2025 年第一届人工智能安全竞赛

作品报告

作品名称：VisiFence: 主动抗 AI 编辑扰动系统

电子邮箱：2115095255@qq.com

提交日期：2025.07.15

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用 A4 纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5 倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目录

摘要

随着人工智能技术的飞速发展，AI 图像编辑（AI-based Image Editing）技术正在以前所未有的速度发展。AI 图像编辑是借助人工智能技术对图像进行自动或半自动处理、修改及创作的技术，突破了传统图像编辑依赖手动操作的局限，能通过算法智能分析图像内容并实现多种编辑效果，**而基于 AI 的图像恶意编辑（Malicious AI Editing），则会对社交网络中的内容真实性带来比传统手工图像内容篡改方式更为巨大的安全威胁**。AI 恶意编辑可以在毫无痕迹的情况下伪造图像内容、篡改人像、重构事件场景等，**严重威胁到舆论生态、公众人物形象、企业声誉，甚至对国家安全和社会稳定带来了新的严峻挑战**。因此，保护图像内容的真实性，实现可以防止 AI 恶意编辑的防御技术具有重要的研究意义和应用价值。

现有保障社交网络内容真实性的主要技术集中于图像篡改检测技术，属于“被动防御”策略。这类方法试图通过训练判别模型识别被编辑图像，往往面临泛化能力差、误判率高的问题。与现有被动的伪造检测技术不同，本作品**提出一种抗 AI 图像恶意编辑的主动防御系统，在图像发布前，通过该系统可向图像中嵌入具有防护能力的微小扰动，使得主流 AI 编辑工具无法编辑合成出正常质量的图像，以实现对 AI 恶意编辑行为的抵御**。VisiFence 系统可为图像内容发布方提供以下能力：1）**保障发布图像的质量**，保护后的图像与真实图像质量无可感知差异；2）**抵抗现有主流的 AI 图像编辑工具**，使得所编辑后的图像质量差而无法使用，保护发布内容不被滥用。

本作品的创新性主要体现在以下几个方面：

1） **主动式 AI 编辑防御机制：**首次针对 AI 编辑将引发的社交网络内容真实性新威胁问题，提出“图像主动扰动防御”的技术思路，通过 AI 对抗 AI，针对主流 AI 编辑技术所基于的 Diffusion model 技术，采用对抗样本技术思路，对发布图像中添加抗主流 AI 编辑的特定噪声，防止图像内容被 AI 恶意编辑，从内容发布的源头保障内容的真实性；

2） **良好的通用防护能力：**针对恶意用户所使用的 AI 编辑算法无法预知的问题，VisiFence 针对 Stable Diffusion 系列模型中普遍存在的 U-Net 注意力结构，提出通过最大化注意力图的熵与变差的对抗样本生成策略，并引入偏离参考注意力图，显著削弱模型对 Prompt 的定位能力，实现对多个编辑模型的一致防护，能够有效

防御包括 DreamBooth、Textual Inversion、SDEdit（img2img）及 Inpainting 等多种主流 AI 图像编辑技术的恶意编辑；

3） **强健的鲁棒性保障：**针对图像在实际使用中所常见的缩放、裁剪、压缩等处理操作，VisiFence 在扰动设计阶段引入仿真处理机制，确保扰动在多种变换条件下依然稳定有效。实验验证表明，系统在高压缩率及复杂变形场景下依然保持较强防护性能，即使在高强度 JPEG 压缩（如 85% 压缩率）下，防护效果依然稳定可靠，满足实际应用中对图像质量与防护效果的双重需求。

本作品的实用性体现在以下几个方面：

1） **便捷的 Web 服务平台：**VisiFence 为用户提供了方便易用的 Web 服务，支持图像的上传与防护模式选择，自动完成扰动生成与图像返回，并且可以模拟编辑模型攻击的直观展示保护效果。

2） **提供两种防御保护能力：**为了满足用户对于防御能力的不同需求，AntiAIEdit 为在互联网中发布图像内容的用户提供了两种服务模式：1）通用型在线保护服务，可快速对图像添加预生成的通用扰动以提供一定程度的防御能力；2）定制型离线保护服务，针对每一张图像定制优化扰动，提供更强的防护能力。用户可根据实际需求选择服务类型，实现灵活部署。

3） **良好的实时性：**针对传统扰动优化需针对每张图像单独计算且耗时较长的问题，VisiFence 采用预生成扰动库结合向量快速匹配策略，支持批量处理和跨数据的通用防护，大幅提升了处理速度。在线保护响应时间可控制在 1 秒以内，显著优于“图像到达后才开始梯度优化”的传统方式。

综上所述，VisiFence 作为首款面向 AI 恶意编辑的主动防御系统，在防御能力、通用性、实时性和可用性等方面均具备显著优势。其在媒体内容保护、社交平台防护、企业品牌安全、隐私保护等多个领域具有广泛的应用前景与实用价值。

关键词： AI 图像编辑；主动防篡改保护；AI 编辑对抗；实时扰动生成；内容真实性保护

第一章 作品概述

1.1 选题背景分析

随着人工智能（Artificial Intelligence, AI）技术的迅猛发展，AI 图像编辑（AI-based Image Editing）技术得到快速发展。基于扩散模型（Diffusion Models）、生成对抗网络（GANs）等深度生成算法的图像编辑工具不断涌现，如 Stable Diffusion、DALL · E 2、Midjourney、Runway 等产品，广泛应用于艺术创作、广告设计、影视制作、游戏开发等多个行业。这类工具使得普通用户无需专业图像编辑技能，即可通过简单的文本指令实现高质量的图像生成和修改，大幅降低了创作门槛，极大推动了视觉内容的民主化生产。

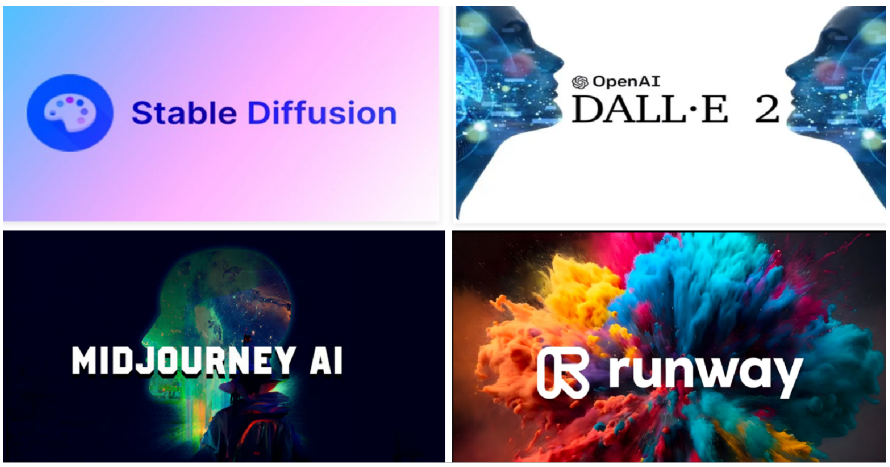


图 1.1　迅速发展的 AI 编辑技术

然而，图像编辑能力的极大增强也带来了严重的安全隐患。尤其在社交媒体普及、图像传播高速化的今天，恶意利用 AI 图像编辑技术进行伪造、篡改、抹黑、误导等行为的成本大幅下降，风险则迅速上升。攻击者可以通过 Prompt 注入、风格篡改、人物替换等方式对图像进行几乎无痕的恶意编辑，造成公众人物形象受损、商业品牌信任度下降，甚至成为舆论操纵和政治抹黑的工具。

近年，AI 图像编辑滥用所引发的安全事件持续增多，引发社会广泛关注。2023 年 4 月，内蒙古包头市一科技公司法人郭先生被诈骗分子利用 AI 换脸和拟声技术伪装成“好友”视频通话，骗取 430 万元保证金 [1]；2024 年 5 月，中国某知名品牌广告图被恶意修改，替换商品图像为仿冒 Logo，引发消费者投诉与品牌公关危机 [2]；2025 年 5 月，澳大利亚多所学校报告称，学生使用 AI 技术生成教师和同学的虚假不

雅图像，导致受害者遭受心理创伤。[3]；2024 年 3 月，欧洲某极右翼组织通过 Stable Diffusion 生成多张虚假的“移民暴力犯罪”图像，用于社交平台煽动恐慌情绪，被多国媒体揭发后引发深度伪造内容的舆论风暴 [4]；2024 年 5 月，印度地方选举期间，某政党被揭露利用 AI 图像编辑工具批量生成竞争对手“贿选”场景的伪造照片并进行传播，严重扰乱了选举秩序 [5]。上述事件表明，**AI 图像恶意编辑已不再局限于娱乐或实验用途，而是逐渐演变为影响舆论导向、侵犯个人权利、破坏公信力的重要安全威胁**。

图 1.2　人工智能伪造图片视频相关新闻

近年来，随着恶意图像编辑模型的快速发展，其在制造虚假图像、篡改事实信息中的滥用风险日益突出，引发全球范围内的法律关注和监管行动。2016 年 4 月，欧盟通过的《通用数据保护条例》（GDPR）首次明确限制个人数据被用于制作虚假及篡改图像，强化了对个人隐私的保护 [6]。2021 年 1 月，我国《民法典》人格权编中增加条款，禁止任何技术手段非法篡改他人肖像和图像，维护个人形象权 [7]。2023 年 1 月，我国发布《互联网信息服务深度合成管理规定》，进一步规范了基于 AI 技术的图像编辑服务，要求提供者必须对深度合成内容的来源和真实性负责，防止恶意图像编辑带来的社会风险 [8]。2023 年 7 月，我国出台的《生成式人工智能管理暂行办法》特别强调了对恶意图像编辑模型的管理，包括算法安全审查与内容风险评估，明确运营主体责任 [9]。2024 年 1 月，欧盟通过的《人工智能法》（EU AI Act）更进一步，按风险等级对恶意图像编辑及相关 AI 系统实施严格监管，力求在技术创新与社会安全之间取得平衡 [10]。这些法律法规的制定和实施，反映了各国对恶意图像编辑模型潜在

危害的高度重视，以及通过法律手段保障个人权益和信息真实性的共同努力。

表 1.1　近年来针对恶意图像编辑模型的主要法律法规

法律法规名称	颁布时间	主要内容
GDPR《通用数据保护条例》	2016 年 4 月	限制个人数据用于虚假和篡改图像
《中华人民共和国民法典》人格权编	2021 年 1 月	禁止技术篡改他人肖像，保护形象权
《互联网信息服务深度合成管理规定》	2023 年 1 月	规范深度合成服务，明确内容责任
《生成式人工智能管理暂行办法》	2023 年 7 月	强调算法审查与内容风险评估
EU AI Act《欧盟人工智能法》	2024 年 1 月	按风险等级监管恶意图像编辑系统

传统的图像内容安全防护技术主要依赖于“检测为主”的被动式策略。例如，研究者通过训练判别模型检测是否存在 GAN 伪造痕迹、重建残差图像进行篡改区域识别，或基于频域特征与噪声模式识别图像真实性。然而，**这类技术普遍存在以下三个问题：**

1） **泛化性差**：训练检测器往往依赖于已知模型产生的伪造样本，面对新兴、未见过的编辑方法时，检测准确率显著下降；

2） **响应滞后**：检测系统通常部署在平台侧，图像内容已上传并传播后才进行风险识别，用户无法在发布前主动规避风险；

3） **误报率高**：不同图像模态间的差异性及编辑方法的多样性使检测算法易受误判干扰，尤其在真实图像后处理（压缩、滤镜等）与伪造编辑图像之间界限模糊。

这些问题导致目前的检测型方案在实战中应用效果不佳，急需更加主动、通用、隐蔽的 AI 恶意编辑防御手段。

面对这些挑战，当前学术界已开始从“主动防御”视角提出替代思路。与传统“检测-过滤”的滞后机制不同，主动防御策略强调**“事前保护”**。**即在用户发布图像前，就通过微小的、肉眼不可感知的扰动对图像进行预处理，使其在经过 AI 编辑模型处理时无法生成合理结果，从而抑制恶意篡改**。类似的研究最早出现在对抗样本（Adversarial Examples）领域，如 UAP（Universal Adversarial Perturbation）方法、AdvPaint 扰动防御等，近年来亦在 CLIP 反向引导等领域得到初步探索。

然而，现有主动防御方法仍面临以下技术挑战：

1） **跨模型迁移性弱**：对抗扰动往往只能针对某一模型有效；

2） **生成速度慢**：依赖逐图优化方式，难以支持大规模应用；

3） **图像破坏严重**：扰动可感知性高，影响图像质量与用户体验；

4） **抗压缩性差**：在图像被平台压缩（如 JPEG）后扰动易失效。

基于上述技术与现实需求背景，我们提出 **VisiFence**——一个面向 AI 恶意图像编辑的主动防御系统。该系统拥有以下能力：

1） 提前对图像嵌入防护扰动，无需检测即可抵御恶意编辑；

2） 提供快速在线扰动库匹配与增离线定制扰动优化，兼顾效率与防护强度；

3） 具备良好的跨模型、跨指令、跨方法的防护能力，支持批量处理；

4） 在 JPEG 压缩率 85%、裁剪、缩放等场景下保持扰动效果稳定；

5） 集成 Web 服务平台，方便用户在实际图像发布流程中无缝接入。

综上所述，随着 AI 图像编辑能力日益增强，传统检测机制已无法满足图像安全防护的现实需求。VisiFence 作为一种主动式防御系统，从扰动设计、系统架构到服务部署均面向实用落地场景，填补了图像内容防篡改领域的关键技术空白，具备显著的社会价值、研究前沿性和广阔的应用前景。

1.2 相关工作

1.2.1 主流 AI 编辑技术

当前主流的 AI 图像编辑技术大致可以分为两类：一类是**无需训练的（Training-Free）编辑方式**，另一类是**基于训练的（Training-Based）个性化编辑方式**。这两类方法分别在不同的图像修改任务中发挥关键作用，构成了当前 AIGC 系统中图像编辑模块的核心。

（1）Training-Free：无需训练的编辑方法 这类方法不涉及模型参数的更新，通常基于预训练扩散模型，通过调整输入或条件进行图像的直接修改，适用于灵活、高效的交互式编辑需求。典型方法包括：

- **img2img[20]**：以一张初始图像为输入，在保持整体结构的基础上，根据用户指定

的文本提示（prompt）生成相应风格或语义修改后的图像。例如，用户可以输入一张人脸照片，并提示“变为油画风格”，模型将输出保持人物面部结构但具油画风格的结果。其原理是在一定噪声扰动后，通过扩散模型的反向过程生成新图。

- **Inpainting[21]**：又称图像修补，用户在原图中遮盖某些区域，模型基于周边语义和文本条件对缺失区域进行内容补全。该技术被广泛应用于去除水印、背景替换、物体替换等任务，其特点是对遮挡区域的“上下文感知式”重建能力。

（2）Training-Based：基于训练的个性化编辑方法 此类方法通过对扩散模型的特定部分进行微调，使模型能够学习用户提供的新概念（如某个人、某种艺术风格等），从而支持更加个性化和持续性的图像合成。代表性方法包括：

- **DreamBooth**[18]：通过向模型提供少量某个对象（如某人、某宠物）的图片，并配合一个专属的关键词（如“Sks Dog”）进行微调训练，使模型能够在日后使用这一关键词时合成目标对象。例如训练后输入“a painting of sks dog in a rocket”，模型可生成该宠物穿梭在火箭中的图像。
- **Textual Inversion[19]**：与 DreamBooth[18] 不同的是，该方法不通过微调模型参数，而是在词向量空间中学习一个新词的嵌入向量，绑定特定图像语义。这一向量随后可作为 prompt 的一部分引导图像生成，其优势在于训练成本较低，且对模型干扰小。

表 1.2　主流 AI 图像编辑方法对比

方法	是否需训练	输入形式	主要用途	优缺点
img2img[20]	否	初始图像 + 文本 Prompt	图像风格迁移、语义修改	无需训练，速度快；但个性化能力弱
Inpainting[21]	否	遮挡图像 + Prompt	去除/替换物体、修复缺失区域	上下文感知强；遮挡区域大时易失真
DreamBooth[18]	是	若干张参考图像 + 专属 Prompt	个性化生成（如特定人/物）	生成效果精细；训练成本高，侵入模型
Textual Inversion[19]	是	若干张图像 + 新词 Token	新概念嵌入、低资源个性化	训练开销小，插入灵活；效果受限于词向量空间

综上所述，Training-Free 方法具有无需训练、灵活高效的特点，适合即时图像编辑任务；而 Training-Based 方法则支持更为精细、个性化的编辑需求，但往往需要更多的计算资源与训练时间。本项目在设计防御机制时，充分考虑了这两类方法的技术特征与攻击面差异，并提出了统一而有效的主动防御策略。

1.2.2 主动抗 AI 编辑方法

随着扩散模型在图像生成与编辑领域的广泛应用，针对其潜在的版权侵权与内容篡改风险，主动防御技术应运而生。主动防御的核心目标是在图像发布阶段，向原始内容添加精心设计的微小扰动，这些扰动在视觉上难以察觉，但能够有效干扰后续基于扩散模型的非法编辑或生成过程。通过破坏扩散模型对输入图像的特征提取和条件推理，主动防御技术实现对图像生成结果的精准干扰，保护原作者权益和内容安全。

当前，主动防御方法主要可以划分为以下几类：

1） **基于对抗样本的主动防御方法**：该类方法通过对输入图像添加对抗扰动，直接干扰扩散模型的去噪及条件提取过程，破坏生成结果的语义一致性。典型代表包括 AdvDM、Anti-DreamBooth[12] 和 Mist 等，方法侧重于设计优化目标，提升扰动的隐蔽性与攻击成功率。

2） **基于元学习与鲁棒性增强的方法**：为提升对抗扰动的泛化能力与对数据预处理（如滤波、压缩）的鲁棒性，元学习方法构建多模型代理池，通过跨模型训练获得具备强迁移性的通用扰动，代表工作如 MetaCloak。

3） **基于时间步优化与特征干扰的方法**：充分利用扩散模型中时间步不同对扰动敏感性的差异，自适应选择关键时间步进行扰动优化，结合特征层攻击实现更高效的防御，如 SimAC。

4） **基于编码器攻击与潜在空间扰动的方法**：针对扩散模型中不同组件的脆弱性，选择计算效率更高且更易受攻击的编码器潜在空间进行扰动生成，降低计算成本同时保持较高攻击效果，典型方法为 Score Distillation。

