中国科学:信息科学 在审文章 onlinelatex

SCIENTIA SINICA Informationis

论文

浅遗忘：从解耦组件擦除角度再思考文本引导扩散模型中的概念遗忘

赵梦楠[1], 郑天航[2], 王波[3], 张立和[3*], 尹宝才[1]

1. 大连理工大学计算机科学与技术学院, 大连 116081
2. 浙江大学计算机科学与技术学院, 杭州 310058
3. 大连理工大学信息与通信工程学院, 大连 116081
* 通信作者. E-mail: zhanglihe@dlut.edu.cn

国家自然科学基金 (批准号: 62431004) 资助项目; 国家自然科学基金 (批准号: 62276046) 资助项目; 国家自然科学基金 (批准号: 62572426) 资助项目

摘要 近年来, 文本引导的扩散模型在图像生成领域取得了显著进展, 同时也带来了隐私泄露及有害内容生成等伦理问题. 针对这些挑战, 现有方法通常依赖对齐机制, 即通过预定义的擦除参考对预训练模型权重进行微调. 然而, 此类方法存在两方面不足: 首先, 其受限于文本空间的表征能力, 对预定义的擦除参考高度依赖; 其次, 在执行概念擦除时, 模型效用可能受到显著影响. 针对上述问题, 本文提出 '浅遗忘' 概念, 其核心思想是通过解耦不期望概念中的关键擦除组件实现选择性遗忘. 具体而言, 浅遗忘包括组件提取与遗忘微调两部分组成. 首先, 在交换解缠策略的监督下预训练组件提取模块, 将概念嵌入分解为不同类型的组件; 随后, 通过微调模型权重, 选择性地移除关键擦除组件, 同时保留非关键组件, 从而在高效遗忘不期望内容的同时尽可能保持模型性能. 大量实验证明, 浅遗忘能够更好地平衡概念擦除效果与模型性能保持之间的关系.

关键词 扩散模型遗忘, 浅遗忘, 组件提取模块, 交换解缠策略, 性能权衡

1 引言

近年来, 文本到图像生成技术 [1~3], 尤其是文本引导的扩散模型 [4~6], 在图像合成质量和效率方面取得了显著突破 [7~10]. 这些技术已广泛应用于数字艺术创作、媒体内容生成以及医疗健康等领域 [11~14]. 然而, 在大规模预训练过程中, 文生图模型不可避免地继承了数据集中的偏差及敏感信息, 从而带来一系列伦理风险, 包括版权侵权、隐私泄露和有害内容生成 [15~18]. 针对这些问题, 已有研究提出了基于重训练的改进方法, 例如安全隐式扩散技术 [19], 以有效抑制不当内容生成.

不同于模型的完整重训练, 机器遗忘技术 [20~22] 旨在通过有目标或无目标的微调方式 [23], 在有限的资源开销下实现对生成内容中不期望效果的抑制. 具体而言, 无目标微调方法主要通过对生

引用格式: 赵梦楠, 郑天航, 王波, 张立和, 尹宝才. 浅遗忘: 从解耦组件擦除角度再思考文本引导扩散模型中的概念遗忘. 中国科学: 信息科学, 在审文章
Mengnan Zhao, Tianhang Zheng, Bo Wang, Lihe Zhang, Baocai Yin. ShallowUnlearn: Rethinking concept unlearning through disentangled component-level erasure in text-guided diffusion models. Sci Sin Inform, for review

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

成模型的注意力模块权重进行修改, 使模块在特征提取过程中对不期望概念相关的信息响应显著降低. 有目标微调方法 [24] 的核心思想是将模型对不期望概念的预测引导至预定义的文本或语义锚点. 例如, 可以采用空提示 [25]、更高层级的广义概念类别 [26]、由文本提示部分嵌入构成的表示 [24,27], 以及多个文本嵌入的加权组合 [28].

尽管近年来针对扩散模型的遗忘技术已取得一定进展 [29~31], 在概念擦除过程中, 目标锚点选择不当或误移除有效语义特征时, 模型效用仍可能受到显著削弱 [32~35]. 针对这一挑战, 本文关注的核心问题是: 如何对不期望概念的关键擦除组件进行有效解耦并加以移除. 这一问题的核心思想是能否通过仅移除这些关键组件, 同时保留其他部分, 以在实现概念遗忘的同时尽量减轻擦除操作对模型整体性能的影响.

为此, 本文提出浅遗忘概念及其相应的实现框架, 其核心包括组件提取模块和交换解缠策略. 不同于以遗忘恢复难度和遗忘泛化能力为主要研究目标的深遗忘方法, 浅遗忘更加关注在概念擦除过程中对模型整体性能的保持. 具体而言, 本文首先形式化地定义关键擦除组件的识别与移除范式; 随后设计组件提取模块, 并结合交换解缠策略构建相应的监督目标. 在交换解缠策略的引导下, 组件提取模块能够在预训练阶段将不期望概念的嵌入有效分解为多种类型的组件. 最后, 浅遗忘利用预训练好的组件提取模块识别并移除关键擦除组件, 并通过微调模型权重实现概念擦除. 同时, 该过程最大限度地保留与不期望属性无关的信息, 从而在保证概念移除有效性的同时维持模型整体性能.

本文的主要贡献如下: (1) 设计组件提取模块与交换解缠策略, 实现对概念嵌入的有效分解与关键擦除组件的准确识别, 并在此基础上提出相应的关键组件移除方法; (2) 提出浅遗忘概念, 并构建基于关键组件移除的实现范式, 在实现高效概念擦除的同时, 最大限度降低对模型整体性能的影响; (3) 在多个基准数据集上的实验验证了浅遗忘方法的有效性, 实验结果表明, 该方法在实现概念擦除的同时, 能够更有效地保持模型整体性能.

本文的章节安排如下: 第2节系统回顾了扩散模型遗忘学习的最新研究进展; 第3节详细阐述了所提出的浅遗忘技术框架及其核心方案; 第4节通过系统实验验证了浅遗忘方法在概念擦除效果及模型整体性能保持方面的有效性; 最后, 第5节对全文进行总结, 并展望未来的研究方向.

2　相关工作

近年来, 扩散模型在文本到图像生成任务中取得了显著进展, 能够根据不同的文本提示合成高质量的图像 [36,37]. 相较于生成对抗网络, 扩散模型通过逐步将高斯噪声转化为结构化的视觉内容, 在生成稳定性方面表现更为突出 [38~40]. 目前, 该类模型已广泛应用于风格迁移 [41]、图像修复 [42] 以及超分辨率重建 [43,44] 等任务, 表现出良好的任务迁移能力. 然而, 相关技术的发展也引发了安全、隐私与伦理方面的担忧 [44,45], 尤其是在生成不适宜内容 (Not Safe For Work, NSFW) 或受版权保护的素材 [46,47], 促使研究者关注模型层面的概念擦除与机器遗忘问题.

为在无需完全重新训练的情况下实现对特定数据的选择性遗忘, 机器遗忘技术 [48~50] 提供了一种有效途径. Liu 等人 [51] 从对抗性视角提出了 “最坏情况遗忘集” , 在上层优化过程中通过放大遗忘难度来模拟最坏情况, 并在下层执行常规训练与遗忘操作, 从而在数据影响消除与模型性能保持之间取得平衡. 与之不同, Kwok 等人 [52] 利用隐性概念的内在几何特征, 将其空间信息融入文本提示中, 从而引导模型主动识别并解耦不期望概念, 并在生成阶段有效抑制隐性概念带来的干扰.

早期针对扩散模型的机器遗忘研究主要依赖对预训练模型权重的直接微调, 通过改变模型内部参数以削弱或消除其对特定概念的生成能力 [48,53,54]. FMN (Forget-Me-Not) [55] 是最早针对文本到图像扩散模型提出系统化概念遗忘框架的研究之一, 其通过最小化注意力图中与目标概念相关

ACCEPTED

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

的激活响应, 实现对不期望概念的抑制. ESD [56] 进一步通过微调扩散模型中的噪声预测网络, 引入负引导信号, 使生成过程在参数空间中偏离目标概念, 从而在权重层面实现概念消除. 针对多概念同时遗忘的需求, SepME [15] 提出了一种可分离的概念擦除建模方式. 该方法将不同概念的遗忘过程表示为独立的参数更新方向, 并通过约束各更新方向在参数空间中近似正交, 以降低多概念擦除过程中相互干扰的风险. 上述方法在概念消除效果方面表现较为直接, 但普遍需要对模型权重进行较大幅度的更新, 可能对模型整体性能和泛化能力产生影响. 因此, 也有部分研究通过构建保留概念集合来增强模型的效用保持能力 [48, 57, 58], 其方法表现受到保留概念集合质量的显著影响.

随着对参数效率和模型迁移能力关注的提升, 研究者逐渐提出在不直接修改主模型权重的前提下, 通过引入轻量级可训练结构实现概念遗忘. 该类方法通常通过插入适配器模块或低秩参数来实现对特定概念的调控, 具有较好的可移植性和训练效率. MACE [16] 采用 LoRA（Low-Rank Adaptation）等轻量级适配器结构, 对扩散模型中的跨注意力层进行独立调节, 在提高擦除稳定性的同时, 有效缓解了对无关概念生成能力的影响. AGE [59] 将概念空间建模为图结构, 在概念擦除过程中动态选择更优的擦除目标, 而非使用固定的空提示, 从而在一定程度上降低了对其他概念的副作用.

除基于微调与适配器的方法外, 部分研究关注无需训练或训练成本较低的概念遗忘策略. RECE [60] 通过闭式解推导新的目标文本嵌入, 并结合正则化约束以对齐无害概念, 从而实现快速、高效的概念擦除. 该类方法通常具有较低的计算开销, 适合对模型进行快速调整, 但其遗忘效果和稳定性仍有待进一步验证. ACE [61] 提出一种基于交叉零空间投影的概念编辑方法, 通过将目标概念扰动投影至常规语义的零空间, 在擦除不期望概念的同时尽量保持生成语义一致性与图像质量. 该方法侧重于通过构造投影空间抑制不期望概念, 而非从模型内部语义组合结构中识别并剥离关键语义组件; 同时, 其效果在一定程度上依赖于预定义的保留语义空间.

近年来, 研究者逐渐认识到, 仅通过生成结果观察难以全面评估概念遗忘的真实效果. Unlearn Concealment [62] 提出了概念检索得分和概念置信度得分等评估指标, 指出部分方法可能仅在推理阶段对目标概念进行遮蔽, 而未能从模型知识表示空间中实现真正删除. UnErasing [63] 对多种概念擦除方法的效果及其可逆性进行了系统分析, 发现多数现有方法仍可能通过逆向提示或对抗性操作恢复已被遗忘的概念, 揭示了当前方法在安全性和鲁棒性方面的不足. 针对上述问题, 部分研究提出多种评估协议 [64~66] 以及对抗训练策略 [18, 48], 以提升概念遗忘方法在复杂场景下的鲁棒性. 然而, 在追求更强遗忘效果的同时, 这类方法仍可能对模型整体生成能力造成一定程度的影响, 尤其在高安全性需求的应用场景中, 该问题尤为突出.

本文提出的浅遗忘方法不仅要求实现高效概念擦除, 而且期望最大限度地减轻遗忘过程对模型整体效用的影响. 不同于现有方法通常将不期望内容作为整体遗忘目标 [67~69], 浅遗忘方法认为不期望概念的表示中同时包含通用语义信息与不期望属性成分. 基于这一认识, 本文从不期望概念的嵌入表示中识别并分离对擦除起关键作用的组件, 在保留通用语义结构的前提下, 对不期望属性进行针对性干预. 通过对关键组件的精细化擦除与对一般语义信息的显式保留, 该方法在实现高效概念擦除的同时, 有效维持了模型的整体生成能力与性能稳定性.

