面向异构生成式威胁的架构无关特征协同通用防御

Bingxue Zhang Yang Gao Feida Zhu zhangbingxue@usst.edu.cn gaoyang@st.usst.edu.cn fdzhu@smu.edu.sg 上海理工大学 新加坡管理大学 新加坡，新加坡 上海，中国 上海理工大学 中国上海

Yanyan Shen shenyy@sjtu.edu.cn 上海交通大学 中国上海

Yang Shi shiyang1121@st.usst.edu.cn 上海理工大学 中国上海

arXiv:2603.14860v1 [cs.CR] 16 Mar 2026

摘要

生成式人工智能的广泛部署带来了内容安全和隐私方面前所未有的挑战。然而，现有防御机制通常针对特定模型架构量身定制（例如，仅针对扩散模型或生成对抗网络），由此形成脆弱的“防御孤岛”，难以泛化应对现实中异构且往往架构未知的生成式威胁。本文指出，朴素像素空间集成策略存在一项根本性的优化障碍：由于目标函数本质上存在差异，来自异构生成器的像素级梯度在统计上趋于正交，从而在联合优化过程中造成破坏性干扰。为突破这一瓶颈，我们基于实证观察提出一种思路：尽管像素空间中的优化景观各不相同，不同架构生成内容的高层特征表示却呈现出对齐现象。基于这一观察，我们设计了架构无关定向特征协同（ATFS）框架。通过引入目标引导图像，ATFS 将多模型防御重新定义为统一的特征空间对齐任务，从而无需复杂的校正即可在特征层面实现内在的梯度对齐。大量实验表明，ATFS 在异构场景（例如，扩散模型+生成对抗网络）中实现了最先进水平的防护。该方法收敛迅速，仅需 40 次迭代即可达到 >90% 的性能；即使扰动预算严格限制在 ϵ=2/255，仍能保持强大的攻击效力。只需更换特征提取器，该框架即可无缝扩展至未见过的架构（例如 VQ-VAE），并且在应对 JPEG压缩和缩放等干扰时表现出卓越的鲁棒性。ATFS 计算高效且轻量，为打破防御孤岛、构建通用生成式安全框架提供了可行路径。代码和模型均已开源，以确保结果可复现并促进应用。

关键词

生成式人工智能安全、主动防御、模型无关防御、跨模态特征同构、特征协同

1 引言

生成式人工智能的爆炸式增长从根本上改变了内容创作和数字媒体编辑。从文生图扩散模型 [1]到基于生成对抗网络的深度伪造 [2] ，各类技术为创作者带来了前所未有的逼真效果和易用性。然而，创造力的普及也同时对内容安全、隐私和社会信任构成了严重风险 [3]。恶意行为者可以轻易利用这些工具篡改肖像 [4], 、伪造数字证据并大规模生成虚假信息，从而对信息生态系统的完整性构成系统性风险 [5]。本文聚焦于防御异构生成模型对人脸图像进行未经授权的编辑和伪造。

为此，人们开发了主动防御机制，以预先保护视觉内容 [6, 7]。其核心原理是在原始图像中嵌入不可感知扰动，从而干扰恶意模型的生成过程。尽管已取得显著进展，当前防御范式仍存在一项关键局限：架构特异性。现有方法通常针对单一类别的生成模型进行优化 [8]。例如，针对扩散模型的防御会扰动去噪轨迹 [9], ，而针对生成对抗网络的防御则会攻击判别边界 [10]。这种有针对性的优化造成了彼此孤立的“防御孤岛”：一种方法即使对某种架构有效，面对另一种架构也会彻底失效。

然而，现实世界的威胁态势复杂多样，且不受特定架构限制。防御者无法预判攻击者会使用扩散模型、生成对抗网络，还是完全未知的系统。如图1所示，为扩散模型量身打造的扰动 [1] 面对生成对抗网络时会失效 [11], ，反之亦然。这种防御孤岛现象严重削弱了现有技术在开放环境[12], 中的实际可靠性，尤其是在保护人物肖像等敏感视觉内容方面。

弥合这些孤岛的一种简单策略，是直接结合针对多个模型的攻击。本文首先揭示了这一看似直观的方法失效的根本原因。异构生成模型受不同优化目标驱动：扩散模型通过分数匹配引导梯度，以进行密度估计

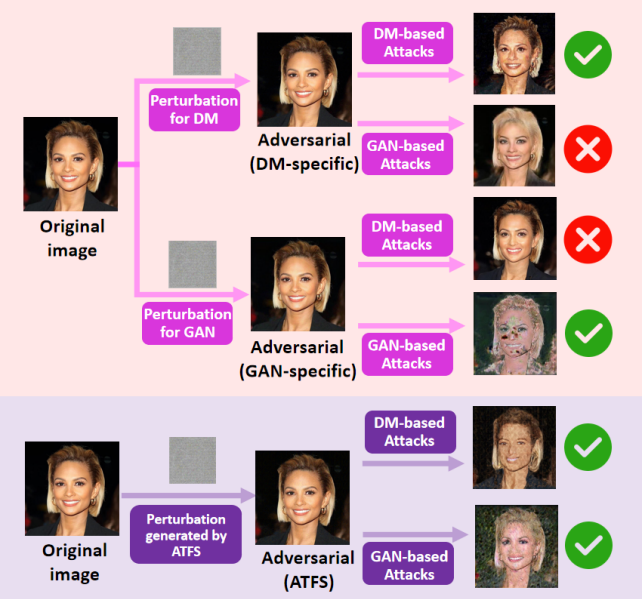


图1：防御孤岛现象：特定架构攻击无法跨模型迁移。该图展示了针对一种模型架构优化的扰动为何无法迁移到另一种架构。对勾（✓）表示防御成功（编辑失败），而叉号（×）表示防御失败（编辑成功）。由扩散模型专用方法保护的人脸图像仍容易受到基于生成对抗网络的编辑攻击，反之亦然。

[13], 而生成对抗网络则将梯度引向对抗性判别边界 [2]。我们证明，这种内在的目标差异会导致它们在高维像素空间中的梯度向量在统计上相互正交 [14]。因此，联合优化带来的不是协同效应，而是破坏性干扰。尽管存在 PCGrad 等梯度缓解算法 [15]，但它们计算成本高昂，也无法解决这些范式之间更深层次的几何不兼容性 [16]。