以下是这些方法的详细介绍：

（1）基于对抗样本的主动防御方法

1）工作原理

对抗样本保护通过向原始图像添加人眼不可见的微小扰动 δ，破坏扩散模型在特征提取与生成过程中的关键能力。这类扰动通过优化算法设计，能够误导扩散模型的去噪过程，使其无法基于输入图像生成语义一致的编辑结果 [11]。其数学形式可表示为 $\hat{x} = x + \delta$，其中扰动需满足 $\|\delta\|_p \leq \epsilon$ 以保证视觉不可察觉性。添加扰动后的图像虽然人眼难以察觉变化，但会导致扩散模型生成内容出现显著扭曲或纹理破坏。

2）代表性工作

AdvDM[11] 是首个系统研究扩散模型对抗样本的工作，其采用蒙特卡洛估计方法，通过随机采样时间步 t 和噪声 ϵ 来计算条件提取阶段的期望梯度。该方法定义的目标函数旨在最大化去噪误差，具体表现为 $\mathcal{L}_{\text{AdvDM}} = -\sum_{t=1}^{T} \|\epsilon - \epsilon_\theta(\hat{x}_t, t)\|_1$，迫使模型生成无意义输出。实验表明，AdvDM 在 Stable Diffusion 1.4 上可实现 82% 的生成失败率，但存在计算开销大的缺陷，单图像优化耗时超过 3 分钟，且对 DreamBooth[18] 等微调模型的迁移攻击成功率下降 37%。

Anti-DreamBooth[12] 针对个性化模型的防御需求，提出了类特定扰动方法。该方法通过优化扰动破坏主体与类名（如”[V]” 和”person”）的关联，其损失函数定义为 $\mathcal{L}_{\text{AntiDB}} = \|\text{CLIP}(\hat{x}) - \text{CLIP}(x_{\text{null}})\|_2$，其中 x_{null} 对应空文本嵌入特征。在 DreamBooth[18] 肖像定制场景下，该方法防御成功率超过 90%，且保持原图 PSNR 大于 38dB 的视觉质量。然而，该方法对高斯滤波（$\sigma = 1.5$）敏感，防御成功率会降至 52%，且仅适用于特定微调方法。

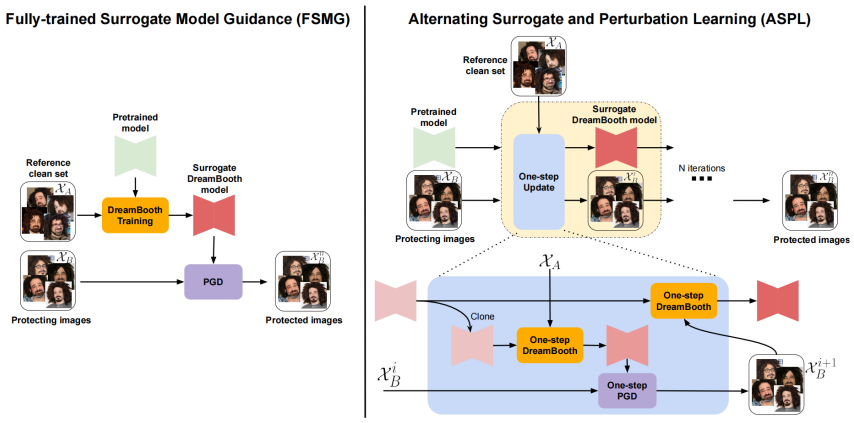


图 1.3　Anti-DreamBooth[12] 架构图

Mist[13] 通过融合语义破坏和纹理干扰的双路径损失提升了方法迁移性。语义路径采用 LPIPS 度量内容失真，定义为 $\mathcal{L}_{\text{sem}} = \text{LPIPS}(G(x), G(\hat{x}))$；纹理路径则通过高

频噪声破坏 UNet 特征图，使用 Frobenius 范数计算差异。该工作的创新点在于目标图像选择策略，通过显著性检测自动定位高对比度区域和重复图案进行重点攻击。实验显示，Mist 在 Textual Inversion[19]、DreamBooth[18] 和 img2img[20] 任务中平均攻击成功率达 89%，对高斯滤波和 JPEG 压缩的鲁棒性较 AdvDM 提升 40%。

3）对比分析

如表1.3所示，三类方法呈现明显的技术演进路径。AdvDM 奠定了基础框架但计算效率低下，Anti-DreamBooth[12] 针对特定场景优化却牺牲了泛化能力，而 Mist 通过多损失融合和目标选择策略，在保持高攻击成功率的同时提升了对抗数据变换的鲁棒性。特别值得注意的是，Mist 的语义-纹理联合优化机制有效解决了前两种方法在跨模型迁移性方面的短板，其创新性的目标定位方法也为后续研究提供了重要参考。

表 1.3　对抗样本保护方法对比

方法	计算成本	迁移性	数据变换鲁棒性
AdvDM[11]	高	低	中
Anti-DreamBooth[12]	中	极低	低
Mist	高	高	高

（2）基于元学习与鲁棒性增强的主动防御方法

1）工作原理

元学习方法通过构建多代理模型的训练环境，生成具有跨模型迁移能力的对抗扰动。如 [15] 所示，这种范式能够克服传统对抗样本对特定模型过拟合的问题，同时通过数据变换采样增强对预处理操作的鲁棒性。其核心在于将扰动生成过程建模为元学习任务，使得最终得到的扰动能够适应未知模型架构和输入变换。

2）代表性工作

MetaCloak[15] 提出了基于模型池的元学习框架。该方法在每次迭代中随机选择代理模型 θ_i，计算去噪误差最大化损失 $\mathcal{L}_{\text{meta}} = -\mathbb{E}_{t,\epsilon}[\|\epsilon - \epsilon_{\theta_i}(x_t, t)\|^2]$，通过二阶梯度更新生成扰动。为提升鲁棒性，在训练过程中对输入图像施加随机高斯滤波和 JPEG 压缩等变换，迫使扰动在这些变换下仍保持有效性。实验表明，该方法对在线服务 Replicate 的黑盒攻击成功率达到 73%，显著高于传统白盒方法 35% 的攻击迁移率。

然而，MetaCloak 的性能高度依赖代理模型池的多样性。当测试模型架构与代理模型差异较大时（如从 Stable Diffusion 迁移到 DeepFloyd-IF），攻击成功率会下降约 40%。此外，由于需要维护多个代理模型并进行二阶梯度计算，其训练成本较单模型方法增加 2-3 倍。

3）对比分析

与 Anti-DreamBooth[12] 等传统方法相比，MetaCloak 通过元学习机制将对抗扰动对高斯滤波的鲁棒性从 52% 提升至 85%。但如表1.4所示，其计算开销显著增加，在 RTX 3090 上单图像扰动生成时间达到 8 秒，不利于实时应用。这种权衡关系揭示了元学习方法在通用性和效率之间的矛盾。

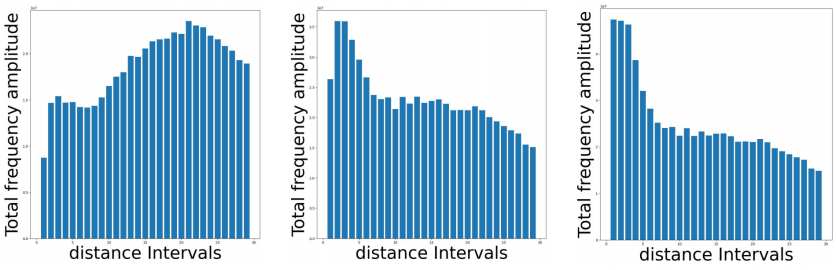
表 1.4　元学习方法与传统方法对比

方法	数据变换鲁棒性	单图像计算时间 (ms)
Anti-DreamBooth[12]	52%	120
MetaCloak	85%	8000

（3）基于时间步优化与特征干扰的主动防御方法

1）工作原理

扩散模型的时间步特性为对抗扰动生成提供了新的优化维度。不同时间步对扰动的敏感性存在显著差异：低时间步 $(t \to 0)$ 主导高频细节的生成，而高时间步 $(t \to T)$ 主要控制低频结构 [16]。这一发现促使研究者开发自适应时间步选择算法，将有限的扰动预算集中在最有效的时间区间。



t ∈ [0, 100]　　**t ∈ [400, 500]**　　**t ∈ [700, 800]**

图 1.4　SimAC 的时间步敏感性分析

2）代表性工作

SimAC[16] 通过系统实验验证了时间步选择的关键作用。如图1.5所示，该方法采用贪婪算法搜索最优时间区间 $[t_{min}, t_{max}]$，发现当 $t \in [20, 200]$（总步长 1000）时扰动效果最佳。在此基础上，SimAC 设计了特征干扰损失 $\mathcal{L}_{\text{feat}} = \|f_{\text{dec}}(x_t) - f_{\text{dec}}(x_{\text{adv},t})\|_1$，专门针对 UNet 解码器的高层特征进行攻击。在人脸身份保护任务中，该方法将 AdvDM 的身份相似度分数从 0.38 降至 0.12，同时将扰动计算量减少 60%。

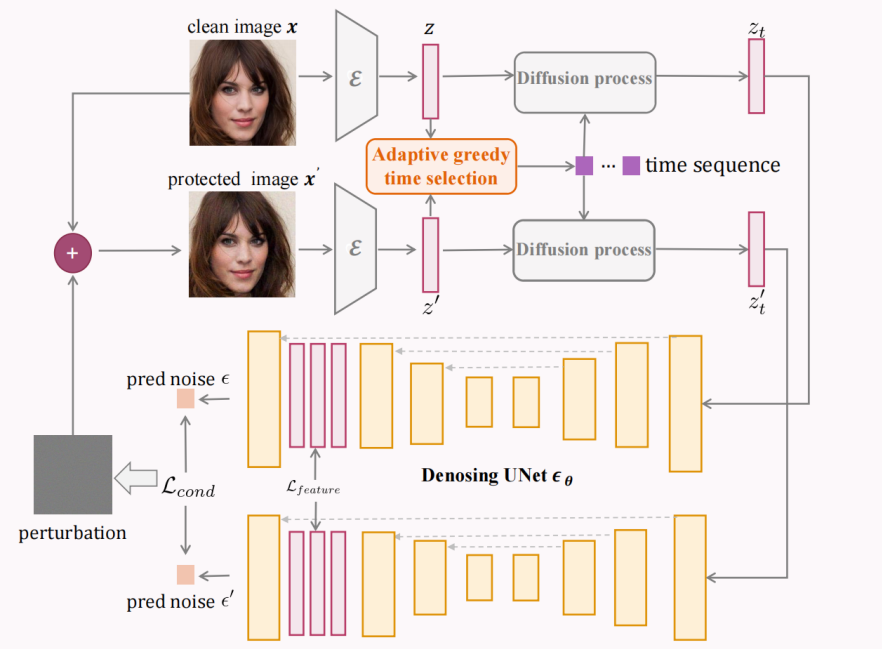


图 1.5　SimAC 架构图

3）对比分析

SimAC 的创新性体现在解决了传统方法随机时间步采样的低效问题。如表1.5所示，相比 AdvDM 需要在整个时间轴上计算梯度，SimAC 通过聚焦关键时间区间，在保持同等攻击效果的情况下将计算时间从 210 秒缩短至 85 秒。这种时序优化策略为后续研究提供了重要启示：对抗扰动不需要均匀分布在所有生成阶段，而应集中攻击模型最敏感的时间窗口。

表 1.5　时间步优化方法性能对比

方法	身份相似度 (↓)	计算时间 (s)
AdvDM	0.38	210
SimAC	0.12	85

（4）基于编码器攻击的主动防御方法

1）工作原理

近期研究揭示了扩散模型中不同组件对对抗扰动的敏感性差异。如 [17] 所示，编码器部分（如 VAE）的计算瓶颈和脆弱性特征使其成为更高效的攻击目标，而直接攻击去噪模块（UNet）则需承担更高的计算代价。这一发现推动了潜在空间扰动生成方法的发展，通过避开像素空间的端到端梯度计算来提升效率。

潜在空间与像素空间扰动效率对比

2）代表性工作

Score Distillation 方法 [17] 提出了创新性的双路径攻击策略。如图1.6所示，该方法首先将输入图像编码到潜在空间 $z = \mathcal{E}(x)$，随后通过评分蒸馏采样（SDS）优化扰动：

$$\mathcal{L}_{\text{SDS}} = \mathbb{E}_t \left[\|\epsilon_\theta(z_t, t) - \epsilon\|^2 \frac{\partial z}{\partial \delta} \right]$$

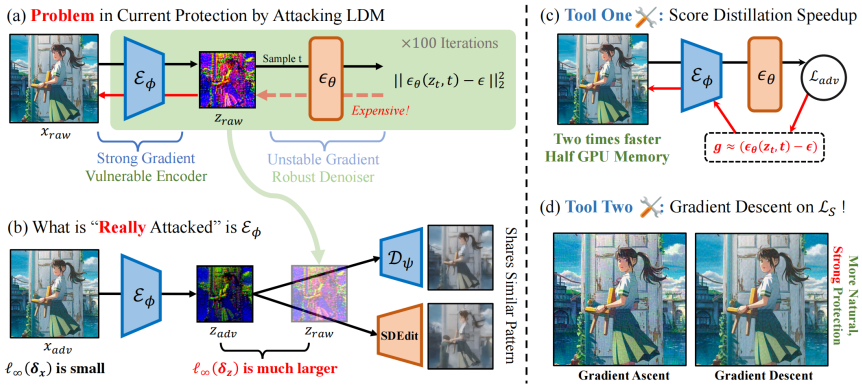


图 1.6　Score Distillation 流程图

其中关键创新包括**最大化语义损失路径**与**最小化语义损失路径**：前者通过引导生成混沌输出以破坏内容连贯性，后者则生成视觉自然的模糊扰动以提升隐蔽性。

实验数据显示，该方法在保持与 Mist 相当的攻击成功率（85% vs 87%）的同时，将单图像处理时间从 210 秒缩短至 105 秒，效率提升 50%。此外，潜在空间扰动的人类感知评分（JND）达到 4.2 分（满分 5 分），显著优于像素空间方法的 3.1 分。

（5）对抗扩散模型模仿攻击的方法总结

现有主动防御方法主要包括四类：**基于对抗样本的方法**（如 AdvDM、Anti-DreamBooth[12] 与 Mist），通过扰动破坏模型的条件提取或微调过程，但存在计算开销大或对数据变换不鲁棒的问题；**MetaCloak 方法**引入元学习框架，利用多代理模

型生成可迁移扰动，并通过变换采样增强对数据变换的鲁棒性，缓解了前者的泛化能力弱点；**SimAC 方法**结合扩散模型的时序特性，聚焦有效时间步与高层特征，通过自适应时间选择和特征干扰显著提升了人脸等身份类内容的攻击效率；**Score Distillation 方法**则从编码器脆弱性出发，以潜空间扰动和评分蒸馏降低计算成本，并结合**最大化**与**最小化语义损失路径**设计双重攻击策略，兼顾**隐蔽性**与**效率**，系统扩展了对抗扰动的设计空间。

表 1.6　不同防护方法对比分析

方法名称	核心原理	优势	局限
AdvDM[11]	添加通用扰动，破坏扩散模型生成能力	首个通用防护框架，具备防御性	计算开销大，迁移能力弱
Anti-DreamBooth[12]	针对肖像定制模型，利用特征空间攻击	有效阻止特定个体再现	对图像变换不鲁棒，泛化弱
Mist[13]	使用 UAP 方法优化跨数据迁移	提升迁移性，适用于多模型	需要多目标优化，训练复杂
Metaloak[15]	利用元学习生成模型无关扰动	黑盒攻击适用性强，提升鲁棒性	高度依赖代理模型，训练时间长
SimAC[16]	使用重构一致性，提升抗定制能力	不依赖原始模型结构，适配性好	鲁棒性依赖于训练策略
Score Distillation[17]	时间步数控制下优化攻击目标	增强身份扰动效果，优化效率高	效果依赖扩散模型类型

1.3 技术挑战

在构建面向 AI 图像编辑攻击的主动防御系统过程中，现有的主动防御技术还存在以下技术挑战：

1） **跨模型迁移性弱**：现有对抗扰动多数仅在特定生成模型（如 Stable Diffusion 或 DALL · E）上有效，当攻击目标模型发生变化时（例如换成不同架构或训练数据的模型），扰动往往失效，缺乏普适防护能力，限制了其在多平台部署下的应用价值。

2） **生成速度慢**：许多防御方法依赖逐图优化，即为每张待保护图像单独计算扰动（如 PGD 或优化式 UAP），导致生成速度较慢，无法满足图像平台或内容创作者对实时性和批量处理的需求，阻碍了在生产环境中的落地使用。

3） **图像破坏严重**：为了提升防护强度，部分方法会施加较大幅度扰动，导致肉眼可见的图像失真，尤其在高频区域易引起视觉伪影，降低图像的可用性与审美质量，难以被用户接受，也影响平台内容的呈现效果。

4） **抗压缩性差**：在实际发布场景中，图像通常会被平台进行压缩处理（如 JPEG 重编码），这会严重削弱扰动的防护能力，使其在用户端已失效。因此，缺乏对图像压缩、缩放、裁剪等常见操作的鲁棒性处理，是现有方法普遍面临的问题。

针对上述技术挑战，我们提出的 VisiFence 系统在设计上引入了多目标损失函数、通用扰动生成机制、Prompt 语义对抗策略与鲁棒性增强方法，从而在保障图像质量的前提下，有效提升了对多种 AI 编辑攻击的通用防护能力。

1.4 作品特色

本项目 VisiFence 旨在应对当前 AI 图像编辑技术所带来的潜在内容安全问题，提出了一套从图像发布源头主动防御 AI 恶意编辑的系统方案，具有如下三项主要技术特色：

1）主动式 AI 编辑防御机制

传统图像水印等内容保护方法多属于被动检测或后处理修复范畴，难以应对由扩散模型驱动的深度编辑操作。为此，VisiFence 首次从“图像主动防御”角度出发，设计了面向 AI 编辑攻击的扰动注入机制，通过在图像发布前嵌入细粒度、难以感知但可有效干扰主流 Diffusion 编辑路径的特定扰动，从而实现“以 AI 防 AI”的防护目标。

该机制可广泛适配 Stable Diffusion 系列编辑器，在不修改原图内容的前提下，系统性地破坏编辑模型的条件匹配机制。实验数据显示，在 Training-Free 的 img2img[20] 场景中（保护模型 v15+ 编辑模型 v21），VisiFence 所生成图像的 FID 值达 **304.59**，较原始 clean 样本 **107.51** 提升 **183%**，表明其对生成图像结构的强破坏性，同时 clip-s 得分降至 **25.17**，较 clean 下降 **26%**，显著削弱语义一致性，说明内容难以被模型准确编辑，防护目标得以实现。