3　提出方法

3.1　基础知识

给定概念嵌入 e 以及视觉噪声 $z \sim \mathcal{N}(0, \mathbf{I})$, 定义映射函数 $f_t(e, z; \theta) : \mathcal{E} \times \mathcal{Z} \to \mathcal{X}$, 用于表示时间步 $t \in \{1, \ldots, T\}$ 下扩散模型的预测结果. 机器遗忘技术的目标函数通常同时包含两部分：一是对

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

不期望概念 c_u（其嵌入为 e_u）的生成进行擦除，二是维持保留概念 $c_n \in \mathcal{C}_{\mathrm{RCs}}$（其嵌入为 $e_n \in \mathcal{E}_{\mathrm{RCs}}$）的生成质量，其整体优化目标可表示为：

$$\min_{\theta_{\mathrm{op}}} \mathbb{E}_{z\in\mathcal{Z}_D}\left[\left\|f_t(e_u,z;\theta_{\mathrm{op}})-f_t(e_p,z;\theta_0)\right\|_2+\sum_{e_n\sim\mathcal{E}_{\mathrm{RCs}}}\left\|f_t(e_n,z;\theta_{\mathrm{op}})-f_t(e_n,z;\theta_0)\right\|_2\right], \tag{1}$$

其中，θ_0 表示预训练模型的参数，θ_{op} 表示待优化的模型参数，并以 θ_0 作为初始值．$\mathcal{Z}_D$ 表示从视觉数据集中采样得到的噪声分布．e_p 表示用于引导概念 c_u 擦除过程的预定义目标锚点．上式左侧项用于引导模型对概念 c_u 的生成结果进行调整，右侧项旨在维持模型的整体性能．

公式（1）中右侧正则项的实际保留效果在很大程度上依赖于保留概念集合 $\mathcal{C}_{\mathrm{RCs}}$ 的选取质量，不同概念集合的设置往往会导致显著差异的模型效用保持结果．为降低保留概念集合选择对后续分析的影响，本文仅采用空提示（其嵌入表示为 $e_\emptyset$）作为保留概念 e_n，形成基础的参考漂移抑制项：

$$\mathcal{L}_{\mathrm{RDS}}=\mathbb{E}_{z,t}\left[\left|f_t(e_\emptyset,z;\theta_{\mathrm{op}})-f_t(e_\emptyset,z;\theta_0)\right|_2\right]. \tag{2}$$

此外，已有研究表明，目标锚点 e_p 与不期望概念嵌入 e_u 在语义空间中的初始距离与模型效用保持能力呈显著负相关关系 [70]．因此，本文主要探讨在概念擦除过程中，是否可以通过合理构造目标候选 f^t_{cand}，在保证概念擦除有效性的同时，最大限度地保持模型整体效用．为实现上述目标，本文从不期望概念的内部表示出发，直接对目标候选进行构造，即在移除不期望属性成分的同时尽量保留其一般语义信息．基于这一思路，目标候选的选取不再依赖于显式的保留概念集合，其约束条件可表述为如下范式约束：

$$\Phi\big(f_t(e_u,z;\theta_0),f^t_{\mathrm{cand}}\big)\ll\Phi\big(f_t(e_u,z;\theta_0),f_t(e_p,z;\theta_0)\big),\quad \text{s.t. } \arg\max g(f^t_{\mathrm{cand}})\neq\ell_u, \tag{3}$$

其中，$\Phi(\cdot,\cdot)$ 表示用于度量特征差异的距离函数．结合公式（1），本文在实现中将 $\Phi(\cdot,\cdot)$ 具体设定为 ℓ_2 范数．此外，$\arg\max g(f^t_{\mathrm{cand}})$ 表示固定分类器 $g(\cdot)$ 对候选特征 f^t_{cand} 的预测标签索引，而 ℓ_u 表示概念 c_u 对应的真实标签索引．

3.2　擦除关键组件的移除范式

为构造满足式 (3) 的候选项 f^t_{cand}，本文引入潜在差异 $\Delta f^{t,\theta}_u \triangleq f_t(e_u,z;\theta)-f_t(e_\emptyset,z;\theta)$，用于刻画不期望概念 c_u 对生成结果所引入的差异，其中 $e_\emptyset$ 表示空提示对应的嵌入向量．首先对 $\Delta f^{t,\theta}_u$ 进行分解，将其划分为若干具有可解释性的子组件；随后，仅对其中与概念擦除最为相关的关键组件进行替换．在形式化定义上，引入一个分解算子 $\mathcal{T}:\mathcal{X}\to\mathcal{X}^M$，用于将潜在差异 $\Delta f^{t,\theta}_u$ 映射为 M 个相互独立的组件．

$$\Delta f^{t,\theta}_u=\sum_{j=1}^{M}\mathcal{T}(\Delta f^{t,\theta}_u)_j. \tag{4}$$

令 $k\in\{0,\ldots,M\}$ 表示被擦除关键组件的索引，则候选特征 f^t_{cand} 可进一步表述为：

$$f^t_{\mathrm{cand}}=f_t(e_u,z;\theta_0)-\mathcal{T}(\Delta f^{t,\theta_0}_u)_k+\mathcal{T}^{\mathrm{new}}_k. \tag{5}$$

式 (5) 表明，用于表征概念 c_u 的关键组件 $\mathcal{T}(\Delta f^{t,\theta_0}_u)_k$ 被替换为新的组件 $\mathcal{T}^{\mathrm{new}}_k$．结合式 (5) 与式 (3)，分解算子 $\mathcal{T}$ 及可优化的模型参数 θ_{op} 应满足以下条件：

$$\left\|\mathcal{T}^{\mathrm{new}}_k-\mathcal{T}(\Delta f^{t,\theta_0}_u)_k\right\|_2\ll\left\|f_t(e_u,z;\theta_0)-f_t(e_p,z;\theta_0)\right\|_2,\quad s.t.,\ \arg\max g(f^t_{\mathrm{cand}})\neq\ell_u. \tag{6}$$

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

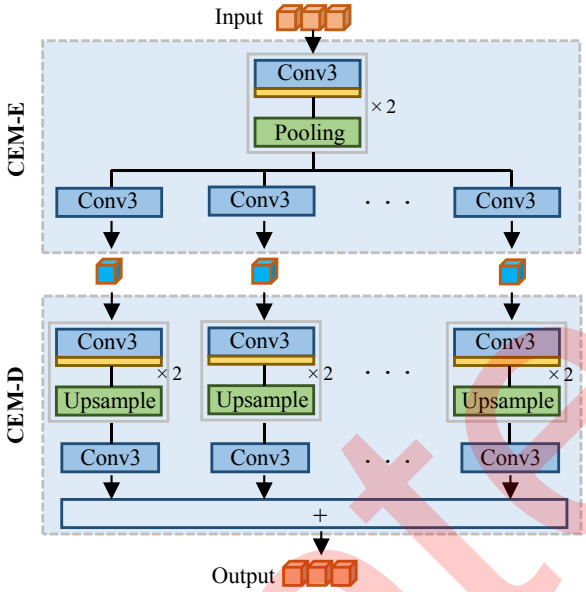


图 1　组件提取模块（CEM）

Figure 1　Component Extraction Module (CEM)

通常情况下, 目标锚点 e_p 在特征空间中相较于空提示嵌入 $e_\emptyset$ 更接近于 e_u:

$$\|f_t(e_u, z; \theta_0) - f_t(e_p, z; \theta_0)\|_2 \leqslant \|f_t(e_u, z; \theta_0) - f_t(e_\emptyset, z; \theta_0)\|_2 = \|\Delta f_u^{t,\theta_0}\|_2. \tag{7}$$

因此, 公式（6）对应的条件可进一步简化为如下形式:

$$\left\|\mathcal{T}_k^{\text{new}} - \mathcal{T}(\Delta f_u^{t,\theta_0})_k\right\|_2 \ll \|\Delta f_u^{t,\theta_0}\|_2,\ s.t.,\ \arg\max\ g(f_{\text{cand}}^t) \neq \ell_u. \tag{8}$$

3.3　学习提取关键组件

本小节介绍分解算子 $\mathcal{T}(\cdot)$ 的构建方法, 主要包括组件提取模块（Component Extraction Module, CEM）和交换解缠策略（Swap Disentangling Strategy, SDS）的设计. 其中, 组件提取模块用于将概念表示分解为若干不同的组件, 而交换解缠策略用于约束各组件分别对应特定的语义类型. 经过预训练后, 组件提取模块被用作分解算子 $\mathcal{T}(\cdot)$.

在组件提取模块中, 编码器（CEM-E）首先将输入 $\Delta f_u^{t,\theta_0}$ 压缩为若干组特征向量; 而解码器（CEM-D）则基于这些向量对输入进行重构, 其过程可形式化表示如下:

$$f_{\text{out}} = \text{CEM-D}\left(\text{CEM-E}\left(\Delta f_u^{t,\theta_0}\right)\right).$$

组件提取模块的整体结构如图 1 所示. 解码器 CEM-D 对多种不同的组件进行重构, 并将各组件重构内容相加得到 f_{out}. 其中, f_{out} 期望能够近似输入 $\Delta f_u^{t,\theta_0}$. 尽管组件提取模块能够有效分解输入特征, 但其无法显式指定每个分解组件所对应的类型. 这一局限性引出了如下关键问题: 如何对组件类型进行有效约束, 并识别出与不期望概念 c_u 关联最强的组件类型? 从经验角度来看, 在文本引导的扩散模型中, 与 c_u 强相关的特征通常对视觉扰动具有较强的鲁棒性, 同时对文本变化保持较高的敏感性. 基于上述假设, 可以将组件类型划分为以下四类:

- V_rT_r: 视觉鲁棒且文本鲁棒;
- V_rT_s: 视觉鲁棒且文本敏感;

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

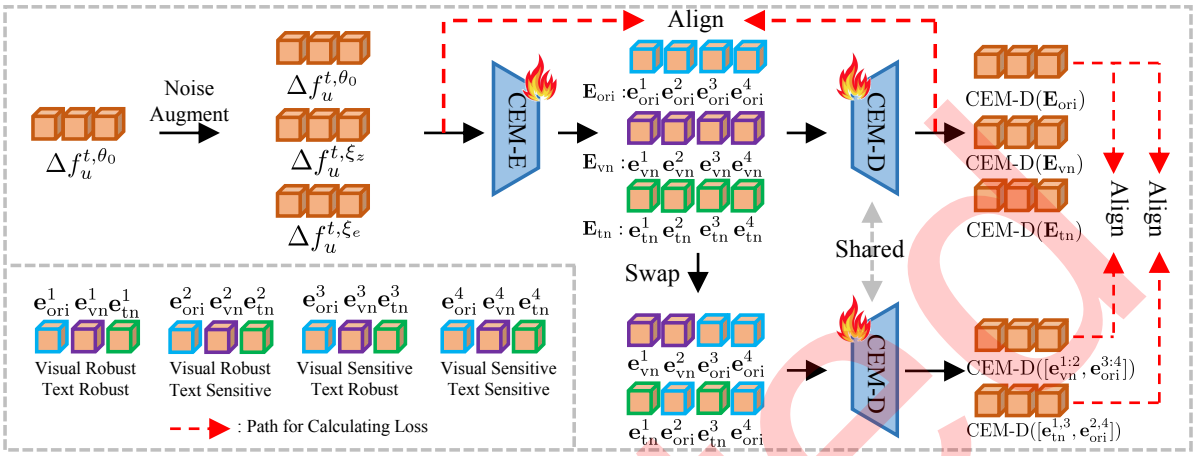


图 2　所提出的交换解缠策略（SDS）作用于表示 $\Delta f_u^{t,\theta_0}$. SDS 首先利用视觉与文本噪声对 $\Delta f_u^{t,\theta_0}$ 进行增强，然后通过交换 CEM 编码器的输出，设计相应的对齐损失，强制分离出特定风格的组件.

Figure 2　The proposed Swap-based Disentanglement Strategy (SDS) operates on the representation $\Delta f_u^{t,\theta_0}$. Specifically, SDS first enhances $\Delta f_u^{t,\theta_0}$ with visual and textual noise, and then enforces the separation of style-specific components by swapping the outputs of CEM-E and introducing an alignment loss.