为了突破这一优化障碍，我们提出了一种范式转变：从像素级冲突转向特征级共识。基于实证证据，我们提出了一项观察：尽管低层生成机制各不相同，但所有能够生成感知上合理内容的模型，都会将输入映射到高层特征空间，而对于相似的感知内容，这些特征空间呈现出对齐现象。特征空间中的这种对齐为统一的模型无关防御提供了实践基础。

受这一洞见启发，我们提出了架构无关的定向特征协同（ATFS）框架。ATFS 利用一个共享目标特征方向——该方向由引导图像导出——将多模型防御问题转化为统一的特征空间对齐任务。理论分析表明，在这一 formulation 下，原本相互冲突的像素级梯度在经过特征提取器投影后会变得方向一致，从而产生建设性干扰。这种对齐使高效的协同防御成为可能，而无需进行计算开销高昂的梯度手术 [17]。

本文的主要贡献总结如下：

1) 我们通过分析表明，异构生成模型在像素空间中的梯度冲突是根本性的，这限制了现有集成方法的有效性。基于对不同架构间特征空间对齐的实证观察，我们提出了一种构建模型无关防御框架的新视角。

2）受这一视角启发，我们设计了 ATFS 框架。通过借助目标特征方向在共享特征空间中对齐模型，该框架将多模型防御统一为单一优化任务，实现梯度协同，而无需复杂的梯度操作。该框架轻量且可扩展；若要将保护扩展至新模型，只需集成其特征提取器。

3）我们针对多个异构威胁场景下的人脸图像保护进行了全面评估。实验表明，ATFS 的防御有效性显著优于所有基线方法，并且对信号干扰具有很强的鲁棒性。尤为重要的是，我们将设计即公平分析纳入防御研究，证明其保护效果在不同人口统计属性群体中始终一致，从而避免算法偏见。

4）为推动可信人工智能社区的发展，我们已将代码和模型完全开源。这不仅确保了研究可复现性，也为后续研究者提供了可审计的防御工具，有助于对生成模型安全性进行标准化评估。

2 相关工作

本节系统回顾生成式人工智能的主动防御领域及相关技术。我们首先回顾针对特定生成式架构量身定制的防御方法，然后分析对抗迁移性与集成策略方面的研究，最后讨论解决多目标优化中梯度冲突的经典方法。

2.1 架构特定的主动防御

随着深度生成模型的广泛应用，主动防御已成为保护图像版权和隐私[18]的前沿技术。现有方法根据其目标模型架构，大致可分为两类：

针对扩散模型的防御：这类工作主要旨在干扰扩散模型（如 Stable Diffusion[1]）的去噪生成过程。Glaze[18]率先提出风格伪装这一概念，通过在特征空间中缩小与目标风格的距离，误导模型对艺术家风格的学习。Mist[9]进一步利用模型对纹理和内容的依赖，生成兼具高迁移性和难以察觉特性的对抗样本。PhotoGuard [18]专注于防御图像编辑（如图像修复），通过攻击潜在空间编码过程来引发生成错误。

针对生成对抗网络（GAN）的防御：这一研究方向旨在破坏生成对抗网络的编辑和伪造能力（例如，用于人脸交换的模型[4]）。Anti-Forgery[10]提出了一种基于感知的扰动攻击，旨在干扰生成对抗网络的图像编辑能力。CMUA-Watermark[19]尝试向人脸图像中添加跨模型通用水印，以防御

来自多个人人脸替换模型的攻击。此外，FakeTagger[3] 通过将可恢复的对抗信息嵌入图像中，实现了对GAN生成内容的被动追踪。

尽管这些方法在针对特定目标时表现出色，但仍存在严重的架构特异性。针对扩散模型（例如，Mist[9]）设计的扰动难以迁移到生成对抗网络[10],，反之亦然。这种“防御孤岛”现象使现有技术难以应对复杂的现实威胁环境，在这些环境中，模型类型可能未知，或多种架构并存[6]。

2.2 对抗迁移性与集成策略

在图像分类等判别任务中，对抗样本的迁移能力是一个核心研究问题[7]。由于高性能分类器（例如，ResNet 和视觉 Transformer）往往会学习相似的决策边界，针对这些分类器生成的对抗梯度方向通常高度一致。经典的集成攻击策略[7]利用了这种梯度对齐特性，只需汇总多个模型的损失或梯度，便能显著增强跨模型迁移性。

然而，将这种集成策略直接迁移到生成模型的防御中，面临着根本性挑战。生成模型（例如扩散模型、生成对抗网络）的核心目标是构建复杂的数据分布，而非执行简单分类。研究表明，采用不同生成范式的模型具有本质不同的优化目标和损失函数[8]。例如，扩散模型通过分数匹配学习数据分布的梯度场，而生成对抗网络则通过极小极大博弈优化生成器和判别器[2]。这种目标函数上的根本差异，使得它们在像素级梯度空间中的几何关系并不兼容[16]。因此，将在判别任务中取得成功的集成策略直接应用于生成模型防御，其有效性仍不明确，并且这类方法面临优化冲突的风险。如何实现针对异构生成模型的有效联合攻击，仍然是一个开放问题。

2.3 多目标优化与梯度冲突

从优化角度来看，同时防御多个异构生成模型本质上是一个典型的多目标优化问题，其中不同模型的损失函数可能产生冲突梯度。多任务学习领域已提出多种算法来解决此类梯度冲突。PCGrad[15] 是应用最广泛的冲突消除算法，它通过将冲突任务的梯度投影到另一任务的法平面上，消除负面干扰。GradNorm[20] 通过动态调整各任务的损失权重来解决梯度幅度不平衡问题。多梯度下降算法[16] 则通过求解二次规划问题，寻找帕累托最优公共下降方向。

尽管这些方法已被证明能够有效训练具有共享参数的多任务神经网络，但将其直接应用于生成模型主动防御这一特定场景仍存在局限。首先，对抗攻击通常需要数十次迭代；PCGrad 和多梯度下降算法等算法所引入的额外计算开销（例如投影、求解二次规划问题）可能会损害实时性能。更重要的是，这些方法通常假设各任务在参数空间中具有潜在兼容性。相比之下，异构生成模型在像素空间中的梯度冲突，可能源于其优化景观存在深层的几何不兼容性[15]。因此，在像素空间强行实施梯度手术难以从根本上解决冲突，还可能使优化过程转向效率低下的折中解。

