多要素协同的文生图扩散模型反定制对抗样本

叶登攀[1,2],唐 龙[2*],陈思润[2],刘梓毅[2],吕芸娜[2],石绣文[2]
(1. 广州大学网络空间安全学院,广东广州 510799;2. 武汉大学国家网络安全学院,湖北武汉 430072)

摘 要: 基于文生图扩散模型的微调技术有助于实现高质量的图像定制化生成效果,但也存在隐私泄露和被用于操纵舆论的风险. 当前研究主要聚焦于构造基于提示词级别或图像级别的对抗样本来实现对生成特定人物或风格定制化图像的反制,然而却忽略了这两个不同模态层面对抗样本之间的关联性,以及模型内部功能模块之间对抗性的关联. 这些不足导致现有方法生成的对抗样本在实际场景中的反定制化性能受到限制. 为此,本文提出了双重反扩散对抗样本生成方法(Dual Anti-Diffusion,DADiff),这是一种反制扩散模型定制化的两阶段对抗样本生成框架,将提示词级别的对抗攻击融入图像级别对抗样本的生成过程中. 在第一阶段,DADiff生成提示词级别的对抗向量,以文本层面的对抗扰动信息引导后续的图像层面对抗扰动生成;第二阶段,除了对扩散UNet模型进行端到端对抗攻击外,DADiff还对其自注意力和交叉注意力模块进行干扰,旨在打破图像像素之间的相关性,并使图像利用实例提示词向量和对抗提示词向量计算得到的交叉注意力结果保持一致. 此外,DADiff还引入了局部随机时间步长梯度集成策略,通过整合多个分段时间步长的随机梯度来更新对抗扰动. 在主流人脸图像数据集和艺术风格图像数据集上的实验结果表明,与现有方法相比,DADiff在跨提示词,关键词不匹配和跨模型的反定制化任务上的平均性能提升了20%.

关键词: 文本生成图像;扩散模型;模型微调;对抗样本;模型反定制
基金项目: 国家自然科学基金(No.62472325)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)10-3730-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250596

Anti-Customization Adversarial Examples Against Text-to-Image Diffusion Models with Multi-Element Collaboration

YE Deng-pan[1,2], TANG Long[2*], CHEN Si-run[2], LIU Zi-yi[2], LÜ Yun-na[2], SHI Xiu-wen[2]
(1. *School of Cyberspace Security*, *Guangzhou University*, *Guangzhou*, *Guangdong* 510799, *China*;
2. *School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan*, *Hubei* 430072, *China*)

Abstract: Fine-tuning text-to-image diffusion models enables high-quality customized image generation, yet it also introduces risks of privacy leakage and potential misuse for opinion manipulation. Current research primarily focuses on prompt- or image-level adversarial attacks to counter model customization; however, it overlooks the inter-modal correlation between prompt- and image-level adversarial perturbations, as well as the adversarial interplay among the model's internal functional modules. This limitation restricts the practical effectiveness of existing anti-customization methods. To address this, we propose dual anti-diffusion (DADiff), a two-stage framework that integrates prompt-level adversarial attacks into the generation of image-level adversarial examples. In the first stage, DADiff generates adversarial prompt vectors to guide the subsequent image-level perturbation. In the second stage, beyond performing an end-to-end attack on the diffusion UNet, DADiff further perturbs its self-attention and cross-attention modules—aiming to break pixel-wise correlations and enforce consistency by aligning the cross-attention maps derived from the original instance prompt and those from the adversarial prompt vector. Additionally, DADiff introduces a local-random timestep gradient ensemble strategy, which updates adversarial perturbations by aggregating stochastic gradients sampled from multiple segmented timestep intervals. Experimental results on mainstream facial and artistic style datasets show that DADiff achieves an average performance improvement of 20% over existing methods across cross-prompt, keyword-mismatch, and cross-model anti-customization scenarios.

收稿日期:2025-07-07;录用日期:2025-10-24;责任编辑:覃怀银
*通讯作者:唐龙

Key words: text-to-image generation; diffusion models; model fine-tuning; adversarial examples; model anti-customization

Foundation Item(s): National Natural Science Foundation of China (No.62472325)

1 引言

扩散模型(Stable Diffusion,SD)的快速发展显著提高了图像生成质量,从而在图像恢复、多模态合成和图像定制等方面实现了根本性的创新. 研究表明[1],大型预训练模型具有较稀疏的内在特征,这意味着在低维参数空间中进行微调可以产生类似于微调整个模型的结果. 因此,通过微调技术进行模型定制化训练的技术受到广泛关注,它们可通过少量地输入图像进行微调训练来执行定制化的高质量生成任务. 用于文生图模型[2-4]的典型定制化图像生成技术,如DreamBooth[5]、低秩适应(Low-Rank Adaptation,LoRA)[6]和文本反演[7]等,只需要特定人物或风格的少量图像进行微调训练,或对隐空间与文本特征进行优化,就可以获得能够生成该人物或风格图像的定制化扩散模型,如图1中的第一行所示,其中白色的推理提示是用于进行DreamBooth训练的提示词,灰色的推理提示是与DreamBooth无关的提示词. 这些技术降低了艺术创作与定制化生成的成本,推动了人工智能生成技术的发展.

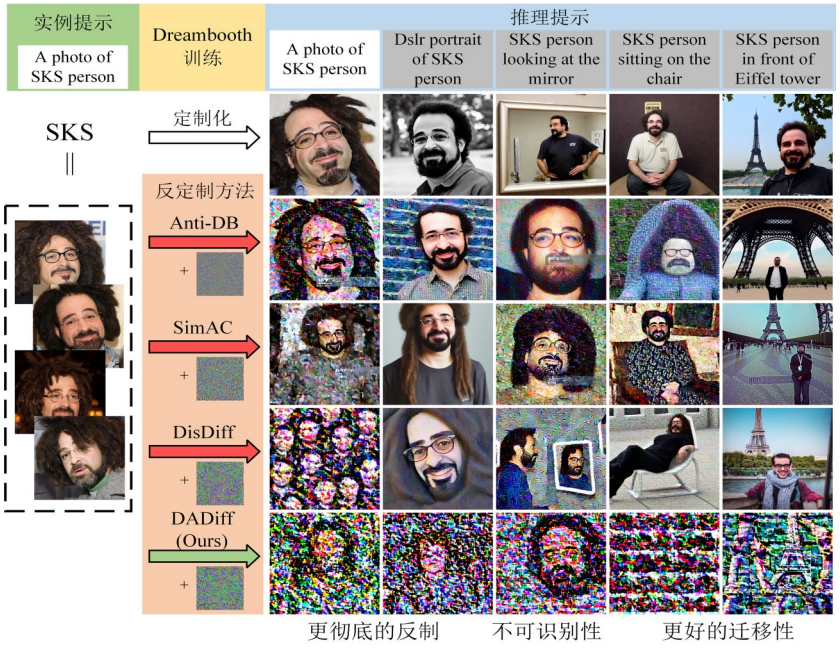


图1　DreamBooth定制化、现有反定制方法和提出DADiff反定制方法的可视化比较

然而,恶意用户可以通过获取目标用户的个人图像并使用文生图模型定制化微调技术来生成大量虚假图像,从而导致隐私泄露. 这些虚假图像在成人产业、诈骗和政治宣传等领域被利用牟利,可能对目标用户造成巨大的经济损失和名誉伤害. 与依赖生成对抗网络进行图像编辑或人脸替换的传统深度伪造方法不同,基于微调扩散模型的深度伪造利用目标用户的图像进行模型训练,这允许恶意用户通过自定义的关键词指代目标用户,并使用不同的提示词生成各种虚假图像. 此外,扩散模型的图像生成机制不同于基于生成对抗网络的伪造方法,这使得针对传统深度伪造的主动防御方法难以直接迁移应用. 为了解决这一问题,防止图像被滥用于定制化微调扩散模型,迫切需要研究保护措施.

当前主流的扩散模型反定制方法是利用对抗样本将不可察觉的扰动引入目标用户的受保护图像中. 恶意用户用受保护的图像进行扩散模型定制化微调,只能生成质量很低的虚假图像. 然而,现有的相关研究要么针对扩散模型的非核心模块,要么只使用单个随机时间步长迭代攻击扩散UNet模型. 此外,现有方法通常侧重于增强图像层面扰动的对抗性,忽略了文本层面的对抗性对图像对抗样本的潜在影响. 这些限制导致在处理涉及跨关键字或跨提示词等实际环境时,现有方法难以达到足够满意的效果. 扩散模型微调训练的输入包括三个关键要素:图像、提示词和时间步长. 本文的核心思想认为,充分利用扩散模型的输入信息

可以显著提高基于对抗样本的反定制技术的有效性.