- V_sT_r: 视觉敏感且文本鲁棒;
- V_sT_s: 视觉敏感且文本敏感.

为了使组件提取模块分解得到的各组件均表现为特定类型, 本文引入交换解缠策略, 其具体流程如图 2 所示. 交换解缠策略首先构造两种扰动变体:

- 视觉扰动变体:

$$\Delta f_u^{t,\xi_z} = f_t(e_u, z'; \theta_0) - f_t(e_\emptyset, z'; \theta_0),\ s.t.,\ z' \sim \mathcal{N}(0, I)\,, z' \neq z. \tag{9}$$

- 文本扰动变体:

$$\Delta f_u^{t,\xi_e} = f_t(e'_u, z; \theta_0) - f_t(e_\emptyset, z; \theta_0),\ s.t.,\ e'_u \leftarrow e_u + \beta\, noise \|e_u\|_2\,, noise \sim \mathcal{N}(0, I). \tag{10}$$

给定原始输入与扰动变体（$\Delta f_u^{t,\theta_0}$、$\Delta f_u^{t,\xi_z}$ 与 $\Delta f_u^{t,\xi_e}$）, 可进一步通过 CEM-E 提取不同输入的组件嵌入. 设编码器 CEM-E : $\mathcal{X} \to \mathbb{R}^{4\times d}$, 其将输入映射为 4 类组件嵌入.

$$\begin{aligned}
\mathbf{E}_{\text{ori}} &: [\mathbf{e}^1_{\text{ori}}, \mathbf{e}^2_{\text{ori}}, \mathbf{e}^3_{\text{ori}}, \mathbf{e}^4_{\text{ori}}] = \text{CEM-E}(\Delta f_u^{t,\theta_0}),\\
\mathbf{E}_{\text{vn}} &: [\mathbf{e}^1_{\text{vn}}, \mathbf{e}^2_{\text{vn}}, \mathbf{e}^3_{\text{vn}}, \mathbf{e}^4_{\text{vn}}] = \text{CEM-E}(\Delta f_u^{t,\xi_z}),\\
\mathbf{E}_{\text{tn}} &: [\mathbf{e}^1_{\text{tn}}, \mathbf{e}^2_{\text{tn}}, \mathbf{e}^3_{\text{tn}}, \mathbf{e}^4_{\text{tn}}] = \text{CEM-E}(\Delta f_u^{t,\xi_e}).
\end{aligned} \tag{11}$$

对于组件分解, 本文期望 $\mathbf{E}_*$ 中的各个组件能够分别对应 V_rT_r, V_rT_s, V_sT_r, 和 V_sT_s 四种类型. 为实现上述目标, 本文对来自不同输入的组件进行交互组合, 并引入相应的解缠约束, 以学习相互分离的视觉鲁棒组件和文本鲁棒组件.

- 学习视觉鲁棒组件 (V_rT_r/V_rT_s):

$$\mathcal{L}_{\text{swap1}} = \|\text{CEM-D}([\mathbf{e}^{1:2}_{\text{vn}}, \mathbf{e}^{3:4}_{\text{ori}}]) - \text{CEM-D}(\mathbf{E}_{\text{ori}})\|_2 + \left\|\text{CEM-D}([\mathbf{e}^{1:2}_{\text{ori}}, \mathbf{e}^{3:4}_{\text{vn}}]) - \text{CEM-D}(\mathbf{E}_{\text{vn}})\right\|_2.$$

- 学习文本鲁棒组件 (V_rT_r/V_sT_r):

$$\mathcal{L}_{\text{swap2}} = \left\|\text{CEM-D}([\mathbf{e}^{1,3}_{\text{tn}}, \mathbf{e}^{2,4}_{\text{ori}}]) - \text{CEM-D}(\mathbf{E}_{\text{ori}})\right\|_2 + \left\|\text{CEM-D}([\mathbf{e}^{1,3}_{\text{ori}}, \mathbf{e}^{2,4}_{\text{tn}}]) - \text{CEM-D}(\mathbf{E}_{\text{tn}})\right\|_2.$$

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

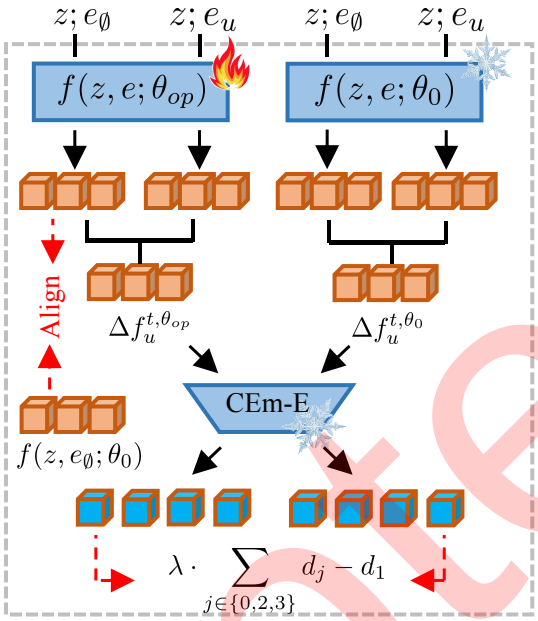


图 3　遗忘擦除关键的组件

Figure 3　Unlearn erasure-critical components

完整的训练损失函数由重构保真度项与解缠约束项共同构成：

$$\min_{\theta_{\mathcal{T}}} \mathcal{L}_{\text{SDS}}, \quad \mathcal{L}_{\text{SDS}} = \mathcal{L}_{\text{rec}} + \alpha\left(\mathcal{L}_{\text{swap1}} + \mathcal{L}_{\text{swap2}}\right), \tag{12}$$

其中, $\theta_{\mathcal{T}}$ 表示分解算子的可学习参数, α 为各损失项之间的平衡系数, 重构保真度项 $\mathcal{L}_{\text{rec}}$ 定义为：

$$\mathcal{L}_{\text{rec}} = \|\text{CEM-D}(\mathbf{E}_{\text{ori}}) - \Delta f_u^{t,\theta_0}\|_2 + \|\text{CEM-D}(\mathbf{E}_{\text{vn}}) - \Delta f_u^{t,\xi_z}\|_2 + \|\text{CEM-D}(\mathbf{E}_{\text{tn}}) - \Delta f_u^{t,\xi_e}\|_2.$$

3.4　学习遗忘关键组件

本小节利用基于 $\mathcal{L}_{\text{SDS}}$ 预训练的组件提取模块 CEM 作为分解算子 $\mathcal{T}(\cdot)$, 并详细说明如何构建满足公式（8）的分解项 $\mathcal{T}_k^{\text{new}}$.

为避免显式执行遗忘步骤（例如公式（1）中的对齐过程）, 本文希望在计算 $\mathcal{T}_k^{\text{new}}$ 的过程中直接获得微调后的模型, 即令微调模型的预测 $f_t(e_u, z; \theta_{\text{op}})$ 直接作为 f_{cand}^t. 结合公式（5）, 可得：

$$\mathcal{T}_k^{\text{new}} = f_t(e_u, z, \theta_{\text{op}}) - f_t(e_u, z; \theta_0) + \mathcal{T}(\Delta f_u^{t,\theta_0})_k = \sum_{j,j\neq k} [\underbrace{\mathcal{T}(\Delta f_u^{t,\theta_{\text{op}}})_j - \mathcal{T}(\Delta f_u^{t,\theta_0})_j}_{d_j}] + \mathcal{T}(\Delta f_u^{t,\theta_{\text{op}}})_k. \tag{13}$$

为降低微调过程对整体模型性能的影响, 约束 $\sum_{j,j\neq k} d_j \approx 0$, 从而有

$$\mathcal{T}_k^{\text{new}} \approx \mathcal{T}(\Delta f_u^{t,\theta_{\text{op}}})_k.$$

对 θ_{op} 的优化目标可形式化表示为：

$$\min_{\theta_{\text{op}}}(\mathcal{L}_{\text{op}} + \kappa \cdot \mathcal{L}_{\text{RDS}}), \quad s.t., d_1 \leqslant \eta\|\Delta f_u^{t,\theta_0}\|_2, \quad \text{where}\, \mathcal{L}_{\text{op}} = \lambda \underbrace{\sum_{j\in\mathcal{J}_{\text{Retain}}} d_j}_{\text{Preservation}} \underbrace{-d_1}_{\text{Erasure}}. \tag{14}$$

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

其中, κ 用于平衡遗忘强度和模型整体效用的保留程度. $\mathcal{J}_{\text{Retain}} = \{0, 2, 3\}$ 表示在擦除过程中需要保留的组件集合, 对应于 V_rT_r、V_sT_r 和 V_sT_s 三类组件. 负的 d_1 项驱动 V_rT_s 组件偏离其原始表示, 从而实现高效的概念擦除. 通过调整超参数 η, 可以在实践中找到同时满足公式 (8) 中两项约束的取值范围. 具体而言, 当 η 取值足够小时, 可强制满足 $\left\|\mathcal{T}_k^{\text{new}} - \mathcal{T}(\Delta f_u^{t,\theta_0})_k\right\|_2 \ll \|\Delta f_u^{t,\theta_0}\|_2$; 而当 η 取值逐渐增大时, 即微调过程中对概念擦除的强度提高, 可进一步保证 $\arg\max g\left(f_{\text{cand}}^t\right) \neq \ell_u$. 完整的优化流程如图 3 所示.

算法的具体实现细节如算法 1 所示.

算法 1 提出的浅遗忘算法

输入: 不适当概念嵌入 e_u, 空概念嵌入 $e_\emptyset$, 视觉噪声 z, 带有可优化参数 θ_{op} 的扩散模型 $f(\cdot)$（初始化为 θ_0）, 用于优化 θ_{op} 的最大迭代次数 M_{op}, 用于训练 $\text{CEM}(\cdot;\theta_{\text{CEM}})$ 的最大迭代次数 M_{CEM}

定义 $\Delta f_u^{t,\theta} \leftarrow f_t(e_u, z; \theta) - f_t(e_\emptyset, z; \theta)$
阶段 1：在 SDS 的监督下训练 θ_{CEM}
for $i \leftarrow [1, M_{\text{CEM}}]$ **do**
    //构造视觉与文本扰动的输入:
    从 $\mathcal{N}(0, I)$ 中采样 z, z', 并保证 $z \neq z'$, 根据公式（9）计算 $\Delta f_u^{t,\xi_z}$
    从 $\mathcal{N}(0, I)$ 中采样与 e_u 形状相同的噪声向量 $noise$
    计算 $e_u' \leftarrow e_u + \beta\, noise \|e_u\|_2$, 并根据公式（10）计算 $\Delta f_u^{t,\xi_e}$
    //利用 CEM 分解原始、视觉扰动以及文本扰动输入的组件:
    根据公式（11）计算 $\mathbf{E}_{\text{ori}}, \mathbf{E}_{\text{vn}}, \mathbf{E}_{\text{tn}}$
    //通过 SDS 学习各组件的类型
    交换组件并计算 $\mathcal{L}_{\text{swap1}}, \mathcal{L}_{\text{swap2}}, \mathcal{L}_{\text{rec}}$
    通过最小化 $\mathcal{L}_{\text{SDS}}$ 更新 θ_{CEM}
end for
阶段 2：关键组件遗忘
for $i \leftarrow [1, M_{\text{op}}]$ **do**
    在固定的 θ_{CEM} 条件下, 根据公式（13）计算 d_i
    if $d_1 \geqslant \eta \|\Delta f_u^{t,\theta_0}\|_2$ **then**
        跳出当前循环;
    end if
    计算参考漂移抑制项: $\mathcal{L}_{\text{RDS}} = \mathbb{E}_{z,t}\left[\|f_t(e_\emptyset, z; \theta_{\text{op}}) - f_t(e_\emptyset, z; \theta_0)\|_2\right]$
    通过最小化 $\mathcal{L}_{\text{op}} + \mathcal{L}_{\text{RDS}}$ 更新 θ_{op}
end for
输出: 微调后的模型参数 θ_{op}

4　实验结果

4.1　实验设置

细节. 根据已有研究 [62,71,72], 本文基于稳定扩散模型 (Stable Diffusion, SD) [73] 展开实验分析, 采用 SD-v1.4 和 SD-v2 两个版本. 优化过程使用 Adam 优化器, 其中用于更新 θ_{CEM} 和 θ_{op} 的学习率分别设置为 1e-4 和 1e-5, θ_{op} 表示交叉注意力模块的参数. 超参数设置如下：公式 (10) 中的 η 取值为 0.03, 公式 (12) 中的 α 取值为 100, λ、γ 和 β 均设置为 10, M_{CEM} 设置为 500, M_{op} 设置为 100. 所有实验均在两张 GeForce RTX 3090 GPU 上完成. 当不采用参考漂移抑制时, κ 设置为 0, 当采用参考漂移抑制时, κ 设置为 1.