3 方法

3.1 动机与问题表述

正如引言中所述，生成式人工智能的快速发展导致特定架构的保护方法中出现了明显的“防御孤岛”。由于优化目标存在根本性差异，这些方法通常无法针对异构生成模型提供通用保护。本节旨在形式化跨模型防御问题，明确其潜在挑战，并为我们的方法奠定理论基础。

3.1.1 异构防御问题的形式化。给定 K 异构生成模型 $\{M_k\}_{k=1}^{K}$（例如扩散模型、生成对抗网络等），每个模型都具有各自独特的生成机制。令 $L_k(x)$ 表示针对模型 M_k的攻击损失函数，其目标是破坏该模型的正常生成或编辑能力。一般的跨模型防御方法旨在寻找一种最小扰动 δ，使其能够同时对所有模型生效：

$$\max_{\|\delta\|_\infty \le \epsilon} \sum_{k=1}^{K} \omega_k L_k(x+\delta) \tag{1}$$

其中， ω_k 表示分配给每个模型的权重， ϵ 表示扰动预算，而 $\|\delta\|_\infty < \epsilon$ 确保扰动在视觉上难以察觉。

3.1.2 像素空间中的梯度冲突分析。虽然直接优化公式（1）看似简单，但我们发现，这种方法在实践中面临着根本性挑战。不同生成模型的优化目标本质上各不相同：扩散模型通过分数匹配学习数据分布的梯度场，而生成对抗网络则通过极小极大博弈优化生成器和判别器。这种目标差异会导致它们在像素空间中的梯度方向几乎正交：

$$\mathbb{E}[\langle \nabla_x L_D, \nabla_x L_G \rangle] \approx 0 \tag{2}$$

其中， L_D 和 L_G 分别表示扩散模型和生成对抗网络的攻击损失。这种梯度正交性会在联合优化过程中造成破坏性干扰，常导致优化过程停滞或振荡，进而阻碍有效扰动的累积。

3.1.3 启发式观察：特征空间对齐。为克服像素空间中的优化瓶颈，我们提出一项关键的经验观察：尽管底层机制各不相同，

所有生成模型都必须从输入图像中捕获高层特征信息，才能生成合理的内容。这一观察基于一项广受认可的设计原则：现代生成模型的编码器层经过专门设计，旨在提取驱动后续生成过程的图像特征。

具体而言，对于任何能够生成可被人类感知的内容的生成模型 M_k，我们都可以确定一个特征提取函数 $\Phi_k: \mathrm{X} \to \mathrm{F}_k$，该函数将输入图像映射到一个低维特征空间 F_k。我们观察到，对于视觉上相似的内容，不同模型提取出的特征表示表现出一定程度的一致性。这一观察结果表明了一种新的防御范式：与其在像素空间中协调相互冲突的优化方向，不如在更高层次的特征空间中寻求共识。

3.2 共享特征空间对齐框架

基于上述观察，我们提出了架构无关的定向特征协同（ATFS）框架。ATFS 的核心思想是避免在相互冲突的像素空间中进行直接优化。相反，它在共享目标特征方向的引导下，将针对异构模型的防御问题统一为高层特征空间内的一项协调式优化任务。

3.2.1 特征提取与共享特征空间。我们首先为每个需要防御的生成模型 M_k 定义一个可访问的特征提取函数 $\Phi_k: \mathrm{X} \to \mathrm{F}_k$。该函数将输入图像 x 映射到模型特定的低维特征空间 F。该函数的选择对应于模型中负责编码高层信息的层:

1) 对于扩散模型（例如 Stable Diffusion）：我们选取其 VAE 编码器的输出特征。该层会压缩图像的分布信息，并且对内容和风格都高度敏感。

2) 对于生成对抗网络（例如 StarGAN）：我们选取生成器编码器的深层激活图。这些特征直接决定了生成图像的高层属性和结构。

3) 对于其他架构（例如 VQ-VAE）：我们选取编码器输出的特征表示。

这一选择基于一个简单的工程事实：为了生成合理的内容，生成模型必须在其架构的某处形成对输入的高层理解。因此，尽管它们的数值形式各不相同，函数 Φ_k仍可作为一种代理表示，体现每个模型如何“理解”输入图像。我们的实证观察表明，对于相同的感知内容，来自不同模型的特征 $\Phi_k(x)$ 往往会在各自的空间中编码相似的概念。这表明它们共同处于一个经验证的抽象特征空间中。

3.2.2 特征空间协同优化机制。 为了在这一共享特征空间中实现协同，我们引入一张目标引导图像 x_{tgt}。该图像无需与原始图像 x具有任何特征相关性；其作用是为特征变化提供明确且一致的目标。

对于每个模型 M_k，我们预先计算其对该目标图像的特征表示，称为目标特征向量:

$$t_k = \Phi_k(x_{\mathrm{tgt}}) \tag{3}$$

所有 t_k 在优化过程中均保持固定，构成一个全局对齐目标集。

因此，我们的防御目标也随之改变：我们要寻找一个最小扰动δ，使扰动后图像 $x+\delta$在每个模型的特征空间中的表示同时偏离原始位置，并朝各自的目标特征向量 t_k移动。这定义了一个统一的目标函数:

$$\min_{\|\delta\|_\infty \le \epsilon} J(\delta) = \sum_{k=1}^{K} \omega_k \cdot \underbrace{\|\Phi_k(x+\delta) - t_k\|_2^2}_{L_k^{\mathrm{align}}(x+\delta)} \tag{4}$$

其中， ω_k 表示模型权重（通常设为相等），ϵ 表示扰动预算。优化这一目标函数，本质上会迫使所有模型对扰动后的图像形成一致的特征级误解，使其表示朝 x_tgt对齐。

3.2.3 特征协同的工作机制。 该方法之所以能够规避像素空间中的梯度冲突，源于其在特征空间中的内在协同。损失函数 L_k^{align} 相对于输入的梯度为:

$$\nabla_x L_k^{\mathrm{align}} = 2 \cdot \mathbf{J}_{\Phi_k}^{\top}(x) \cdot (\Phi_k(x) - t_k) \tag{5}$$

其中，$\mathbf{J}_{\Phi_k}(x)$ 是 x 中特征提取函数的雅可比矩阵。

梯度对齐主要由两个因素促成：1）所有模型的优化目标都是缩短与其固定目标 t_k 之间的距离。由于各模型的特征空间在经验上具有对齐性，差异向量 $\Phi_k(x) - t_k$意味着高层特征变化的方向相似。2）$\mathbf{J}_{\Phi_k}(x)$ 将变化从特征空间映射回像素空间。由于 Φ_k 的设计对高层特征变化十分敏感，其雅可比矩阵的主方向对应于图像中与特征表达最相关的结构变化。

