arXiv:2602.01513v1 [eess.IV] 2 Feb 2026

MarkCleaner：通过不可感知的微几何扰动实现高保真水印去除

孔晓曦 [1] 袁洁宇 [2] 陈鹏迪 [1] 张源林 [2] 李崇义 [2] 李斌 [1]

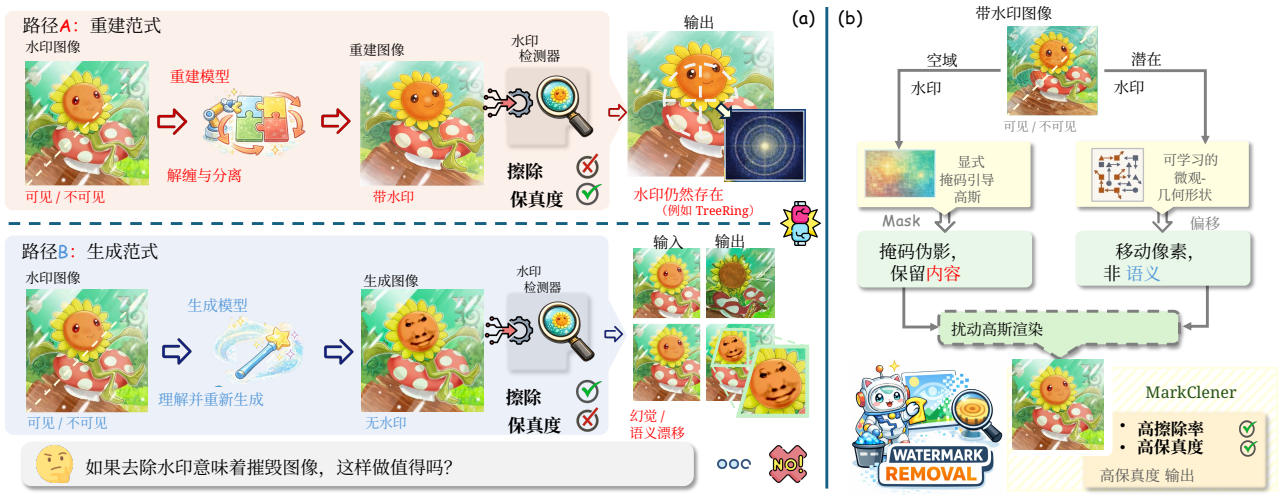


图 *1.* **MarkCleaner 在不损害视觉保真度的情况下去除水印。** 现有范式面临一个根本困境：只能实现高保真度或高擦除率，无法兼得。(a) 现有方法在保真度与擦除率之间的权衡：基于重建的方法（路径 *A*）保真度高，但无法去除语义水印；生成式方法（路径 *B*）虽能成功擦除水印，却会造成严重的语义漂移、产生幻觉内容或改变细节。(b) 我们的 MarkCleaner 通过掩码引导编码和几何扰动解决这一问题，为可见水印和不可见水印提供统一方案，从而干净地去除水印且不扭曲内容。

摘要

语义水印对传统图像空间攻击具有很强的鲁棒性。在这项工作中，我们表明，这种鲁棒性在微几何扰动下不复存在：空间位移可通过破坏相位对齐来去除水印。受此观察启发，我们提出 MarkCleaner，这是一种水印去除框架，可避免基于再生成的水印去除方法所导致的语义漂移。具体而言，MarkCleaner 使用微几何扰动监督进行训练，促使模型将语义内容与严格的空间对齐分离，并能够在细微的几何位移下实现鲁棒重建。该框架采用掩码引导的编码器学习显式空间表征，基于二维高斯泼溅的解码器则对几何扰动进行显式参数化，同时保留语义内容。大量实验表明，MarkCleaner 在水印去除效果和视觉保真度方面均表现出色，同时支持高效的实时推理。我们的代码将在论文录用后公开。

[1]广东省智能信息处理重点实验室，深圳大学 [2]南开大学计算机科学系 VCIP。通讯作者：袁洁宇 <jieyuyuan.cn@gmail.com>，李斌 <libin@szu.edu.cn>。

1. 引言

数字水印是图像版权归属和内容溯源的基础技术。随着生成式人工智能的快速发展（Rombach 等人，2022年；Achiam 等人，2023；Labs 等人，2025），人类创作内容与人工智能生成内容之间的界限日益模糊，可靠的水印嵌入需求也愈发迫切。近年来，水印嵌入技术已从可见叠加层转向不可见水印（Sharma 等人，2024年；Gowal 等人，2025）。在不可见水印中，

近期的生成式方法通过干预采样过程或调制初始潜在分布来嵌入信号。TreeRing 等方法（Wen 等人，2023）通常被归类为**语义水印**（M¨uller 等人，2025），因为它们通过与生成图像的高层语义结构相融合来嵌入水印信息，而不是直接操纵像素或频域。此类深度耦合显著增加了移除难度，使语义水印成为评估水印净化方法极限的极具挑战性的基准测试。为了评估其实际可靠性，有必要探索潜在的移除机制。