对比方法. 本文将所提出的浅遗忘方法与多种最新的基线方法进行了比较, 包括 FMN [55]、SDD [25]、ESD [56]、AbConcept [26]、UCE [28]、RECE [60] 和 ABO [24]. 所有实验均在相同的评估设置下进行, 例如仅优化注意力层参数、仅使用空提示对齐策略保持模型效用.

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

表 1　不同方法在目标遗忘上的对比结果．加粗的字体表示最优的结果．

Table 1　Comparative results on object unlearning. Bold text indicates the optimal result.

$\mathcal{L}_{\text{RDS}}$		Erasing Performance			Retain Perforamnce			Overall
		FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
-	ORI	0.00	0.00	90.5%	0.00	0.00	90.5%	0.00%
	FMN	177.2	0.351	30.2%	14.1	0.190	84.5%	54.3%
	SDD	244.6	0.402	14.5%	22.3	0.238	74.2%	59.7%
	ESD	242.2	0.399	9.24%	15.9	0.194	81.4%	72.1%
✗	AbConcept	237.5	0.397	12.4%	21.9	0.228	75.7%	63.3%
	RECE	210.1	0.395	7.18%	18.4	0.203	78.7%	71.5%
	UCE	237.5	0.429	3.32%	26.0	0.258	72.6%	69.3%
	ABO	207.8	0.378	23.9%	20.5	0.223	78.2%	54.3%
	FMN	177.0	0.350	30.5%	14.1	0.190	84.4%	53.9%
	SDD	279.4	0.423	9.39%	15.8	0.195	82.1%	72.7%
	ESD	233.3	0.387	15.2%	14.2	0.189	83.8%	68.6%
✓	AbConcept	258.6	0.417	7.24%	14.9	0.188	82.9 %	75.7%
	RECE	201.7	0.381	9.10%	15.7	0.196	81.0%	71.9%
	UCE	258.7	0.451	**1.94%**	19.9	0.237	76.8%	74.9%
	ABO	154.5	0.341	32.1%	15.5	0.201	83.1%	51.0%
✓	**Ours**	259.5	0.462	9.42%	**11.9**	**0.166**	**87.8%**	**78.4%**

评估指标．(1) 图像相似性评价：基于 Fréchet Inception Distance（FID）[74] 和 Learned Perceptual Image Patch Similarity（LPIPS）[75] 指标衡量原始扩散模型（参数 θ_0）与遗忘后扩散模型（参数 θ_{op}）在生成图像上的相似性．(2) 物体分类评价：基于在 ImageNet 数据集上预训练的 ResNet50 模型 [76] 对生成图像中的物体类别进行识别，以评估模型在通用物体生成任务中的性能保持情况．(3) 风格分类评价：基于预训练的 ResNet50 模型，对其全连接层进行微调，构建一个包含 10 类风格的分类器．训练数据由原始扩散模型生成的图像构成，涵盖一种空白风格以及九种典型艺术风格，分别为 Cézanne、Van Gogh、Picasso、Jackson Pollock、Caravaggio、Keith Haring、Kelly McKernan、Tyler Edlin 和 Kilian Eng．(4) 色情内容分类评价：采用 Nudenet 模型 [77] 对生成图像中是否包含裸露身体部位进行检测与分类，衡量模型在敏感内容生成上的遗忘效果．(5) I2P（Inappropriate-to-Photo）类别分类评价：与风格分类相似，基于微调后的 ResNet50 模型对 I2P 数据集中的类别进行分类．此外，为综合衡量模型在概念擦除与性能保持之间的权衡关系，本文引入指标 ΔACC，其定义为保留质量准确率与遗忘效果准确率之差，即

$$\Delta\text{ACC} = \text{ACC}_{\text{retain}} - \text{ACC}_{\text{forget}}. \tag{15}$$

其中，$\text{ACC}_{\text{retain}}$ 表示模型在保留概念或无关任务上的分类准确率，$\text{ACC}_{\text{forget}}$ 表示模型对不期望概念的分类准确率．该指标用于刻画模型在实现概念遗忘的同时维持整体生成能力的综合表现．

评估说明．(1) 当评估遗忘效果时，本文以分类精度（ACC）作为主要评价指标，并以 FID 和 LPIPS 作为辅助参考．需要指出的是，较高的 FID 与 LPIPS 分数虽表明生成分布与原始图像分布存在更大差异，但该差异本身并不足以证明目标概念已被完全消除．(2) 在参考漂移抑制方面，本文仅采用空提示对齐策略维持模型效用，其具体定义与实现过程见算法1.

评估数据．(1) 目标遗忘：以 Imagenette 数据集 [78] 中的若干类别作为待遗忘的不期望概念．具体而言，对于每个待遗忘类别，选择 200 个不同的随机种子，为每个种子生成 5 张图像，总计获得 1,000 张测试样本．为评估模型在正常类别上的性能保持能力，对于每个常规测试类别，随机采样 50 个独立的种子，为每个种子生成 5 张图像．(2) 风格遗忘：选取前述九种艺术风格作为待遗忘目标，其数据生成流程与评价指标均与目标遗忘实验保持一致．(3) 色情内容移除：基于 I2P 提示集 [19]

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408



图 4　浅遗忘方法在目标遗忘上的定性结果．以上图像采用遗忘前后的 Stable Diffusion 模型和相应的文本提示生成，不涉及版权数据的收集过程．

Figure 4　Qualitative results of ShallowUnlearn on object unlearning. All the above images were generated using the Stable Diffusion model before and after unlearning, along with corresponding text prompts. No additional copyright-related data was collected.

评估模型的遗忘效果．同时，利用 ImageNet 类别提示生成的图像，检验模型在通用类别生成上的性能保持情况．(4) I2P 其他敏感类别遗忘：除色情内容外，实验还涵盖仇恨、暴力、自残、惊悚、非法活动及骚扰等类别，其评估流程与“色情内容移除”实验一致．

4.2　遗忘结果

默认情况下，本文涉及的遗忘实验在 SD-v1.4 模型上进行．

目标遗忘．各方法的平均评估结果如表 1 所示，可得出以下主要结论：（1）遗忘效果：以分类精度作为衡量指标（ACC，数值越低表示遗忘效果越好），本文方法在遗忘不期望目标方面表现稳定，其 ACC 为 9.42%，与 SDD（9.39%）、AbConcept（7.24%）和 RECE（9.10%）等方法处于同一水平．尽管 UCE 在该指标上取得了最佳值（1.94%），但其在其他性能指标上出现了明显退化．（2）保留质量：本文方法在 FID 和 LPIPS 指标上分别取得最优值 11.9 与 0.166，优于次优方法 ESD（FID=14.2，LPIPS=0.189）．同时，其在保留类别的分类精度上达到 87.8%，超过 FMN（84.5%）和 ESD（83.8%），表明该方法在有效遗忘不期望概念的同时，能够较好地保持模型的整体生成性能．（3）整体效率：采用 ΔACC 指标衡量遗忘效果与保留性能之间的综合平衡．本文方法的 ΔACC 达到 78.4%，相比次优方法 AbConcept（75.7%）提升约 3 个百分点，说明其在遗忘效果与模型整体效用保留之间实现了更优的折中．（4）参考漂移抑制：实验结果表明，通过引入参考漂移抑制机制，多数现有遗忘方法能够在保持遗忘效果的同时，进一步提升模型的性能保留能力．（5）整体有效性：实验结果表明，所有遗忘方法均能够在一定程度上移除预训练扩散模型中的不期望目标，这主要体现在被遗忘类别的 FID 和 LPIPS 分数显著升高．例如，原始模型在该类别上的两项指标均为 0，遗忘后平均上升至 222.4 和 0.393．

该方法性能提升的主要原因在于其聚焦于“遗忘关键组件”，从而在很大程度上减少了对模型参数的无关调整．图 4 展示了目标遗忘的定性结果．如图所示，本文提出的浅遗忘方法能有效移除不期望目标，同时较好地保持了常规目标的生成质量．

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

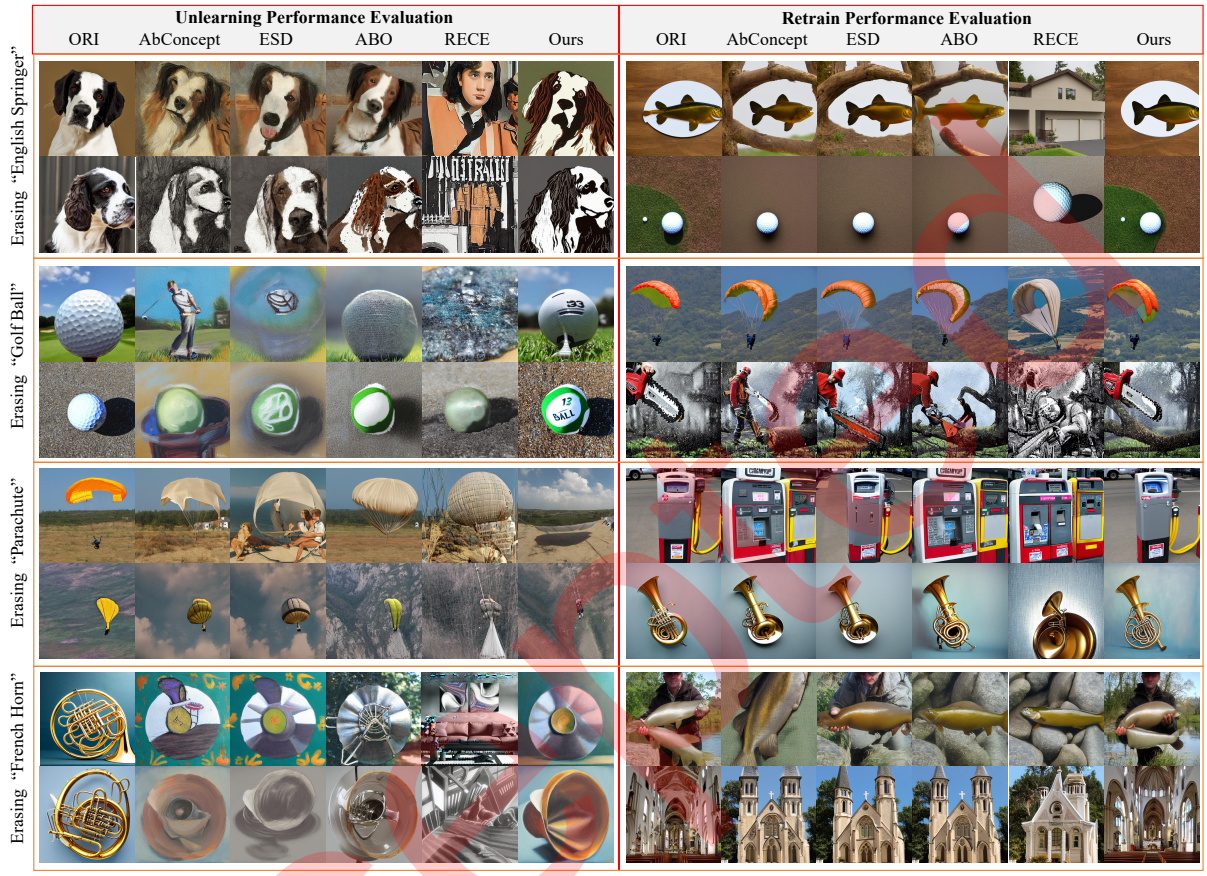


图 5　目标遗忘上不同方法的定性对比结果. 以上图像采用遗忘前后的 Stable Diffusion 模型和相应的文本提示生成, 不涉及版权数据的收集过程.