因此，尽管不同模型的雅可比矩阵 $\mathbf{J}_{\Phi_k}$ 在数学上各不相同，但它们都会将相似的“特征失真意图”映射到像素空间。这使最终的像素空间梯度 $\nabla_x L_k^{\mathrm{align}}$ 方向趋于一致，从而产生建设性叠加，而非相互抵消。这一机制使我们能够高效生成可同时对抗多个异构模型的对抗性扰动，而无需复杂的梯度手术。

3.3 算法实现与细节

本节详细介绍 ATFS 框架的具体实现步骤、优化流程和关键设计选择，以确保结果可复现。

3.3.1 优化算法。 算法1概述了 ATFS 的完整工作流程。其核心目标是迭代最小化第 3.2.2 节定义的特征对齐损失，从而生成能够同时对多个异构模型奏效的对抗性扰动。流程图如图2所示。

该算法分为两个阶段。第一阶段（第 2–5 行）涉及预计算目标特征。在此阶段，固定的目标特征

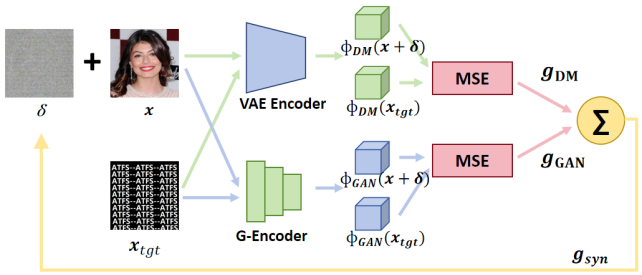


图2：通用特征锚定：通过共享特征空间统一异构模型

算法1 架构无关的定向特征协同（ATFS）

输入： 目标模型 $\{M_k\}_{k=1}^{K}$；干净图像 x；目标图像 x_{tgt}；特征提取器 $\{\Phi_k\}_{k=1}^{K}$；预算 ϵ，步数 T，步长 α。
输出： 对抗样本 x_{adv}。

1: **初始化：** $x^{(0)} \leftarrow x$, $\delta_0 \leftarrow 0$
2: ▷ **阶段1：预计算目标特征**
3: **对于** $k = 1$ 至 K **执行**
4: $\mathbf{t}_k \leftarrow \Phi_k(x_{tgt})$ ▷ 提取目标特征方向
5: **结束循环**
6: ▷ **第 2 阶段：协同优化循环**
7: **对于** $t = 0$ 到 $T-1$ **执行**
8: $g_{syn} \leftarrow 0$ ▷ 初始化协同梯度
9: **对于** $k = 1$ 到 K **执行**
10: $\mathbf{f}_k \leftarrow \Phi_k(x+\delta_t)$ ▷ 前向传播
11: $\mathcal{L}_k \leftarrow \|\mathbf{f}_k - \mathbf{t}_k\|_2^2$ ▷ 特征对齐损失
12: $g_k \leftarrow \nabla_\delta \mathbf{L}_k$ ▷ 计算梯度
13: $\hat{g}_k \leftarrow g_k/(\|g_k\|_2 + \xi)$ ▷ 梯度归一化
14: $g_{syn} \leftarrow g_{syn} + \omega_k \cdot \hat{g}_k$ ▷ 累积梯度
15: **结束循环**
16: ▷ 通过投影梯度下降更新扰动
17: $\delta_{t+1} \leftarrow$裁剪$_{[-\epsilon,\epsilon]}\left(\delta_t - \alpha \cdot \text{符号}(g_{syn})\right)$
18: **结束循环**
19: $x_{adv} \leftarrow$剪裁$_{[0,1]}(x+\delta_T)$
20: **返回** x_{adv}

向量 t_k 从目标图像 x_{tgt} 中为每个模型提取，作为优化的共同指引。第二阶段（第 7–19 行）是协同优化循环。其目标是找到一种扰动 δ，使扰动图像的特征 f_k 同时向各自的目标 t_k靠近。在每次迭代中，算法都会计算每个模型的损失梯度，对其进行归一化，然后将这些梯度线性聚合为协同梯度g_{syn}。最后，使用投影梯度下降（PGD）更新扰动。

3.3.2 关键实现细节。目标图像选择：如算法输入所示，x_{tgt} 可以是任何与原始 x无关的图像。为了最大限度地扩大特征空间中的失真，我们选择具有强烈且结构化视觉模式的目标图像，以便提供与原始特征明显不同的方向。具体而言，借鉴图像感知和对抗模式设计领域的经典研究，我们采用三种公认能够引发强烈视觉反应的模式：（1）高对比度条纹；（2）莫尔条纹；（3）过于繁杂的纹理。这些模式的特征分布与自然图像明显不同，能够有效引导优化过程，使特征空间产生显著且稳定的偏移。

梯度归一化：此步骤对于实现有效协同至关重要。不同模型之间的特征维度、激活幅度和损失曲面尺度存在显著差异，如果直接对原始梯度 g_k 求和，优化可能会被梯度幅度最大的模型主导。我们对每个模型的梯度应用 L_2 归一化：

$$\hat{g}_k = \frac{g_k}{\|g_k\|_2 + \xi} \tag{6}$$

其中，ξ是一个较小的常数（例如，10^{-8}），用于保证数值稳定性。这样可以确保在梯度聚合过程中，每个模型的贡献由其梯度方向而非幅度决定，从而公平、有效地融合来自异构模型的优化信号。

