494.

[6] CHEN J, YU J, GE C, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis[C]//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria: OpenReview, 2024.

[7] CHEN X, WU Z, LIU X, et al. Janus-Pro: Unified multimodal understanding and generation with data and model scaling[EB/OL]. (2025-01-29). https://arxiv.org/abs/2501.17811.

[8] HIDREAM-AI. HiDream-i1-Dev[EB/OL]. (2025-01-26). https://huggingface.co/HiDream-ai/HiDream-I1-Dev.

[9] JAIMES R. Stable diffusion statistics and user trends 2026[EB/OL].[2026-03-24]. https://www.quantumrun.com/consulting/stable-diffusion-statistics/.

[10] ZHANG C, ZHANG C, ZHANG M, et al. Midjourney statistics: Users, polls, & growth[EB/OL]. (2023-10-23). https://approachableai.com/midjourney-statistics/.

[11] QU Y, SHEN X, HE X, et al. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen, Denmark: ACM, 2023: 3403-3417.

[12] YANG Y, HUI B, YUAN H, et al. SneakyPrompt: Evaluating robustness of text-to-image generative models' safety filters [C]//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, California, USA: IEEE, 2024.

[13] ZHANG C, HU M, LI W, et al. Adversarial attacks and defenses on text-to-image diffusion models: A survey[J]. Information Fusion, 2024, 114: 102701.

[14] GANDIKOTA R, MATERZYNSKA J, FIOTTO-KAUFMAN J, et al. Erasing concepts from diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 2426-2436.

[15] KUMARI N, ZHANG B, WANG S Y, et al. Ablating concepts in text-to-image diffusion models[C]//Proceedings of ICCV. Paris, France: IEEE, 2023: 22691-22702.

[16] GANDIKOTA R, ORGAD H, BELINKOV Y, et al. Unified concept editing in diffusion models[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, Hawaii, USA: IEEE, 2024: 5111-5120.

[17] ORGAD H, KAWAR B, BELINKOV Y. Editing implicit assumptions in text-to-image diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 7053-7061.

[18] SCHRAMOWSKI P, BRACK M, DEISEROTH B, et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 22522-22531.

[19] LI H, SHEN C, TORR P, et al. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2024: 12006-12016.

[20] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of Advances in Neural Information Processing Systems. Montreal, Canada: Curran Associates, Inc., 2014: 2672-2680.

[21] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, Califonia, USA: IEEE, 2019: 4401-4410.

[22] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]//Proceedings of International Conference on Machine Learning. New York, USA: PMLR, 2016: 1060-1069.

[23] ZHANG H, XU T, LI H, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on ComputerVision. Venice, Italy: IEEE, 2017: 5907-5915.

[24] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Conditional image generation with PixelCNN decoders [C]//Proceedings of Advances in Neural Information Processing Systems 29. Barcelona, Spain: Curran Associates, Inc., 2016: 4790-4798.

[25] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//Proceedings of International Conference on Machine Learning.[S.l.]: PMLR, 2021: 8821-8831.

[26] SOHL-DICKSTEIN J, WEISS E, MAHAESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015: 2256-2265.

[27] Stability AI. Stable-diffusion-xl[EB/OL]. (2024-10-10). https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0.

[28] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Proceedings of Advances in Neural Information Processing Systems 34. Virtual: Curran Associates, Inc., 2021: 8780-8794.

[29] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of CVPR. New Orleans, Louisiana, USA: IEEE, 2022: 10674-10685.

[30] NICHOL A, DHARIWAL P, RAMESH A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models[EB/OL]. (2021-12-20). https://arxiv.org/abs/2112.10741.

[31] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. (2022-04-13). https://arxiv.org/abs/2204.06125.

[32] ESSER P, ROMBACH R, OMMER B. Taming Transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2021: 12873-12883.

[33] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of International Conference on Machine Learning.[S.l.]: PMLR, 2021: 8748-8763.

[34] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[35] HO J, SALIMANS T. Classifier-free diffusion guidance[EB/OL]. (2022-07-26). https://arxiv.org/abs/2207.12598.

[36] YANG Y, GAO R, WANG X, et al. MMA-Diffusion: Multimodal attack on diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2024: 7737-7746.