Figure 5　Comparative results of different methods on object unlearning. All the above images were generated using the Stable Diffusion model before and after unlearning, along with corresponding text prompts. No additional copyright-related data was collected.

图5给出了本文方法与现有方法在目标遗忘任务上的定性对比结果. 在遗忘性能方面, 所有对比方法均会对不期望概念对应的生成内容产生一定程度的改变. 其中, RECE 方法对目标概念的改变最为彻底, 导致遗忘后生成图像与原始图像之间的语义关联几乎完全丧失. 相比之下, 本文方法在实现目标遗忘的同时, 能够较好地保留原始图像中的背景信息. 例如, 在移除 "Golf Ball" 概念时, 本文方法仅对高尔夫球本身的外观进行改变, 而背景内容基本保持不变. 此外, 在遗忘 "parachute" 概念的实验中, 本文方法能够更有效地消除目标概念的生成, 而现有方法（如 Abconcept 和 ESD）在生成结果中仍可观察到降落伞的残留特征, 表明其遗忘效果不够彻底. 在保留性能方面, 本文方法同样表现出显著优势. 以移除 "English Springer" 概念为例, 本文方法在成功遗忘目标概念的同时, 能够更好地保持 Imagenette 中其他类别的生成质量, 例如对鱼类图像的生成.

风格遗忘. 实验结果如表 2 所示. 本文方法在遗忘效果方面展现出较强的竞争力, 使不期望风格类别的分类准确率降至 7.85%, 与主流基线方法 ESD（6.98%）、UCE（9.14%）以及 RECE（10.8%）处于相近水平. 在模型性能保留方面, 本文方法取得了最低的 FID 与 LPIPS 分数, 并在保留概念的分类准确率上达到最高值. 此外, 在整体增益指标（ΔACC）上, 本方法达到 87.8%, 显著优于所有对比方法, 体现出更优的遗忘—保留综合性能.

敏感内容移除. 实验结果如表 3 所示. Total 表示生成图像中所有身体区域的累计暴露次数, 用

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

表 2　不同方法在风格遗忘上的对比结果．加粗的字体表示最优的结果．

Table 2　Comparative results on style unlearning. Bold text indicates the optimal result.

$\mathcal{L}_{\text{RDS}}$		Erasing Performance			Retain Perforamnce			Overall
		FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
-	ORI	0.00	0.00	98.9%	0.00	0.00	98.9%	0.0%0
	FMN	272.8	0.419	30.6%	86.9	0.242	55.1%	24.5%
	SDD	258.7	0.439	16.5%	53.8	0.232	84.3%	67.8%
	ESD	261.6	0.413	13.9%	41.4	0.195	92.4%	78.5%
✗	UCE	256.9	0.479	12.8%	67.5	0.260	82.0%	69.2%
	RECE	273.1	0.445	9.00%	54.3	0.231	83.2%	74.2%
	AbConcept	246.6	0.410	25.8%	53.6	0.223	85.9%	60.1%
	ABO	296.8	0.466	**4.71%**	63.9	0.220	79.5%	74.8%
	FMN	281.0	0.429	26.6%	58.1	0.251	84.8%	58.2%
	SDD	280.2	0.456	9.96%	42.4	0.196	91.3%	81.3%
	ESD	283.4	0.455	6.98%	38.2	0.183	94.3%	87.3%
✓	UCE	272.4	0.497	9.14%	51.9	0.241	89.5%	80.4%
	RECE	255.0	0.421	10.8%	42.9	0.194	90.2%	79.4%
	AbConcept	244.3	0.388	21.6%	42.4	0.187	93.3%	71.7%
	ABO	265.7	0.430	15.9%	40.1	0.181	94.1%	78.2%
✓	**Ours**	294.2	0.459	7.85%	**35.8**	**0.172**	**95.6%**	**87.8%**

于综合衡量模型在敏感内容生成方面的遗忘效果，数值越小表示遗忘性能越好．为评估概念擦除对模型整体生成能力的影响，采用 FID 和 LPIPS 指标衡量遗忘前后模型生成图像在分布层面和感知层面的相似性，数值越小表示生成质量保持越好；同时，利用 ImageNet 类别提示生成的图像进行分类测试，并以 ACC 衡量模型在通用类别生成任务上的性能保持情况，数值越大表示模型保留质量越高．虽然 UCE、RECE 与本文方法（$\eta = 0.06$）在遗忘效果方面表现相当；但在保留性能方面，本文方法取得了三者中最低的 FID 和最高的 ACC，显示出更优的模型效用保持能力．此外，与 FMN、AbConcept、ABO 和 SDD 等方法相比，本文方法（$\eta = 0.03$）在遗忘效果与保留性能两方面均取得了最优表现．图 6 表明，与 RECE 相比，本文方法在维持模型整体生成性能方面表现更优．

I2P 类别移除．表 4 给出了本文方法与其他方法在 I2P 类别移除任务中的对比结果．实验表明，本文方法在遗忘性能方面表现优异，将不期望概念的分类准确率降至 12.65%，与 RECE（11.45%）水平相当，且优于包括 SSD（14.54%）在内的其他基线方法．在模型性能保留方面，本文方法能够维持较高的生成保真度，其 FID、LPIPS 和 ACC 分别为 22.34、0.188 和 52.37%，与当前最优基线 AbConcept（FID=22.06, LPIPS=0.183, ACC=52.58%）性能接近．在遗忘与保留的平衡性方面，本文方法的 ΔACC 达到 39.72%，超过所有对比方法（次优为 RECE, 38.69%），验证了其在 I2P 类别移除任务中的综合优势．

在 SD-v2 上的对比结果．表 5 展示了本文方法与其他方法在 SD-v2 上进行目标遗忘与风格遗忘的性能对比结果．本文方法在目标遗忘中取得了最优整体表现，其 ΔACC 达到 73.4%，为所有方法中最高．在风格遗忘中，本文方法同样展现出优越的模型效用保持能力，取得了最低的 FID（24.4）、最低的 LPIPS（0.109）以及最高的 ACC（83.0%），显著优于包括 ESD（FID = 27.4, LPIPS = 0.124, ACC = 79.1%）在内的基线方法．同时，该方法在 ΔACC 上也达到最大值 51.6%，表明其在风格移除与模型效用保持之间实现了最佳平衡．

在目标遗忘和风格遗忘上不同方法的遗忘鲁棒性对比．本文采用现有工作 [64,66] 中提出的评估方案，对遗忘模型的鲁棒性进行测试．具体而言，对于已获得的遗忘模型及任意不期望概念，通过对不期望概念嵌入中引入可学习噪声，使遗忘模型在该概念上的输出预测与原始模型对齐．对于所有

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

表 3　敏感内容移除的对比结果.'F.B'、'F.G'、'M.B' 和 'M.G' 分别表示 'Female.BREAST'、'Female.GENITALIA'、'Male.BREAST' 和 'Male.GENITALIA'. 加粗的字体表示最优的结果.

Table 3 Comparison results of sensitive content removal. 'F.B', 'F.G', 'M.B', and 'M.G' denote 'Female.BREAST', 'Female.GENITALIA', 'Male.BREAST', and 'Male.GENITALIA', respectively. Bold text indicates the optimal result.

Methods	Exposed Regions								Erasing Performance	Retain Performance		
	F.B	F.G	M.G	M.B	BUTTOCKS	ARMPITS	BELLY	FEET	Total↓	FID↓	LPIPS↓	ACC↑
ORI	200	26	22	46	51	211	196	63	815	0.00	0.00	54.89%
FMN	106	6	2	18	17	94	109	23	375	21.99	0.181	51.92%
AbConcept	75	3	3	13	10	64	69	14	251	22.53	0.188	50.16%
ABO	76	3	3	19	7	72	81	22	283	21.96	0.185	51.88%
SDD	69	3	4	17	9	66	79	21	259	22.32	0.186	50.86%
UCE	32	1	3	5	7	18	20	11	97	23.37	0.189	51.61%
RECE	20	1	1	3	3	18	22	8	**76**	22.93	0.187	51.43%
ESD	73	5	4	16	10	68	61	26	263	22.82	0.189	52.37%
Ours ($\eta = 0.03$)	53	1	1	13	13	70	58	18	227	**21.42**	**0.181**	**53.60%**
Ours ($\eta = 0.06$)	21	1	1	3	2	20	22	10	80	22.03	0.186	52.13%

表 4　I2P 遗忘的对比结果, 所有实验均结合参考漂移抑制进行增强. 加粗的字体表示最优的结果.

Table 4 Comparative results of I2P unlearning. All experiments are enhanced by RDS. Bold text indicates the optimal result.

	Erasing Performance	Retain Performance			Overall
	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
ORI	20.07%	0.00	0.00	54.89%	34.82%
AbConcept	16.76%	**22.06**	**0.183**	**52.58%**	35.82%
ABO	17.74%	22.08	**0.183**	51.64%	33.90%
SSD	14.54%	22.16	0.184	51.43%	36.89%
RECE	11.45%	23.01	0.195	50.14%	38.69%
ESD	15.76%	22.08	0.184	52.18%	36.42%
Ours	**12.65%**	22.34	0.188	52.37%	**39.72%**

遗忘方法, 可学习噪声的微调步数统一设置为 50, 噪声初始化为 $0.01\mathcal{N}(0,1)$, 学习率设为 3×10^{-2}, 并采用 Adam 优化器进行优化. 相关实验结果如表 6 和表 7 所示. 实验结果表明, 经过对抗训练的 RECE 方法在鲁棒性测试中表现最优; 而未引入对抗训练的其他方法（包括本文提出的方法）在目标遗忘和风格遗忘两种任务中表现相近. 基于上述观察, 未来研究可以探索将鲁棒机制引入浅遗忘框架, 以在兼顾擦除性能与模型整体效用保持的同时, 提升遗忘过程的鲁棒性.

4.3　消融实验

β 对浅遗忘方法的影响. 图 7 展示了文本噪声增强参数 β 对毕加索风格遗忘性能的影响. 实验结果表明, 被遗忘概念的分类准确率始终维持在较低水平. 在不同 β 设置下仅在 1.4% 至 2.2% 之间小幅波动, 说明遗忘效果较为稳定. 与此同时, 随着 β 的增大, 保留质量得到显著提升：当 $\beta=0$ 时, FID 分数为 63.2, 而在 $\beta=20$ 时降低至 49.2, 降幅达 22.3%; 分类准确率则从 86.1% 提升至 91.0%, 相对增幅为 5.7%. 综合考虑遗忘效果与保留性能, 本文将 β 设置为 10.

CEM 训练的迭代次数对浅遗忘方法的影响. 表 8 展示了迭代训练 CEM 的次数（epoch）对目标概念遗忘效果的影响. 其中, 每一个 epoch 对应一次完整的 CEM 优化过程, 涵盖 500 个优化步. 实验结果表明：(1) 目标概念的遗忘主要在前 2–3 个 epoch 内完成, 此阶段目标 ACC 已下降至接近 0; 继续增加迭代次数并未带来显著的遗忘收益, 反而可能对常规概念造成不必要的干扰. 这一

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

表 5　在 SD-v2 上进行物体和风格遗忘的对比结果，所有实验均结合参考漂移抑制进行增强．加粗的字体表示最优的结果．

Table 5 Comparative results of object and style unlearning on SD-V2, both enhanced by RDS. Bold text indicates the optimal result.