4 实验

本节通过实证评估所提出的架构无关的定向特征协同（ATFS）框架。我们设计了一系列定性和定量实验，以回答五个核心研究问题（RQs），验证该方法的有效性和实用价值：

RQ1：ATFS 能否有效打破“防御孤岛”，在异构生成模型上实现通用防御，并优于专用方法和现有集成策略？

RQ2：ATFS 的特征空间对齐机制是否真正架构无关，能否保护更多种类的生成模型（例如 VQ-VAE）？

RQ3：ATFS 的计算效率、收敛速度和扰动不可感知性如何？这些都是实际部署的关键指标。

RQ4：ATFS 生成的对抗样本能否抵御图像传输和处理过程中常见的信号失真（例如压缩、缩放）？

RQ5：ATFS 的防御有效性在不同人口统计属性（例如性别、种族）下是否保持一致，从而避免算法偏见？

4.1 实验设置

为了全面评估负责任的人工智能框架下的防御方法[21]，我们采用了多维评估设置，涵盖以下方面：

数据集与模型：实验在广泛使用的 VGG-Face2 数据集子集上进行[22]。为模拟异构威胁环境，我们构建了两个核心场景：

场景 A（扩散模型 +生成对抗网络）：结合 Stable Diffusion v1.5（基于概率密度建模）和 StarGAN v2[11]（基于判别边界）。

场景 B（扩散模型 + VAE）：包括 Stable Diffusion v1.5 和 VQ-VAE（基于离散码本重建）。

基线方法：我们选取了四类具有代表性的基线方法：（1）最先进的架构专用方法：适用于扩散模型的*Mist*[9]、适用于生成对抗网络的*Anti-Forgery*[10]，以及适用于 VQ-VAE 的标准 PGD攻击；（2）朴素联合方法：在像素空间中通过简单求和汇总不同模型的损失梯度；（3）梯度校正方法：应用*PCGrad*[15] 算法进行梯度协调（称为 PCGrad 协调式），以投影并校正来自异构模型的冲突梯度。

评估指标：对于图像生成和模型微调任务（例如 DreamBooth[23]），核心目标是防止模型学习特定的视觉概念（例如某人的面部）。我们使用 FID[24]量化生成图像与真实分布之间的距离，并使用 IL-NIQE[25] 评估感知质量。数值越高，表示与原始概念的偏差越大、质量越低，说明保护效果越好。对于图像编辑和图像转换任务（例如 SDEdit、StarGANEdit[11], VAE 重建），重点是破坏编辑准确性。我们使用 MS-SSIM 和 PSNR[25] 衡量受攻击结果与理想（未受攻击）编辑结果之间的相似度，并结合 IL-NIQE 进行评估。在此情况下，较低的 MS-SSIM/PSNR 和较高的 IL-NIQE 表明编辑结果受到严重失真，证明防御有效。

实现细节：所有实验均在 NVIDIA RTX 3090 GPU 上完成。ATFS 的默认设置为：迭代次数$T = 100$、扰动预算 $\epsilon = 6/255$、步长 $\alpha = \epsilon/10$，并使用结构化噪声图像作为特征锚点。

4.2 核心有效性（RQ1）

本实验验证 ATFS 在异构扩散模型 + 生成对抗网络场景中的核心防御能力。我们从两个方面进行评估：1）验证对抗训练样本在其目标编辑任务上的直接效果（扩散模型对应 SDEdit，生成对抗网络对应 StarGANEdit）；2）验证这些样本在未经任何重新训练的情况下，能否迁移至另一种基于扩散的模型 DreamBooth。

4.2.1 SDEdit 和 *StarGANEdit* 上的防御性能。如表1所示，现有方法普遍表现出“防御孤岛”效应。MIST 对 SDEdit 有效（多尺度结构相似性：0.6387），但其对 StarGANEdit 的防御性能急剧下降（多尺度结构相似性：0.8011）；Anti-Forgery 则呈现相反趋势。这表明架构专用优化存在局限。尽管朴素联合攻击和 PCGrad 校正方法取得了一定的跨模型效果，但其性能显著低于 ATFS。例如，在 StarGANEdit 任务中，ATFS 的多尺度结构相似性（0.5228）远低于 PCGrad（0.6204）。ATFS 在两项编辑任务的所有指标上均取得最佳结果，证明其能够通过特征级协同实现有效的通用防御。

值得注意的是，ATFS 不仅在跨模型防御方面表现出色，甚至在各自的单模型防御任务中也超越了最先进的专用方法。这一现象源于对抗范式的根本转变：尽管 Mist 在扩散模型的 VAE 潜在空间（一个特征空间）中运作，但其攻击目标本质上是特定于模型的去噪轨迹或编码器特征距离。相比之下，ATFS 借助目标引导图像干扰跨模型共享高层特征对齐空间，从而实现更彻底的特征级失真。

表1：针对 DM+GAN 异构防御的定量结果

方法	SDEdit			StarGANEdit		
	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑
原始	0.8516	31.26	28.1654	0.9989	49.94	30.3326
MIST	0.6387	20.65	39.6495	0.8011	24.43	32.1648
Antiforgery	0.7534	26.42	33.7946	0.5464	15.29	42.4615
朴素联合	0.6812	21.22	35.3013	0.6503	17.78	38.7418
PCGrad 协调式	0.6563	20.85	37.6615	0.6204	16.46	39.3195
Ours	**0.5122**	**18.69**	**41.2712**	**0.5228**	**13.63**	**43.3251**

类似地，尽管 Anti-Forgery 通过在像素空间中添加扰动来攻击生成对抗网络的输出层——这是一种特定于模型的对抗博弈——ATFS 却直接在特征空间层面破坏生成器的编码核心。这一实证结果验证了，攻击跨模型共享高层特征对齐空间（一个更高层次、架构无关的表示层）本质上比针对各个模型特有的低层机制更具破坏性。

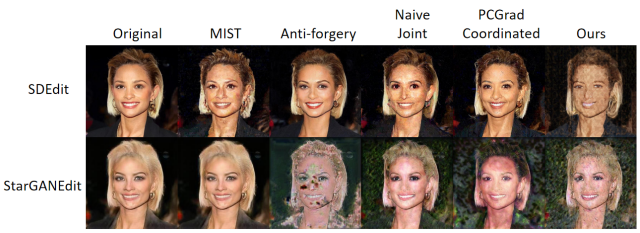


图3：跨架构防御效果的定性可视化

如图3所示，视觉结果与定量分析一致。经 MIST 保护的图像无法被 SDEdit 编辑，但 StarGANEdit 仍能成功改变头发颜色；经 Anti-Forgery 保护的图像则呈现相反趋势。相比之下，经 ATFS 保护的图像经两种编辑器处理后都会产生完全失真的结果，且不再保留任何面部身份特征。这从视觉上验证了 ATFS 通过特征协同机制，仅凭一次扰动便能实现对异构模型的可靠双重防护，从根本上打破“防御孤岛”。