从技术实现角度来看，现有水印去除方法可分为两种范式，如图1（a）所示，但这两种范式在应对语义水印时都存在固有局限。**基于重建的去除方法**尝试通过去噪和压缩技术重建干净图像来消除水印，这些技术优先保证像素级保真度，从而有效抑制嵌入的水印信号（Dabov等人，2007；B´egaint等人，2020；Alam等人，2025）。然而，语义水印是通过修改初始潜在噪声嵌入的，因此不可避免地与潜在结构表示纠缠在一起。因此，重建保真度越高，保留的水印信息就越多，这与水印去除的目标直接矛盾。**基于生成的方法**利用生成模型重新生成图像内容，有效破坏原有的潜在结构，从而很好地擦除语义水印（Zhao等人，2024年；Liu等人，2025b；Qiu等人，2025）。由于生成过程具有迭代性，通常计算开销较大；要获得高质量结果，需要投入大量资源和时间。最重要的是，去噪过程固有的随机性不可避免地会引发幻觉、语义漂移，并导致细粒度细节丢失。这揭示了现有方法面临的根本困境：高保真度或高擦除率，二者不可兼得。如果去除水印必须以破坏图像内容为代价，这种权衡真的可以接受吗?

在本研究中，我们发现语义水印从根本上容易受到几何扰动的影响。这一关键见解源自频域视角。研究已充分证实，相位谱主要编码结构信息，而幅度谱则捕捉纹理属性（Oppenheim & Lim，1981年；Yang & Soatto，2020；Xu et al.，2021年）。TreeRing 水印通过调制傅里叶变换潜在空间中的频谱嵌入，使水印信号与图像结构紧密耦合，并在信号处理攻击下保持稳定。然而，几何变换在频域中表现为相移。即使是极小的几何扰动，也足以破坏精确的相位对齐，同时对人眼几乎不可察觉。因此，水印去除可以从内容再生重新定义为几何扰动，从而在不损害视觉内容一致性的前提下实现高保真度去除。

基于这一发现，我们提出**MarkCleaner**，一种新颖的水印移除框架，如图1(b) 所示。其核心原理是训练网络学习微几何扰动，同时保持视觉一致性，使输出在视觉上与输入一致，但在几何上发生偏移，从而移除水印。具体而言，我们提出一种掩码引导的编码策略，在频率和空域层级上抑制水印模式，同时提取语义表征。在解码阶段，我们提出一种新颖的 2D 高斯泼溅（2DGS）渲染器，通过显式参数化微几何扰动来打破空间对齐。训练过程中，模型以经过几何扰动的目标而非原始输入作为监督，从而引导解码器应用这些变换，同时使编码器学习扰动不变语义特征。为保持语义一致性，我们还引入自监督视觉特征对齐作为额外约束。

总而言之，我们的贡献如下：

• **揭示语义水印的几何脆弱性**。我们揭示了语义水印中一种被忽视的相位敏感性，并证明即使是不可感知的微几何扰动也能破坏水印检测。这一发现揭示了其根本性的几何脆弱性，并促使水印移除范式从内容再生转向基于几何的水印移除。

• **一种新颖的水印移除框架**。 我们提出了 MarkCleaner，它结合掩码引导编码与几何扰动监督，训练模型在空域位移下保持语义一致性。该框架在 12 种不同的水印方案上均取得最先进的性能，且无需预先了解水印类型或检测机制。

• **扰动二维高斯渲染**。 我们提出了一种基于 2DGS 的解码器，采用可微光栅化来学习空域位移，从而实现实时推理，同时保持像素级结构。

2. 相关工作

图像水印。 图像水印嵌入可见和不可感知的信号，从而实现来源验证。现有的不可见水印可根据水印嵌入的时机以及其在表征中的位置进行分类。 后处理水印在生成图像合成后对其进行处理。传统方法（Cox 等人，2007年；Navas 等人，2008年）嵌入水印-

标记通过小波变换和离散余弦变换嵌入频域。基于深度学习的方法（Zhang 等，2019；Fernandez 等，2022；Tancik 等，2020）采用编码器-解码器结构。近期工作（Liu 等，2025a；Lu 等，2025）引入渐进式训练，并利用生成先验来提升对部分图像盗用和基于人工智能的编辑的鲁棒性。处理中水印将水印嵌入整合到生成过程中，从而形成内在的语义耦合。模型微调方法（Fernandez 等，2023）对 VAE 解码器进行微调，而潜变量操纵方法（Wen 等，2023；Ci 等，2025；Lee 和 Cho，2025；Yang 等，2024；2025）在傅里叶变换空间中构造扩散噪声模式，或保留高斯噪声的统计特性。这种耦合使模型对压缩和噪声具备更优异的鲁棒性。