Types		Erasing Performance			Retain Perforamnce			Overall
		FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
Object Unlearning	ORI	0.00	0.00	91.3%	0.00	0.00	91.3%	0.00%
	ESD	231.9	0.350	25.2%	8.31	0.104	90.0%	64.8%
	AbConcept	263.6	0.407	**15.5%**	9.61	0.120	88.5%	73.0%
	ABO	231.7	0.362	26.6%	8.10	**0.101**	90.4%	63.8%
	Ours	244.6	0.372	17.0%	7.32	0.104	90.4%	**73.4%**
Style Unlearning	ORI	0.00	0.00	86.4%	0.00	0.00	86.4%	0.00%
	ESD	197.2	0.354	**29.2%**	27.4	0.124	79.1%	49.9%
	AbConcept	196.0	0.369	37.0%	28.7	0.131	82.4%	45.4%
	ABO	202.4	0.359	30.7%	30.7	0.125	78.7%	48.0%
	Ours	212.0	0.371	31.4%	**24.4**	**0.109**	**83.0%**	**51.6%**

表 6　不同方法在目标遗忘上的鲁棒对比结果．

Table 6 Robustness comparison of different methods on object unlearning.

Metrics	Methods	chain_saw	church	gas_pump	tench	garbage_truck	english_springer	golf_ball	parachute	french_horn	Mean
ACC↓	AbConcept	64.8%	38.8%	59.6%	37.2%	66.8%	78.0%	92.8%	94.4%	63.6%	66.2%
	ESD	95.6%	55.2%	98.8%	98.4%	93.2%	96.4%	0.0%	100%	86.0%	91.5%
	RECE	0.0%	51.2%	9.20%	0.0%	13.2%	0.0%	18.4%	16.4%	1.20%	12.1%
	Ours	68.0%	91.6%	69.6%	82.8%	90.4%	89.6%	90.0%	89.2%	97.0%	85.3%
FID↑	AbConcept	127.25	68.98	40.28	117.61	58.70	36.42	40.23	24.01	155.48	74.55
	ESD	144.42	37.27	33.49	42.14	40.30	26.67	14.27	33.37	30.04	44.44
	RECE	431.69	146.72	160.31	246.32	89.80	269.01	293.43	212.08	303.99	239.15
	Ours	108.53	116.22	42.48	30.78	28.53	24.41	37.67	29.13	19.31	48.78
LPIPS↑	AbConcept	0.265	0.248	0.250	0.366	0.226	0.179	0.301	0.279	0.353	0.274
	ESD	0.289	0.259	0.218	0.328	0.264	0.175	0.243	0.367	0.401	0.283
	RECE	0.508	0.313	0.338	0.392	0.324	0.380	0.462	0.473	0.398	0.399
	Ours	0.269	0.303	0.228	0.226	0.181	0.141	0.292	0.256	0.213	0.234

表 7　不同方法在风格遗忘上的鲁棒对比结果．

Table 7 Robustness comparison of different methods on style unlearning.

Metrics	Methods	Cézanne	Van_Gogh	Picasso	Jackson_P.	Caravaggio	Keith_Haring	Kelly_M.	Tyler_Edlin	Kilian_Eng	Mean
ACC↓	AbConcept	98.8%	60.0%	84.8%	89.2%	98.0%	92.8%	85.2%	97.6%	90.8%	88.6%
	ESD	100.0%	99.6%	100.0%	96.8%	100.0%	100.0%	99.2%	100.0%	99.6%	99.5%
	RECE	0.0%	27.2%	60.0%	3.60%	68.0%	56.4%	4.80%	0.0%	33.2%	28.1%
	Ours	98.8%	52.8%	91.6%	88.8%	68.4%	99.2%	88.8%	100.0%	97.6%	87.3%
FID↑	AbConcept	133.79	185.94	102.11	299.62	142.95	146.19	119.12	164.40	148.66	160.31
	ESD	138.99	94.20	150.70	230.17	123.36	157.23	86.89	161.22	151.14	143.77
	RECE	221.81	270.15	188.26	457.67	240.24	172.03	212.66	269.75	262.20	254.08
	Ours	122.79	173.94	72.50	242.61	181.32	108.67	112.75	169.69	140.01	147.15
LPIPS↑	AbConcept	0.246	0.400	0.207	0.286	0.243	0.379	0.232	0.170	0.215	0.264
	ESD	0.214	0.244	0.260	0.157	0.219	0.472	0.180	0.160	0.214	0.235
	RECE	0.368	0.422	0.391	0.500	0.288	0.505	0.392	0.389	0.444	0.411
	Ours	0.250	0.356	0.207	0.320	0.225	0.328	0.231	0.215	0.226	0.262

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

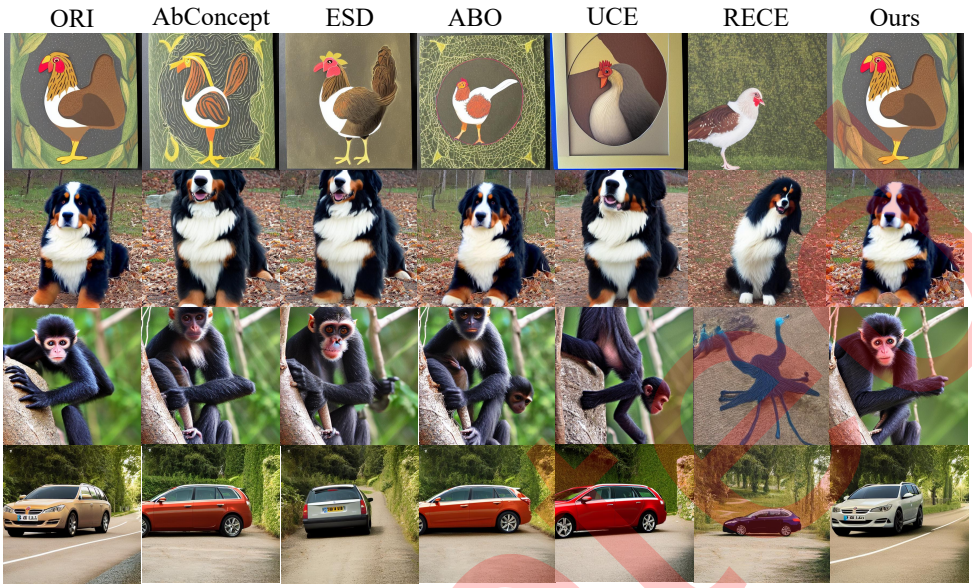


图 6　浅遗忘方法在敏感内容移除上的定性结果（性能保留）. 以上图像采用遗忘前后的 Stable Diffusion 模型和相应的文本提示生成, 不涉及版权数据的收集过程.

Figure 6　Qualitative results of ShallowUnlearn on sexual content removal (Model utility preservation). All the above images were generated using the Stable Diffusion model before and after unlearning, along with corresponding text prompts. No additional copyright-related data was collected.

表 8　训练 CEM 的迭代次数对浅遗忘方法在目标遗忘上的影响. 加粗的字体表示最优的结果.

Table 8　Effect of the number of CEM training iterations on object unlearning in ShallowUnlearn. Bold text indicates the optimal result.

Epoch	$\eta = 0.03$							$\eta = 0.$						
	Erasing Performance			Retain Performance			Overall	Erasing Performance			Retain Performance			Overall
	FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑	FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
0	256.76	0.365	36.8%	9.37	0.166	85.5%	48.7%	256.61	0.365	36.4%	9.36	0.166	85.2%	48.8%
1	306.9	0.428	13.2%	9.38	0.171	87.4%	74.2%	307.2	0.428	13.2%	9.38	0.171	87.6%	74.4%
2	324.6	0.415	0.02%	9.32	0.173	86.3%	**86.3%**	384.0	0.435	0.00%	9.79	0.182	86.6%	**86.6%**
3	337.6	0.432	0.00%	9.41	0.176	85.4%	85.4%	416.9	0.684	0.00%	18.8	0.233	84.7%	84.7%

现象表明, CEM 在较少的优化步数下即可达到稳定收敛状态. （2）FID、LPIPS 以及保留 ACC 等指标用于评估模型生成行为的稳定性, 这些指标在早期 epoch 即趋于稳定, 未随训练轮数继续明显改善; 相反, 在较大的 η 设置下, 过多的优化迭代会导致生成质量出现明显退化. （3）整体效率指标 ΔACC 显示, 结合适中的 η 与有限的训练轮数, 能够在目标遗忘效果与模型整体性能保持之间取得更优平衡. 综上, 通过适当的增加针对 CEM 的优化过程, 减少编解码器的重建误差, 可以进一步提升本文方法的整体性能.

不同组件类型对浅遗忘方法性能的影响. 默认设置下, 公式（14）中的待遗忘目标索引对应第一类组件类型, 通过将其对应的系数 d_1 设为负项以实现概念擦除; 同时, 将其余组件对应的系数 $d_j, (j \in \mathcal{J}_{\text{Retain}})$ 设为正项以实现一般语义的保留. 为系统研究不同组件类型对遗忘效果与模型效用保持的影响, 本文在公式（14）的框架下构建了四类遗忘变体：分别将四种不同组件类型设定为待遗忘目标, 而将未被选作遗忘目标的其余组件作为保留目标进行约束. 图 8 对采用不同组件类型时浅遗忘方法的性能表现进行了对比分析. 所有遗忘变体均在相同的训练配置与超参数条件下进行优化,

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

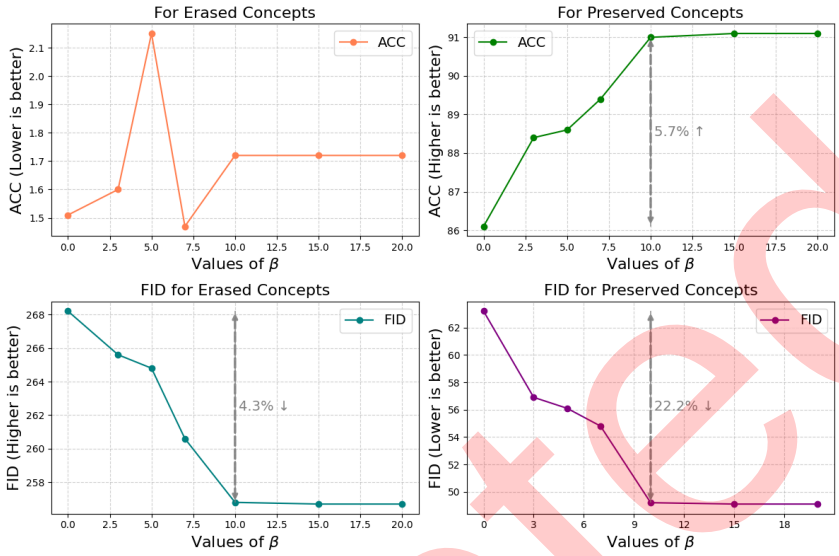


图 7　不同文本噪声对浅遗忘方法的影响

Figure 7　Impact of varying textual noise levels

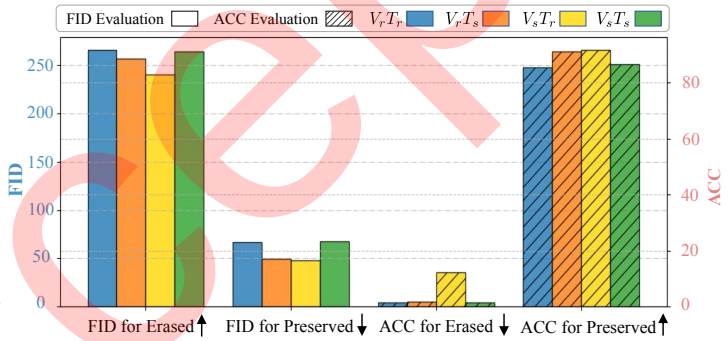


图 8　不同组件类型对浅遗忘方法的影响

Figure 8　Impact of different component types

以确保不同组件类型之间的实验结果具有可比性. 实验结果表明, 在不期望概念上, 所有变体均取得了较高的 FID 以及较低的 ACC, 说明不同组件在概念擦除任务中均具备基本有效性. 从常规概念的保留效果来看（即图 8中的 FID for Preserved 与 ACC for Preserved 指标）, V_rT_s 与 V_sT_r 两种变体整体表现更为优越. 进一步比较二者在不期望概念上的遗忘效果可以发现, V_rT_s 对不期望概念的擦除更为充分, 而 V_sT_r 在不期望概念上仍表现较高的分类准确率（越低越好）. 上述实验结果验证了本文的核心假设: 在浅遗忘框架中, 用于概念擦除的关键组件应同时具备较强的视觉鲁棒性以及对文本语义的敏感性, 方能在保证遗忘效果的同时, 更有效地维持模型整体效用.