本文提出了一种针对扩散模型的反定制两阶段对抗攻击方法,名为双重扩散反制(Dual Anti-Diffusion,DADiff). DADiff充分利用扩散模型训练过程中的各项输入,以及扩散UNet模型的内部功能模块,旨在利用多要素协同提高图像对抗扰动对扩散反定制的有效性和可扩展性. 具体来说,DADiff首先在提示词层面进行对抗攻击,生成对抗提示词向量,为后续的图像层面对抗攻击提供跨模态对抗先验知识. 然后,在图像层面的对抗攻击中,除了考虑UNet模型的输出端损失外,DADiff还为模型内部的自注意力和交叉注意力模块[8]构建损失函数,以破坏模型对图像像素之间注意关系的处理能力,并确保由正常提示词和对抗提示词向量引导的交叉注意力之间保持对齐. 此外,现有方法忽略了不同时间步长下的梯度相关性,因此DADiff提出了局部随机时间步长梯度集成策略(Local Random Timestep Gradient Ensemble,LRTGE),通过在等间隔的时间步长组内随机选择时间,并集成多步长的对抗梯度来更新对抗样本,从而更有效地捕获时间维度上的梯度相关性,并进一步增强对抗样本的反定制能力.

2 相关工作

2.1 扩散模型与定制化

扩散模型[9]利用连续马尔可夫链将高斯噪声逐步添加到图像数据中,随后反转此过程以重建图像样本. Stable Diffusion[10]将此机制应用于文本引导图像生成,在隐空间中嵌入文本特征,并结合交叉注意力模块以适应不同的用户定义提示词. 基于扩散模型的模型定制化技术已经提出,可以根据特定需求定制图像生成过程. LoRA[6]使用附加参数对扩散模型进行微调,以将图像表示与描述性提示词进行关联. 文本反演[7]利用文本编码器的灵活单词嵌入空间将新概念映射到一个或多个随机伪单词. DreamBooth[5]仅使用3~5张图像和相应的标识符即可对预训练的Stable Diffusion模型进行微调,使模型能够“记忆”并在新的上下文中再现图像对应的特定概念. DreamBooth叠加LoRA的定制化方法进一步优化了其效率,并激发了许多扩散模型定制化项目的落地应用. 然而,伦理考虑和保障措施对于防止滥用此类技术至关重要,这些技术可能会用于产生非法或有害的内容.

2.2 针对扩散模型的对抗攻击方法

对文生图扩散模型的对抗攻击可分为文本层面对抗和图像层面对抗. 在文本层面,Zhuang等人[11]为扩散模型提出了一种文本层面的对抗字符攻击. 通过最小化对抗文本和正常文本编码之间的余弦相似性,攻击者只需使用5个额外的对抗字符即可显著更改生成图像的内容. Yang等人[12]通过图像梯度增强文本模态下的对抗攻击,使对抗文本能够绕过扩散模型的敏感词安全检查,生成有害图像. 在图像层面,PhotoGuard方法[13]攻击扩散模型中包含的VAE编码器和UNet模型,并以灰度图像作为优化目标生成对抗样本. AdvDM[14]直接使用投影梯度下降(Projected Gradient Descent,PGD)方法[15]攻击UNet模型,以防止文本反演. Anti-DreamBooth(Anti-DB)[16]使用交替训练和扰动学习(Alternating Surrogate and Perturbation Learning,ASPL)的方式来模拟模型的真实训练过程,并交替执行DreamBooth训练和对抗样本攻击. Simple Anti-Customization(SimAC)[17]在Anti-DB的基础上利用贪心算法选择具有最高梯度分数的时间步长来更新对抗样本. Disrupting Diffusion(DisDiff)[18]进一步设计了交叉注意力擦除损失,以在执行Anti-DB攻击的同时擦除对关键词的交叉注意力.

然而,尽管已有研究试图在文本层面对抗攻击中利用图像梯度,但图像层面的对抗攻击尚未与文本层面的对抗攻击建立联系. 此外,现有研究已经注意到了时间步长的影响,但他们只随机使用单个时间步长在多轮迭代中生成梯度,忽略了不同噪声条件下图像中梯度的多样性. 因此,在图像层面的扩散模型反定制技术仍有很大的发展空间.

3 先验知识与公式定义

3.1 扩散定制化

扩散模型是一种生成模型,通常分为前向扩散过程(向结构化数据添加噪声)和反向去噪过程(通过去噪恢复或生成数据). 具体来说,对于一张给定的图像$x_0 \sim q(x)$,正向扩散过程根据预设的噪声调度系数$\{\beta_t: \beta_t \in (0,1)\}(t=1,2,\cdots,T)$逐渐添加噪声,生成一系列中间扩散噪声图像$\{x_1, x_2, \cdots, x_T\}$,直到图像完全转换为高斯噪声$\epsilon \sim \mathcal{N}(0, I)$. 正向扩散过程中,在$t$时刻的加噪图像$x_t$如式(1)表示,其中,$\alpha_t = 1-\beta_t$,$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. 在常见的扩散模型研究中,扩散过程通常包含$T=1\,000$步.

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon \tag{1}$$

在反向去噪过程中,加入图像x_T中的噪声被逐渐去除,直到初始图像x_0被最终还原. 在这个过程中,扩散去噪UNet模型ϵ_θ通常学习如何通过$t+1$时刻的噪声图像x_{t+1}预测在正向过程中加入到t时刻噪声图像x_t中的噪声ϵ. 而对于允许提示词输入的扩散模型来说,为了确保提示词能够有效控制扩散模型输出的多样性,在模型ϵ_θ的内部结构中加入了文本-图像的交叉注意力模块来融合图像与文本提示词信息,并最终返回预测的在正向扩散过程中加入的采样噪声. 因此,训练扩散

模型的损失函数旨在最小化模型输出与采样高斯噪声 ϵ 之间的 L_2 距离,模型的输入包括 $t+1$ 时刻的噪声图像 x_{t+1}、条件提示词 P 和当前的时间步长 t. 当无输入条件时,P 为空:

$$L_{\text{cond}}(\theta,x_0,P)=E_{x_0,t,\epsilon}\left\|\epsilon-\epsilon_\theta(x_{t+1},t,P)\right\|_2^2 \tag{2}$$

在实际撰写公式的过程中,时间步长 t 通常在损失函数中被省略,因为它是一个离散的随机输入,它决定了在扩散过程中将有多少噪声添加到图像中.

以下以 DreamBooth 为例,形式化地介绍基于 DreamBooth 的扩散模型定制化技术. DreamBooth 是一种通过使用少量自定义主题或对象的参考图像对预训练的扩散模型进行个性化微调的技术,可以生成具有特定主题特征的新图像. 假设有一组人物[A]的照片,为了使模型能够生成与[A]相关的自定义图像,可以利用实例提示词 P_{new}="一张[A]的照片"和基础提示词 P_{base}="一张人的照片"的组合来设计损失函数,以微调扩散模型:

$$\begin{aligned}L_{\text{DB}}(\theta,x_0)=&E_{x_0,t,\epsilon}\left\|\epsilon-\epsilon_\theta(x_{t+1},t,P_{\text{base}})\right\|_2^2\\&+\lambda\left\|\epsilon'-\epsilon_\theta(x'_{t'+1},t',P_{\text{new}})\right\|_2^2\end{aligned} \tag{3}$$

其中,ϵ 和 ϵ' 是高斯噪声;t 和 t' 是两段不同的随机时间步长;x_{t+1} 是人物[A]的噪声图像;$x'_{t'+1}$ 是通过预训练的 ϵ_θ 生成的"一张人的照片"的图像在 $t'+1$ 时间下的效果.[A]可以是任何自定义的关键词,例如"sks, asdf"等等. DreamBooth 的更多技术细节可参阅文献[5].