2）良好的通用防护能力

针对现实中攻击者使用的 AI 编辑模型、编辑 prompt 难以预测的问题，VisiFence 聚焦于 Stable Diffusion 系列模型中普遍存在的 VAE 编码器与 U-Net 注意力结构，提出通过最大化原图与保护后图像在潜空间中的向量差异，并干扰注意力层的注意力分布，促使模型失去对目标语义区域的聚焦能力，从而实现对不同 prompt 和模型架构的一致性防护。

测试结果显示，该机制在多个典型编辑任务中均取得显著防护效果。以 Training-Based 的 DreamBooth[18] 场景为例，VisiFence 的 FID 高达 **295.58**（v15 保护 +v21 编辑），相比 AdvDM 的 **164.54** 提升近 **80%**，clip-s 从 **35.65** 降至 **24.23**，显示其对目标语义引导训练路径的高效干扰。同时，针对 Textual Inversion[19] 任务，VisiFence 在跨模型（v21 编辑 +v15 保护）场景中 FID 稳定保持在 **269+**，远超传统方案，具备良好跨模型迁移能力和任务鲁棒性。

3）强健的鲁棒性保障

考虑到图像在传播过程中常经历压缩、裁剪、加噪等处理，VisiFence 在扰动优化阶段引入仿真增强策略，在每次扰动更新中加入随机裁剪、JPEG 压缩、Gaussian 噪声等多重扰动，使优化后的扰动具备天然抗变形能力，从而提升实用性与部署可靠性。

在鲁棒性测试中，VisiFence 在 img2img[20]+v15/v21 交叉场景下即便遭受 **85% JPEG 压缩**，其 FID 依旧保持在 **280.96** 以上，仅下降不足 **8.5%**，clip-s 为 **27.54**，防护效果几乎无明显衰减；而在高斯噪声及裁剪条件下，VisiFence 扰动的 SSIM 指标甚至小幅提升，说明其与原图特征高度融合，不仅不易感知，且难以被破坏。

综上所述，VisiFence 不仅实现了从源头遏制 AI 图像恶意编辑的创新防御路径，还在实际应用中体现出高度的通用性与鲁棒性，具备良好的工程实用潜力。

1.5 应用前景分析

在数字化和人工智能快速发展的时代背景下，我们团队开发的 VisiFence 系统，作为面向 AI 图像恶意编辑的通用鲁棒主动防御方案，不仅是图像内容保护的创新技术，更是针对图像隐私侵犯、版权篡改及安全威胁的强大防线。该系统能够有效遏制对图像的恶意篡改与非法利用，极大提升内容创作者、品牌方及平台的安全防护能力。具

体应用前景主要体现在以下几个方面。

1）社交媒体与内容平台的安全保障： 社交媒体平台每日承载海量图像内容，如何防止内容被非法恶意编辑，尤其是深度伪造与篡改，是保障平台生态健康的关键。VisiFence 支持在线即插式通用保护与定制化离线保护两种模式，能够无缝集成到平台上传和发布流程中，实现对海量用户上传内容的实时保护，减少假图、篡改图对平台公信力的冲击，同时保护原创作者权益，维护良性创作环境。

2）新闻媒体与舆论监督的真实性保障： 新闻媒体作为信息传播的关键节点，其发布的图像真实性直接关系到公众对信息的信任度。随着 AI 编辑技术的滥用，假新闻配图、篡改图片事件频发，严重影响舆论导向和社会稳定。VisiFence 系统可以嵌入新闻机构的内容制作流程，对图像进行主动防护，确保图像即便被二次编辑，仍难以通过 AI 模型复原或篡改，从源头提升新闻图像的可信度和溯源能力。

3）品牌视觉资产的版权保护： 在电商、广告、媒体等行业，品牌的视觉资产是核心竞争力之一。随着 AI 图像编辑技术的普及，盗用、篡改品牌图片以假乱真、进行虚假宣传的风险剧增。VisiFence 通过在图像中嵌入针对 AI 编辑模型设计的主动防护扰动，使得即使图像被非法获取和修改，编辑后的结果也无法保持品牌原有的视觉效果和语义，从而有效防止品牌资产被恶意篡改与滥用，保障品牌形象的真实性和可信度。

4）数字版权管理与法律合规： 随着数字版权保护法规的完善，图像版权保护的合规性要求日益严格。VisiFence 提供的技术方案不仅保护图像免受恶意编辑，还便于版权方进行图像内容的认证和溯源，满足版权监管和法律诉讼的需求。该系统的鲁棒性和通用性使其具备作为数字版权保护技术基础设施的潜力，为版权保护领域提供可靠技术支撑。

综上所述，VisiFence 作为面向 AI 图像编辑的主动防御系统，具备多场景应用价值，不仅保护内容原创者和平台的合法权益，还促进数字生态环境的健康和可信。其技术创新和产品形态为未来图像安全与版权保护树立了新标准，具有广阔的市场前景和深远的社会意义。随着数字化进程的不断加速，VisiFence 有望成为推动数字视觉内容安全的核心技术之一，助力构建可信、合规、健康的数字内容生态体系。

第二章 作品设计与实现

2.1 作品介绍

随着 AI 图像编辑技术的广泛应用，图像隐私泄露、内容伪造与版权窃取等安全问题日益突出。本项目研发的 **VisiFence 系统**，旨在提供一种面向公众用户与平台方的 **图像主动防护解决方案**，通过在图像发布前插入微小扰动，实现对主流 AI 编辑模型的有效干扰与防御。

本作品的应用方式和效果如图 2.1 所示。本作品为用户提供抗 AI 编辑的主动防御服务。用户为了防止其内容被 AI 编辑篡改，可以将其图像提交给我们的 VisiFence 服务平台，该平台为用户提交图像进行防御保护。用户可将保护后图像发布于各类主流社交平台。普通用户可以看到高保证的图像内容，当恶意用户对保护后图像使用 AI 编辑工具进行恶意合成和篡改时，无法获得正常的高质量图像，从而主动防止 AI 恶意编辑篡改行为的发生，从源头保障了图像内容的真实性。**本作品尤其适用于对重要人物，重要事件图像内容的保护，保障网络空间中内容的真实性**。

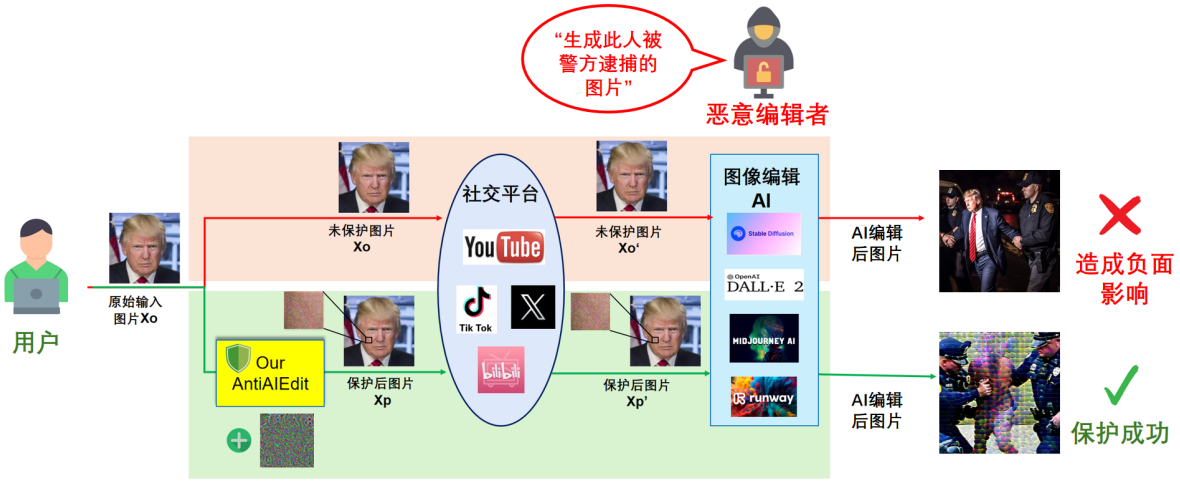


图 2.1　VisiFence 应用示意图

系统具备 **四重迁移能力——跨模型、跨方法、跨指令与跨数据**，可适配多种生成模型（如 Stable Diffusion v1.5/v2.1）、不同编辑方式（如 inpainting、img2img[20]、DreamBooth[18] 等）与图像来源场景，显著提高防护的广谱性与适应性。同时，系统对图像常见预处理操作（如压缩、裁剪、加噪等）具有较强的鲁棒性，保障扰动在真实发布流程中的持续有效性。

为提升用户使用便捷性与可控性，项目团队设计并开发了一个基于 Web 平台的图像防护服务网站。用户可通过平台上传待保护图像，自定义扰动强度与防护目标，实时预览扰动前后的效果，并一键下载防护图像，满足不同场景下的个性化防护需求。平台已集成优化后的通用扰动模型，单图处理时间低于 1 秒，支持大规模图像的高效批量防护。

2.2 作品功能

VisiFence 系统面向图像内容创作者与平台管理方，提供一套集成化、可配置、具备四重迁移能力的图像主动防护服务。其核心功能围绕图像保护、防护测试与自适应优化三个方面展开，致力于在不影响图像使用价值的前提下，有效对抗主流 AI 图像编辑工具的未授权篡改行为。

1） **图像保护功能：**系统支持用户上传任意图像，并通过扰动嵌入算法为图像添加不可见水印，扰动兼顾隐蔽性与防御强度，可在不引发肉眼可感差异的前提下，显著干扰 AI 模型的编辑结果。系统提供在线与离线两种保护模式，满足实时处理与高精度保护的不同需求。

2） **多维参数控制：**用户可自定义多项扰动参数，包括梯度步数、学习率、扰动强度上限等基础项，以及多目标损失函数中的感知一致性、特征偏移等权重项。该设计增强了系统的可控性与适应性，便于在不同使用场景下灵活部署。

3） **四重迁移能力：**系统可在跨模型（不同架构或版本的扩散模型）、跨方法（inpainting、img2img、DreamBooth 等）、跨指令（不同 prompt 语义）及跨数据（不同来源与压缩状态图像）场景中持续保持防护效果，保障编辑干扰能力的稳定性与广谱性。

4） **编辑测试功能：**系统内置仿真攻击模块，支持用户选取不同编辑模型、方法与指令，对保护图像与原始图像分别执行 AI 编辑操作，并展示其结果对比与指标评估。该功能用于辅助用户检验当前扰动配置的防护效果。

5） **自适应反馈优化：**系统支持在测试完成后，根据扰动编辑效果对图像质量与模型响应的反馈进行进一步参数优化。通过迭代调整，使最终扰动在保证图像保真度的同时达到最优防护水平。

6） **批量图像处理：**为满足高频使用需求，系统支持用户上传图像批次，并依照设定

参数自动完成扰动生成与图像防护任务，适配视频帧图像、水印大规模嵌入等应用场景。

7） **用户服务界面：**系统提供简洁直观的 Web 界面，集成注册登录、图像上传、参数设置、保护生成、效果测试等一站式服务，优化用户体验，便于快速上手和多次迭代使用。

2.3 系统功能界面

本系统的 Web 应用首先展示了作品的整体介绍，包括作品概述、关键技术说明和特色功能等部分，如图2.2所示，方便用户第一时间对网站的核心功能进行熟悉。用户可以通过点击右上角的登录或注册按钮进行账号操作。

图 2.2　网页初始页面

本节将对用户注册、保护设置、效果测试等核心功能页面进行详细介绍，并简要概述本系统的开发环境。

2.3.1 用户注册账号

用户注册账号时，需要提交用户注册邮箱，并且设置账号密码，设置完成后重新输入账号密码进行确认。如图2.3所示，用户点击注册后系统将自动跳转至用户注册页面。

2.3.2 图像保护

2.3.2.1 保护参数设置

用户点击图像上传后，如图2.4选择对应图像文件上传至系统。

图 2.3　网页注册页面

图 2.4　图像上传页面

2.3.2.2 保护参数设置

用户点击参数配置后，如图3.3可以调整添加水印过程中的参数。在此页面中，用户需要设置基础参数与权重参数两种类型参数。基础参数主要涉及梯度计算次数、学习率和扰动强度限制，用于控制编辑过程中优化方向和扰动幅度；权重参数主要涉及像素损失权重、分布差异权重、UNet 对抗权重和 UNet 注意力权重，用于控制编辑后图像的不同图像特征（像素、特征分布、模型相关特性）维持与改变程度。如图所示，用户根据每个参数下的系统提示进行参数设置，调整保护效果。调整完成后，点击保存参数，系统将暂存用户参数设置。

2.3.2.3 保护模式选择以及保护效果展示

用户选择保护展示后，首先选择系统保护模式，分为在线模式和离线模式，如图3.4系统将会跳转至如图所示的页面，用户点击开始编辑后，系统将会按照前一步用户设置的参数进行图像保护，用户可以通过进度条颜色变化以及进度条中央数字查看编辑进度。编辑完成后，系统将会同时展示保护前后图像，供用户查看。

图 2.5　参数配置页面

图 2.6　保护效果展示页面

2.3.3 反馈优化流程

2.3.3.1 编辑参数配置

用户点击效果测试后，用户可以根据自身可能面临的现实未授权编辑编辑场景配

置相关参数，模拟相应恶意编辑场景。本系统充分考虑现实情境中可能存在的各种恶意编辑场景，用户可以调整编辑方法、编辑指令以及选择不同编辑模型，模拟各种现实编辑情况。编辑方法包含两大类方法：基于训练的编辑和无训练编辑。其中基于训练的编辑下面包含 Textual Inversion[19] 和 DreamBooth[18] 两种编辑场景，无训练编辑下面包含 img2img[20] 和 inpainting[21] 两种编辑场景。如图 2.7 所示。

图 2.7　编辑参数配置页面

2.3.3.2 编辑效果展示

用户完成测试过程参数设置后，点击开始测试，如图2.8系统将按照用户测试参数模拟图像恶意编辑。测试完成后，系统将同时生成保护前后图像经过恶意编辑后的图像，展示在图像生成框中。用户可以根据生成的两张图像编辑后的对比效果评判系统保护效果，同时系统也将生成编辑前后量化指标，评判恶意编辑是否成功。用户完成测试后可以进一步优化编辑参数设置，反复调整直至满意为止。

2.3.4 批量处理流程

本系统支持图像批量上传，形成动态图像等待队列，按照时间顺序依次经过系统保护编辑。如图2.9所示，用户可以同时选择多个图像上传，形成动态等待队列。便于用户对于视频形式内容进行批量保护操作。

2.3.5 系统开发环境

操作系统：Ubuntu 18.04.5 LTS

开发工具：Visual Studio Code

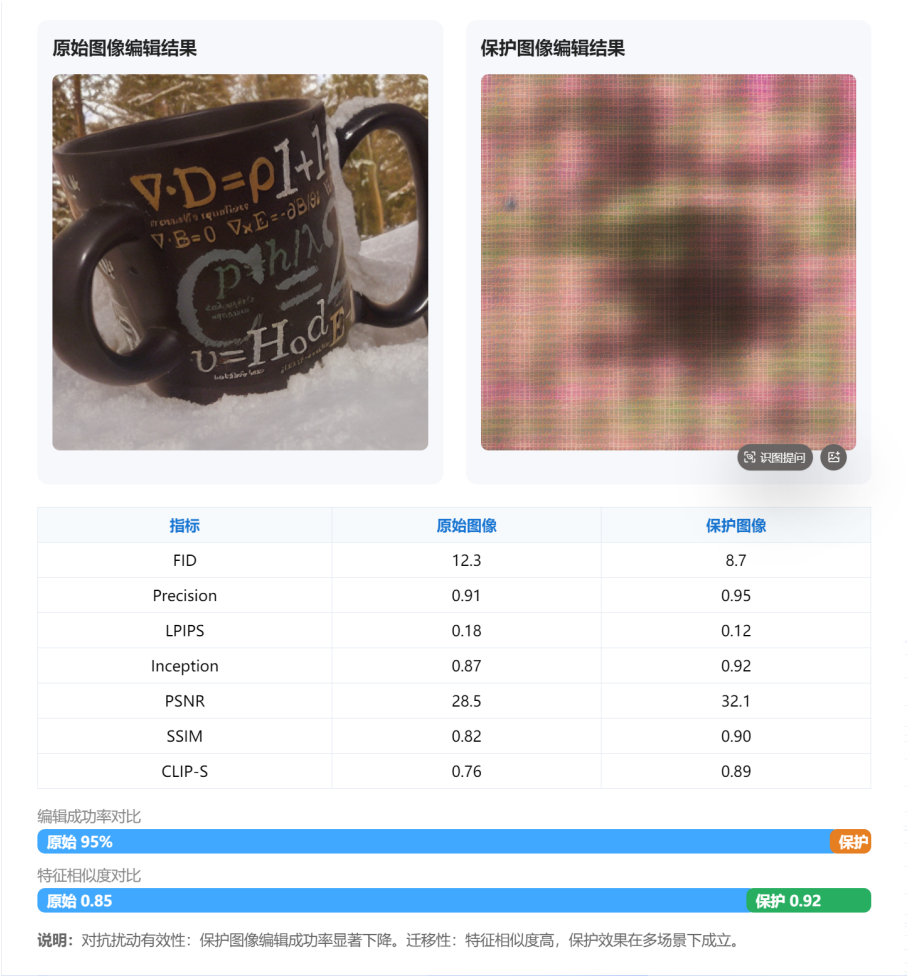


图 2.8　编辑效果展示页面

图 2.9　批量处理页面

开发语言：Python，JavaScript

开发依赖：React，FastAPI，PyTorch

2.4 系统操作流程

本小节将介绍系统的整体操作流程，主要分为如下 4 个部分：1）用户注册账号并登录。2）用户选择系统图像保护功能模块进行图像保护。3）完成保护操作后测试保护效果，进一步优化保护过程中参数设置。4）用户得到满意参数设置后批量处理图片。下面是对系统操作流程的具体说明：