[37] ZHANG C, WANG L, LIU A. Revealing vulnerabilities in stable diffusion via targeted attacks[EB/OL]. (2024-01-16). https://arxiv.org/abs/2401.08725.

[38] CHIN Z Y, JIANG C M, HUANG C C, et al. Prompting debugging: Red-teaming text-to-image diffusion models by finding problematic prompts[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024.

[39] DONG Y, MENG X, YU N, et al. Fuzz-testing meets LLM-based agents: An automated and efficient framework for jailbreaking text-to-image generation models[C]//Proceedings of 2025 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2025: 373-391.

[40] DANG P, HU X, LI D, et al. DiffZOO: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization[C]//Proceedings of Findings of the Association for Computational Linguistics, NAACL 2025.

Albuquerque, New Mexico: Association for Computational Linguistics, 2025: 17-31.

[41] ZHANG C, WANG L, MA Y, et al. Reason2Attack: Jailbreaking text-to-image models via LLM reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Singapore: AAAI Press, 2026: 36030-36038.

[42] YANG F, ZHANG C, WANG L. Culture-based adversarial attack on text-to-image models[C]//Proceedings of IEEE International Conference on Multimedia and Expo. Nantes, France: IEEE, 2025.

[43] MEHRABI N, GOYAL P, DUPUY C, et al. FLIRT: Feedback loop in-context red teaming[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics, 2024: 703-718.

[44] HUANG Y, LIANG L, LI T, et al. Perception-guided jailbreak against text-to-image models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania, USA: AAAI Press, 2025: 26238-26247.

[45] BA Z, ZHONG J, LEI J, et al. SurrogatePrompt: Bypassing the safety filter of text-to-image models via substitution[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, Utah, USA: ACM, 2024: 1166-1180.

[46] ZHANG C, MA Y, WANG L, et al. Metaphor-based jailbreaking attacks on text-to-image models[EB/OL]. (2025-03-23). https://arxiv.org/abs/2503.17987.

[47] PENG D, KE Q, HUANG M H, et al. Unified prompt attack against text-to-image generation models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4816-4834.

[48] LI G, CHEN K, ZHANG S, et al. ART: Automatic red-teaming for text-to-image models to protect benign users[C]//Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2024: 91184-91219.

[49] WANG Z, ZHENG X, WANG X, et al. GenBreak: Red teaming text-to-image generators using large language models[EB/OL]. (2025-06-11). https://arxiv.org/abs/2506.10047.

[50] SHAO Z, WANG P, ZHU Q, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models [EB/OL]. (2024-02-05). https://arxiv.org/abs/2402.03300.

[51] TSAI Y L, HSU C Y, XIE C, et al. Ring-a-bell! how reliable are concept removal methods for diffusion models[C]//Proceedings of International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.

[52] WU Y, ZHOU S, YANG M, et al. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, PA, USA: AAAI Press, 2025: 8496-8504.

[53] GAO S, JIA X, HUANG Y, et al. RT-Attack: Jailbreaking text-to-image models via random token[EB/OL]. (2024-08-27). https://arxiv.org/abs/2408.13896v2.

[54] MAUS N, CHAO P, WONG E, et al. Black box adversarial prompting for foundation models[EB/OL].(2023-06-09). https://arxiv.org/abs/2302.04237.

[55] MA J, CAO A, XIAO Z, et al. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models[C]//Proceedings of Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico: Association for Computational Linguistics, 2025: 3141-3157.

[56] GAO S, JIA X, HUANG Y, et al. HTS-Attack: Heuristic token search for jailbreaking text-to-image models[EB/OL]. (2024-02-19). https://arxiv.org/abs/2402.12100.

[57] LYU X, LIU Y, LI Y, et al. PLA: Prompt learning attack against text-to-image generative models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, HI, USA: IEEE, 2025: 16851-16860.

[58] LIU Y, YANG G, DENG G, et al. Groot: Adversarial testing for generative text-to-image models with tree-based semantic transformation[EB/OL]. (2024-12-15). https://arxiv.org/abs/2408.13896v3.

[59] DENG Y, CHEN H. Divide-and-conquer attack: Harnessing the power of LLM to bypass the censorship of text-to-image generation model[EB/OL]. (2024-11-23). https://arxiv.org/abs/2312.07130.