3.2　反定制对抗攻击

下面介绍生成对抗样本以进行反定制的基本方法. 给定一组人物[A]的图像 X,将不可察觉的扰动 δ 添加到 X 中,基础的反定制优化目标是最大化扩散 UNet 模型的输出与高斯噪声 ϵ 之间的 L_2 距离,使恶意用户无法利用 X 进行 DreamBooth 训练:

$$\delta\leftarrow\arg\max_\delta L_{\text{cond}}(\theta,x_0+\delta,P_{\text{new}}) \tag{4}$$

其中,$x_0\in x$,$L_{\text{cond}}(\cdot)$ 为式(2)表示的扩散模型 L_2 损失,P_{new} 是新的实例提示. 在确定损失函数后,通常用 PGD 攻击方法[15]更新对抗样本,这些对抗样本通过多个符号化的梯度迭代更新扰动:

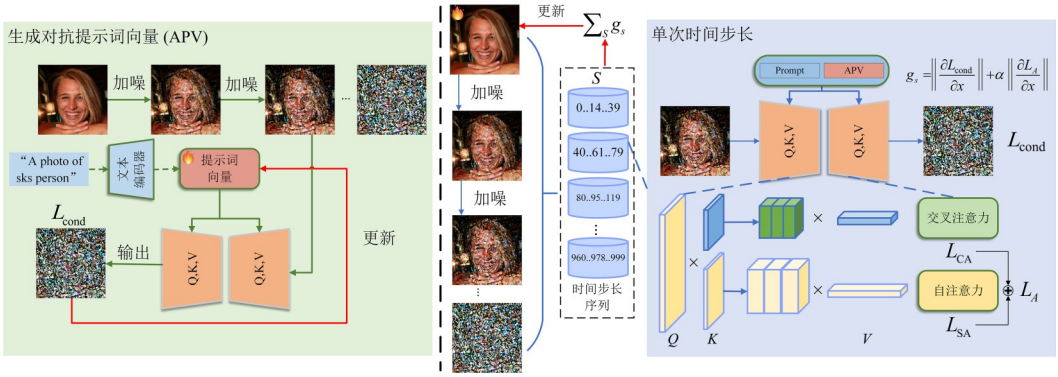
$$\delta_{\text{adv}}^{r+1}=\text{clip}_{\delta,\omega}\left(\delta_{\text{adv}}^r+\eta\,\text{sign}\left(\nabla_\delta L\left(\theta,x_0+\delta_{\text{adv}}^r,P_{\text{new}}\right)\right)\right) \tag{5}$$

其中,$r\in R$ 是 PGD 攻击的步数;η 是学习率;ω 是扰动约束,通常使用 L_∞ 范数进行衡量.

4　基于双重反定制的扩散模型对抗样本生成方法

4.1　整体框架

本节展示了 DADiff 实现的整体框架. 如图 2(a)所示,DADiff 首先使用式(2)计算的对抗损失迭代更新文本层面对抗提示词向量(Adversarial Prompt Vector, APV). 随后,APV 被用于执行图 2(b)中的图像层面对抗攻击. 在图像层面对抗攻击过程中,DADiff 同时关注注意力模块的输出和模型的端到端输出. 对于某个时间步长的梯度计算,DADiff 采用两重策略:一方面,要求 UNet 模型的输出远离叠加的高斯噪声;另一方面,对于自注意模块,DADiff 最大化对抗样本和干净图像的自注意输出差异;对于交叉注意力模块,DADiff 要求基本提示和 APV 计算出的交叉注意力输出尽可能一致. 在得到各项损失后,DADiff 分别计算输出损失和注意力损失的梯度. 然后,DADiff 采用局部随机时间步长梯度集成策略 LRTGE,将整个时间步长序列分为等长的若干组,并从每组中随机选择一个时间步长进行上述梯度计算. 最后,将所有时间步长组的梯度叠加为总梯度,用于在每次迭代中更新对抗样本,并执行类似 Anti-DB 的交替训练和攻击策略.



(a) 对抗提示词向量生成流程　　(b) 图像层面对抗样本生成流程

注:每一环节仅更新带有火焰图标的部分.DADiff 首先执行阶段(a)以获取对抗提示向量 APV,然后使用 APV 和实例提示在阶段(b)中生成图像层面的对抗样本.

图 2　DADiff 的总体流程

4.2 对抗提示词向量

首先，从实例提示的CLIP编码特征初始化，即$P_{\text{adv}}=P_{\text{new}}$，DADiff根据式(2)通过最大化UNet输出和高斯噪声之间的损失来迭代优化提示词向量，以获得APV. 因此，可以将式(4)修改为式(6)：

$$\begin{aligned}P_{\text{adv}} &\leftarrow \arg\max_{P_{\text{adv}}} L(\theta, x_0, P_{\text{adv}}) \\ &\leftarrow \max_{P_{\text{adv}}} E_{x_0,t,\epsilon}\left\| \epsilon-\epsilon_\theta(x_{t+1},t,P_{\text{adv}})\right\|_2^2\end{aligned} \tag{6}$$

这种方法的灵感来自空提示词反演技术[19]. 在这项工作中，作者更新提示词向量以提高扩散模型反演的生成质量. 由于在此过程中APV不需要扰动上限约束，为了缓解不同时间步长下梯度的数值差距，DADiff借鉴MI-FGSM对抗攻击[20]，对梯度进行归一化，并使用式(7)更新对抗提示词向量P_{adv}，其中$\|g\|=\dfrac{g}{\|g\|_1}$：

$$P_{\text{adv}}^{r'+1}=P_{\text{adv}}^{r'}+\eta\left\|\nabla_{P_{\text{adv}}} L(\theta,x_0,P_{\text{adv}}^{r'})\right\| \tag{7}$$

在攻击过程中，随机选择$t\in\{0,1,\cdots,999\}$进行$r'\in\mathbf{R}'$次迭代的攻击. 之后，文本层面的对抗攻击完成，并为后续对抗攻击提供了P_{adv}.

为了证明APV具有的对抗攻击性能，图3分别使用随机噪声和去噪扩散隐式模型(Denoising Diffusion Implicit Model，DDIM)图像反演[21,22]的噪声作为起点，结合不同迭代轮的原始提示和APV生成图像并展示. 初始提示是“一张女人的照片”. 第一行显示从随机高斯噪声生成的图像，第二行显示从对初始图像进行DDIM反转后的噪声开始生成的图像. APV分别通过10、50、100和500轮迭代生成. 可以看到，只需要10轮迭代攻击，扩散模型就无法生成具有初始意义的图像. 随着迭代次数的增加，生成的图像包含的原始图像的原始结构和纹理信息越来越少，最终在500次迭代中形成稳定的对抗性生成效果. 这证明了应用于条件向量的攻击确实具有显著的对抗效果，可以通过利用对抗文本来进一步增强对抗图像的攻击性能.

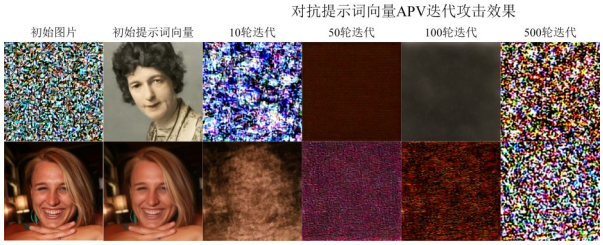


图3 由原始提示词和APV引导的生成图像

4.3 攻击扩散模型注意力模块

在图像层面对抗攻击中，早期研究已证明，对抗样本可有效使UNet模型无法预测扩散过程中前一个时间步添加的高斯噪声. 为了将APV的效果与图像层面对抗样本相结合，并进一步破坏图像中的像素级关联性，DADiff提出重点破坏UNet模型的注意力模块的思路，并分别针对扩散模型的自注意力模块和交叉注意力模块设计对抗攻击策略.

4.3.1 自注意力模块对抗攻击

自注意力模块计算图像像素与像素之间特征的相关性，这可以帮助UNet捕捉图像像素位置之间的全局依赖关系，从而提高生成图像的准确性. 站在对抗攻击的角度，为了降低图像生成的质量，DADiff希望UNet无法计算对抗样本的像素相关性. 假设UNet模型总共包含S个自注意力模块，则第s个自注意力模块的输出为$f_s(\cdot)$. 自注意力损失可按式(8)计算：

$$L_{\text{SA}}=\sum_{s=1}^{S} J\left(f_s(x_0+\delta_{\text{adv}}),f_s(x_0)\right) \tag{8}$$

由于自注意力损失计算不需要输入提示词向量，并且时间步长相等，此处省略了式(8)中的提示词和时间步长输入，并对原始图像和对抗样本使用相同的实例提示. DADiff使用均方误差作为$J(\cdot)$来衡量并最大化两个结果之间的距离.

4.3.2 交叉注意力模块对抗攻击

交叉注意力模块计算像素特征和提示词向量之间计算的注意力权重，反映图像像素和提示词之间的关联程度. DADiff计算了对抗样本和实例提示的交叉注意力输出，以及干净图像和APV的交叉注意力输出，并最小化它们之间的距离，以实现APV的对抗性先验与图像层面对抗样本的对齐. 假设UNet模型总共包含C个交叉注意力模块，则第c个交叉注意模块的输出为$f_c(\cdot)$. 交叉注意力损失可按式(9)计算. 此处同样省略了时间步长输入，并选择余弦相似性距离为$J(\cdot)$来衡量和最大化两个结果之间的相似性.

$$L_{\text{CA}}=\sum_{c=1}^{C} J\left(f_c(x_0+\delta_{\text{adv}},P_{\text{new}}),f_c(x_0,P_{\text{adv}})\right) \tag{9}$$

在获得L_{SA}和L_{CA}后，DADiff使用超参数α_1来平衡这两个损失，并通过式(10)获得注意力损失L_A.

$$L_A=\alpha_1 L_{\text{CA}}+(1-\alpha_1)L_{\text{SA}} \tag{10}$$

4.4 局部随机时间步梯度集成

先前的研究表明[23,24]，时间步长会影响攻击扩散模型的有效性. SimAC为攻击选择梯度得分最高的时间步，而DisDiff根据不同时间步的混合质量得分[25]分配学习率. 然而在PGD的单次迭代中，它们只计算单个时间步的梯度. 随机选择时间步长会导致损失和梯度得分的波动，从而削弱优化效果. 尽管不同时间步长的梯度重要性不同，但它们都包含有利于优化对抗样本的信息. 梯度分数随着时间步长的增加而降低，相邻时间步长的梯度之间存在相关性. 在扩散过程中在多个时间步长添加噪声类似于图像分类中的噪声增强.