4.2.2 对未见过的基于扩散的模型（*DreamBooth*）的迁移能力。为了进一步验证我们的方法在同一生成范式内跨模型的泛化能力，我们将针对 DM+GAN 场景训练的对抗样本直接应用于 DreamBooth（另一种专为主体驱动生成设计的基于扩散的模型），不进行任何重新训练或适配。如图4所示，ATFS 生成的对抗样本能有效破坏 DreamBooth 学习和重建目标身份的能力。该模型无法生成受保护主体连贯的图像，生成的反而是失真或无关的内容。

这种可迁移性表明，ATFS 生成的扰动不仅能有效抵御训练期间使用的特定扩散模型



图4：对 DreamBooth 的可迁移性。ATFS 生成的对抗样本能有效干扰 DreamBooth 中的主体驱动生成

（SDEdit），还能成功泛化到另一种基于扩散的架构（DreamBooth）。这一能力源于 ATFS 的核心机制：它攻击的是共享的高层特征空间，而非特定于模型的低层机制，因此生成的扰动在同一范式下的不同模型中都能保持有效。

4.3 泛化能力验证（RQ2）

为验证我们的框架是否受限于特定的模型组合，我们在“扩散模型 + VQ-VAE”场景中评估 ATFS。VQ-VAE 依赖离散码本重建，其机制与扩散模型和生成对抗网络都存在本质差异。我们仅切换相应模型的特征提取器（编码器）来执行攻击，而不改变底层算法逻辑。

4.3.1 对 *SDEdit* 和 *VAE*重建的直接防御性能。如表 2 所示，MIST 等专用方法几乎无法抵御 VQ-VAE（多尺度结构相似性：0.9194），这证实了现有方法在架构上的局限性。相比之下，ATFS 在 VAE重建上的表现显著更优（多尺度结构相似性：0.7443），同时在 SDEdit 上也保持了出色的效果，展现出其架构无关能力。定性结果见图 5。

表2：扩散模型+VQ-VAE 架构泛化的定量结果

	SDEdit			VAE重建		
方法	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑
MIST	0.6177	**17.86**	40.5124	0.9194	29.74	31.3451
PGD	0.6828	23.17	35.2631	0.7668	19.41	35.2115
朴素联合	0.6614	22.03	35.5144	0.7961	20.31	34.2142
PCGrad 协调式	0.6536	20.67	37.1154	0.8183	21.85	33.8145
Ours	**0.5215**	18.78	**41.0214**	**0.7443**	**16.51**	**36.7187**

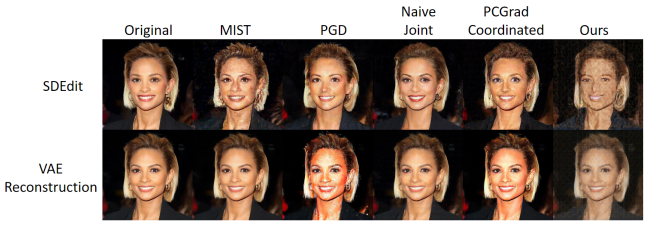


图 5：扩散模型 + VQ-VAE 场景下的定性结果

4.3.2 对多种异构模型的可扩展性。为严格评估 ATFS 的可扩展性，我们同时测试了其对三种异构模型的防御能力：扩散模型（Stable Diffusion）、生成对抗网络（StarGAN）和 VQ-VAE。这一具有挑战性的场景评估了该框架在统一优化中应对更高架构多样性的能力。

图 6 展示了定性结果，其中 ATFS 生成的单个扰动可同时干扰三个生成过程。对抗样本会导致：（1）SDEdit 输出中的身份信息丢失；（2）StarGANEdit 结果中的结构失真；（3）VQ-VAE 重建中的感知质量受损。这表明，特征空间对齐能够有效扩展至多模型防御。

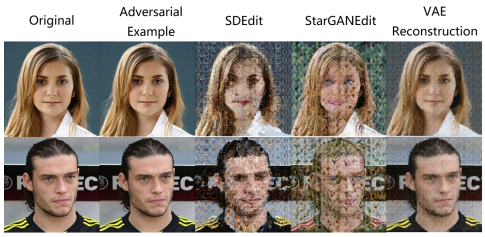


图 6：跨架构防御有效性。每行代表一个测试样本。

值得注意的是，ATFS 无需修改算法即可实现这一点。若要防御额外的模型，只需将其特征提取器 Φ_k 和目标特征 t_k 纳入现有的优化目标（式 7）即可。这种即插即用的可扩展性——无需重新设计攻击或进行超参数调优——凸显了我们的特征空间方法在包含多种生成式架构的真实环境中的实用性。

4.4 实用性分析（RQ3）

我们系统评估了 ATFS 在不同扰动预算 $\epsilon \in 2/255$、4/255、8/255、16/255（表3）下的防御有效性。结果表明，即使在严格的不可感知扰动预算 $\epsilon = 2/255$ 下，ATFS 仍能保持强劲的效果，显著扰乱 SDEdit（多尺度结构相似性：0.646）和 StarGANEdit（多尺度结构相似性：0.5915）。值得注意的是，随着 ϵ 增大，性能提升的边际收益逐渐递减，这表明只需极小的像素级失真即可实现大部分防御效果。这种效率与我们的理论洞见相符：通过在共享特征空间中运作，ATFS 能够以单位像素变化造成最大的特征级失真。因此，ATFS 在攻击效力和视觉隐蔽性之间实现了良好的权衡，适用于扰动可见性至关重要的现实世界部署场景。

表 3：不同扰动预算下的性能比较

	SDEdit			StarGANEdit		
扰动预算	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑	多尺度结构相似性↓	峰值信噪比↓	IL-NIQE↑
2/255	0.646	20.56	38.6471	0.5915	15.61	39.4461
4/255	0.5834	19.67	39.5461	0.554	15.9	40.3215
8/255	0.4807	18.26	40.8471	0.426	15.09	42.5147
16/255	0.4047	17.08	42.7214	0.2914	13.24	46.1541