[60] CAO Y, MIAO Y, GAO X S, et al. Red-teaming text-to-image systems by rule-based preference modeling[C]//Proceedings of Advances in Neural Information Processing Systems. San Diego, CA, USA: Curran Associates, Inc., 2025.

[61] CHEN R, GUO H, WANG L, et al. TRCE: Towards reliable malicious concept erasure in text-to-image diffusion models [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, HI, USA: IEEE, 2025.

[62] ZHANG E, WANG K, XU X, et al. Forget-Me-Not: Learning to forget in text-to-image diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, WA, USA: IEEE, 2024: 1755-1764.

[63] LI X, YANG Y, DENG J, et al. SafeGen: Mitigating sexually explicit content generation in text-to-image models[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, Utah, USA: ACM, 2024: 4807-4821.

[64] POPPI S, POPPI T, COCCHI F, et al. Safe-CLIP: Removing NSFW concepts from vision-and-language models[C]//Proceedings of the European Conference on Computer Vision. Milan, Italy: Springer, 2024: 340-356.

[65] ZHANG Y, CHEN X, JIA J, et al. Defensive unlearning with adversarial training for robust concept erasure in diffusion models [C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver, Canada: Curran Associates, Inc., 2024: 36748-36776.

[66] LEE B H, LIM S, LEE S, et al. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate[C]//Proceedings of International Conference on Learning Representations. Singapore: OpenReview.net, 2025.

[67] HENG A, SOH H. Selective amnesia: A continual learning approach to forgetting in deep generative models[C]//Proceedings of Advances in Neural Information Processing Systems. New Orleans, LA, USA: Curran Associates, Inc., 2023: 6.

[68] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526.

[69] GANDIKOTA R, ORGAD H, BELINKOV Y, et al. Unified concept editing in diffusion models[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, Hawaii, USA: IEEE, 2024: 5111-5120.

[70] LU S, WANG Z, LI L, et al. MACE: Mass concept erasure in diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2024: 6430-6440.

[71] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C]//Proceedings of International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2022.

[72] CHAVHAN R, LI D, HOSPEDALES T M. ConceptPrune: Concept editing in diffusion models via skilled neuron pruning [C]//Proceedings of International Conference on Learning Representations. Singapore: OpenReview.net, 2025.

[73] SCHRAMOWSKI P, BRACK M, DEISEROTH B, et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 22522-22531.

[74] KIM S, JUNG S, KIM B, et al. Towards safe self-distillation of internet-scale text-to-image diffusion models[C]//Proceedings of ICML 2023 Workshop on Challenges in Deployable Generative AI. Honolulu, HI, USA: ICML, 2023.

[75] YOON J, YU S, PATIL V, et al. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation [C]//Proceedings of International Conference on Learning Representations. Singapore: OpenReview.net, 2025.

[76] WANG Y, LI O, MU T, et al. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, Tennessee, USA: IEEE, 2025: 28759-28768.

[77] QI P, TANG K, ZHOU W, et al. SafeGuider: Robust and practical content safety control for text-to-image models[C]//Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. Taipei, China: ACM, 2025: 2818- 2832.

[78] ZHANG L, XIE Y, FU Y, et al. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, Tennessee, USA: IEEE, 2025: 8172-8181.

[79] KIM C, MIN K, YANG Y. Race: Robust adversarial concept erasure for secure text-to-image diffusion model[C]//Proceedings of the European Conference on Computer Vision. Milan, Italy: Springer, 2024: 461-478.

[80] GONG C, CHEN K, WEI Z, et al. Reliable and efficient concept erasure of text-to-image diffusion models[C]//Proceedings

of European Conference on Computer Vision. Milan, Italy: Springer, 2024: 73-88.

[81] BISWAS S D, ROY A, ROY K. CURE: Concept unlearning via orthogonal representation editing in diffusion models[C]//Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems. San Diego, California, USA: Curran Associates, Inc., 2025.

[82] Lexica. Lexica[EB/OL]. (2025-02-10). https://lexica.art/.

[83] MIAO Y, ZHU Y, YU L, et al. T2VSafetyBench: Evaluating the safety of text-to-video generative models[C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc., 2024: 63858-63872.