整合不同时间步长的梯度可以产生更鲁棒和稳定的对抗梯度,从而改善攻击效果.

因此,DADiff结合梯度集成和随机梯度采样的优点,提出了局部随机时间步梯度集成策略LRTGE. LRTGE将扩散模型的时间步长划分为B个相等的段,并在每个段内随机选择一个时间步长来计算梯度. 然后将每个分段的梯度相加,并用作单个PGD攻击的更新梯度. 式(11)将该过程进行形式化,其中$\|g\| = \frac{g}{\|g\|_1}$. DADiff分别用$L_{\text{cond}}$和$L_A$计算图像的对抗攻击梯度,因为分析表明两者获得的梯度大小存在显著差异. 由于L_{cond}是核心优化函数,DADiff采用以下策略:首先分别对两个梯度进行正则化,然后使用超参数α_2对它们求和.

$$g^r = \left\| \sum_{b=1}^{B} \nabla_\delta L_{\text{cond}} \right\| + \alpha_2 \left\| \sum_{b=1}^{B} \nabla_\delta L_A \right\| \tag{11}$$

之后,DADiff使用集成梯度g^r在每次PGD攻击[15]迭代中更新图像层面的对抗样本,如式(12)所示:

$$\delta_{\text{adv}}^{r+1} = \text{clip}_{\delta,\omega}\left(\delta_{\text{adv}}^{r} + \eta \text{sign}(g^r)\right) \tag{12}$$

图4分别使用随机单时间步长和LRTGE实现DADiff,并在梯度正则化之前利用SimAC提到的梯度分数计算方法来观察优化过程,其中随机单时间步长用蓝线表示,LRTGE用红线表示. 第一行的两张图展示了执行攻击时损失函数值变化情况,第二行的两张图展示了执行攻击后计算求得的梯度分数变化情况. PGD攻击以ASPL方式[16]迭代共计300次,其中包含50次ASPL迭代,每次包含6轮PGD迭代. 第二行的结果表明,注意力损失的梯度分数大于UNet损失,因此需要对两个梯度进行归一化后再叠加. 图中结果还表明,当使用随机单时间步长计算梯度时,损失表现出明显的不稳定性,梯度分数波动大且不规则,这不利于对抗样本的优化. 相比之下,当使用集成的多时间步长梯度进行优化时,损失显示出稳定的上升趋势,梯度分数总体保持稳定,这更有利于优化对抗样本.

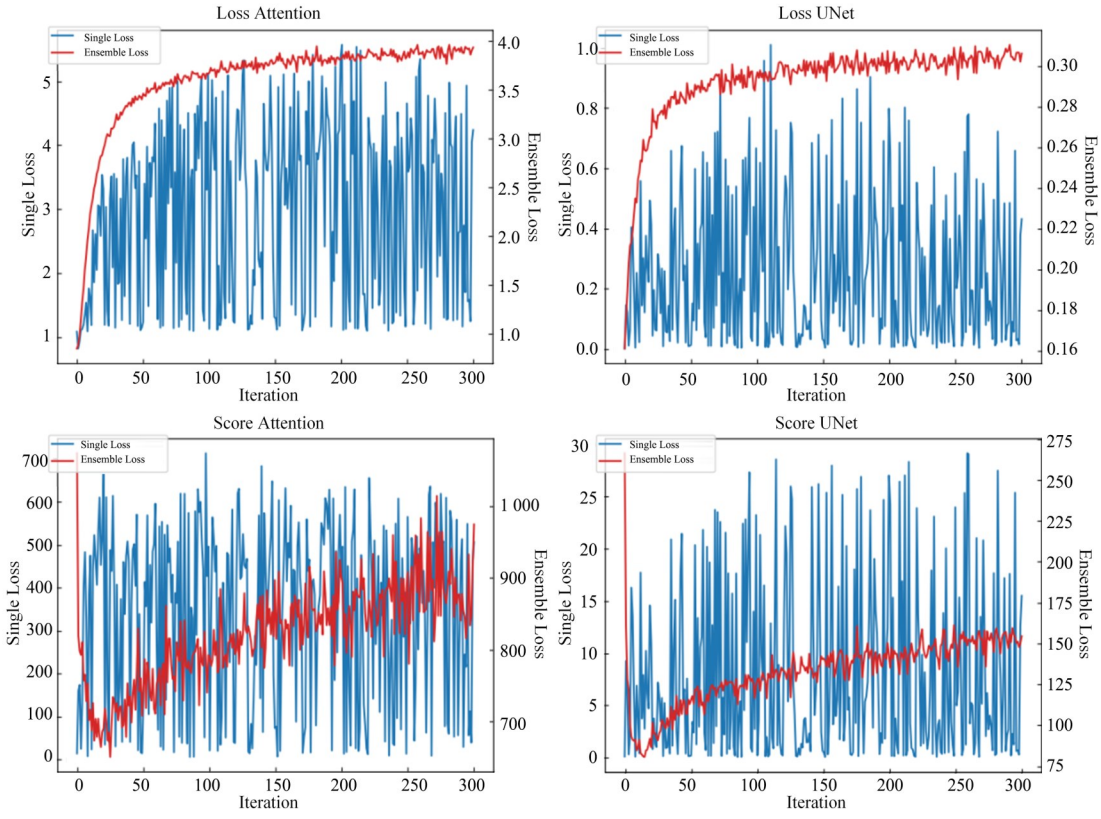


图4　使用随机单时间步长和LRTGE执行PGD攻击时的损失和梯度分数对比

5　实验结果与分析

5.1　实验设置

5.1.1　数据集与对比方法

本节在两个广泛使用的人脸数据集上选取人物图像进行扩散模型定制化,以评估DADiff:CelebA-HQ数据集[26]和VGGFace2数据集[27]. 在每个数据集中随机选择16个身份,并为每个身份准备8张不同的图像. 图像被平均分为干净集和扰动集. 所有图像都被调整为512×512的大小. 本节实验将DADiff与最相关的Ad-

vDM、Anti-DB、SimAC 和 DisDiff 方法进行了比较，实验均在单张 NVIDIA A800 GPU 上进行.

5.1.2 评价指标

为评估 DADiff 的有效性，本节采用了一系列指标评估定制化微调训练失败的 DreamBooth 模型产生的图像的质量. 首先，使用 Retina 人脸检测器[28]评估人脸检测失败率（Face Detection Failure Rate，FDFR）判定扩散模型是否生成了人脸；然后，假如成功检测到了人脸，进一步利用 ArcFace 人脸识别模型[29]来判定生成图像身份向量与用户整个干净图像数据集的平均身份向量的余弦距离，这一过程称为身份评分匹配（Identity Score Matching，ISM）. 此外，还使用基于随机嵌入鲁棒性的人脸图像质量评估（Face Image Quality based on Stochastic Embedding Robustness，SER-FIQ）[30]、无参考图像空间质量评估器（Blind/Referenceless Image Spatial QUality Evaluator，BRISQUE）[31]和弗雷歇初始距离（Fréchet Inception Distance，FID）[32]三个指标评估生成图像视觉质量. SER-FIQ 通过计算人脸识别模型生成的特征嵌入置信度来评估质量，分数越低说明生成图像质量越低；BRISQUE 是一种基于自然场景统计的无参考图像质量评估指标，通过分析图像空间域特征偏离自然统计规律的程度来量化失真，数值越高表示图像质量越差；FID 是利用神经网络中间层特征评估图像相似性的指标，数值越高表示图像失真越大. 在后续的量化实验中，以加粗强调最佳的实验结果.