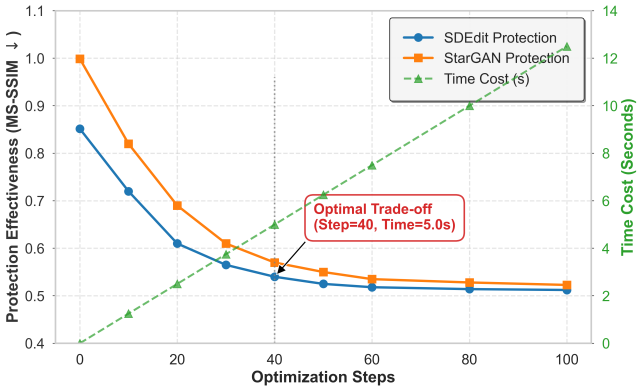


图7：效率分析：保护有效性与时间开销

收敛与权衡：为直观量化性能与开销之间的权衡，图7采用双轴可视化。实线（保护有效性）在初始阶段急剧下降，而虚线（时间开销）则线性增长。我们确定了一个“最佳权衡”点：ATFS 仅需 40 次迭代（约 5.0 秒），即可达到最终收敛性能的 90% 以上。这种效率得益于目标引导图像提供了明确的全局优化方向，大幅减少了异质梯度场中的振荡。与需要高开销投影的 PCGrad 不同，ATFS 的线性聚合使其非常适用于实时隐私保护系统。

4.5 鲁棒性分析（RQ4）

在现实世界的社交媒体传播中，图像不可避免地会经历有损压缩或噪声干扰。我们模拟了三种常见的信号失真：高斯噪声、JPEG压缩和缩放。

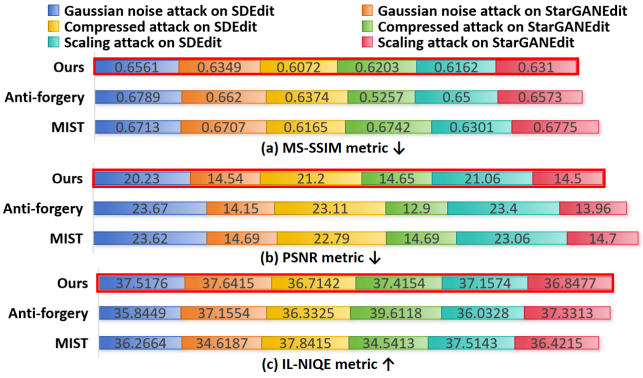


图8：鲁棒性分析：不同攻击场景下的性能

如图8所示，尽管所有方法的效果都有所下降，ATFS在所有场景下仍保持最强的防护能力。尤其是在颇具挑战性的 JPEG压缩场景中，基线方法（如 Anti-Forgery）的效果显著下降（多尺度结构相似性升至 0.6374）。这是因为传统的像素级攻击通常依赖细微的高频噪声模式，而这些模式在 JPEG 的离散余弦变换（DCT）和量化过程中很容易被滤除。相比之下，ATFS 的多尺度结构相似性仍保持在较低的 0.6072。这一优势源于我们从攻击浅层像素空间转向高层特征空间这一根本性改变。由于特征表示本身具有尺度不变性和抗噪能力，通过特征协同生成的对抗性扰动也成功继承了这些鲁棒特性。

4.6 公平性分析（RQ5）

为确保 ATFS 不会引入算法偏见，我们评估了其在人脸图像上的表现，这些图像具有不同的人口统计属性（性别、年龄、种族）。

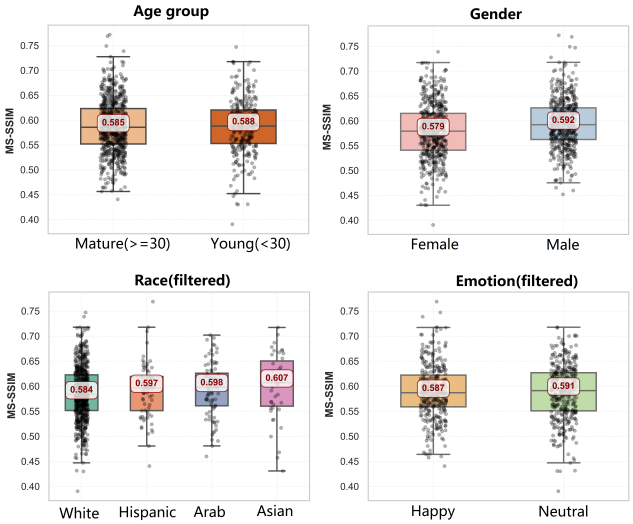


图9：不同人口统计群体的性能变化分析

如图7所示，不同子群体（例如男性与女性、不同种族群体）的防御性能（多尺度结构相似性）箱线图高度重叠，中位数也非常接近。Kruskal-Wallis 检验进一步证实，不同属性群体之间的性能分布不存在统计学上的显著差异（$p > 0.05$）。这表明 ATFS 能够提供公平且一致的保护，避免因人口统计属性而造成防御效果不均，这符合负责任的人工智能的伦理要求。

5 结论

面对现实环境中异构生成模型带来的威胁，现有主动防御方法因其架构特异性而陷入“防御孤岛”。本文指出，根本原因在于异构模型在像素空间中的优化目标存在根本冲突，从而阻碍了梯度协同。为解决这一问题，我们基于不同架构间特征空间对齐这一经验观察，提出一种新型防御范式。我们设计了架构无关的

定向特征协同（ATFS）框架。通过利用目标引导图像，将异构防御统一为特征空间对齐任务，ATFS 从理论上实现了梯度对齐，并在实践中实现高效、轻量的扰动生成（例如，每张图像耗时 ~2 秒，而梯度投影基线需要 > 15 秒）。实验表明，ATFS 在异构场景中达到最先进水平，并且能够轻松扩展至其他架构。该方法在低扰动预算下依然有效，能够抵御常见信号失真，并对不同人口群体公平，因而为可信生成式人工智能安全提供了一种兼顾效率与可靠性的实用方案。未来工作将聚焦于三个方向：1）动态搜索目标特征方向，以实现最大程度的干扰；2）扩展至三维时空特征空间，用于视频保护。