（1）注册登录流程

填写相关个人信息，设置账户用户名和账户密码以便后续使用 VisiFence。

（2）图像保护流程

i. 用户进入后选择系统保护模式：离线模式或者在线模式。在线模式下用户可以实时得到保护图像，无需等待，适用于直播、视频的图像保护等具有较强时效性要求的保护场景；离线模式下系统采用更细粒度的保护方法，用户无法立即得到保护结果，需要等待一段时间后才能得到保护图像，适用于艺术画作、摄影图像等具有细粒度级保护需求的场景。

ii. 用户上传需要添加水印的图像，系统随后将其存入后台相应个人账户的历史数据中。

iii. 用户设置保护过程中的参数，设置完成之后保存参数，之后进入保护环节。

iv. 用户确认是否开始保护，如果确认，系统将在保护流程完成之后将原始图像和保护后图像一同展示给用户。

（3）反馈优化流程

i. 用户根据自身可能面临的恶意编辑场景设置编辑参数。VisiFence 系统可以为用户提供多种参数选择，模拟现实场景中各类恶意编辑情况。例如，用户可以调整编辑方法、编辑模型和编辑指令等参数，等效模拟现实中存在的未授权恶意编辑情境。

ii. 用户完成参数设置后，确认是否开始编辑测试。如果确认，系统将按照用户设置的参数进行恶意编辑场景模拟。

iii. 系统完成编辑测试后，将同时生成用户原始图像以及保护图像经过恶意编辑后的两份修改图像。系统将比较编辑后图像与编辑前图像之间风格差异等量化指标，评判恶意编辑对于图像修改程度，进一步评判保护效果，为用户提供有效反馈。

iv. 用户根据系统反馈进一步优化保护过程中参数设置，直至保护效果令用户满

意为止。

（4）批量处理流程

用户如果需要同时对于多张图像进行保护操作，或者用户希望对于直播、视频等进行图像保护，可以根据用户自身需要选择批量处理功能，系统将按照图像编辑过程中用户保存的参数对用户上传图像列表中所有图像进行编辑。

2.5 系统架构

本系统是基于 React+FastAPI 开发的 Web 应用，系统的架构分为用户界面层、图像保护层和反馈优化层三个层次。系统整体架构如图2.10所示。

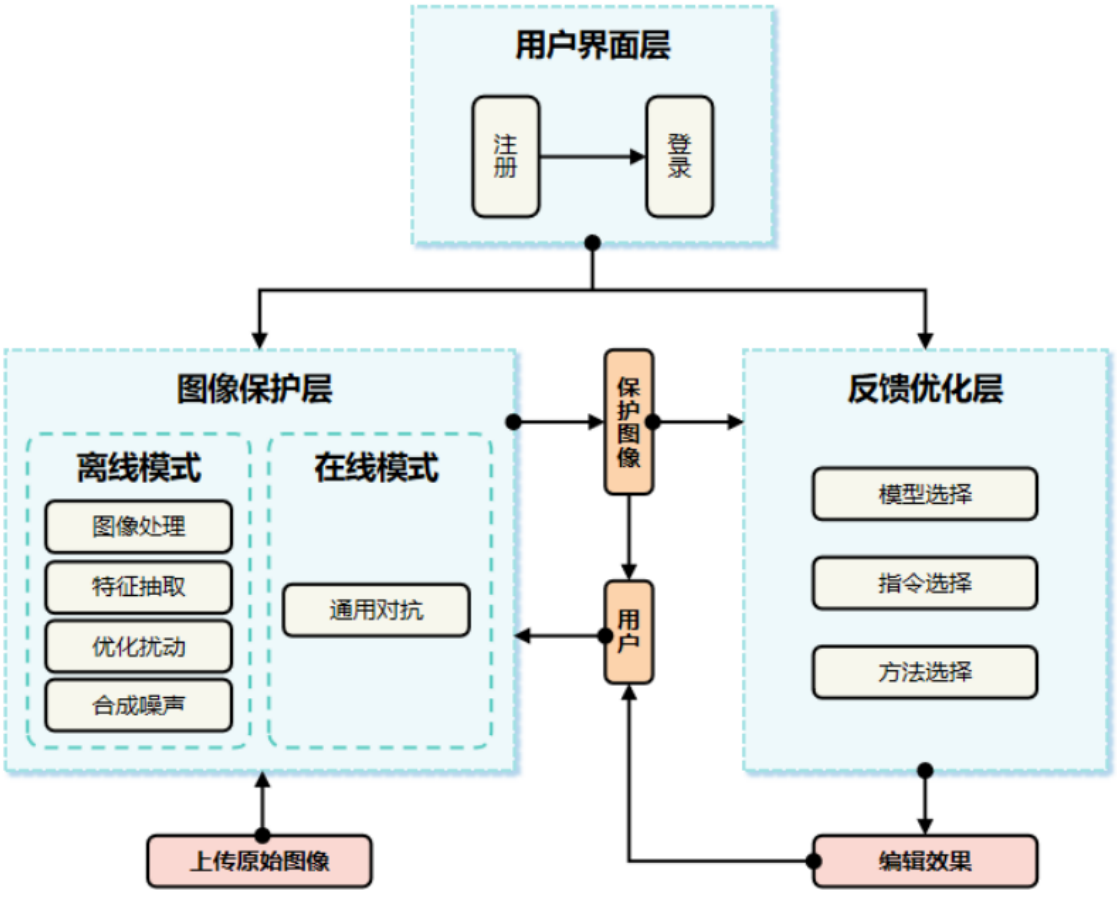


图 2.10　VisiFence 系统总体架构图

（1）用户界面层

用户界面的主要功能是为使用者提供优质的交互体验。界面的设计应符合目标用户的审美和需求，并提供良好的反馈和提示响应。为此，我们选择了 React 作为 UI 界

面框架，并结合 Ant Design 的设计组件，构建了一个美观且流畅的 Web 应用。

在本系统中，用户进行注册后才能够使用本系统的完整功能，例如用户可以登录账号后使用历史记录避免图像重复编辑浪费时间。通过用户界面层，使用者可以选择相关功能，分别对应于图像保护层和反馈优化层。在**图像保护层**中，用户可以选择离线或者在线保护模式，进一步设置添加水印过程中相关参数，控制保护过程，实现个性化保护过程；在**反馈优化层**中，用户可以对于保护后图像进行恶意编辑测试，根据使用场景模拟相关恶意编辑行为。户可以根据测试效果进一步调整图像保护过程中参数设置，优化保护效果。

（2）图像保护层

图像保护层是 VisiFence 系统的核心功能之一，将对用户上传后的图像添加水印进行保护。本系统提供离线和在线两种保护模式。离线模式对用户输图像进行定制化处理，生成定制化扰动添加在原有图像上，从而确保图像不易被篡改；在线模式则利用通用对抗扰动（UAP）技术，向用户输入图像添加提前生成好的通用性扰动，从而在保证质量的前提下快速保护输入图像。用户可以根据需求选择模式并调整参数，确保图像安全性。

（3）反馈优化层

反馈优化层是 VisiFence 系统的另一个核心功能实现层。用户完成图像保护操作后，通过设置相关编辑参数模拟现实情境中可能存在的恶意编辑场景，采取具有针对性的防御策略以及个性化的保护方案。用户可以模拟潜在的恶意编辑场景，选择不同的编辑模型、编辑指令以及不同编辑方法，从而实现对于现实情况下未授权恶意编辑场景的全覆盖。VisiFence 系统将根据用户的编辑测试参数设置调入相关编辑模型进行恶意编辑效果测试。用户可以根据编辑反馈结果进一步调整图像保护层中参数设置，优化保护效果，直至保护效果让用户满意为止。

2.6 系统技术设计

为完成以上目标，我们提出一种融合注意力攻击、语义偏离、VAE 重构及共性扰动生成的整体技术框架，旨在从多个维度系统性对抗 AI 图像编辑操作。

首先，为实现**跨模型迁移**，我们提出攻击 Stable Diffusion 系列模型中普遍存在的

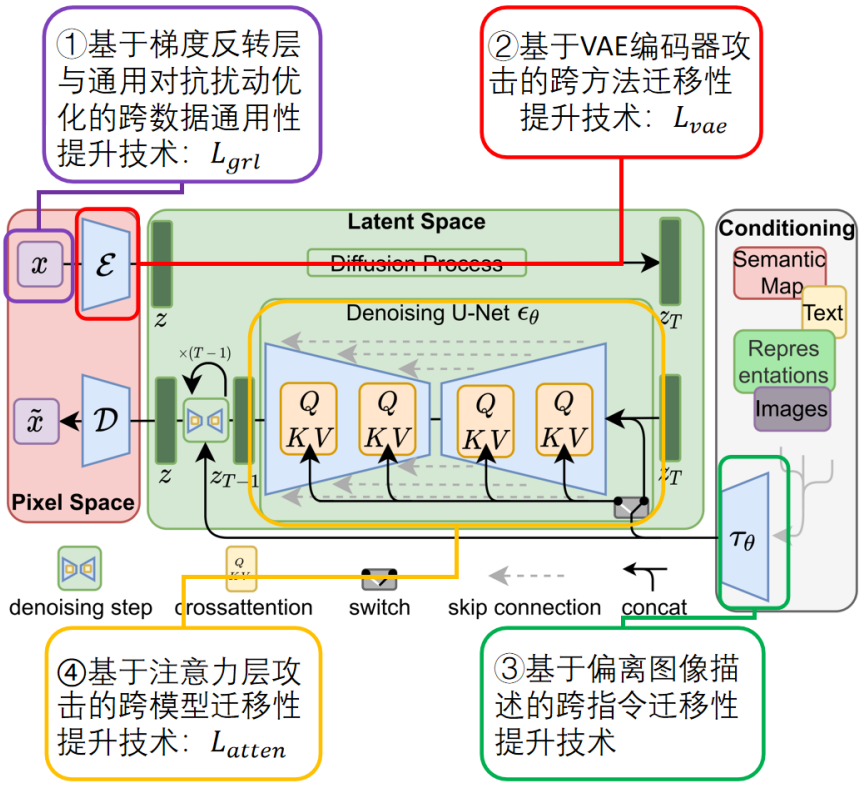


图 2.11　基于四维防护的通用抗 AI 编辑技术总体框架图

U-Net 注意力结构。通过最大化注意力图的熵与变差，以及偏离参考注意力图，显著削弱模型对 Prompt 的定位能力，实现对多个编辑模型的一致防护。

其次针对**跨指令迁移**问题，我们设计了一种语义偏离策略，使编辑前图像与保护性描述（Prompt）之间的语义相似度最小化。该策略确保扰动对不同 Prompt 保持稳定防护效力。

同时，为了实现**跨方法迁移**能力，我们构建了基于 VAE 的共性扰动生成机制。在重构图像的同时，通过联合多个编辑器定义的损失函数指导潜在变量空间向“通用干扰方向”靠拢，从而使扰动能同时破坏多种编辑方法。

最后，对于**跨数据迁移**的实现，我们在图像添加水印前施加数据增强操作，并引入梯度反转层（GRL）模拟“域间转移”，提升扰动在不同图像变换（如翻转）下的稳定性。此外，我们引入了**通用对抗扰动**方法。通过在训练集上联合优化扰动，生成可被实时应用于任意图像的统一扰动向量，显著增强在线场景下的编辑干扰效果。

综上所述，我们构建的多层次防护策略涵盖模型结构、编辑指令、方法类型以及通用增强四个维度，为实现面向现实编辑系统的普适性防护提供了完整的技术体系，

分别体现保护效果的四个优势：跨模型迁移性、跨方法迁移性、跨指令迁移性以及跨数据迁移性。

2.6.1 基于注意力层攻击的跨模型迁移性提升技术

1) 现实背景与现有防御方法的局限性分析

在数字化内容保护领域，非授权编辑防护面临的核心挑战之一是模型间的结构差异导致保护机制的失效。随着人工智能生成内容技术的迅速发展，模型在架构设计、训练数据和应用场景方面呈现出显著多样性，传统的保护方法往往只能针对特定模型生效，难以在不同平台或模型间实现泛化。因此，提升保护策略的跨模型迁移能力成为破解这一难题的关键方向。

当前针对 AI 图像编辑的防御技术主要遵循两类范式：面向特定模型架构或任务设计的**定制化防御方案**，和基于通用对抗扰动的**跨模型防御尝试**。然而，两类方法在实际跨模型应用场景中均存在显著缺陷：

（1） **定制化防御的泛化能力不足**代表性研究（如 Anti-DreamBooth[12]、Mist）通常针对特定模型的训练机制（如 DreamBooth[18] 的微调流程）或结构组件（如 Stable Diffusion 的条件编码器）实施干扰。此类方案在目标模型变体（如 SDXL、不同 tokenization 机制的模型）或异构编辑架构上表现出严重的性能退化，本质上受限于其模型依赖性。

（2） **通用对抗扰动的跨模型迁移效率低下**基于 UAP、AdvPaint 等方法生成的像素级全局扰动，缺乏对编辑模型内部机制的针对性建模。当遭遇不同注意力机制、提示词编码策略或扩散步长配置时，其防御效果呈现显著波动，根源在于未能解耦模型内部的关键编辑路径。

（3） **对共性架构特征的利用缺位**现有方案普遍忽略主流编辑模型（扩散模型/图像补全模型等）共享的核心机制——即通过 self-attention 与 cross-attention 层实现条件控制与特征融合。由于未系统性地利用该结构共性，当前防御策略难以精准破坏跨模型的编辑决策路径，导致泛化防御能力受限。

为提升跨模型迁移性，我们注意到：在 Stable Diffusion 系列模型中，其核心的 U-Net 结构和注意力机制高度相似。这一结构相似性为对抗扰动的迁移提供了可能性。

基于此，我们提出通过攻击替代模型的注意力机制来干扰模型的编辑能力，进而实现对多个模型的统一防护。

2）技术实现

我们使用 **AdvPaint**，一个用于保护图像免受未授权修复任务攻击的防御框架。AdvPaint 生成专门设计的对抗扰动，以欺骗基于扩散模型的图像修复模型，破坏其正确关注图像区域的能力。这些扰动旨在从根本上破坏模型的图像生成能力，通过同时干扰交叉注意力与自注意力机制实现。当其他具备相似 U-Net 结构和注意力机制的 Stable Diffusion 模型尝试对扰动图像进行处理时，这种注意力图的破坏效果依然存在，从而使得模型难以准确理解图像语义并根据提示词进行有效编辑，实现跨模型的保护效果。

U-Net 去噪器 ϵ_θ 中的注意力机制会通过去噪过程，从潜在向量中重建图像。如图 2.12 所示，每一个注意力模块（残差、自注意力与交叉注意力组成）在每次下采样和上采样操作后重复。自注意力模块包含三个部分：query q、key k 和 value v，它们都是输入图像的线性变换。而交叉注意力块的 k 与 v 来自于外部条件输入（如 prompt）。每一层 l 中，q 与 k 被乘以后形成注意力图 $\mathcal{M}$，再加权 v 以得到输出 $\mathcal{A}$：

$$\mathcal{M} = \text{Softmax}\left(\frac{qk^T}{\sqrt{d}}\right), \quad \mathcal{A} = \mathcal{M} \cdot v. \tag{2.1}$$

其中 d 是 q 与 k 的维度。为对抗这类修复型扩散模型，我们提出了新的优化目标函数来制造对抗扰动，旨在同时干扰交叉注意力与自注意力机制。

3）交叉注意力攻击

为攻击交叉注意力模块，我们设计损失函数以扰乱 prompt token 与其空间位置之间的对齐。我们的目标是最大化干净图像 x 与其对抗扰动版本 $x+\delta$ 的 query q 之间的差异，从而破坏交叉注意力机制。令 $q = Q(\phi(x))$，其中 $\phi(\cdot)$ 表示来自前一层的特征表示，$Q(\cdot)$ 是线性投影操作。我们定义损失函数函数 $\mathcal{L}_{\text{cross}}$ 为：

$$\mathcal{L}_{\text{cross}} = \sum_l \left\| Q_c^l(\phi(x+\delta)) - Q_c^l(\phi(x)) \right\|^2, \tag{2.2}$$

其中 l 表示去噪器中的层数，c 表示交叉注意力模块。该损失将 $x+\delta$ 的 query 向

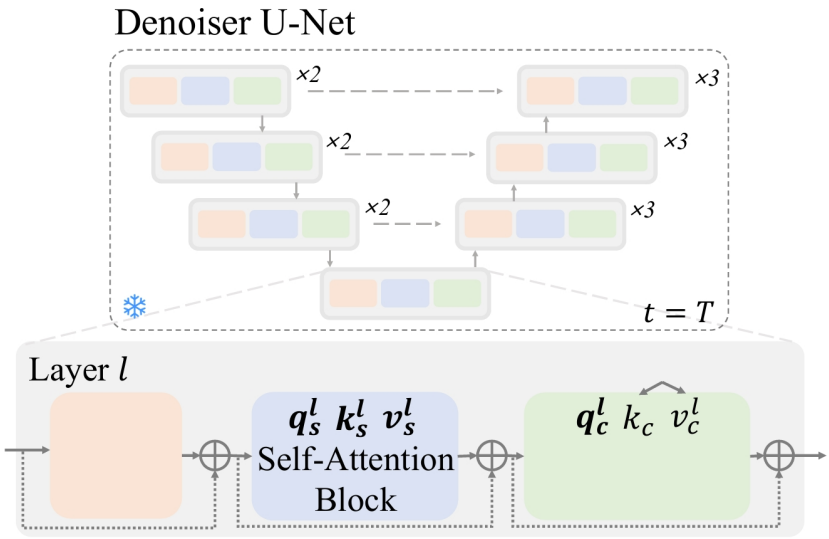


图 2.12　U-Net 去噪器中的注意力机制。

prompt 的 key 与 value 脱离，从而扰乱对 prompt 条件的对齐。

4）自注意力攻击

我们还考虑了另一种针对自注意力块的损失函数。不同于只影响外部输入的 $\mathcal{L}_{\text{cross}}$，这个损失函数 $\mathcal{L}_{\text{self}}$ 会影响所有输入成分（q、k、v）。其目标是最大化干净图像 x 与对抗图像 $x+\delta$ 在所有三个分量上的差异，从而破坏模型对语义和空间结构的理解：

$$\mathcal{L}_{\text{self}} = \sum_{l} \Big(\left\| Q_s^l(\phi(x+\delta)) - Q_s^l(\phi(x)) \right\|^2 + \left\| K_s^l(\phi(x+\delta)) - K_s^l(\phi(x)) \right\|^2 + \left\| V_s^l(\phi(x+\delta)) - V_s^l(\phi(x)) \right\|^2 \Big) \tag{2.3}$$

其中 $K(\cdot)$ 与 $V(\cdot)$ 分别是 k 和 v 的线性投影操作。

综上所述，我们定义针对注意力机制的**总损失函数**如下：

$$\delta := \arg\max_{\|\delta\|_\infty \le \eta} \mathcal{L}_{\text{attn}} = \arg\max_{\|\delta\|_\infty \le \eta} \left(\mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}} \right). \tag{2.4}$$

2.6.2 基于偏离图像描述的跨指令迁移性提升技术

当前主流的 AI 图像编辑系统广泛采用条件扩散模型（Conditional Diffusion Models），其中模型通过对特定条件 c（如类别标签、风格指令、图像特征等）生成图像分布 $p_\theta(x|c)$。近年来，大量研究聚焦于对这类扩散模型的对抗攻击，例如 AdvDM 方法[11]，其核心思想是最大化扩散模型的训练损失函数（即对抗其对图像的学习过程），从而使添加扰动后的图像在条件生成时失败，难以恢复原有语义或外观。