5.1.3 使用模型与参数设置

本节使用 HuggingFace[33]推出的三个版本的预训练开源扩散模型：SD-v1.4、SD-v1.5 和 SD-v2.1 进行实验. 对于 DreamBooth 的训练，单个批次输入的图像数量设置为 2，DreamBooth 训练迭代次数设置为 1 000，学习率设置为 5×10^{-7}. 除非额外声明，否则都使用 SD-v1.5 模型作为生成对抗样本的代理模型，且所有的对抗攻击，包括 APV 生成，都以“一张 sks 的照片”作为生成对抗样本的实例提示词. DADiff 中对抗样本的学习率 η 都设置为 0.005. 在 APV 生成中，迭代次数 $R'=500$. APV 不需要受到约束，但图像层面对抗样本都被约束在 $\omega=0.05$. 在图像层面对抗攻击中，DADiff 对 UNet 模型上采样模块中的所有自注意力和交叉注意力层计算损失. 在 LRTGE 中，将总时间步长 $T=1\,000$ 等分为 $B=25$ 段，以计算集成梯度. 由于在扩散模型反制中，使 UNet 输出尽可能远离是首要目标，且希望尽量均衡地破坏自注意力与交叉注意力模块，因此设置 $\alpha_1=0.5$ 和 $\alpha_2=0.4$. DADiff 遵循与 Anti-DB 中提到的 ASPL 攻击方法相同的设置，总共执行 300 轮攻击，其中包含 50 轮的“模型训练—对抗样本生成”迭代和 6 轮的 PGD 对抗攻击.

5.2 实验结果评估

5.2.1 跨提示词量化结果对比

表 1 展示了 DADiff 方法在 SD-v1.5 上与主流方法的量化对比结果. 在白盒实例提示“一张 sks 的照片”上，与基线 AdvDM 和 Anti-DB 相比，更先进的 SimAC 和 DisDiff 方法产生了更好的效果. SimAC 在 CelebA-HQ 数据集上产生了具有竞争力的 BRISQUE 值，但在 VGGFace2 数据集上的性能明显下降. DADiff 在两个数据集上都对生成图像质量造成了更大的破坏，不仅导致模型难以生成可识别的人脸（更高的 FDFR 和更低的 ISM），而且还会更加恶化生成图像的背景纹理等，进一步扰乱了其视觉质量（更低的 SER-FIQ、更高的 BRISQUE 和 FID）.

而在黑盒推理提示“一张 sks 的单反肖像照”上，DADiff 在两个数据集上都显示出显著的优势，无法检测到人脸的图像比例增加了 32% 以上，生成的图像质量大幅下降了 30% 以上，这说明 DADiff 同时攻击文本提示词、注意力模块和不同时间步长的策略对提高对抗样本的黑盒反定制效果带来了显著提升，表明 DADiff 更彻底地破坏了关键词“sks”和目标人物图像之间的相关性.

5.2.2 跨提示词可视化效果展示

图 1 显示了不同提示词下主要对比方法和 DADiff 的反定制生成质量. 可以看出，DADiff 对生成质量造成了更彻底的破坏，使生成的图像更难识别，并在不同的黑盒推理提示之间表现出更好的迁移性. 此外，图 5 和图 6 利用 VGGFace2 和 CelebA-HQ 数据集中的其他 ID 进一步显示了更多可视化实验结果对比. 可以看出，DADiff 在不同数据集和推理提示词的不同 ID 图像上均取得了更好的反定制生成效果. 现有的方法在“一张 sks 的单反肖像”和“一张 sks 在埃菲尔铁塔前的照片”的黑盒推理提示词中的生成质量没有明显破坏，但 DADiff 仍然取得了显著的结果，不仅难以识别生成图像中的人脸，而且严重破坏了背景语义的纹理.

为了进一步分析不同对抗样本对 UNet 模型内部模块的影响，图 7 显示了生成的图像在 UNet 模型上的交叉注意力和自注意力显著图. 第一行显示了从未加噪图像进行 DreamBooth 训练模型生成图像的显著图，其余的是用对抗样本进行 DreamBooth 训练模型上生成图像的显著图. 选择 16×16 大小的交叉注意力和自注意力模块的输出来绘制显著图，并在时间步长 $t=500$ 时进行展示. 自注意力显著图是使用所有像素点之间的平均注意力绘制的，从总体上反映了图像的高热度区域. 从图中可看出，用 DADiff 生成的对抗样本训练的 DreamBooth 模型很难在提示关键词和图像之间建立关联，也无法有效地捕捉像素的全局相关性. 对自注意力模块的攻击破坏了模型学习图像像素之间相关性的能

表1　跨提示词量化结果评估

Dataset	Method	"a photo of sks person"					"a dslr portrait of sks person"				
		FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
VGGFace2	w/o Protect	0.00	0.72	0.77	10.96	185.61	0.00	0.53	0.65	1.27	218.03
	AdvDM	0.33	0.17	0.46	15.33	341.05	0.00	0.23	0.54	15.15	274.12
	Anti-DB	0.78	0.10	0.19	40.06	373.48	0.10	0.22	0.46	21.18	329.76
	SimAC	0.72	0.15	0.17	34.88	399.45	0.22	0.38	0.51	20.46	266.95
	DisDiff	0.97	0.02	0.01	50.45	427.12	0.20	0.30	0.55	20.92	281.64
	DADiff(Ours)	**0.99**	**0.00**	**0.00**	**53.50**	**526.90**	**0.91**	**0.04**	**0.08**	**43.79**	**462.63**
CelebA-HQ	w/o Protect	0.00	0.77	0.84	27.35	139.33	0.07	0.49	0.79	2.79	226.92
	AdvDM	0.67	0.06	0.53	14.41	294.45	0.10	0.06	0.68	16.40	258.36
	Anti-DB	0.77	0.04	0.17	54.97	336.57	0.17	0.12	0.70	17.08	291.99
	SimAC	0.97	0.03	0.03	**60.15**	478.24	0.47	0.16	0.55	19.50	228.09
	DisDiff	**0.98**	0.03	0.01	58.15	471.06	0.56	0.12	0.54	20.92	356.36
	DADiff(Ours)	**0.98**	**0.01**	**0.00**	59.77	**479.66**	**0.88**	**0.03**	**0.15**	**36.41**	**383.60**

注:加粗结果为最佳结果.

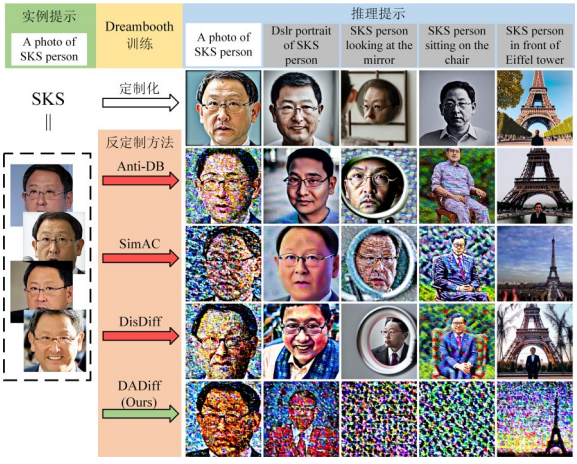


图5　VGGFace2数据集ID图像的扩展可视化对比

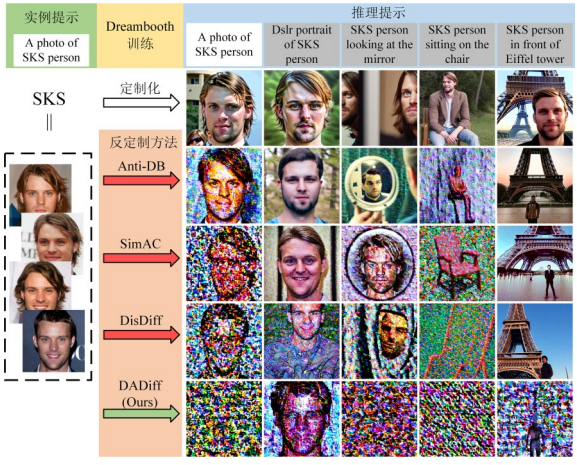


图6　CelebA-HQ数据集ID图像的扩展可视化对比

力,而交叉注意力模块中APV的引导也加剧了对模型学习提示词和目标人图像之间相关性的破坏.这反映了对自注意力和交叉注意力模块的攻击的有效性.

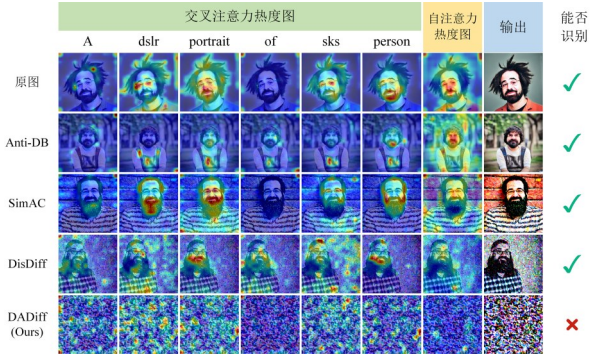


图7　从常规DreamBooth训练和不同反定制方法生成图像的注意力显著图

5.2.3　关键词失配实验效果对比

实际情况下,添加对抗噪声的保护方通常难以预测恶意用户将使用哪些关键词[*A*]来指代目标用户.为此,表2评估了在DreamBooth训练中使用失配的关键词后模型的生成质量.具体而言,选择"sks"作为生成对抗样本的关键词,并使用"asdf"作为训练和评估DreamBooth生成质量时的关键词.结果表明,DADiff在关键词失配的情况下仍然表现出了更好的性能.在黑盒推理提示中,DADiff仍然优于现有方法.这说明对自注意力模块的干扰可能削弱了对抗样本对特定提示词的依赖,从而在关键字失配的情况下表现出更强的对抗攻击效果.