参考文献

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser 和 B. Ommer，《基于潜在扩散模型的高分辨率图像合成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10684–10695页，2022年。[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville 和 Y. Bengio，《生成对抗网络》，载于神经信息处理系统进展，第2672–2680页，2014年。[3] L. Zheng, N. Yu, A. Saberi, M. Backes, M. Fritz 和 Y. Zhang，《Unganable：防御基于生成对抗网络的人脸操纵》，载于第*32*届*USENIX*安全研讨会（*USENIX Security 23*），第7213–7230页，2023年。[4] I. Korshunova, W. Shi, J. Dambre 和 L. Theis，《使用卷积神经网络快速交换人脸》，载于*IEEE*国际计算机视觉会议论文集，第3697–3705页，2017年。[5] J. Smith 和 B. Johnson，《在线虚假信息行动中扩散网络的图结构》，载于万维网大会（*WWW*）论文集，第200–212页，2024年。[6] J. C. Costa, T. Roxo, H. Proença 和 P. R. M. Inacio，《深度学习如何看待世界：对抗攻击与防御综述》，*IEEE Access*，第12卷，第61113–61136页，2024年。[7] J. Liu, Y. Li, Y. Guo, Y. Liu, J. Tao 和 Y. Niu，《视觉领域对抗样本的生成与对策：综述》，人工智能评论，第57卷，第8期，第199页，2024年。[8] H. Chen, Q. Xiang, J. Hu, M. Ye, C. Yu, H. Cheng 和 L. Zhang，《图像生成中的扩散模型全面探索：综述》，人工智能评论，第58卷，第4期，第99页，2025年。[9] C. Liang 和 X. Wu，《Mist：迈向更优的扩散模型对抗样本》，*arXiv*预印本 *arXiv:2305.12683*，2023年。[10] R. Wang, Z. Huang, Z. Chen, L. Li, J. Chen 和 L. Wang，《Anti-Forgery：通过对抗性知觉感知扰动实现隐蔽且稳健的深度伪造破坏攻击》，*arXiv*预印本 *arXiv:2206.00477*，2022年。[11] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim 和 J. Choo，《StarGAN：用于多域图像到图像转换的统一生成对抗网络》，载于*IEEE*计算机视觉与模式识别会议论文集，第8789–8797页，2018年。[12] W. Zhang 和 H. Li，《生成式人工智能驱动的网络攻击全面威胁模型》，载于*ACM SIGSAC*计算机与通信安全会议论文集，第55–70页，2023年。[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon 和 B. Poole，《通过随机微分方程进行基于分数的生成建模》，国际学习表征大会（*ICLR*），2021年。[14] R. Vershynin，《高维概率：数据科学导论与应用》，第47卷。剑桥大学出版社，2018年。[15] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman 和 C. Finn，《多任务学习的梯度手术》，载于神经信息处理系统进展，第33卷，第5824–5836页，2020年。[16] O. Sener 和 V. Koltun，《作为多目标优化的多任务学习》，载于神经信息处理系统进展，第31卷，2018年。[17] S. Lin, X. Zhang, W. Susilo, X. Chen 和 J. K. Liu，《GDR-GMA：通过方向校正和幅值调整梯度实现机器遗忘》，载于第*32*届*ACM*国际多媒体会议论文集，第9087–9095页，2024年。[18] S. Shin, J. Chen, E. Wexler, H. Zhang, R. Han 和 B. Y. Zou，《Glaze：保护艺术家免受文本到图像模型的风格模仿》，*arXiv*预印本 *arXiv:2302.04222*，2023年。[19] B. Zhang, Y. Song, F. Zhang 和 W.-C. Lee，《生成式人工智能去中心化数据保护协议的设计与实现》，载于第*30*届数据库系统高级应用国际会议论文集，2025年。

[20] C. Zhao, B. Vijay, C.-Y. Lee 和 A. Rabinovich，《GradNorm：深度多任务网络中用于自适应损失平衡的梯度归一化》，载于国际机器学习大会，第794–803页，PMLR，2018年。[21] T. Lin, M. Yen, C. Meng, Y. Mao, R. Jia, A. Raghunathan 和 L. Schmidt，《文本到图像模型的整体评估》，载于神经信息处理系统进展，第36卷，第49181–49201页，2023年。[22] Q. Cao, L. Shen, W. Xie, O. M. Parkhi 和 A. Zisserman，《Vggface2：用于跨姿态和年龄识别人脸的数据集》，载于*2018*年*IEEE*第*13*届国际自动人脸与手势识别会议（*FG 2018*），第67–74页，IEEE，2018年。[23] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein 和 K. Aberman，《DreamBooth：用于主体驱动生成的文本到图像扩散模型微调》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第22500–22510页，2023年。[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler 和 S. Hochreiter，《通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡》，神经信息处理系统进展（神经信息处理系统大会），第30卷，2017年。[25] C. Ma, Z. Sun, Z. Lv, S. Xu, F. Chen 和 Y. Sun，《图像质量评估综述：见解、分析与未来展望》，*arXiv*预印本 *arXiv:2502 08540*，2025年。

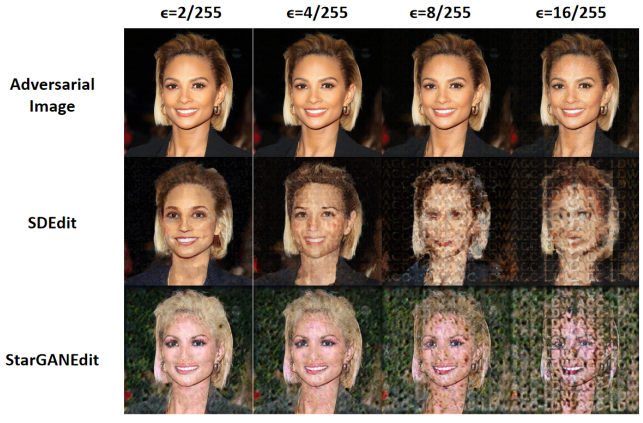
不同扰动预算下的视觉效果

本附录对不同扰动预算（ϵ）下对抗样本及其防御效果进行了扩展的视觉分析，以补充正文第4.4节。

图A.1系统展示了扰动预算 $\epsilon \in 2/255$，4/255，8/255，16/255对防御效果的影响。每一列对应一个特定的 ϵ 值。

视觉结果与正文表4中的定量数据高度一致。随着 ϵ 增加，SDEdit和StarGANEdit的输出质量都急剧下降。当扰动预算达到 $\epsilon = 16/255$时，两个编辑器的输出都完全丧失了原始身份信息，变得无法辨认，成为抽象图案或噪声，从而成功实现内容保护。

即使在 $\epsilon = 2/255$时，对抗样本的视觉质量仍然很高，却已对编辑结果造成明显干扰。这为实际应用提供了灵活性：对于不可感知性要求严格的场景，可以选择较小的预算；而对于安全需求较高的场景，则可以使用较大的预算。



图A.1：不同扰动预算下 ATFS 防御的可视化效果