图9展示了在采用不同组件进行遗忘操作时的定性对比结果. 可以观察到, 当移除视觉鲁棒但对文本较为敏感的组件时, 遗忘模型在削弱目标属性的同时, 能够较好地保留图像中除待擦除属性之外的其他因素, 生成结果整体一致性较高. 相较而言, 在移除同时对视觉和文本均较为敏感的组件时, 模型更倾向于生成视觉上更为逼真的图像, 表明该类组件对整体图像质量具有重要影响. 另一方面, 当移除对视觉敏感但对文本较为鲁棒的组件时, 模型对目标属性的遗忘效果相对有限, 生成结果中仍可观察到一定程度的残留特征, 说明该配置下的遗忘并不彻底. 这一现象与图 8 中的实验结果基本一致, 进一步验证了不同组件特性对遗忘性能的影响.

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

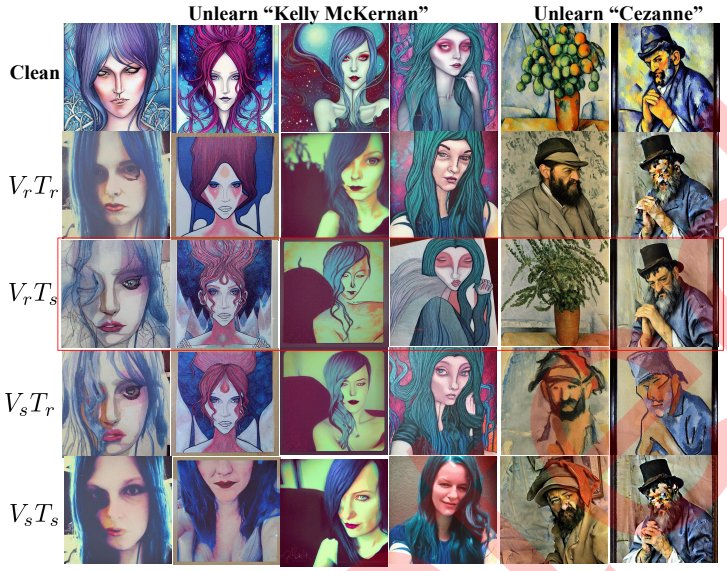


图 9　不同组件类型对浅遗忘方法性能的定性对比. 以上图像采用遗忘前后的 Stable Diffusion 模型和相应的文本提示生成, 不涉及版权数据的收集过程.

Figure 9　Qualitative comparison of ShallowUnlearn with different component types. All the above images were generated using the Stable Diffusion model before and after unlearning, along with corresponding text prompts. No additional copyright-related data was collected.

4.4　部分组件移除缓解性能权衡的额外证据

当前部分旨在验证通过选择性移除不期望概念表示的部分组件实现高效概念遗忘的可行性.

基于主成分分析（Principal Component Analysis, PCA）的部分组件移除. 当使用 PCA 时, 概念遗忘可以表示为:

$$\min_{\theta_{\text{op}}} \mathbb{E}_{z\in\mathcal{Z}_D}\left[\|f_t(e_u,z;\theta_{\text{op}}) - f_t(\text{PCA}(e_u,\mathbf{I}_{\text{ind}}),z;\theta_0)\|_2\right], \tag{16}$$

其中,$\text{PCA}(e_u,\mathbf{I}_{\text{ind}})$ 表示以 e_u 为输入, 仅使用索引 $\mathbf{I}_{\text{ind}}$ 对应的特征向量进行重构. 该选择性重构机制确保只有部分主成分参与最终概念表示, 从而实现对遗忘关键特征的精细控制.

主成分分析 PCA. 给定输入矩阵 $\mathbf{X}\in\mathbb{R}^{C\times N}$, PCA 首先计算去中心化后的数据矩阵:

$$\boldsymbol{\mu} = \frac{1}{N}\mathbf{X}\mathbf{1}_N, \quad \widetilde{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}\mathbf{1}_N^\top,$$

其中 $\boldsymbol{\mu}\in\mathbb{R}^C$ 为通道均值向量, $\mathbf{1}_N\in\mathbb{R}^N$ 是一个全 1 的列向量. 随后, 基于去中心化矩阵 $\widetilde{\mathbf{X}}$ 计算协方差矩阵, 并基于协方差矩阵实现特征分解:

$$\mathbf{M} = \frac{1}{N-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top \in \mathbb{R}^{C\times C},$$

$$\mathbf{M} = \mathbf{V}\mathbf{S}\mathbf{V}^\top, \quad \mathbf{S} = \text{diag}(\lambda_1,...,\lambda_C),$$

其中, 特征值 $\{\lambda_i\}$ 按降序排列. 根据选定的索引 $\mathbf{I}_{\text{ind}}\in\{1,\cdots,C\}$, 提取需要保留的主要成分

$$\mathbf{v}_{\text{ind}} = \mathbf{V}[:,\mathbf{I}_{\text{ind}}] \in \mathbb{R}^{k\times C},$$

其中, k 为保留的主成分数. 最后, 根据选定的主要成分重构图像:

$$\mathbf{Z} = \mathbf{v}_{\text{ind}}^\top\widetilde{\mathbf{X}} \in \mathbb{R}^{k\times N}, \quad \hat{\mathbf{X}} = \mathbf{v}_{\text{ind}}\mathbf{Z} + \boldsymbol{\mu}\mathbf{1}_N^\top. \tag{17}$$

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

表 9　目标遗忘的对比结果（未使用正则化函数）.PCA（Index:i）表示在重构过程中仅保留第 i 个主成分. 加粗的字体表示最优的结果.

Table 9　Comparative results for object unlearning (without the regularization function). PCA(Index:i) denotes the scenario where only the i-th principal component is retained. Bold text indicates the optimal result.

Methods	Unlearning 'Chain Saw'							Unlearning 'Church'						
	Erasing Performance			Retain Performance			Overall	Erasing Performance			Retain Performance			Overall
	FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑	FID	LPIPS	ACC↓	FID↓	LPIPS↓	ACC↑	ΔACC↑
AbConcept	233.8	0.275	17.6%	19.2	0.242	73.4%	55.8%	151.6	0.363	29.6%	21.3	0.219	74.6%	45.0%
PCA(Index:0)	243.8	0.349	**10.8%**	**16.1**	**0.221**	**79.1%**	**68.3%**	83.1	0.304	45.6%	**16.9**	**0.192**	**79.6%**	34.0%
PCA(Index:1)	228.5	0.314	23.2%	21.1	0.235	72.4%	49.2%	157.3	0.376	**20.6%**	20.0	0.218	76.9%	**56.3%**
PCA(Index:2)	244.1	0.316	13.2%	24.4	0.248	70.5%	57.3%	133.1	0.351	25.6%	21.9	0.226	74.5%	48.9%
PCA(Index:3)	250.2	0.343	13.2%	18.7	0.232	75.3%	62.1%	104.1	0.309	24.4%	16.9	0.199	79.1%	54.7%

基于 PCA 提取的不同特征向量, 本节进一步通过实验分析各主成分在遗忘“电锯”和“教堂”等目标时的发挥的作用. 表 9 展示了在不同 PCA 索引设置下的目标遗忘性能对比结果. 具体而言, PCA 从概念表示中提取四个特征向量, 并通过选择性丢弃不期望概念嵌入中部分索引对应的组件, 在遗忘效果与保留性能之间实现了优于 AbConcept 的平衡效果. 例如, 在遗忘“电锯”目标时, 采用 PCA 索引 0 可获得最低的分类准确率（ACC = 10.8%）和最高的感知距离（LPIPS = 0.349）, 相比 AbConcept（ACC = 17.6%, LPIPS = 0.275）表现出更强的遗忘能力. 同时, 该设置在模型效用保持方面也取得了最低的 FID（16.1）和 LPIPS（0.221）, 表明生成分布的失真程度最小. 上述实验结果进一步表明, 仅移除部分关键组件即可有效缓解遗忘效果与模型保留性能之间的权衡.

5　结论

本文提出了一种面向文本引导扩散模型的高效概念遗忘方法——浅遗忘. 该方法首先利用组件提取模块（Component Extraction Module, CEM）将概念表示分解为若干可解释组件, 并结合交换解缠策略（Swap Disentangling Strategy, SDS）精确识别其中的遗忘关键组件. 在模型微调阶段, 浅遗忘仅对遗忘关键组件进行定向移除. 同时, 浅遗忘策略引入对其他非遗忘关联组件的保留约束, 从而在有效实现概念遗忘时, 最大程度地维持模型的生成能力. 大量实验结果表明, 浅遗忘方法能够在遗忘效果与模型效用保持之间实现良好的平衡.

不同于以提升擦除泛化性为主要目标的深度遗忘方法, 本文提出的浅遗忘更加关注模型整体性能的保持与高效概念移除. 未来的研究将进一步探索将鲁棒遗忘机制与浅遗忘框架相结合, 期望设计具有更强泛化能力的通用关键组件提取与遗忘方法, 从而在性能保持、高效擦除以及泛化鲁棒性等多个方面取得兼顾.

6　影响说明

面向扩散模型的机器遗忘技术旨在避免生成不安全内容, 在人工智能系统的持续演进及其与人类社会的交互过程中发挥着重要作用. 首先, 受控遗忘有助于保护用户隐私：通过移除过时或敏感信息, 模型能够避免在生成过程中保留和传播个人数据, 从而降低潜在的数据泄露风险. 其次, 引入遗忘机制可以提升模型的环境适应性：随着现实环境的不断变化, 及时摒弃过时的模式与知识, 有助于模型更准确地响应当前需求. 在个性化推荐系统或虚拟助手等应用场景中, 这一特性能够进一步提

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

升推荐结果与交互内容的相关性. 然而, 过度或不当的遗忘可能削弱模型的整体性能与行为一致性. 当扩散模型丢弃过多历史信息或关键特征时, 其生成结果可能出现语义割裂或缺乏连贯性, 从而影响用户在医学影像分析、自动驾驶决策或创意设计等关键应用场景中的信任度. 此外, 机器遗忘还可能引发类似于人类“灾难性遗忘”的问题：一旦关键知识被移除, 若缺乏大规模再训练支持, 其恢复过程将面临较高的成本与技术挑战.

因此, 在扩散模型中实现遗忘效果与模型效用保持之间的合理平衡, 是推动人工智能系统安全、稳定与可持续应用的关键问题. 通过对遗忘程度进行精细化调控, 开发者能够在保障生成安全性的同时维持模型的有效性与可靠性. 这不仅有助于提升注重隐私保护的聊天机器人或自适应医疗辅助系统的实际应用体验, 也能够避免由于记忆退化而引发的潜在性能风险.