图8对DADiff在关键词失配情况下的生成质量进行了可视化.在关键字失配的情况下,虽然在个别推理提示下的对抗视觉效果有所下降,但DADiff始终保持着良好的反定制效果.这在"一张[*A*]坐在椅子上的照片"和"一张[*A*]在埃菲尔铁塔前的照片"的提示词下尤为明显.即使是使用"一张sks的照片"生成对抗样本,

表2 关键词失配情况下的量化对比实验

Dreambooth Prompt	Method	“a photo of asdf person”				
		FDFR ↑	ISM ↓	SER-FIQ ↓	BRIS-QUE ↑	FID ↑
“A photo of asdf person”	AdvDM	0.08	0.18	0.53	13.25	339.24
	Anti-DB	0.47	0.14	0.35	23.32	347.25
	SimAC	0.59	0.04	0.35	29.02	348.97
	DisDiff	0.88	**0.02**	0.10	37.53	417.05
	DADiff(Ours)	**0.91**	**0.02**	**0.08**	**44.58**	**443.06**
		“a dslr portrait of asdf person”				
	AdvDM	0.11	0.15	0.60	15.75	284.37
	Anti-DB	0.11	0.17	0.60	12.68	277.85
	SimAC	0.08	0.17	0.70	5.59	227.91
	DisDiff	0.13	0.23	0.60	19.32	253.16
	DADiff(Ours)	**0.25**	**0.09**	**0.14**	**24.80**	**289.75**

注:加粗结果为最佳结果.

当使用“一张 asdf 的照片”进行 DreamBooth 训练时,仍然无法生成清晰的图像.

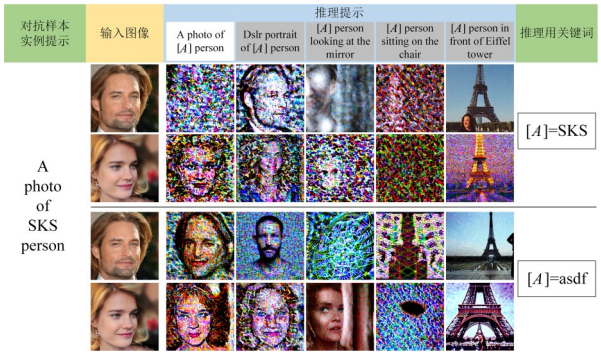


图8 关键词失配情况下的DADiff生成质量可视化

5.2.4 跨模型可迁移性实验效果对比

表3展示了分别使用SD-v1.5、SD-v1.4、SD-v2.1模型,分别使用VGGFace2和CelebA-HQ数据集和“一张sks的照片”作为实例提示词来生成对抗样本,并随后使用“一张sks的照片”、“一张sks的单反肖像”、“一张sks在埃菲尔铁塔前的照片”、“一张sks照镜子的照片”和“一张sks坐在椅子上的照片”作为推理提示来生成图像并计算这五个推理提示词生成图像质量评价的平均结果. 在几乎所有跨模型场景下的所有指标中,DADiff都表现出更强的反定制效果. 尤其是在跨模型设置下,与现有方法相比,DADiff将模型间迁移性平均值提高了2%~40%. 考虑到每个值表示的是在两个数据集上的五个提示词中对生成的32个ID图像进行的平均质量评估,结果表明,DADiff在不同模型和不同提示词之间都具有更广泛的可迁移性,而且对不同的人脸数据集均有效. 该实验进一步证明,在针对扩散模型的对抗攻击中,更全面、更彻底地利用扩散模型的关键组件和输入,可以实现更有效的反定制效果. 图9展示了DADiff在跨模型情况下的可视化实验效果,可以看出,DADiff仍然表现出有效性,但与关键字不匹配时的情况相比,破坏效果在不同模型上视觉质量下降程度直观上稍微减弱了. 这表明,模型层面的差异对对抗样本的性能的影响比提示词层面的差异更大. 因此,这更加强调了深入挖掘攻击UNet模型内部细节对提升反定制对抗样本可迁移性的必要性.

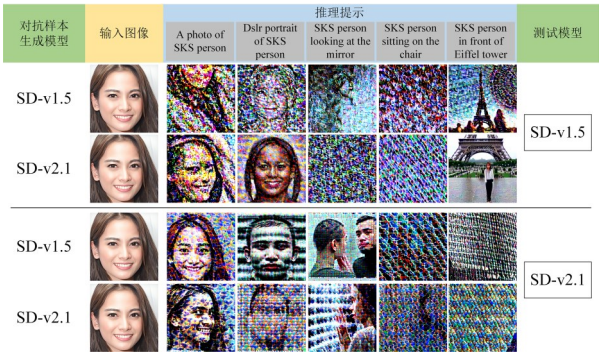


图9 DADiff跨模型反定制效果可视化

5.2.5 消融实验

DADiff由几个损失组成,即UNet输出损失L_{cond}[式(2)]和注意力损失L_A[式(10)]. 表4记录了为所有损失构建的消融实验,其中的数值表示在五个推理提示下从两个数据集生成的32个ID的图像的平均结果,类似于表3. 当只使用一个注意力损失时,其相应的超参数(α_1或$1-\alpha_1$)设置为1. 当UNet输出损失和任何注意力损失的梯度组合在一起时,超参数α_2始终设置为0.4.

结果表明,L_{cond}确实是反定制的主要损失. 排除L_{cond}会导致反定制的有效性显著降低. 将L_{SA}或L_{CA}与L_{cond}组合,都会增强对抗样本的反定制效果. 计算自注意力模块的损失会使生成的图像更难识别以检测到人(FDFR更高),这表明破坏自注意力模块会破坏定制模型学习像素相关性的能力. 此外,在此基础上添加交叉注意力模块的损失会进一步降低生成图像的质量(SER-FIQ更低,BRISQUE和FID更高),这表明引入的文本层面对抗先验可以指导图像层面对抗样本对抗属性的进一步增强. 总而言之,所有提出的损失都是有效的.

此外,DADiff还提出了使用局部随机时间步长梯度集成(LRTGE)策略来整合在多个时间步长上计算的梯度,然后使用集成的梯度在使用PGD攻击的对抗样本生成过程中进行扰动更新. 在使用所有损失的情况下,表5对使用单步更新扰动和使用LRTGE更新扰动进行了消融实验. 由于消融实验中没有考虑最佳时间步长选择(SimAC)和基于时间步长分配学习率(DisDiff),单时间步长设置下的性能仅与Anti-DB相当. 这

表3　跨模型可迁移性对比实验

Surrogate	Method	SD-v1.5					SD-v1.4					SD-v2.1				
		FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
w/o Protect		0.07	0.55	0.62	11.97	232.98	0.09	0.39	0.58	15.73	310.72	0.06	0.41	0.71	10.04	207.54
v1.5	Anti-DB	0.37	0.15	0.27	24.49	377.66	0.39	0.12	0.29	21.10	391.21	0.09	0.32	0.58	17.99	225.15
	SimAC	0.26	0.25	0.29	20.55	356.03	0.04	0.37	0.50	18.22	343.94	0.79	0.04	0.23	30.45	284.75
	DisDiff	0.62	0.13	0.25	25.45	399.73	0.31	0.22	0.36	22.75	374.05	0.78	0.05	0.30	32.71	294.28
	DADiff (Ours)	**0.93**	**0.06**	**0.03**	**40.59**	**464.56**	**0.93**	**0.06**	**0.10**	**42.21**	**444.20**	**0.81**	**0.03**	**0.22**	**36.88**	**311.12**
v1.4	Anti-DB	0.50	0.10	0.26	26.00	396.45	0.48	0.09	0.32	22.18	397.46	0.16	0.24	0.36	34.74	340.39
	SimAC	0.19	0.31	0.43	18.52	344.69	0.29	0.20	0.36	28.90	395.83	0.59	0.12	0.13	42.60	407.08
	DisDiff	0.61	0.10	0.16	31.20	409.17	0.19	0.23	0.31	22.22	344.71	0.41	0.14	0.28	**43.33**	389.85
	DADiff (Ours)	**0.86**	**0.08**	**0.12**	**33.25**	**441.07**	**0.94**	**0.04**	**0.04**	**38.72**	**478.57**	**0.64**	**0.11**	**0.13**	40.42	**414.60**
v2.1	Anti-DB	0.10	0.26	0.42	20.68	349.72	0.25	0.20	0.33	21.54	368.05	0.16	0.22	0.45	25.63	315.55
	SimAC	0.44	0.18	0.33	29.00	394.69	0.24	0.29	0.41	24.23	374.57	0.70	0.14	0.17	43.02	389.57
	DisDiff	0.39	0.13	0.32	26.27	397.12	0.46	0.12	0.27	27.27	386.26	0.70	0.11	0.16	43.06	411.66
	DADiff (Ours)	**0.63**	**0.12**	**0.18**	**32.15**	**413.28**	**0.78**	**0.11**	**0.12**	**33.54**	**419.26**	**0.72**	**0.10**	**0.14**	**44.43**	**424.20**

注:加粗结果为最佳结果.