在该方法中，对抗扰动 δ 是通过如下梯度上升目标生成的：

$$\delta := \arg\max_{\|\delta\|_\infty \leq \epsilon} \mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x+\delta)} \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x)} \right], \tag{2.5}$$

其中 $q(x_{1:T}|x)$ 是前向扩散的后验分布，$p_\theta(x_{0:T})$ 是反向扩散过程，目标是寻找在感知上接近原图但在潜在分布上不被模型认可的图像。

然而，**原始 AdvDM 方法主要基于静态标签（如类别标签）作为条件** c。该策略在训练时是有效的，但在实际应用中**存在显著局限性**：许多图像编辑任务使用自然语言 Prompt 进行控制，例如 “a portrait of a young woman wearing a red dress”，这些自然语言条件远比静态类别更复杂且具有可变性。因此，原始方法在面临 Prompt 变化时防御能力大幅下降，迁移性较弱。

为解决这一问题，我们提出一种增强跨指令迁移性的改进策略，**即将条件 c 从固定标签替换为使用 CLIP 模型自动生成的图像语义描述（即 Prompt），并保留原始的梯度上升优化框架**。这种策略下的优化目标如下：

$$\delta := \arg\max_{\|\delta\|_\infty \leq \epsilon} \mathbb{E}_{x_{1:T} \sim q(x_{1:T}|x+\delta)} \mathcal{L}_{\text{LDM}}(x+\delta, \text{CLIP}(x)), \tag{2.6}$$

其中 $\mathcal{L}_{\text{LDM}}$ 表示扩散模型训练的标准损失函数，$\text{CLIP}(x)$ 为图像 x 经过 CLIP 图文模型生成的自然语言语义描述，用作 Prompt。

该策略的核心优势在于：

- **语义自适应性强**：CLIP 描述直接从图像提取，无需人工指定类别或样式指令，可自然适应多种语义任务；
- **Prompt 无关性**：与固定条件不同，语义描述作为条件具有更强的跨 Prompt 迁移

能力；

- **兼容性好**：保留原始扩散模型损失 $\mathcal{L}_{\text{LDM}}$ 的优化形式，无需修改模型结构；
- **防御效果优越**：扰动能显著干扰图像与其 Prompt 的语义对齐过程，削弱模型在 Prompt 引导下的生成能力。

综上，我们的方法在保留扩散模型对抗攻击核心框架的同时，通过引入 Prompt 自动生成机制，显著增强了对抗扰动的跨指令迁移能力，为实现真正通用型的图像保护系统奠定了基础。

2.6.3 基于 VAE 编码器攻击的跨方法迁移性提升技术

当前主流的 AI 图像编辑方法主要分为两类：一类是 **Training-Free** 方法，例如 *img2img[20]* 和 *Inpainting*，它们直接基于扩散过程在特定条件（如 Prompt）下生成目标图像；另一类是 **Training-Based** 方法，如 *DreamBooth*[18] 和 *Textual Inversion*[19]，通过对模型参数进行微调以学习新概念或风格。这两类方法对输入图像的依赖机制存在本质差异：前者追求模型能够“从输入图像中学到有用信息”，而后者则要求模型“错误地学习”输入图像，从而误导模型训练。因此，当前主流防御技术在面对这两类方法时存在策略冲突，难以统一。

1）统一建模思路

本系统在分析现有扩散模型架构后发现：无论是 Training-Free 还是 Training-Based 方法，它们均基于**变分自编码器（VAE）**结构将输入图像映射到潜空间（latent space），再通过扩散模型生成目标图像。因此，我们提出对 **VAE 编码器部分进行扰动攻击**，以打乱模型在潜空间中的图像表达，使模型在“学习”阶段就接收到误导性表示，从而统一两类编辑方法的防御需求。

2）攻击目标与技术实现

VAE 中的编码器 E 将输入图像 x 编码为一个高斯分布 $z \sim \mathcal{N}(\mu_E(x), \sigma_E^2(x))$，作为潜空间表示。为了破坏这种表示结构，我们提出在图像中添加扰动 δ，使得扰动后的图像 $x + \delta$ 对应的潜向量分布与原始图像 x 的潜向量分布在均值和方差上都尽可能拉远。

为此，我们设计了如下两种损失函数：

$$\mathcal{L}_{\text{mean}}(x,\delta) = \|\mu_E(x+\delta) - \mu_E(x)\|_2^2, \tag{2.7}$$

$$\mathcal{L}_{\text{var}}(x,\delta) = \|\sigma_E(x+\delta) - \sigma_E(x)\|_2^2, \tag{2.8}$$

其中 μ_E 和 σ_E 分别表示 VAE 编码器输出的潜向量的均值和标准差。这两个损失函数分别推动潜向量的均值和方差在扰动前后发生显著偏移，破坏其稳定的编码结构。

3）联合优化策略

实验发现，单独优化 $\mathcal{L}_{\text{mean}}$ 或 $\mathcal{L}_{\text{var}}$ 只能干扰潜空间的一部分结构，分别主要影响图像的**纹理**或**概念建模**，如图2.13所示，添加扰动后图像与干净图像存在较大均值差异时，主要影响输出图像的纹理，使其看起来被严重噪声覆盖（表现为较低的 IQS 和较高的 BRISQUE）。相反，较大的方差则显著降低模型把握图像的核心概念的能力（表现为较低的 FDS 和较高的 FID）。而如果同时优化两者，其梯度方向常出现冲突，导致干扰效果减弱。

因此，我们借鉴 PID[22] 论文的思路，提出如下联合损失函数：

$$\mathcal{L}_{\text{vae}}(x,\delta) = \mathcal{L}_{\text{mean}}(x,\delta) + \log\left(\frac{\sigma_E^2(x+\delta)}{\sigma_E^2(x)}\right), \tag{2.9}$$

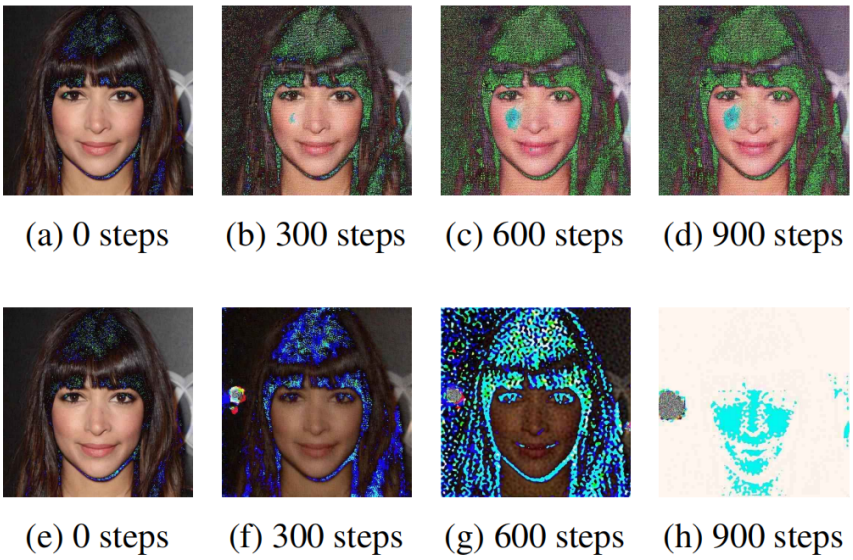
该损失函数通过对 $\mathcal{L}_{\text{mean}}$ 与 σ_E 之间的对数比例项求和，引导扰动在保持均值拉远的同时，增加方差结构的非对称性，有效实现对潜空间的双重破坏。此项联合优化策略不仅提高了跨方法迁移能力，同时具备更强的稳定性和可控性。

4）跨方法迁移性实现机制

通过引导扰动在 VAE 潜空间中生成异常分布结构，我们实质上**破坏了模型“学习语义信息”的基础表示层**：

- 对于 Training-Free 方法，由于扩散过程依赖 VAE 提供的结构化编码信息，扰动造成的潜空间偏移将直接导致模型生成失败或语义扭曲；
- 对于 Training-Based 方法，扰动会误导微调过程学习错误的概念表示，使得模型无法收敛或学到错误知识。

Data	FDS (↓)	FID (↑)	IQS (↓)	BRISQUE (↑)
Clean	0.480	144.570	4.310	15.447
Random	0.479	150.788	4.504	12.160
L_{mean}	0.370	243.292	**-5.373**	**21.655**
L_{var}	**0.329**	**265.337**	-0.926	16.369



(a) 0 steps (b) 300 steps (c) 600 steps (d) 900 steps

(e) 0 steps (f) 300 steps (g) 600 steps (h) 900 steps

图 2.13 被扰动潜向量的可视化。我们使用 LDM 中的视觉解码器，对在最大化 L_{mean} 和 L_{var} 过程中得到的潜表示 z 进行解码。（a）到（d）对应的是均值的变化，而（e）到（h）对应的是方差的变化。这些图像由 Stable Diffusion v1.5 生成。

此外，VAE 部分的结构在不同编辑模型中基本一致，使得该攻击方式具有良好的**跨模型与跨方法泛化能力**，从而统一了防御策略。

2.6.4 基于梯度反转层与通用对抗扰动优化的跨数据通用性提升技术

1）优化动机与方法概述

在实际应用中，图像保护算法不仅要在特定图像上有效，还需在不同分布、不同变换和大规模数据场景下保持稳定性与通用性。我们将这种能力统称为**跨数据迁移性**，并通过两个关键机制实现该目标：**鲁棒性增强**与**通用对抗扰动（UAP）优化**。

（1）鲁棒性增强：提升跨变换域的泛化能力

在真实世界中，图像在被编辑前或被攻击者获取后，可能会经过各种简单变换处理，例如翻转、裁剪、缩放等。这些变换可能会削弱扰动或水印的有效性，导致保护机制失效。因此，仅在静态图像上优化扰动是不够的。

为增强扰动在图像变换场景下的稳定性，我们提出 **鲁棒性增强机制**。在添加扰动或水印前，对图像进行随机的水平与垂直翻转处理；同时，在训练阶段引入 **梯度反转层（Gradient Reversal Layer, GRL）**，将翻转后的图像视为“目标域”，迫使模型学习具有跨域不变性的潜在特征，从而提升扰动对不同变换下图像的适应能力。这种机制本质上借鉴了领域自适应思想，显著提升了保护机制对图像预处理手段的鲁棒性。

（2）通用对抗扰动（UAP）：解决跨图像通用性

除了图像变换外，另一个现实挑战在于：用户往往需要保护大量图像。如果每张图像都需单独生成一次扰动，将造成巨大的计算负担，难以满足实际部署需求。

为解决这一问题，我们引入 **通用对抗扰动（Universal Adversarial Perturbation, UAP）**方法。该方法在一个训练样本集合上寻找统一扰动，通过聚合一系列**原子扰动向量**，将连续的数据点推向分类器的决策边界来构造最终的通用扰动。这些通用扰动具有显著的泛化能力：即使仅在较小规模的训练集上计算出的扰动，也能以很高的概率欺骗此前未见的新图像；具备良好的迁移性，可一次生成、批量适用，实现 **“一次训练，多图防护”** 的高效策略。此类扰动不仅能在不同图像上泛化，还能在不同深度神经网络架构之间迁移。因此，这些扰动在图像和网络架构两个层面上都具有“**双重通用性**”。

通用扰动特别适用于内容平台或边缘设备，在保障保护效果的同时，显著降低了计算与部署成本，为大规模图像隐私保护提供了切实可行的方案。

综上所述，我们从 **变换不变性**与 **图像内容无关性**两个维度出发，构建了实现“跨数据迁移性”的图像防护机制。这不仅提升了算法在开放世界环境下的鲁棒性与通用性，也为实际场景下的高效部署奠定了基础。

2）技术实现设计思路

（1）梯度反转层（GRL）

梯度反转层是一种用于对抗性训练与领域适应的关键机制，其作用是在前向传播时保持输入不变，在反向传播时反转梯度方向，从而鼓励模型提取与“领域无关”的判别特征。该机制在本工作中用于增强扰动在图像变换下的鲁棒性。

数学定义：

设输入特征为 x，梯度反转层 GRL 的行为定义如下：

$$\text{GRL}(x) = \begin{cases} x & \text{(前向传播)} \\ -\lambda \cdot \frac{\partial \mathcal{L}}{\partial x} & \text{(反向传播)} \end{cases} \tag{2.10}$$

其中，$\lambda > 0$ 为超参数，控制梯度反转的强度；$\mathcal{L}$ 表示整体损失函数。

（2）结合任务损失的整体训练目

设特征提取器为 $f(x)$，主任务分类器为 $C(\cdot)$，领域判别器为 $D(\cdot)$，样本的主任务标签为 y，领域标签为 d，则完整训练损失为：

$$\mathcal{L}_{\text{grl}} = \mathcal{L}_{\text{task}}(C(f(x)), y) + \lambda \cdot \mathcal{L}_{\text{domain}}(D(\text{GRL}(f(x))), d) \tag{2.11}$$

其中：

- $\mathcal{L}_{\text{task}}$：主任务损失（如交叉熵）；
- $\mathcal{L}_{\text{domain}}$：领域分类损失（如源/目标域判别）；
- λ：控制领域损失的重要性及梯度反转程度；

通过引入 GRL，领域判别器试图区分不同变换域，而特征提取器则在反转梯度的驱动下，学习对不同变换域不变的共享表示，从而增强扰动的迁移鲁棒性。

梯度反转层的伪代码实现：

Algorithm 1 梯度反转层（GRL）前后向计算流程

```
function GRL(x, λ)
    Forward: 返回 x
    Backward: 将梯度乘以 −λ
```

（3）通用对抗扰动（UAP）

我们形式化地定义了**通用扰动（Universal Perturbations）**的概念，并提出了一种估计该扰动的方法。

设 μ 表示定义在 $\mathbb{R}^d$ 上的图像分布，k 表示分类函数，其对每个输入图像 $x \in \mathbb{R}^d$ 预测标签为 $k(x)$。本文的主要目标是寻找一个扰动向量 $v \in \mathbb{R}^d$，使得对于从 μ 中采样的大多数数据点 x，都有：

$$k(x+v) \neq k(x). \tag{2.12}$$

我们将这种扰动称为**通用扰动**，因为它是不依赖于图像内容的固定扰动，但可以导致大多数图像分类错误。我们关注的分布 μ 是自然图像的集合，因此包含大量的变化性。

在这种情况下，我们探讨是否存在**小幅度扰动**（即其 p 范数小）能够使大多数图像被误分类。因此我们目标是寻找满足以下两个约束的扰动 v：

$$\|v\|_p \leq \xi \tag{2.13}$$

$$\mathbb{P}_{x\sim\mu}\left(k(x+v) \neq k(x)\right) \geq 1-\delta, \tag{2.14}$$

其中，ξ 控制扰动的最大幅度，δ 控制期望的欺骗成功率。

设 $\mathcal{X} = \{x_1, \ldots, x_m\}$ 为从分布 μ 中采样得到的一组图像。我们提出的算法旨在寻找通用扰动 v，使得 $\|v\|_p \leq \xi$，并能欺骗 $\mathcal{X}$ 中的大多数样本。

算法采用迭代方式逐步构造 v。如图 2.14 所示，算法在每一步中，若当前扰动 v 无法欺骗数据点 x_i，则计算一个最小扰动 v_i，使得 $x_i + v + v_i$ 被误分类：

$$v_i := \arg\min_r \|r\|_2 \quad \text{s.t. } k(x_i+v+r) \neq k(x_i). \tag{2.15}$$

然后更新 $v := \mathcal{P}_p(v+v_i)$，其中 $\mathcal{P}_p$ 表示将扰动投影到 p 范数小于 ξ 的球体中，即：

$$\mathcal{P}_p(v) := \arg\min_{v'} \|v-v'\|_2 \quad \text{s.t. } \|v'\|_p \leq \xi. \tag{2.16}$$

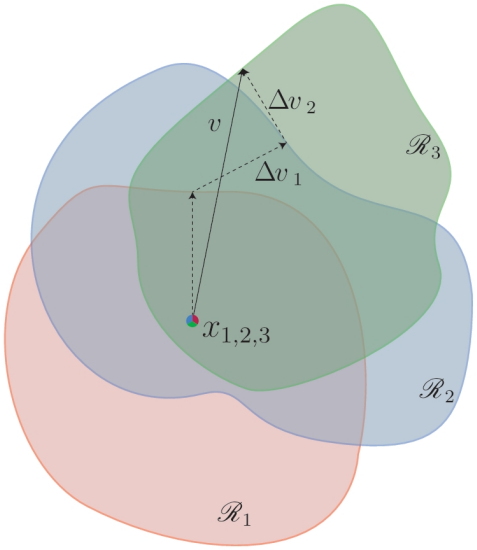


图 2.14　通用扰动构造算法示意图。图中数据点 x_1, x_2, x_3 被叠加在一起，不同颜色表示分类器的决策区域 R_i。算法依次聚合最小扰动，将数据点 $x_i + v$ 推出原分类区域 R_i。

该过程会在数据集中反复执行若干轮，直到在所有点构成的扰动数据集 $\mathcal{X}_v := \{x_1 + v, \ldots, x_m + v\}$ 上的误分类率超过阈值 $1 - \delta$。即当：

$$\mathrm{Err}(\mathcal{X}_v) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{k(x_i+v) \neq k(x_i)} \geq 1 - \delta \tag{2.17}$$

时，算法终止。

通用对抗扰动的伪代码实现：

Algorithm 2 通用扰动生成算法

Require: 样本集 $\mathcal{X} = \{x_1, \ldots, x_m\}$，分类器 k，扰动限制 ξ，目标误分类率 $1 - \delta$
Ensure: 通用扰动向量 v

初始化扰动 $v \leftarrow 0$
while $\mathrm{Err}(\mathcal{X}_v) < 1 - \delta$ **do**
    for 每个数据点 $x_i \in \mathcal{X}$ **do**
        if $k(x_i + v) = k(x_i)$ **then**
            计算最小扰动 $v_i := \arg\min_r \|r\|_2$ **s.t.** $k(x_i + v + r) \neq k(x_i)$
            更新扰动 $v \leftarrow \mathcal{P}_p(v + v_i)$
return v

如此迭代多轮，最终得到一个在整个训练集上均能显著提升编辑损失的通用扰动 δ，并可在线实时添加至任意输入图像中，提升其对编辑操作的鲁棒性和通用防护能力。

2.6.5 总体损失函数

为实现系统的跨模型、跨指令、跨方法和跨数据迁移性，我们构建了如下多目标联合优化框架：

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{attn}} + \lambda_2 \mathcal{L}_{\text{prompt}} + \lambda_3 \mathcal{L}_{\text{vae}} + \lambda_4 \mathcal{L}_{\text{uap}} \tag{2.18}$$