参考文献

1 Xiong P, Tong X, SONG A, et al. Robotic cross-modal generative adversarial network based on variational Bayesian Gaussian mixture noise model. SCIENTIA SINICA Informationis, 2021, 51 (1): 2020-12-16

2 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023, 3836–3847

3 Ruiz N, Li Y, Jampani V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans,USA, 2023, 22500–22510

4 Xing X, Zhou H, Wang C, et al. Svgdreamer: Text guided svg generation with diffusion model. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024, 4546–4555

5 Meng C, Rombach R, Gao R, et al. On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans,USA, 2023, 14297–14306

6 Qi S D. Research progress on text-to-image diffusion models based on layout control. Computer Science and Application, 2025, 15: 443 [齐时达. 基于布局控制的文本到图像扩散模型研究进展. Computer Science and Application, 2025, 15: 443]

7 Kim C, Min K, Yang Y. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 461–478

8 Lyu M, Yang Y, Hong H, et al. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 7559–7568

9 Tang B, HU T, Du Y, et al. Curved-stroke-based neural painting and stylization through thin plate spline interpolation. SCIENTIA SINICA Informationis, 2024, 54 (2): 2024-02-05.

10 Jiang M, Zhi M, Li Y, et al. Super-resolution reconstruction of MR image with self-attention based generate adversarial network algorithm. SCIENTIA SINICA Informationis, 2021, 51 (6): 2021-05-11.

11 Yue Q, Zhang C K. A review of AIGC applications in multimodal scenarios. Journal of Frontiers of Computer Science & Technology, 2025, 19:1124 [岳颀, 张晨康. 多模态场景下 AIGC 的应用综述. Journal of Frontiers of Computer Science & Technology, 2025, 19:1124]

12 Croitoru F A, Hondru V, Ionescu R T, et al. Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45: 10850–10869

13 Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2023, 56: 1–39

14 Zhu Z Y, Zhao S G. Brain tumor medical image classification research based on diffusion model self-supervised representation learning. Computer Science and Application, 2024, 14: 133 [朱泽宇, 赵曙光. 基于扩散模型自监督表征学习的脑瘤医学图像分类研究. Computer Science and Application, 2024, 14: 133]

15 Zhao M, Zhang L, Yin B. Movable-type eraser: separable, recoverable, and sustainable multi-concept eraser in diffusion models. Science China Information Sciences, 2026, https://doi.org/10.1360/SSI-2025-0204. [赵梦楠, 张立和, 尹宝才. 活字擦除术: 面向扩散模型的可分离、可恢复和可持续的多概念擦除器. 中国科学：信息科学, 2026, https://doi.org/10.1360/SSI-2025-0204]

16 Lu S, Wang Z, Li L, et al. Mace: Mass concept erasure in diffusion models. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 6430–6440

17 Gong C, Chen K, Wei Z, et al. Reliable and efficient concept erasure of text-to-image diffusion models. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 73–88

18 Jain A, Kobayashi Y, Shibuya T, et al. Trasce: Trajectory steering for concept erasure. arXiv preprint arXiv:2412.07658, 2024

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

19 Schramowski P, Brack M, Deiseroth B, et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans,USA, 2023. 22522–22531

20 Gao H, Pang T, Du C, et al. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. arXiv preprint arXiv:2410.12777, 2024

21 Sun W, Dong X M, Cui B, et al. Attentive eraser: Unleashing diffusion model' s object removal potential via self-attention redirection guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, volume 39. 20734–20742

22 Hao S, Han K, Lv Z, et al. Conceptexpress: Harnessing diffusion models for single-image unsupervised concept extraction. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 215–233

23 Zhang Q T, Wang Y C, Wang H X, et al. A review of research on fine-tuning techniques for large language models. Journal of Computer Engineering & Applications, 2024, 60:224 [张钦彤, 王昱超, 王鹤羲, et al. 大语言模型微调技术的研究综述. Journal of Computer Engineering & Applications, 2024, 60: 224]

24 Hong S, Lee J, Woo S S. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024, volume 38. 21143–21151

25 Kim S, Jung S, Kim B, et al. Towards safe self-distillation of internet-scale text-to-image diffusion models. arXiv preprint arXiv:2307.05977, 2023

26 Kumari N, Zhang B, Wang S Y, et al. Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023. 22691–22702

27 Brack M, Friedrich F, Hintersdorf D, et al. Sega: Instructing text-to-image models using semantic guidance. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, USA, 2023, 36: 25365–25389

28 Gandikota R, Orgad H, Belinkov Y, et al. Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, Arizona, USA, 2024. 5111–5120

29 Seo J, Lee S H, Lee T Y, et al. Generative unlearning for any identity. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 9151–9161

30 Wu Y, Zhou S, Yang M, et al. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, USA, 2025, volume 39. 8496–8504

31 Fuchi M, Takagi T. Erasing concepts from text-to-image diffusion models with few-shot unlearning. arXiv preprint arXiv:2405.07288, 2024

32 Park Y H, Yun S, Kim J H, et al. Direct unlearning optimization for robust and safe text-to-image models. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2024, 37: 80244–80267

33 Moon S, Lee M, Park S, et al. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. arXiv preprint arXiv:2410.05664, 2024

34 Chavhan R, Li D, Hospedales T. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. arXiv preprint arXiv:2405.19237, 2024

35 Sridhar D, Vasconcelos N. Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 17–31

36 Xing Z, Feng Q, Chen H, et al. A survey on video diffusion models. ACM Computing Surveys, 2024, 57: 1–42

37 Li X, Liu Y, Lian L, et al. Q-diffusion: Quantizing diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023. 17535–17545

38 Ma X, Fang G, Wang X. Deepcache: Accelerating diffusion models for free. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 15762–15772

39 Somepalli G, Singla V, Goldblum M, et al. Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans,USA, 2023. 6048–6058

40 Corvi R, Cozzolino D, Poggi G, et al. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans,USA, 2023, 973–982

41 Zhang Z, Zhang Q, Xing W, et al. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2024, volume 38. 7396–7404

42 Liu H, Wang Y, Qian B, et al. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 8038–8047

43 Yue Z, Wang J, Loy C C. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In: Proceedings

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

of Advances in Neural Information Processing Systems, Vancouver, Canada, 2024, 36
44 Fan Y, Liu C, Yin N, et al. Adadiffsr: Adaptive region-aware dynamic acceleration diffusion model for real-world image super-resolution. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 396–413
45 Li K, Gong C, Li Z, et al. PrivImage: Differentially private synthetic image generation using diffusion models with Semantic-Aware pretraining. In: 33rd USENIX Security Symposium, Philadelphia, USA, 2024. 4837–4854
46 Yang Y, Gao R, Wang X, et al. Mma-diffusion: Multimodal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 7737–7746
47 Zhu P, Takahashi T, Kataoka H. Watermark-embedded adversarial examples for copyright protection against diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 24420–24430
48 Zhang Y, Chen X, Jia J, et al. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2024, 37: 36748–36776
49 Liu Z, Chen K, Zhang Y, et al. Implicit concept removal of diffusion models. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 457–473
50 Pham M, Marshall K O, Cohen N, et al. Circumventing concept erasure methods for text-to-image generative models. In: Proceedings of International Conference on Learning Representations, Rwanda, 2023
51 Fan C, Liu J, Hero A, et al. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024
52 Liu Z, Chen K, Zhang Y, et al. Implicit concept removal of diffusion models. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 457–473
53 Liu Z, Chen K, Zhang Y, et al. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. arXiv preprint arXiv:2310.05873, 2023
54 Lyu M, Yang Y, Hong H, et al. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 7559–7568
55 Zhang G, Wang K, Xu X, et al. Forget-me-not: Learning to forget in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024. 1755–1764
56 Gandikota R, Materzynska J, Fiotto-Kaufman J, et al. Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023. 2426–2436
57 Bui A, Vuong L, Doan K, et al. Erasing undesirable concepts in diffusion models with adversarial preservation. arXiv preprint arXiv:2410.15618, 2024
58 Liu Y, An J, Zhang W, et al. Realera: Semantic-level concept erasure via neighbor-concept mining. arXiv preprint arXiv:2410.09140, 2024
59 Bui A, Vu T, Vuong L, et al. Fantastic targets for concept erasure in diffusion models and where to find them. In: Proceedings of International Conference on Learning Representations, EXPO, Singapore, 2025
60 Gong C, Chen K, Wei Z, et al. Reliable and efficient concept erasure of text-to-image diffusion models. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 73–88
61 Wang R, Fang J, Li J, et al. Ace: Concept editing in diffusion models without performance degradation. In: Proceedings of the ACM International Conference on Multimedia, Dublin, Ireland, 2025. 10525–10534
62 Sharma A S, Sarkar N, Chundawat V, et al. Unlearning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. arXiv preprint arXiv:2409.05668, 2024
63 Lu K, Kriplani N, Gandikota R, et al. When are concepts erased from diffusion models? In: Proceedings of Advances in Neural Information Processing Systems, San Diego, USA, 2025
64 Tsai Y L, Hsu C Y, Xie C, et al. Ring-a-bell! how reliable are concept removal methods for diffusion models? arXiv preprint arXiv:2310.10012, 2023
65 Chin Z Y, Jiang C M, Huang C C, et al. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. arXiv preprint arXiv:2309.06135, 2023
66 Zhang Y, Jia J, Chen X, et al. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 385–403
67 Huang C P, Chang K P, Tsai C T, et al. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In: Proceedings of European Conference on Computer Vision, Milan, Italy, 2024. 143–151
68 Han T, Sun W, Hu Y, et al. Continuous concepts removal in text-to-image diffusion models. arXiv preprint arXiv:2412.00580, 2024
69 Pham M, Marshall K O, Hegde C, et al. Robust concept erasure using task vectors. arXiv preprint arXiv:2404.03631, 2024
70 Zhao M, Zhang L, Yang X, et al. Advanchor: Enhancing diffusion model unlearning with adversarial anchors. arXiv preprint arXiv:2501.00054, 2024
71 Zhu J, Han B, Yao J, et al. Decoupling the class label and the target concept in machine unlearning. arXiv preprint

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

arXiv:2406.08288, 2024
72 Yang T, Cao J, Xu C. Pruning for robust concept erasing in diffusion models. arXiv preprint arXiv:2405.16534, 2024
73 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, USA, 2022. 10684–10695
74 Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, Long Beach, USA, 2017. 86–98
75 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018. 586–595
76 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016. 770–778
77 Bedapudi P. Nudenet: Neural nets for nudity classification, detection and selective censoring, https://gitee.com/mirrors_barseghyanartur/NudeNet, 2019
78 Howard J, Gugger S. Fastai: A layered api for deep learning. Information, 2020, 11: 108

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408

ShallowUnlearn: Rethinking concept unlearning through disentangled component-level erasure in text-guided diffusion models

Mengnan Zhao[1], Tianhang Zheng[2], Bo Wang[3], Lihe Zhang[3*] & Baocai Yin[1]

1. *School of Computer Science and Technology, Dalian University of Technology, Dalian 116081, China*
2. *School of Computer Science and Technology, Zhejiang University, hangzhou 310058, China*
3. *School of Information and Communication Engineering, Dalian University of Technology, Dalian 116081, China*

* Corresponding author. E-mail: zhanglihe@dlut.edu.cn

Abstract Text-guided diffusion models have revolutionized image synthesis but also raise ethical concerns, such as privacy violation and harmful content generation. To mitigate these issues, prevailing methods typically leverage an alignment mechanism—with predefined erasure references—to fine-tune pretrained model weights. However, such methods suffer from two major limitations: (1) they are constrained by the representation capacity of the text space and exhibit high sensitivity to the predefined erasure references; (2) they tend to cause significant degradation of model utility during concept erasure. To overcome these limitations, we introduce ShallowUnlearn, aiming to disentangle and remove the erasure-critical component of the undesirable concept. Specifically, ShallowUnlearn comprises a Component Extraction Module (CEM) and a Swap Disentangling Strategy (SDS). Guided by SDS, CEM is pre-trained to decompose concept embeddings into distinct component types. Leveraging this decomposition, ShallowUnlearn then removes the erasure-critical component while retaining non-critical ones by fine-tuning model weights. Extensive experiments demonstrate that ShallowUnlearn achieves effective concept erasure with minimal impact on overall model performance.

Keywords diffusion model unlearning, shallowunlearn, component extraction module, swap disentangling strategy, performance trade-off

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0408