表4　损失函数量化消融实验

Method	Average				
	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
L_{cond} only	0.59	0.09	0.20	34.21	430.10
L_{A} only	0.22	0.23	0.41	28.36	355.61
$L_{\text{cond}}+L_{\text{SA}}$	0.81	0.08	0.11	34.86	451.48
$L_{\text{cond}}+L_{\text{CA}}$	0.66	0.09	0.16	33.90	449.65
DADiff	**0.93**	**0.06**	**0.03**	**40.59**	**464.56**

注:加粗结果为最佳结果.

表明在单个时间步计算的梯度值不足以反映损失中包含的所有对抗信息,显著变化的梯度分数和损失确实阻碍了对抗样本性能的提升. 相比之下,使用LRTGE来集成每个时间段内的随机步长梯度,可以使单个PGD攻击梯度中包含更丰富的对抗信息. 固定时间分割还可以稳定集成梯度的总体得分,最终显著提高生成的对抗样本的有效性.

表5　梯度集成策略消融实验

Method	Average				
	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
Single Step	0.32	0.19	0.31	25.03	372.65
LRTGE	**0.93**	**0.06**	**0.03**	**40.59**	**464.56**

注:加粗结果为最佳结果.

接下来,表6和表7分别对LRTGE中使用的超参数B和α_2进行消融实验. 表6表示固定$\alpha_2=0.4$,对B取不同值的实验结果. 可以看出,当B越大时,对抗效果更好,但B越大意味着需要采样的时间步长梯度越多,也意味着显著增长的运行时间. $B=50$所需的运行时长约等于$B=25$的两倍,但对抗效果的增长却没有翻倍. 表7表示固定$B=25$,对α_2取不同数值的实验结果. 4.4节指出,α_2的取值决定了破坏注意力模块的梯度与端到端破坏的梯度之间的比例. α_2越高,注意力的破坏越多,对语义信息的破坏越彻底,对于人脸数据来说,意味着对面部结构的关联破坏更彻底,体现为FDFR、ISM、SER-FIQ数值上的更优;α_2越低,端到端破坏比例越高,体现为对像素层面的破坏越明显(FID数值更优)但对语义信息的破坏略弱于注意力模块的破坏. 因此,为了权衡语义破坏与像素破坏效果,同时考虑到时间开销,本文选择了$\alpha_2=0.4$,$B=25$的超参数组合.

表6　超参数B消融实验

$\alpha_2=0.4$	Average				
	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
$B=10$	0.90	0.09	0.10	38.26	433.10
$B=25$ (Ours)	0.93	0.06	0.03	40.59	464.56
$B=50$	**1.00**	**None**	**0.00**	**45.09**	**497.51**

注:加粗结果为最佳结果.

在表4损失函数消融实验的基础上,表8对APV的价值进行了进一步的消融实验. 具体而言,分别设计了不使用APV(w/o APV)、使用与APV同维度的随机高斯噪声向量(w/ Random Vector)和全零向量(w/ Zero Vector)生成对抗样本并验证效果,以量化APV作为“图像

表 7 超参数 α_2 消融实验

$B=25$	Average				
	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
$\alpha_2=0.1$	0.918	0.072	0.040	37.675	**497.572**
$\alpha_2=0.4$ (Ours)	0.930	0.060	0.030	**40.590**	464.560
$\alpha_2=0.8$	**1.000**	**None**	**0.001**	40.147	457.312

注:加粗结果为最佳结果.

层面对抗样本对抗性先验”的贡献. 结果表明,使用了APV引导的DADiff在所有评估指标上都显著优于w/o APV,这直接证明了引入APV作为先验知识,能够有效引导并增强图像层面的对抗攻击,使其破坏效果更强. 此外,DADiff的效果也明显优于“w/ Random Vector”和“w/ Zero Vector”,这验证了APV本身携带了有效的、针对性的对抗语义信息,而非一个任意或无效的向量. APV作为“跨模态先验知识”,其核心价值在于能够引导图像对抗攻击,使其更精准地破坏模型内部的语义绑定,从而获得更强的反定制效果和迁移性.

表 8 对抗提示向量消融实验

Method	Average				
	FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUE ↑	FID ↑
w/o APV	0.61	0.12	0.16	34.21	430.10
w/ Random Vector	0.69	0.13	0.17	35.05	422.35
w/ Zero Vector	0.63	0.10	0.12	34.13	433.02
w/APV (DADiff)	**0.93**	**0.06**	**0.03**	**40.59**	**464.56**

注:加粗结果为最佳结果.

5.2.6 对艺术风格画像反定制的补充实验

现实场景中,艺术品版权等场景同样易受侵害且关注度高. 本节从WikiArt艺术风格数据集中选取15个不同艺术流派的风格图片,以“sks”为关键词进行DreamBooth训练,并执行一系列反定制对抗扰动实验. 表9展示了关键词失配场景下的量化对比效果,由于艺术风格图像没有人脸相关属性,因此不再使用FDFR、ISM和SER-FIQ指标进行衡量. 本实验用sks指代艺术风格,用“sks风格的油画”作为提示词进行对抗样本生成,然后用asdf进行了关键词失配迁移性分析. 量化结果表明,DADiff在大多数量化指标上均优于现有方法.

随后,本节用“巴洛克”风格作为示例进行可视化,并在表9实验的基础上进一步补充了“sks风格的皇家肖像”和“sks风格的天使油画”两个跨提示词图像生成效果,如图10所示. 结果表明,在关键词失配情况下,DADiff对图像中对象的结构信息实现了更彻底的破坏,相比而言,其他方法的对抗效果更多体现在纹理层面的扰动,但仍保留了巴洛克油画风格和人物结构. 在跨提示词情况下,DADiff已无法描绘出巴洛克艺术风格,而其他方法对艺术风格的呈现几乎没有影响,说明对抗效果完全丧失. DADiff在跨关键词和跨提示词情况下的效果均远好于其他方法,这表明DADiff对注意力模块的扰动更彻底地破坏了扩散模型对图像结构关系的理解与推理效果.

表 9 WikiArt数据集关键词失配量化结果评估

Dataset	Method	“an oil painting in sks style”		“an oil painting in asdf style”	
		BRISQUE ↑	FID ↑	BRISQUE ↑	FID ↑
WikiArt	w/o Protect	22.235	341.938	11.137	388.006
	Anti-DB	45.733	554.505	14.995	565.801
	SimAC	**55.686**	552.395	18.698	416.948
	DisDiff	52.104	452.504	15.382	401.065
	DADiff(Ours)	53.644	**637.006**	**19.491**	**577.001**

注:加粗结果为最佳结果.

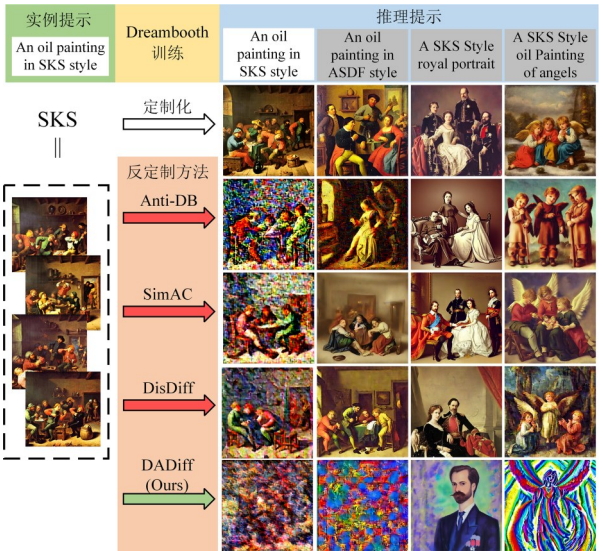


图 10 使用“sks”指代巴洛克风格的对比实验

5.2.7 对抗样本对常规生成能力的影响

为了验证经对抗样本训练后的模型的常规生成能力是否受到影响,本节用“A photo of sks person”生成对抗样本,并进行Dreambooth微调后,使用一组与“sks”语义无关的提示词(包括:“A photo of a person”“A photo of a landscape”“A photo of a cat sitting on the chair”“A cityscape in the style of Van Gogh”)分别驱动原始模型与对抗微调后模型生成图像,结果如图11所示. 对比可见,对抗微调后的模型在这些非目标提示词下的生成质量与原始模型高度一致,未出现语义偏移、结构崩塌或大面积噪点等退化现象. 这表明,尽管相比其他方法,DADiff更充分地利用了扩散模型的各层级结构与输出,但得到的对抗样本只精准削弱提示词与训练图像之间的绑定关系,而不会对模型原有的通用生成能力造成显著干扰或破坏.