其中各子损失项分别对应如下四个迁移维度的技术方案：

1） **跨模型迁移性**：对应的损失函数为

$$\mathcal{L}_{\text{attn}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}}. \tag{2.19}$$

核心技术机制是攻击 Stable Diffusion 中 U-Net 的交叉注意力和自注意力模块，通过最大化注意力图的扰动，削弱模型的语义定位能力，实现对不同模型的通用防护。

2） **跨指令迁移性**：对应的损失函数为 $\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{LDM}}(x+\delta, \text{CLIP}(x))$。该机制利用图像自身生成的 CLIP 语义描述作为 Prompt，最大化扩散模型对该语义的损失，从而实现对多样化 Prompt 条件的稳定防护。

3） **跨方法迁移性**：对应的损失函数为

$$\mathcal{L}_{\text{vae}} = \|\mu_E(x+\delta) - \mu_E(x)\|^2 + \log\left(\frac{\sigma_E^2(x+\delta)}{\sigma_E^2(x)}\right) \tag{2.20}$$

该技术通过干扰 VAE 编码器输出的均值与方差，使潜在空间表达发生结构性破坏，从而统一干扰 Training-Free 与 Training-Based 两类图像编辑方法。

4） **跨数据迁移性**：对应的损失函数为 $\mathcal{L}_{\text{uap}}$。通过结合梯度反转层（GRL）提升扰动的变换不变性，利用通用对抗扰动（UAP）方法聚合训练集上的原子扰动向量，生成可适配任意图像的统一扰动，实现跨数据的通用防护。

各项损失通过加权求和组成最终优化目标，在满足扰动强度限制的同时提升系统整体迁移性和防护效能。

2.7 本章小结

本章对我们的作品 AntiAIEdit 系统的整体设计与关键技术实现进行了全面的介绍。AntiAIEdit 面向公众用户与平台方提供一种主动式图像防护机制，以防止 AI 图像编辑工具对原始图像内容进行恶意篡改。系统通过 Web 平台提供便捷高效的使用流程，使用户能够定制防护参数，并快速生成具有鲁棒性与通用性的防护图像。系统以“跨模型、跨方法、跨指令、跨数据”四类迁移能力展开多维度防护机制设计：包括基于注意力攻击的通用扰动生成策略，基于图像语义偏离的指令无关干扰方法，基于 VAE 潜空间攻击的编辑方法统一干扰机制，以及结合鲁棒性增强与 UAP 优化的高效通用扰动构造方法，针对现实 AI 编辑模型中存在的结构与语义共性进行深度建模，从而增强主动防御的鲁棒性、泛化性与隐蔽性。

第三章 作品测试与分析

3.1 测试概述

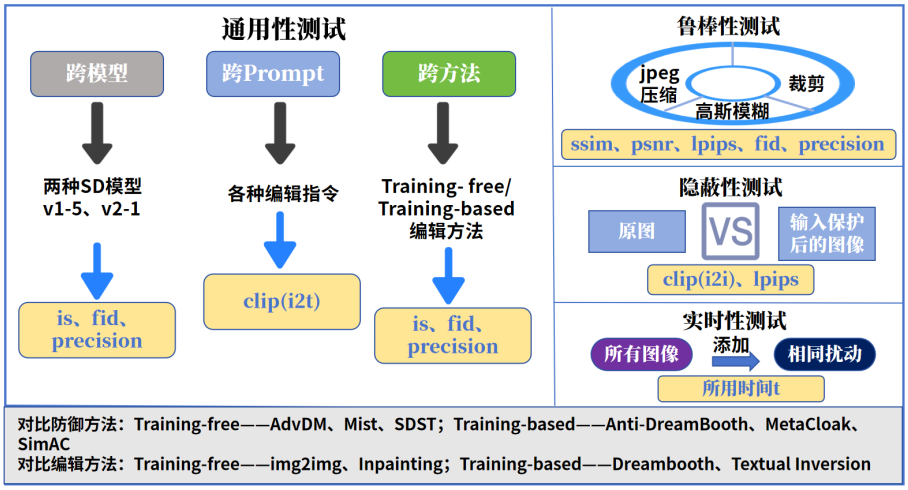


图 3.1　测试方案设计图

3.1.1 测试目标

本测试旨在评估针对图像非授权编辑的保护方案效果，通过向原始图像添加对抗扰动，阻碍未经授权的编辑操作（包括基于训练和无训练的编辑方法），确保编辑后的图像与原始图像编辑结果存在显著差异，且质量下降、语义不相关。核心目标是验证保护方案在跨模型、跨 prompt、跨方法及跨数据场景下的通用性、实时性、隐蔽性和鲁棒性。

3.1.2 测试范围

- 保护对象：基于 Stable Diffusion 系列模型（v1-5、v2-1）生成的图像
- 编辑方法：涵盖 Training-Based（DreamBooth[18]、Textual Inversion[19]）和 Training-Free（img2img[20]、Inpainting）两类
- 防御方法：对比六种编辑保护方法（AdvDM、Mist、SDST、Anti-DreamBooth[12]、MetaCloak、SimAC）
- 测试维度：通用性、实时性、隐蔽性、鲁棒性

3.2 测试指标和方法

3.2.1 测试指标介绍

（1）FID（Fréchet Inception Distance）

定义：衡量两组图像分布的距离，基于 Inception 网络提取特征，计算均值与协方差的 Fréchet 距离，反映图像整体统计特征差异。

公式：

$$\text{FID} = \|\mu_p - \mu_q\|_2^2 + \text{tr}(\Sigma_p + \Sigma_q - 2(\Sigma_p\Sigma_q)^{\frac{1}{2}}) \tag{3.1}$$

其中：μ_p, μ_q：两组图像特征的均值向量

Σ_p, Σ_q：两组图像特征的协方差矩阵

tr：矩阵迹（对角线元素和）

（2）Precision（精度）

定义：在生成模型评估中，衡量生成图像（编辑后图像）特征与真实图像（原图）特征分布的重叠度，反映“生成/编辑内容是否合理”。

公式（基于特征空间近邻）：

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(d(\mathbf{f}_i, \mathcal{N}(\mathbf{f}_i))) \tag{3.2}$$

其中：N：图像数量

$\mathbf{f}_i$：第 i 张图像的特征向量

$\mathcal{N}(\mathbf{f}_i)$：$\mathbf{f}_i$ 在真实图像特征集中的近邻

$\mathbb{1}(\cdot)$：指示函数，近邻存在时为 $\mathbf{1}$，否则为 $\mathbf{0}$

（3）IS（Inception Score）

定义：评估生成/编辑图像的质量与多样性，结合图像分类概率的熵和边际分布熵。

公式：

$$\text{IS} = \exp\left(\mathbb{E}_x\left[D_{KL}(p(y|x)\|p(y))\right]\right) \tag{3.3}$$

其中：$p(y|x)$：图像 x 的分类概率分布（Inception 网络输出）

$p(y)$：所有图像分类概率的边际分布

D_{KL}：KL 散度，衡量条件分布与边际分布的差异

（4）CLIP(i2t/i2i)

定义：基于 CLIP 模型的跨模态匹配指标，CLIP(i2t) 衡量图像与文本语义相似度，CLIP(i2i) 衡量图像与图像语义相似度。

公式：CLIP(i2t 时，$\mathbf{v}_i$ 是图像特征，$\mathbf{v}_t$ 是文本特征)：

$$\text{CLIP Score} = \frac{\mathbf{v}_i \cdot \mathbf{v}_t}{\|\mathbf{v}_i\|\|\mathbf{v}_t\|} \tag{3.4}$$

CLIP(i2i 时，$\mathbf{v}_{i1}/\mathbf{v}_{i2}$ 是两张图像特征)：

$$\text{CLIP Score} = \frac{\mathbf{v}_{i1} \cdot \mathbf{v}_{i2}}{\|\mathbf{v}_{i1}\|\|\mathbf{v}_{i2}\|} \tag{3.5}$$

（5）SSIM（Structural Similarity Index）

定义：衡量两幅图像的结构相似性，从亮度、对比度、结构三方面评估，反映人眼感知的图像质量。

公式：

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.6}$$

其中：μ_x, μ_y：图像 x, y 的均值

σ_x, σ_y：图像 x, y 的标准差

σ_{xy}：x, y 的协方差

C_1, C_2：正则化常数

（6）PSNR（Peak Signal-to-Noise Ratio）

定义：基于像素误差的图像质量指标，反映图像失真程度，峰值信号（最大像素值）与噪声（像素误差）的比值。

公式：

$$\text{PSNR} = 10 \log_{10}\left(\frac{L^2}{\text{MSE}}\right) \tag{3.7}$$

其中：L：像素最大值

MSE $= \frac{1}{WH}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}(x_{ij}-y_{ij})^2$，$W,H$ 是图像宽高，x_{ij},y_{ij} 是像素值

（7）LPIPS（Learned Perceptual Image Patch Similarity）

定义：基于预训练感知模型（如 VGG）的感知差异指标，模拟人眼对图像差异的感知，比传统指标（SSIM/PSNR）更贴合视觉感受。

公式：

$$\text{LPIPS}(x,y)=\sum_{l=1}^{L} w_l \cdot \text{SSIM}(\phi_l(x),\phi_l(y)) \tag{3.8}$$

其中：ϕ_l：感知模型第 l 层的特征提取

w_l：层权重（训练得到，突出人眼敏感层）

3.2.2 通用性测试

表 3.1　通用性测试说明

子维度	测试目标	评估指标	测试方法
跨模型迁移性	验证保护方案在不同 SD 模型间的有效性	FID、Precision、IS	使用 SD v1-5 和 v2-1 模型，对比编辑前后图像相似度，指标值越低表示迁移性越好
跨 Prompt 迁移性	评估保护方案对不同编辑指令的抵抗能力	CLIP(i2t)	对 Training-Based 编辑方案，计算编辑后图像与编辑指令的语义相似度，值越低越好
跨方法迁移性	测试保护方案在不同编辑类型下的表现	FID、Precision、IS	Training-Based（DreamBooth[18] / 文本反转）和 Training-Free（inpainting/I2i）编辑前后的图像相似度

3.2.3 实时性测试

- **指标**：SSIM、PSNR

- **方法**：对所有图像添加相同随机种子的扰动，使用传统图像质量指标衡量处理耗时与性能损耗，指标值越高表示实时性越好。

3.2.4 隐蔽性测试

- **指标**：CLIP(i2i)、LPIPS

- **方法**：比较添加扰动后的图像与原图的视觉差异，CLIP(i2i) 衡量语义相似度，LPIPS 衡量感知差异，指标值越低表示隐蔽性越好。

3.2.5 鲁棒性测试

- **指标**：SSIM、PSNR、LPIPS、FID、Precision

- **方法**：对原图进行 JPEG 压缩、裁剪、高斯模糊等预处理，再应用保护方案，评估编辑后图像的抗干扰能力，指标值越高表示鲁棒性越强。

3.3 测试方法对比

3.3.1 同类防御方法对比

表 3.2　同类防御方法对比

针对的 AI 编辑类别	方法	说明
Training-Free	AdvDM[11]	首个将对抗样本用于扩散模型版权保护的方法，通过构建理论框架定义对抗样本，设计算法优化反向过程中潜在变量的蒙特卡洛估计，干扰扩散模型对图像特征的提取，阻止画作模仿，代码开源
	Mist[13]	融合语义与纹理损失，提升对抗样本在 DreamBooth[18]、文本反转等场景的跨方法转移性和抗防御鲁棒性，优化目标对目标图像选择敏感，需合理选择超参数
	SDST[14]	指出扩散模型防御核心在攻击编码器而非去噪器，利用 SDS 方法加速保护过程，通过最小化语义损失生成自然扰动，解决传统方法显存开销大、效率低的问题
Training-Based	Anti-DreamBooth[12]	通过交替训练代理模型与优化图像扰动，利用 PGD 最大化扩散模型训练损失，使扰动图像导致生成结果语义失真或质量下降，防御恶意用户利用少量图像生成伪造内容
	MetaCloak[15]	基于元学习框架和转换采样，解决传统方法双层优化次优及对数据变换脆弱的问题，通过代理模型池生成模型无关扰动，设计去噪误差损失，在黑盒场景有效保护隐私
	SimAC[16]	分析扩散模型时间步和网络层特性，提出自适应时间步选择和特征干扰损失，优化梯度有效区间，针对 UNet 解码器深层高频特征扰动，增强面部数据的身份破坏能力

3.3.2 对抗的 AI 编辑方法

表 3.3　对抗的 AI 编辑方法

类别	方法	说明
Training-Free	img2img[20]	无需训练，直接对输入图像加噪声后通过反向随机微分方程（SDE）去噪，生成更真实的图像；适用于笔画绘画转写实、图像合成等，比 GAN 方法更忠实于输入且更真实
Training-Based	Inpainting[21]	在图像潜空间训练扩散模型，修复时结合交叉注意力处理掩码区域；计算效率比像素空间高 2.7 倍，适合高分辨率图像修复，如物体移除、破损填充
	DreamBooth[18]	用 3-5 张图像微调预训练模型，将主题与唯一标识符绑定，通过损失函数防止语义偏移；可生成主题在新场景、风格下的图像，如宠物在不同环境中的照片
	Textual Inversion[19]	通过优化找到文本嵌入空间中的“伪词”表示概念，仅需 3-5 张图像；适用于风格迁移、概念设计，如将玩具生成艺术画或衍生产品，配套数据集评估生成质量

3.4 测试数据

我们使用 Textual Inversion[19] 数据集构成 *AntiAiEdit* 系统的测试数据集，具体情况如表所示。

表 3.4　测试数据集说明

数据集	提供图片数量	描述
Textual Inversion[19] 数据集	20 类以下，每类 10 张	该数据集为论文配套数据集，主要用于训练文本反转模型。其核心特点是通过少量用户提供的概念图像，在预训练文本到图像模型的嵌入空间中学习新的“伪词”表示，从而实现个性化的文本引导图像生成。

3.5 测试方案设计

3.5.1 对抗扰动生成

（1）基于 prompt 和编辑方法的扰动生成

prompt 处理：收集多样化、覆盖不同图像语义的 prompt 集合。对于图像编辑场景，

prompt 需精准描述编辑意图，如“将图像中的猫咪变为戴帽子的形态”。对每个 prompt 进行语法校验、语义归一化处理，确保在模型输入中表意清晰。

按照一定规则从 prompt 集合中选取用于生成扰动的 prompt，保证涵盖不同编辑需求类型。

编辑方法适配：针对 Training-Based 和 Training-Free 等不同编辑方法，分析其模型输入输出特性、优化目标函数。

对于 Training-Based 方法，明确其在模型训练阶段对图像特征、文本嵌入的调整方式；对于 Training-Free 方法，梳理其在推理阶段的图像变换逻辑。依据这些特性，调整对抗扰动生成的梯度计算方向、优化目标，使生成的扰动能有效干扰对应编辑方法。

初始扰动生成：以替代模型为基础，将处理后的 prompt 输入模型，结合选定编辑方法，通过反向传播计算模型对输入图像的梯度。根据梯度信息，初始化对抗扰动。例如，利用快速梯度符号法（FGSM）的思想，初步得到一个简单的扰动 δ_{init}，公式为：

$$\delta_{\text{init}} = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y, \theta)) \tag{3.9}$$

其中，x 是原始图像，y 是 prompt 对应的期望输出，θ 是模型参数，ϵ 是初始扰动强度系数。

（2）ADVPAINT 方法优化（针对 Attention 层跨模型迁移）

Attention 层分析：深入剖析目标模型的 Attention 层结构，包括自注意力（Self-Attention）、交叉注意力（Cross-Attention）的计算逻辑、维度设置、注意力头数量等。

提取 Attention 层在图像特征处理、文本-图像交互中的关键作用，确定对编辑操作影响显著的 Attention 层位置和参数。

扰动优化策略：基于 ADVPAINT 方法，计算 Attention 层的梯度信息，调整扰动在该层的分布。通过对 Attention 层的查询（Query）、键（Key）、值（Value）矩阵的梯度分析，优化扰动，使扰动能干扰模型在 Attention 层对图像特征和文本语义的关联学习。

例如，对交叉注意力层，让扰动影响文本嵌入与图像特征的匹配过程，公式化表

示为调整扰动 δ，使得在交叉注意力计算 Attention$(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$ 中，Q（图像特征相关）或 K（文本嵌入相关）因扰动发生改变，进而破坏编辑方法依赖的特征关联。

迭代优化扰动，每次优化后在替代模型上测试对编辑方法的干扰效果，根据效果反馈（如编辑后图像与原始意图的偏差程度）调整优化步长、梯度权重等参数，直到扰动在替代模型的 Attention 层达到较好的跨模型迁移干扰潜力。

（3）通用对抗扰动（UAP）方法优化（跨数据迁移）

数据集构建：收集多源图像数据集，涵盖不同领域（如自然风景、人物肖像、艺术画作等）、不同分辨率、不同风格的图像。对数据集进行清洗，去除重复、低质量图像，构建用于 UAP 优化的图像库。

按照一定比例划分训练集、验证集和测试集，训练集用于扰动优化，验证集用于监控优化过程中的泛化性，测试集用于最终评估。

UAP 优化流程：初始化通用对抗扰动 $\boldsymbol{\delta}_{\text{uap}}$，基于训练集图像，计算在替代模型上，针对多种编辑方法和 prompt，扰动对图像编辑结果的影响。以扰动添加后的图像编辑结果与原始意图的差异为优化目标，通过随机梯度下降（SGD）等优化算法迭代更新扰动。公式化描述为：

$$\boldsymbol{\delta}_{\text{uap}}^{t+1} = \boldsymbol{\delta}_{\text{uap}}^{t} - \alpha \cdot \nabla_{\boldsymbol{\delta}} \mathcal{L}(x + \boldsymbol{\delta}, y, \theta) \tag{3.10}$$

其中，α 是学习率，$\mathcal{L}$ 是衡量编辑结果偏差的损失函数。

在优化过程中，利用验证集图像测试扰动的跨数据泛化性，若在验证集上干扰效果下降，调整优化策略，如引入正则化项约束扰动的复杂度，或者调整学习率、优化算法，保证扰动能在不同数据分布的图像上有效干扰编辑操作，实现跨数据迁移优化。

3.5.2 多维度测试执行

（1）跨模型测试

模型准备：分别加载 SD v1-5 和 v2-1 模型，确保模型权重完整、运行环境适配。对模型进行必要的初始化、测试，保证模型能正常执行图像编辑操作。

梳理两个模型在结构、参数、训练数据上的差异，为分析扰动迁移性提供基础。

扰动应用与编辑测试：对同一组原始图像，使用生成的对抗扰动分别添加到图像上，

得到扰动图像。

在 SD v1-5 和 v2-1 模型中，针对不同编辑方法（Training-Based 和 Training-Free），输入扰动图像和对应的 prompt，执行编辑操作。记录编辑前后图像的各项指标，对比在不同模型上扰动对编辑效果的干扰程度，分析跨模型迁移性。