[84] ZHANG C, ZHANG T, WANG L, et al. T2I-RiskyPrompt: A benchmark for safety evaluation, attack, and defense on text-to-image model[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Singapore: AAAI Press, 2026: 36039-36047.

[85] WikiArt. WikiArt: Visual art encyclopedia[EB/OL]. (2011-06-06). https://www.wikiart.org/.

[86] Giphy. Celeb-detection-oss: Celebrity detection library[EB/OL]. (2019-03-04). https://github.com/Giphy /celeb-detection-oss.

[87] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA: IEEE, 2009: 248-255.

[88] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740-755.

[89] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates, Inc., 2017, 30: 6627-6638.

[90] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL]. (2023-03-15). https://arxiv. org/abs/2303.08774.

[91] DENG C, ZHU D, LI K, et al. Emerging properties in unified multimodal pretraining[EB/OL]. (2025-05-20). https://arxiv.org/abs/2505.14683.

[92] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of Advances in Neural Information Processing Systems 35. New Orleans, LA, USA: Neural Information Processing Systems Foundation, Inc., 2022: 27730-27744.

[93] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[C]//Proceedings of Advances in Neural Information Processing Systems 36. New Orleans, LA, USA: Neural Information Processing Systems Foundation, Inc., 2023: 53728-53741.

作者简介:

刘安安(1982-),**通信作者**,男,教授,研究方向:计算机视觉与机器学习,E-mail: anan0422@gmail.com。

张晨宇(1997-),男,博士研究生,研究方向:视觉大模型安全、对抗攻击与防御。

王岚君(1983-),女,研究员,研究方向:可信人工智能。

李文辉(1990-),男,副教授,研究方向:计算机视觉、跨模态学习及3D模型理解。

(编辑:刘彦东)

A Survey on Risks and Governance of Content Generated by Visual Generation Models

LIU An'an[1*], ZHANG Chenyu[2], WANG Lanjun[2], LI Wenhui[1]

(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; 2. School of New Media and Communication, Tianjin University, Tianjin 300072, China)

Abstract: With breakthroughs in deep generative technologies such as diffusion models, visual generation models have achieved significant leaps in generation quality and semantic consistency, finding extensive applications in fields like artistic creation and industrial design. However, the powerful generative capability has also triggered severe content safety risks. Malicious users can induce models to generate pornographic, violent, or copyright-infringing images, posing an urgent need for the safety governance of generative AI. This paper provides a systematic review that focuses on two core adversarial tasks of T2I models: (1) Jailbreak attacks, which aim to induce models to breach safety guardrails; (2) Concept erasure, which aims to eliminate internal risk knowledge from the models. First, we establish a taxonomy of jailbreak attacks. By analyzing them across four dimensions: Technical category, perturbation strategy, query type, and adversary knowledge, we reveal the evolutionary trend of attack methods shifting from feature-space perturbations to semantic-space reasoning. Second, regarding risk governance, this paper delves into concept erasure technologies, comparatively analyzing three mainstream technical routes: Model fine-tuning, model editing, and inference guidance. We elucidate the trade-offs among erasure effectiveness, computational efficiency, and the preservation of general generation capabilities. Finally, we summarize the commonly used benchmark datasets in this field and identify the current challenges and future directions regarding adversarial robustness and multi-concept joint governance, aiming to provide theoretical references and technical guidance for building safe and controllable T2I systems.

Highlights:

1. Systematically reviews content risks and governance methods for visual generation large models.
2. A taxonomy of jailbreak attacks is presented from four perspectives: Attack knowledge, perturbation type, query type, and technical paradigm.
3. Concept erasure methods are comprehensively summarized into three categories: Model fine-tuning, model editing, and inference-time guidance.
4. Common benchmark datasets, open challenges, and future directions for safer and more controllable generative vision systems are discussed.

Key words: visual generation models; content safety; jailbreak attack; concept erasure; model governance; diffusion models

---

Foundation items: National Natural Science Foundation of China (Nos.62425307, 62572346, U21B2024).

Received: 2026-01-10; **Revised**: 2026-02-25

***Corresponding author**, **E-mail**: anan0422@gmail.com.