图 11　使用与“sks”语义无关的提示词在对抗样本训练之前和之后的模型上生成的图像

5.2.8　局限性分析

DADiff 为了实现更强的跨模态协同攻击效果,在第一阶段引入了对抗提示词向量(APV)的生成,并在第二阶段对自注意力和交叉注意力模块构建了额外的损失函数,该方法会带来显著的计算开销. 为了量化这一开销,本节补充了与主要对比方法在相同实验环境下的单张图像对抗样本平均生成耗时对比,实验所用设备均与以上实验相同. 结果如表 10 所示,DADiff 的计算开销确实明显高于现有方法,约为 Anti-DB 的 15 倍. 然而,在扩散模型安全性研究尚不完善的情况下,优先探索和验证“有效性上限”具有重要的研究价值. 如表 1~表 3 的实验结果所示,DADiff 在跨提示词、关键词失配、跨模型迁移性等核心指标上均取得了显著优于现有方法的效果(提升 20% 以上),这证明了多要素协同攻击思路的巨大潜力. 本文会将“效率优化”列为后续工作的主要研究方向.

表 10　各方法 GPU 运行时间对比　　单位:s

Anti-DB	SimAC	DisDiff	DADiff
155.5	174.5	172.3	2 328.3

6　结束语

本文提出了 DADiff,一种基于提示词层面和图像层面两阶段对抗攻击的扩散模型反定制方法. 在攻击过程中,DADiff 充分利用了扩散模型的所有输入和关键组件,并在多个评价指标的定性和定量评估中取得了比现有方法更好的结果. DADiff 还证实了这样一种观点,即对于具有复杂输入和复杂结构的模型,每个输入和模块都可能存在潜在的对抗样本风险. 这一见解可指导后续研究不断优化和提高针对扩散模型的对抗样本在社交平台和复杂微调算法等更实际的应用场景中的有效性、实用性和鲁棒性.

由于 DADiff 充分考虑了文本、图像和时间层面,以及扩散模型细粒度模块之间的对抗性,因此对抗样本的生成需要较大的时间与性能开销,现阶段的实际应用价值受限. 随着硬件性能的提升以及对扩散模型安全特性认知的加深,DADiff 在时间与性能方面的劣势有望得到改善. 此外,一旦获取反定制对抗样本,便能灵活适配多种模型的反定制需求,这一特性为后续构建更完善的扩散模型安全防护体系奠定了坚实基础. 未来研究可基于 DADiff 的思路,从模型多维度输入进一步探索降低开销的优化策略,从而为保障扩散模型的安全应用开辟新的路径.

参考文献

[1] HUANG Y, HUANG J C, LIU Y F, et al. Diffusion model-based image editing: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(6): 4409-4437.

[2] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[EB/OL]. (2021-02-26) [2025-09-30]. https://arXiv.org/abs/2102.12092.

[3] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[EB/OL]. (2022-05-23)[2025-09-30]. https://arXiv.org/abs/2205.11487.

[4] NICHOL A, DHARIWAL P. Improved denoising diffusion probabilistic models[EB/OL]. (2021-02-18)[2025-09-30]. https://arXiv.org/abs/2102.09672.

[5] RUIZ N, LI Y Z, JAMPANI V, et al. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 22500-22510.

[6] HU E J, SHEN Y L, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[EB/OL]. (2021-10-16)[2025-09-30]. https://arXiv.org/abs/2106.09685.

[7] GAL R, ALALUF Y, ATZMON Y, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion[EB/OL]. (2022-08-02)[2025-09-20]. https://arXiv.org/abs/2208.01618.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2017: 30.

[9] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB/OL]. (2020-12-16)[2025-09-30]. https://arXiv.org/abs/2006.11239.

[10] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 10674-10685.

[11] ZHUANG H M, ZHANG Y H, LIU S J. A pilot study of query-free adversarial attack against stable diffusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2023: 2385-2392.

[12] YANG Y J, GAO R Y, WANG X S, et al. MMA-diffusion: Multimodal attack on diffusion models[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 7737-7746.

[13] SALMAN H, KHADDAJ A, LECLERC G, et al. Raising the cost of malicious AI-powered image editing[EB/OL]. (2023-02-13)[2025-09-30]. https://arxiv.org/abs/2302.06588.

[14] LIANG C M, WU X Y, HUA Y, et al. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples[EB/OL]. (2023-06-06)[2025-09-30]. https://arXiv.org/abs/2302.04578.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2019-09-04)[2025-09-30]. https://arXiv.org/abs/1706.06083.

[16] VAN LE T, PHUNG H, NGUYEN T H, et al. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 2116-2127.

[17] WANG F F, TAN Z T, WEI T Y, et al. SimAC: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 12047-12056.

[18] LIU Y S, AN J Y, ZHANG W Q, et al. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 3587-3596.

[19] MOKADY R, HERTZ A, ABERMAN K, et al. Null-text inversion for editing real images using guided diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 6038-6047.

[20] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9193.

[21] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[EB/OL]. (2022-10-05)[2025-09-30]. https://arXiv.org/abs/2010.02502.

[22] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 8780-8794.

[23] YANG L, ZHANG Z L, SONG Y, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2023, 56(4): 1-39.

[24] CAO Z Y, LI J H, XU X R. DiffusionAAE: Enhancing hyperspectral image classification with conditional diffusion model and Adversarial Autoencoder[J]. Ecological Informatics, 2025, 87: 103118.

[25] WANG L Z, YANG S, LIU S, et al. Not all steps are created equal: Selective diffusion distillation for image manipulation[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 7438-7447.

[26] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2016: 3730-3738.

[27] CAO Q, SHEN L, XIE W D, et al. VGGFace2: A dataset for recognising faces across pose and age[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway: IEEE, 2018: 67-74.

[28] DENG J K, GUO J, VERVERAS E, et al. RetinaFace: Single-shot multi-level face localisation in the wild[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5202-5211.

[29] DENG J K, GUO J, XUE N N, et al. ArcFace: Additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4685-4694.

[30] TERHÖRST P, KOLF J N, DAMER N, et al. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5650-5659.

[31] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708.

[32] 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 the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6629-6640.

[33] VON P, PATIL S, LOZHKOV A, et al. Diffusers: State-of-the-art diffusion models[EB/OL]. (2025-08-20)[2025-09-30]. https://github.com/huggingface/diffusers.

作者简介

叶登攀　男,1975年9月出生于湖北省黄冈市. 现为广州大学网络空间安全学院教授、博士生导师. 主要研究方向为多媒体安全、人工智能与机器学习等. 中国电子学会会员编号:E190161643M.
E-mail: yedp2001@163.com

唐　龙　男,1996年8月出生于四川省攀枝花市. 2025年博士毕业于武汉大学国家网络安全学院. 现就职于华为技术有限公司. 主要研究方向为对抗样本攻防、多模态大模型安全、人脸隐私保护等.
E-mail: l_tang@whu.edu.cn

陈思润　男,2001年9月出生于河南省许昌市. 2023年本科毕业于华南理工大学计算机科学与工程学院,现为武汉大学国家网络安全学院硕士研究生. 主要研究方向为多媒体安全、对抗攻防等. 中国电子学会会员编号:E190160925A.
E-mail: chensirun@whu.edu.cn

刘梓毅　男,1995年7月出生于新疆维吾尔自治区乌鲁木齐市. 2017年本科毕业于重庆交通大学信息科学与工程学院,2022年硕士毕业于桂林电子科技大学计算机与信息安全学院,现为武汉大学国家网络安全学院博士研究生. 主要研究方向为网络流量分析,对抗样本攻击与防御. 中国电子学会会员编号:E190184468A.
E-mail: ziyi_liu@whu.edu.cn

吕芸娜　女,2001年3月出生于福建省泉州市. 2023年本科毕业于湖南大学信息科学与工程学院,现为武汉大学国家网络安全学院硕士研究生. 主要研究方向为对抗样本攻击与防御. 中国电子学会会员编号:E190160924A.
E-mail: lvyunna@whu.edu.cn

石绣文　女,2002年7月出生于湖北省黄冈市. 2024年本科毕业于中国地质大学(武汉),现为武汉大学硕士研究生. 主要研究方向为多媒体信息安全.
E-mail: shixiuwen@whu.edu.cn