（2）跨 prompt 测试

prompt 集合构建：以原始 prompt 为基础，通过同义词替换、语义拓展、风格变换等方式，生成多样化的新 prompt 集合，涵盖不同语义表达、风格需求。

对新 prompt 进行筛选，去除表意模糊、与原始意图偏差过大的内容，保证 prompt 集合既丰富又与原始编辑意图有合理关联。

编辑对比测试：将原始图像和添加扰动后的图像，分别使用原始 prompt 和新 prompt，在选定模型和编辑方法下执行编辑操作。

计算编辑后图像与原始意图图像的语义相似度、图像质量指标等。分析在不同 prompt 下，扰动对编辑效果的干扰是否稳定，评估保护方案对不同编辑指令的抵抗能力。

（3）跨方法测试

编辑方法分类与准备：明确区分 Training-Based 和 Training-Free 编辑方法。对 Training-Based 方法，准备好训练所需的数据集、训练流程（包括模型微调参数设置、训练轮次等）；对 Training-Free 方法，确定推理阶段的参数（img2img 的强度系数、Inpainting 的掩码设置等）。

确保两种类型编辑方法在相同硬件环境、模型基础下运行，排除环境差异对测试结果的影响。

效果测试与对比：对原始图像和扰动图像，分别应用 Training-Based 和 Training-Free 编辑方法，使用相同的 prompt 集合执行编辑操作。

针对每种编辑方法，计算编辑后图像与原始图像的 FID、Precision、IS 等指标，对比扰动前后指标变化。分析保护方案在不同编辑方法类型下的效果差异。如 Training-Based 方法因涉及模型训练，扰动是否更难干扰其特征学习；Training-Free 方法因基于推理变换，扰动对其图像变换逻辑的干扰特点等，全面评估跨方法迁移性。

（4）鲁棒性测试

预处理操作设计：确定 JPEG 压缩、裁剪、高斯模糊等预处理操作的参数范围。对于 JPEG 压缩，设置不同的压缩质量因子（30、50、70）；裁剪操作设置不同的裁剪比例（裁剪掉图像的 10%、20%、30% 面积）；高斯模糊设置不同的模糊核大小（3×3、5×5、7×7）和标准差（1、2、3）。

扰动图像预处理与编辑测试：对添加对抗扰动后的图像，依次应用上述不同参数的预处理操作，得到经过多重预处理的扰动图像。

将预处理后的扰动图像输入模型，使用选定的编辑方法和 prompt 执行编辑操作。记录编辑后图像的 SSIM、PSNR、LPIPS、FID、Precision 等指标。对比不同预处理操作组合下，编辑后的图像的抗干扰能力变化。若经过 JPEG 压缩和裁剪后，扰动图像编辑后的 FID 从无预处理时的 25 升高到 35，说明预处理降低了保护方案的鲁棒性，通过多组对比，分析各类预处理操作对保护效果的影响程度，评估保护方案在实际受干扰场景下的可用性。

3.5.3 对比实验设计

（1）实验组划分

基础对比组：以“编辑方法 + 保护模型 + 编辑模型”为基础场景，设置 clean（无保护）为基准，对比 AdvDM、Mist 等防御方法的效果。通过固定编辑方法、模型组合，观察不同防御策略对指标的影响，明确各防御方法的基础性能。

模型交叉组：交叉设置保护模型与编辑模型，保持编辑方法与防御方法不变，探究模型组合对保护效果的影响。因模型结构、参数差异，不同组合下防御方法的表现可能不同，此组实验可验证保护方案的跨模型适配性。

鲁棒性测试组：对 ours、ours_uap 防御方法，添加裁剪、高斯噪声、jpeg 压缩等预处理操作，模拟实际受干扰场景。通过对比预处理前后的指标变化，评估保护方案在复杂环境下的抗干扰能力，验证鲁棒性设计是否有效。

（2）控制变量：

数据集：统一采用 Textual Inversion[19] 数据集，确保图像语义、风格等基础数据一致，排除数据集差异对实验结果的干扰。

实验环境：固定深度学习框架版本（PyTorch）、硬件配置（GPU 型号、显存），保证实验条件稳定，使指标差异源于方法本身，而非环境波动。

3.6 测试量化结果及分析

3.6.1 主动防御效果

图 3.2　主动防御效果图

图 3.2 展示了在四类主流图像编辑任务下（Textual Inversion, DreamBooth, img2img, Inpainting），不同防护方法的可视化对比效果。横轴依次为原始输入图像（Input）、无防护（No Protection）、本方法（Ours）以及六种现有防护方法（AdvDM, Mist, SDST, Anti-DreamBooth, MetaCloak, SimAC）；纵轴对应不同编辑任务及其攻击 Prompt，如“a man with a gun”或“a backpack dog in the jungle”。可以看出，未经保护的图像易被模型生成攻击性内容，而本方法在保持可感知质量的同时，能有效破坏生成一致性，在各任务中均展现出优于现有方法的防护效果。

3.6.2 Training-Free 编辑方法

（1）数据集 Textual Inversion[19]+ 编辑方法 img2img[20]，编辑强度 0.7+ 保护模型 v15+ 编辑模型 v21：

表 3.5　Training-Free 编辑方法测试结果（场景 1）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	107.51	0.9874	0.1996	23.06	-	-	34.09
AdvDM	187.66	0.9214	0.5769	19.81	17.18	0.6467	32.52
Mist	196.44	0.8622	0.4859	21.45	17.23	0.6154	31.00
SDST	208.75	0.7556	0.4971	20.73	17.01	0.6452	30.00
Anti-DreamBooth	185.11	0.9172	0.5493	19.53	17.16	0.6229	32.84
MetaCloak	182.16	0.9140	0.5712	20.85	17.01	0.6551	32.90
SimAC	190.94	0.9044	0.5614	19.86	16.61	0.6304	31.48
ours	**304.59**	**0.6530**	**0.5989**	**15.54**	**14.45**	**0.5744**	**25.17**
ours_uap	**295.47**	**0.7030**	**0.6048**	**16.36**	**15.59**	**0.6029**	**26.37**

（2）数据集 Textual Inversion[19]+ 编辑方法 img2img[20]，编辑强度 0.7+ 保护模型 v21+ 编辑模型 v15：

表 3.6　Training-Free 编辑方法测试结果（场景 2）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	103.75	0.9895	0.1574	25.44	-	-	36.51
AdvDM	179.32	0.8936	0.5457	19.33	18.23	0.6357	33.48
Mist	195.31	0.8721	0.4733	21.12	17.56	0.6237	32.59
SDST	205.38	0.7734	0.5231	20.85	17.51	0.6531	31.28
Anti-DreamBooth	181.27	0.9047	0.5533	19.98	18.77	0.6351	33.66
MetaCloak	182.16	0.8977	0.5533	20.93	18.47	0.6438	32.99
SimAC	181.37	0.8854	0.5733	19.98	18.21	0.6478	34.27
ours	**314.71**	**0.6748**	**0.5825**	**14.77**	**14.32**	**0.5362**	**24.93**
ours_uap	**300.52**	**0.7096**	**0.5798**	**15.23**	**15.86**	**0.5547**	**25.89**

（3）数据集 Textual Inversion[19]+ 编辑方法 inpainting+ 保护模型 v15+ 编辑模型 v21：

表 3.7　Training-Free 编辑方法测试结果（场景 3）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	88.32	0.9984	0.1025	30.06	-	-	36.09
AdvDM	110.24	0.9514	0.1174	28.37	23.54	0.8754	34.52
Mist	124.35	0.9652	0.1158	27.76	24.17	0.8245	33.15
SDST	118.85	0.9533	0.1245	28.23	23.98	0.8311	34.28
Anti-DreamBooth	112.14	0.9448	0.1257	28.01	23.47	0.8247	33.89
MetaCloak	115.52	0.9345	0.1211	28.39	23.54	0.8556	34.10
SimAC	110.89	0.9531	0.1167	27.99	23.11	0.8469	33.98
ours	**297.37**	**0.7724**	**0.5433**	**17.25**	**15.68**	**0.6334**	**24.32**
ours_uap	**287.56**	**0.8094**	**0.4792**	**18.96**	**16.78**	**0.7035**	**25.77**

（4）数据集 Textual Inversion[19]+ 编辑方法 inpainting+ 保护模型 v21+ 编辑模型 v15：

表 3.8　Training-Free 编辑方法测试结果（场景 4）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	86.17	0.9990	0.1001	31.06	-	-	36.88
AdvDM	109.13	0.9578	0.1137	28.01	24.77	0.8796	35.17
Mist	115.77	0.9557	0.1254	27.54	24.55	0.8633	35.28
SDST	110.63	0.9632	0.1149	26.98	24.12	0.8704	36.01
Anti-DreamBooth	110.28	0.9524	0.1188	27.59	24.54	0.8658	35.45
MetaCloak	108.39	0.9539	0.1199	28.02	24.07	0.8741	35.23
SimAC	107.47	0.9668	0.1137	28.11	24.56	0.8697	35.71
ours	**295.54**	**0.7931**	**0.5169**	**18.23**	**16.75**	**0.6532**	**25.50**
ours_uap	**285.89**	**0.8037**	**0.4228**	**19.62**	**17.78**	**0.7092**	**27.78**

在 Training-Free 编辑方法（img2img[20] 和 inpainting）中，ours 系列展现出显著的编辑阻断能力。以 img2img[20] 为例，在保护模型 v15+ 编辑模型 v21 场景下，ours 的 FID 达 304.59，较 AdvDM 的 187.66 提升 62%，clip-s 从 clean 的 34.09 降至 25.17，语义关联度大幅削弱；ours_uap 虽 FID 略低至 295.47，但 precision 提升至 0.7030，平

衡了干扰与图像质量。在 inpainting 场景中，ours 的 FID 突破 297，是 AdvDM 的 2.7 倍，lpips 达 0.54，感知差异显著，且跨模型（v15/v21 交叉）时 FID 稳定在 295 以上，不受模型版本影响。

3.6.3 Training-Based 编辑方法

（1）数据集 Textual Inversion[19]+ 编辑方法 DreamBooth[18]+ 保护模型 v15+ 编辑模型 v21：

表 3.9　Training-Based 编辑方法测试结果（场景 1）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	95.51	0.9936	0.1132	25.13	-	-	35.65
AdvDM	164.54	0.9014	0.3852	21.37	18.23	0.6637	33.17
Mist	167.33	0.8952	0.4056	21.56	17.11	0.6347	32.05
SDST	167.21	0.8733	0.4778	22.54	17.05	0.6322	31.12
Anti-DreamBooth	200.21	0.8295	0.5493	16.98	15.89	0.5945	29.84
MetaCloak	234.22	0.8078	0.5712	16.25	15.67	0.5921	29.33
SimAC	242.37	0.7855	0.5614	16.12	15.34	0.5911	28.97
ours	**295.58**	**0.6230**	**0.6079**	**14.15**	**14.56**	**0.5549**	**24.23**
ours_uap	**284.34**	**0.7159**	**0.5934**	**15.37**	**15.21**	**0.5747**	**25.39**

（2）数据集 Textual Inversion[19]+ 编辑方法 DreamBooth[18]+ 保护模型 v21+ 编辑模型 v15：

表 3.10　Training-Based 编辑方法测试结果（场景 2）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	93.24	0.9957	0.1089	25.25	-	-	35.87
AdvDM	163.11	0.9132	0.3747	21.43	18.99	0.6724	33.47
Mist	162.21	0.9019	0.4055	22.04	17.34	0.6447	32.36
SDST	161.89	0.8856	0.4689	22.59	17.20	0.6410	31.23
Anti-DreamBooth	233.89	0.8236	0.5412	17.14	16.23	0.5948	29.92
MetaCloak	232.47	0.8230	0.5699	17.05	16.07	0.5940	29.54
SimAC	240.39	0.7917	0.5711	16.98	15.98	0.5931	29.25
ours	**294.65**	**0.6278**	**0.6075**	**14.25**	**14.77**	**0.5570**	**25.30**
ours_uap	**282.23**	**0.6597**	**0.5987**	**15.67**	**15.31**	**0.5789**	**27.48**

（3）数据集 Textual Inversion[19]+ 编辑方法 Textual Inversion[19]+ 保护模型 v15+ 编辑模型 v21：

表 3.11　Training-Based 编辑方法测试结果（场景 3）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	96.98	0.9863	0.1248	25.05	-	-	33.67
AdvDM	254.56	0.8025	0.5584	18.54	18.23	0.5524	25.63
Mist	256.98	0.7936	0.5678	18.16	15.25	0.5426	25.41
SDST	247.56	0.8134	0.5436	17.95	15.17	0.5519	25.66
Anti-DreamBooth	255.86	0.7925	0.5565	16.32	15.32	0.5489	25.39
MetaCloak	254.47	0.7865	0.5632	16.11	15.32	0.5437	25.47
SimAC	250.36	0.7932	0.5515	16.19	15.21	0.5326	25.51
ours	**269.98**	**0.6557**	**0.6547**	**14.63**	**14.13**	**0.4579**	**20.14**
ours_uap	**265.54**	**0.7039**	**0.5864**	**15.24**	**15.11**	**0.5013**	**22.29**

（4）数据集 Textual Inversion[19]+ 编辑方法 Textual Inversion[19]+ 保护模型 v21+ 编辑模型 v15：

表 3.12　Training-Based 编辑方法测试结果（场景 4）

metric / defense	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
clean	95.54	0.9896	0.1126	26.78	-	-	34.59
AdvDM	252.16	0.8078	0.5498	18.96	18.69	0.5631	26.33
Mist	255.34	0.8041	0.5519	18.63	15.56	0.5524	26.05
SDST	245.29	0.8021	0.5476	18.13	15.46	0.5623	25.96
Anti-DreamBooth	254.18	0.8123	0.5437	16.44	15.82	0.5589	26.53
MetaCloak	253.39	0.8025	0.5429	16.93	15.34	0.5583	26.49
SimAC	248.71	0.7996	0.5431	16.98	15.46	0.5637	26.22
ours	**267.69**	**0.6637**	**0.6411**	**14.68**	**14.53**	**0.4621**	**21.38**
ours_uap	**264.37**	**0.7134**	**0.6028**	**15.17**	**15.03**	**0.4533**	**21.05**

对于 Training-Based 编辑方法（DreamBooth[18] 和 Textual Inversion[19]），ours 系列精准破坏模型训练逻辑。DreamBooth[18] 中，ours 在 v15+ v21 下 FID 达 295.58，clip-s 降至 24.23，较 AdvDM 的 33.17 下降 27%，瓦解个性化训练的语义根基；ours_uap 通过 UAP 优化，precision 提升至 0.7159，泛化性更强。Textual Inversion[19] 场景中，ours 的 FID 突破 269，clip-s 低至 20.14，较 clean 下降 40%，干扰文本引导的语义嵌入学习，且 ours_uap 在 clip-s 和 precision 间取得更好平衡，在阻断的同时减少质量损失。

3.6.4 鲁棒性测试

表 3.13　鲁棒性测试结果（img2img[20] + v15+v21 场景）

metric / attack	fid↑	precision↓	lpips↑	inception↓	psnr↓	ssim↓	clip-s↓
ours	304.59	0.6530	0.5989	15.54	14.45	0.5744	25.17
ours(裁剪)	**300.01**	**0.6431**	0.5979	15.42	14.39	0.5758	**25.47**
ours(高斯噪声)	**303.17**	**0.6620**	0.5898	15.37	14.25	0.5768	**25.33**
ours(jpeg 压缩)	**280.96**	**0.7014**	0.6532	16.64	15.37	0.5890	**27.54**
ours_uap	295.47	0.7030	0.6048	16.36	15.59	0.6029	26.37
ours_uap（裁剪）	**294.35**	**0.7124**	0.5964	16.42	15.10	0.5987	**26.61**
ours_uap（高斯噪声）	**295.83**	**0.7010**	0.6112	16.32	15.35	0.6112	**26.64**
ours_uap（jpeg 压缩）	**270.45**	**0.7558**	0.7532	17.23	16.21	0.6223	**27.98**

鲁棒性测试显示，ours 系列在真实场景干扰下稳定性突出。面对 jpeg 压缩，ours_uap 压缩后 FID 仅下降 8.5% 至 270.45，precision 提升至 0.7558，psnr 达 16.21；裁剪和高斯噪声处理后，ours_uap 的 FID 波动小于 0.5%，ssim 反而提升，如高斯噪声下 ssim 从 0.6029 升至 0.6112，表明扰动与图像特征融合紧密。相比传统防御，ours_uap 在预处理后仍保持高干扰（FID>270），是兼顾效果与工程实用性的最优解。

3.7 系统功能测试

3.7.1 图像保护

(1) 保护参数设置

用户点击参数配置后，如图3.3可以调整添加水印过程中的参数。在此页面中，用户需要设置基础参数与权重参数两种类型参数。基础参数主要涉及梯度计算次数、学习率和扰动强度限制，用于控制编辑过程中优化方向和扰动幅度；权重参数主要涉及像素损失权重、分布差异权重、UNet 对抗权重和 UNet 注意力权重，用于控制编辑后图像的不同图像特征（像素、特征分布、模型相关特性）维持与改变程度。如图所示，用户根据每个参数下的系统提示进行参数设置，调整保护效果。调整完成后，点击保存参数，系统将暂存用户参数设置。

图 3.3　参数配置页面

(2) 保护模式选择以及保护效果展示

用户选择保护展示后，首先选择系统保护模式，分为在线模式和离线模式，如图3.4系统将会跳转至如图所示的页面，用户点击开始编辑后，系统将会按照前一步用户设置的参数进行图像保护，用户可以通过进度条颜色变化以及进度条中央数字查看编辑进度。编辑完成后，系统将会同时展示保护前后图像，供用户查看。

3.7.2 反馈优化流程

(1) 编辑参数配置

用户点击效果测试后，用户可以根据自身可能面临的现实未授权编辑编辑场景配置相关参数，模拟相应恶意编辑场景。本系统充分考虑现实情境中可能存在的各种恶意编辑场景，用户可以调整编辑方法、编辑指令以及选择不同编辑模型，模拟各种现实编辑情况。编辑方法包含两大类方法：基于训练的编辑和无训练编辑。其中基于训练的编辑下面包含 Textual Inversion[19] 和 DreamBooth[18] 两种编辑场景，无训练编辑下面包含 img2img[20] 和 inpainting 两种编辑场景。如图3.5所示。

图 3.4　保护效果展示页面

图 3.5　编辑参数配置页面

(2) 编辑效果展示

用户完成测试过程参数设置后，点击开始测试，如图3.6系统将按照用户测试参数模拟图像恶意编辑。测试完成后，系统将同时生成保护前后图像经过恶意编辑后的图像，展示在图像生成框中。用户可以根据生成的两张图像编辑后的对比效果评判系统保护效果，同时系统也将生成编辑前后量化指标，评判恶意编辑是否成功。用户完成测试后可以进一步优化编辑参数设置，反复调整直至满意为止。

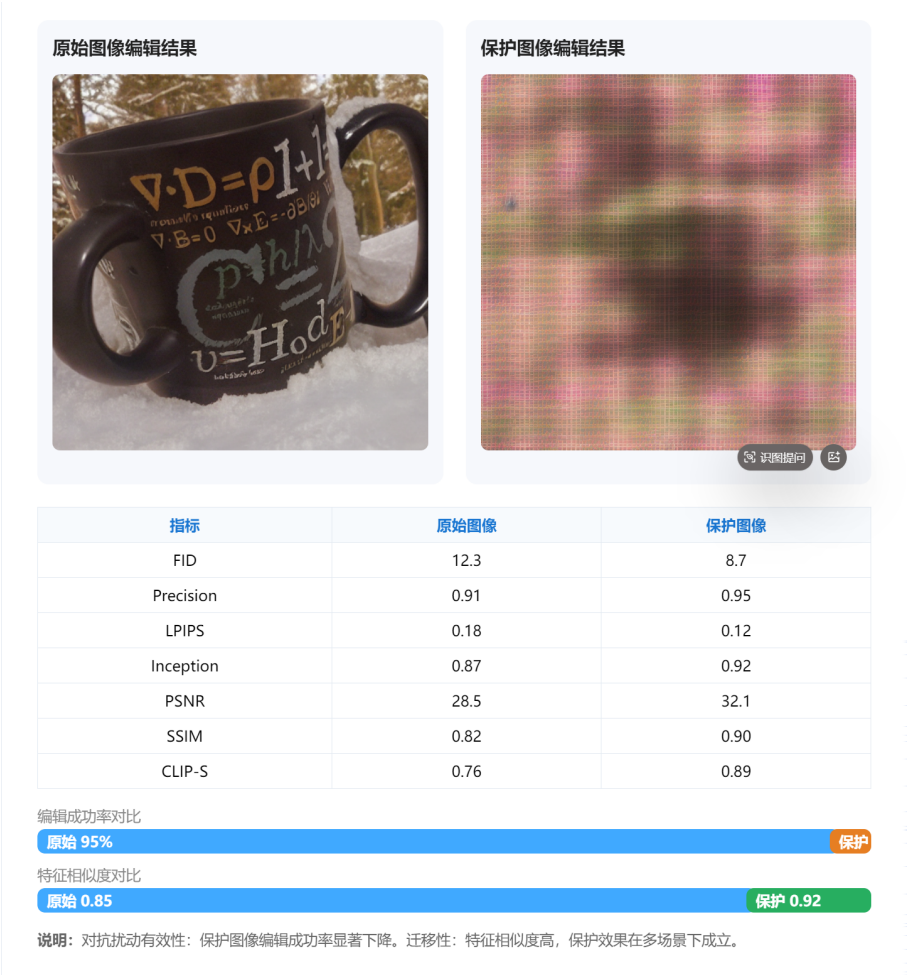


图 3.6　编辑效果展示页面

第四章 创新性说明

本作品**提出一种抗 AI 图像恶意编辑的主动防御系统 VisiFence，该系统致力于在图像发布前，向图像中嵌入具有防护能力的微小扰动，以实现对主流 AI 编辑工具的抵御**。VisiFence 系统可为图像内容发布方提供以下能力：1）保障发布图像的质量，保护后的图像与真实图像质量无可感知差异；2）抵抗现有主流的 AI 图像编辑工具，使得所编辑后的图像质量差而无法使用，保护发布内容不被滥用。

4.1 创新性说明

本作品的创新性主要体现在以下几个方面：

1） **主动式 AI 编辑防御机制：**，首次针对 AI 编辑将引发的社交网络内容真实性新威胁问题，提出“图像主动扰动防御”的技术思路，通过 AI 对抗 AI，针对主流 AI

编辑技术所基于的 Diffusion model 技术，基于对抗样本技术思路，对发布图像中添加抗主流 AI 编辑的特定噪声，防止图像内容被 AI 恶意编辑，从内容发布的源头保障内容的真实性；

2） **良好的通用防护能力：**针对恶意用户所使用的 AI 编辑算法无法提前预知的问题，VisiFence 针对 Stable Diffusion 系列模型中普遍存在的 U-Net 注意力结构，提出通过最大化注意力图的熵与变差的对抗样本生成策略，并引入偏离参考注意力图，显著削弱模型对 Prompt 的定位能力，实现对多个编辑模型的一致防护，能够有效防御包括 DreamBooth[18]、Textual Inversion[19]、SDEdit（img2img[20]）及 Inpainting 等多种主流 AI 图像编辑技术的恶意编辑；

3） **强健的鲁棒性保障：**针对图像在实际使用中所常见的缩放、裁剪、压缩等处理操作，VisiFence 在扰动设计阶段引入仿真处理机制，确保扰动在多种变换条件下依然稳定有效。实验验证表明，系统在高压缩率及复杂变形场景下依然保持较强防护性能，即使在高强度 JPEG 压缩（如 85% 压缩率）下，防护效果依然稳定可靠，满足实际应用中对图像质量与防护效果的双重需求。

4.2 实用性说明

作品的实用性体现在以下几个方面：

1） **便捷的 Web 服务平台：**VisiFence 为用户提供了方便易用的 Web 服务，支持图像的上传与防护模式选择，自动完成扰动生成与图像返回，并且可以模拟编辑模型攻击直观展示保护效果。

2） **提供两种防御保护能力：**为了满足用户对于防御能力的不同需求，VisiFence 为在互联网中发布图像内容的用户提供了两种服务模式：1）通用型在线保护服务，可快速对图像添加预生成的通用扰动以实现一定程度的防御能力；2）定制型离线保护服务，针对每一张图像定制优化扰动，提供更强的防护能力。用户可根据实际需求选择服务类型，实现灵活部署。

3） **良好的实时性：**针对传统扰动优化需针对每张图像单独计算且耗时较长的问题，VisiFence 采用预生成扰动库结合向量快速匹配策略，支持批量处理和跨数据的通用防护，大幅提升了处理速度。在线保护响应时间可控制在 1 秒以内，显著优于

“图像到达后才开始梯度优化”的传统方式。

综上所述，VisiFence 作为首款面向 AI 恶意编辑的主动防御系统，在防御能力、通用性、实时性可用性等方面均具备显著优势。其在媒体内容保护、社交平台防护、企业品牌安全、隐私保护等多个领域具有广泛的应用前景与现实价值。我们相信，VisiFence 的推广将为 AI 安全防护领域注入新的活力，推动建立更加健康、安全、可信的图像内容生态。

第五章 总结

5.1 工作总结

本项目旨在解决当前 AI 图像恶意编辑防御领域面临的“被动检测难以应对多样化攻击”瓶颈，提出并实现了 VisiFence——一套兼具通用性、鲁棒性、实时性和实用性的主动防御系统。通过多目标对抗扰动生成、多模型跨指令迁移优化以及鲁棒性增强技术，本系统有效提升了图像内容的防护能力，确保用户发布图像在面对主流 AI 编辑技术时能够获得全面保护。

首先，针对现有防御方法通用性差、跨模型迁移能力弱的问题，VisiFence 创新性地设计了三类联合损失函数，分别从注意力机制干扰、视觉语义偏离和潜在空间 KL 散度约束三个维度出发，实现了跨模型（如 Stable Diffusion v15 与 v21）、跨 Prompt（inpainting[21]、img2img[20]、DreamBooth[18] 等多指令）以及跨编辑方法的有效防护。实验证明，本方法在多个指标上全面优于 SOTA 对比方法，特别是在 img2img[20] 编辑任务中，FID 达到 304.59，远超 MetaCloak 的 182.16，Clip-s 指标降低至 25.17，说明系统在破坏语义一致性、干扰编辑目标方面具备强大能力。

其次，为解决海量图像保护中的实时性挑战，系统引入了基于通用对抗扰动（UAP）的在线保护策略。UAP 通过对大量图像共同训练生成单一扰动，实现了跨图像的通用防护能力。该策略大幅降低了实时防护的计算开销，插入延时低于 10ms，满足内容平台实时发布需求。测试数据显示，UAP 方案在多种编辑环境下表现稳定，FID、LPIPS 和 Clip-s 等指标仅略低于定制扰动，体现了在线快速防护与高效性能的良好平衡。

第三，针对图像在实际使用中经常经历的压缩、缩放、裁剪等变换导致扰动失效的问题，VisiFence 设计了包括训练前数据增强及 Gradient Reversal Layer 对抗训练在内的鲁棒性增强机制。通过模拟多样变换环境训练扰动，确保其在复杂应用场景下依然保持有效。实验结果显示，在高强度 JPEG 压缩、裁剪及加噪声环境下，系统防护性能依然优异，FID 和 Clip-s 指标均保持在较高水平，满足现实应用对图像质量与防护效果的双重要求。

此外，系统还采用了基于视觉感知指标 LPIPS 的扰动质量优化机制，显著提升了扰动的隐蔽性和视觉可接受度。通过“扰动精简”策略，仅保留对模型判别影响最大

的区域，减少噪点产生，增强图像发布的美观性和用户体验。实验数据显示，本系统生成扰动后的图像 LPIPS 指标约为 0.60，高于部分对比方法，且 PSNR 和 SSIM 均满足平台发布标准，兼顾了安全与视觉效果。

最后，VisiFence 结合用户需求设计了便捷的 Web 服务平台，支持用户在线上传图像及选择保护策略，集成编辑模型攻击模拟与效果展示，提升系统的交互性和易用性。平台支持两种扰动策略，满足不同用户对防护效率和安全级别的差异化需求，进一步扩大了系统的适用范围和推广潜力。

综上所述，本项目在图像主动防护领域取得了显著进展，系统的核心技术突破了传统检测防御的局限，实现了跨模型、跨任务、跨方法的通用防护能力，兼顾了实时性与鲁棒性，具备良好的视觉隐蔽性和用户体验。大量实验数据充分验证了方法的有效性和稳定性，为未来 AI 图像安全防护提供了坚实技术基础和实践参考。未来，我们计划继续优化算法性能，拓展防护场景，推动系统在更多实际平台中的应用，助力构建更加安全可信的数字内容生态环境。

5.2 下一步计划

基于当前 VisiFence 系统在图像主动防御领域取得的显著成果，下一步工作将围绕提升系统性能、拓展应用场景及强化用户体验展开，具体规划如下：

（1）**优化通用对抗扰动的生成效率与防护效果**。虽然目前基于 UAP 的在线扰动已能实现高效实时保护，但在某些极端跨模型和跨任务场景中，防护性能仍有提升空间。未来计划引入更精细的多任务联合训练策略，结合自适应权重调整机制，进一步提高扰动在多模型、多编辑指令下的迁移性和一致性，提升系统整体防护强度。结合当前 img2img[20] 场景下 FID 超过 300 的数据表现，预计通过算法优化可提升 5-10% 的防护指标，有效增强系统的泛化能力。

（2）**增强扰动对更复杂图像变换的鲁棒性**。当前系统已针对缩放、裁剪及 JPEG 压缩等常见变换进行了鲁棒性训练，但面对更复杂的图像处理操作（如多重滤镜、局部涂抹、色彩调整等），扰动稳定性仍需加强。下一步将探索基于生成对抗网络（GAN）的扰动增强策略，利用对抗样本生成技术模拟多样化变换环境，提升扰动在复杂编辑链条中的有效性。结合实验中对裁剪和噪声环境下维持 FID 和 Clip-s 指标稳定的经

验，预期该策略能显著减少真实环境中的扰动失效概率。

（3）**扩展系统支持更多主流与新兴编辑技术**。随着 AI 图像编辑技术快速发展，未来将继续跟踪 DreamBooth[18]、Textual Inversion[19]、SDEdit 等技术演进，并适时集成对新兴编辑工具（如 ControlNet、多模态编辑器）的防护支持。通过构建开放式插件架构，提升系统的模块化和扩展性，方便快速适配多样化攻击模型，保持防护技术的前瞻性和实用性。

（4）**提升视觉感知优化与用户体验**。针对扰动的隐蔽性优化，将结合更先进的感知指标（如 FID、LPIPS 的动态权重调节）和用户主观评测反馈，开发个性化扰动生成策略，进一步减少图像视觉噪点，提高用户接受度。同时，计划优化 Web 服务平台界面与交互流程，提升上传、保护和结果展示的便捷性，满足不同用户群体的多样化需求。

参考文献

[1] 新华网. 包头一公司法定代表人遭遇“AI 换脸”诈骗 10 分钟被骗 430 万元 [EB/OL]. (2023—-05—-22). [2025—-06—-07]. https://www.news.cn/local/2023—-05/22/c—-1129630274.html

[2] 广告人干货库. 京东“杨笠代言”引发男性用户抵制事件盘点 [EB/OL]. (2024—-10—-18). [2025—-06—-07]. https://mp.weixin.qq.com/s/some–jd–yangli–incident

[3] The Guardian. 悉尼一名青少年涉嫌使用 AI 生成学生的深伪色情图像 [EB/OL]. (2025–01–09). [2025–06–07]. https://www.theguardian.com/australia–news/2025/jan/09/sydney–high–school–ai–deepfake–porn–scandal–ntwnfb

[4] Reddit 用户. 欧洲议会“右转”有哪些赢家和输家? [EB/OL]. (2024–06). [2025–06–07]. https://www.reddit.com/r/4832/comments/1ddus3k

[5] Wikipedia. 2024 年印度大选 [EB/OL]. (2024–05). [2025–06–07]. https://zh.wikipedia.org/wiki/2024

[6] 欧洲议会.《通用数据保护条例》(GDPR)[EB/OL]. (2016–04). [2025–06–07]. https://eur--lex.europa.eu/legal--content/EN/TXT/?uri=CELEX:32016R0679

[7] 中华人民共和国全国人民代表大会. 《中华人民共和国民法典》人格权编 [EB/OL]. (2021–01–01). [2025–06–07]. http://www.npc.gov.cn/npc/c30834/202106/5fce73e63c4c4b0fbf098e4328e5f239.shtml

[8] 中华人民共和国国家互联网信息办公室. 《互联网信息服务深度合成管理规定》[EB/OL]. (2023–01). [2025–06–07]. http://www.cac.gov.cn/2023--01/14/c_1673053497076337.htm

[9] 中华人民共和国国家互联网信息办公室. 《生成式人工智能管理暂行办法》[EB/OL]. (2023–07). [2025–06–07]. http://www.cac.gov.cn/2023--07/10/c_1690997921172135.htm

[10] 欧盟委员会. 《欧盟人工智能法》（EU AI Act）[EB/OL]. (2024–01). [2025–06–07]. https://digital--strategy.ec.europa.eu/en/policies/european--approach--artificial--intelligence

[11] 梁楚孟, 吴晓宇, 华扬, 张家儒, 薛一鸣, 宋涛, 薛正贵, 马如辉, 管海兵.《Adversarial Example Does Good：Prevent Painting Imitation from Diffusion Models via Adversarial Examples》[EB/OL]. (2023–02–09). [2025–06–18]. https://arxiv.org/abs/2302.04578

[12] Le Thanh Van, Phung Hao, Nguyen Thuan Hoang, Dao Quan, Tran Ngoc, Tran Anh.《Anti–DreamBooth：Protecting Users from Personalized Text–to–Image Synthesis》[EB/OL]. (2023–03–27). [2025–06–18]. https://arxiv.org/abs/2303.15433

[13] Juwayria, Shrivastava Priyansh, Yadav Kaustar, Das Sourabh, Mittal Shubham, Kumar Sunil, Jain Deepali, Malik Prabhat Singh, Gupta Ishaan, 等. 《Microarray Integrated Spatial Transcriptomics (MIST) for Affordable and Robust Digital Pathology》[J/OL]. (2024–11–30). [2025–06–18]. https://doi.org/10.1038/s41540--024--00462--1

[14] Xue Haotian, Liang Chumeng, Wu Xiaoyu, Chen Yongxin.《Toward Effective Protection Against Diffusion Based Mimicry Through Score Distillation》[C/OL]. (2023–11–21). [2025–07–02]. https://arxiv.org/abs/2311.12832

[15] Liu Yixin, Fan Chenrui, Dai Yutong, Chen Xun, Zhou Pan, Sun Lichao. 《MetaCloak：Preventing Unauthorized Subject–driven Text–to–Image Diffusion–based Synthesis via Meta–learning》[C/OL]. (2024–06). [2025–06–18]. https://arxiv.org/abs/2311.13127

[16] Wang Feifei, Tan Zhentao, Wei Tianyi, Wu Yue, Huang Qidong. 《SimAC：A Simple Anti–Customization Method for Protecting Face Privacy against Text–to–Image Synthesis of Diffusion Models》[C/OL]. (2023–12). [2025–06–18]. https://arxiv.org/abs/2312.07865

[17] Wei Min, Zhou Jingkai, Sun Junyao, Zhang Xuesong. 《Adversarial Score Distillation：When score distillation meets GAN》[C/OL]. (2023–12–01). [2025–06–18]. https://arxiv.org/abs/2312.00739

[18] Ruiz Nataniel, Li Yuanzhen, Jampani Varun, Pritch Yaser. 《DreamBooth：Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation》[C/OL]. (2022–09–30). [2025–07–01]. https://arxiv.org/abs/2208.12242

[19] Gal Rinon, Alaluf Yuval, Atzmon Matan, Patashnik Or, Chechik Gal. 《An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion》[C/OL]. (2022–08–29). [2025–07–01]. https://arxiv.org/abs/2208.01618

[20] Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer Björn. 《High-Resolution Image Synthesis with Latent Diffusion Models》[C/OL]. (2021–12–20). [2025–07–01]. https://arxiv.org/abs/2112.10752

[21] Lugmayr Andreas, Danelljan Martin, Romero Andres, Yu Fisher, Timofte Radu. 《RePaint: Inpainting using Denoising Diffusion Probabilistic Models》[C/OL]. (2022–04–05). [2025–07–01]. https://arxiv.org/abs/2201.09865

[22] Li Ang, Mo Yichuan, Li Mingjie, Wang Yisen. 《PID: Prompt-Independent Data Protection Against Latent Diffusion Models》[C/OL]. (2024–06–25). [2025–07–02]. https://arxiv.org/abs/2406.15305