水印去除攻击。从实现角度来看，现有的移除策略可分为三类（Liu等人，2025b）。信号处理方法采用 JPEG 压缩和滤波等变换（Ding等人，2024年），但主要针对像素级扰动，对语义水印的效果有限。再生成方法利用扩散模型（Liu等人，2025b；Qiu等人，2025）或变分自编码器（Cheng等人，2020）重建图像，但抑制语义水印通常需要大量注入噪声，可能导致语义漂移。对抗攻击通过优化扰动来误导检测器（Ren等人，2024年），但通常需要白盒访问（Lukas等人，2024年）或代理模型训练（Wu等人，2023；Saberi等人，2024年），限制了其实用性。这些范式关注的是修改图像内容，而非空间配置。我们的分析表明，语义水印可能对几何扰动较为敏感，因此可以通过最小幅度的几何变换来抑制水印，同时保持语义保真度。

高斯泼溅。高斯泼溅（GS）（Kerbl等人，2023）使用显式高斯基元表示场景，从而实现兼具高视觉保真度的实时渲染。近期研究将GS扩展至二维图像域，将图像建模为二维高斯分布的连续混合（Hu等人，2025年）。凭借出色的效能，基于图像的2DGS迅速应用于图像压缩（Zhang等人，2024年）、超分辨率（Chen等人，2025年）和图像修复（Li等人，2025年），展现出连续空间建模和高效渲染方面的优势。此外，基于高斯的表示在几何变换下表现出很强的鲁棒性和灵活性（Lee等人，2024年；Peng等人，2025年）。我们认为，GS精确的位置参数化为几何扰动建模提供了一种机制，因此非常适合在保留视觉连贯性的同时破坏刚性的水印模式。受这些特性的启发，我们探索了2DGS在水印去除方面的潜力。

3. 初步分析

本节以TreeRing（Wen等人，2023）为例，对语义水印中的几何脆弱性进行理论和实证分析。

3.1. 水印形式化与验证

考虑一个通过 DDIM采样（Song 等人，2021年）和 VAE解码生成图像的潜扩散模型（LDM），其图像来自噪声 $I \in \mathbb{R}^{H\times W\times 3}$。TreeRing 将签名模式 $k^* \in \mathbb{C}^{|M|}$ 嵌入到 z_T 的低频掩码内的傅里叶系数中 M。具体而言，水印噪声 $\tilde{z}_T$满足 $\mathcal{F}(\tilde{z}_T)[M] = k^*[M]$和_ _$\mathcal{F}(\tilde{z}_T)[M] \sim \mathcal{N}(0,\ 1)$ 其中 M表示其补集。随后，通过去噪和VAE解码器生成带水印图像 I_w。

检测通过 VAE编码和 DDIM反演，从疑似图像 I 中恢复初始噪声潜变量 $\hat{z}_T$。通过比较恢复的傅里叶系数与频率掩码 M上的k^* 之间的归一化 L_1距离，验证水印是否存在：

$$d = \frac{1}{|M|}\sum_{\mathbf{u}\in M}\left|k^*_{\mathbf{u}} - \hat{Z}_{\mathbf{u}}\right|, \tag{1}$$

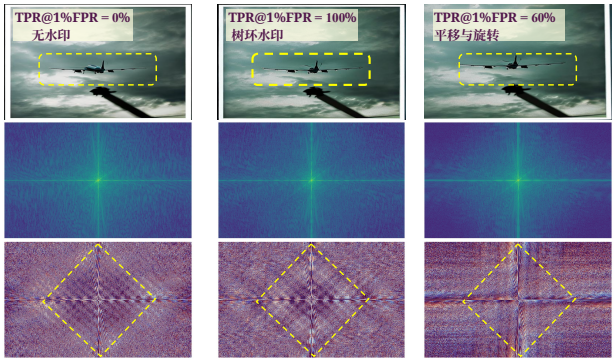
其中，$\hat{Z} = \mathcal{F}(\hat{z}_T)\mathbf{u}$ 表示潜空间傅里叶空间中的频率坐标，且 $d < \tau$ 表示检测结果达到预设阈值 τ。该机制依赖于 k^* 与在 M中恢复的 $\hat{Z}$ 之间的对齐。

3.2. 潜相位的几何脆弱性

变分自编码器的步幅架构造成了一个关键脆弱性：像素域变换$\mathcal{T}_{\boldsymbol{\Delta},\ \theta}$ 会引发一连串相位域扰乱（Zhou 等人，2025）。像素空间中的平移$\boldsymbol{\Delta}$ 会在潜在空间中引入分数偏移 $\boldsymbol{\delta} = \boldsymbol{\Delta}/d$（$d = 8$）。根据傅里叶移位定理，这种空域位移会传递为相位调制：

$$\hat{Z}'_{\mathbf{u}} = \hat{Z}_{\mathbf{u}} \cdot e^{i\Phi(\mathbf{u},\mathcal{T})}, \tag{2}$$

其中 $\Phi(\mathbf{u}\ \mathcal{T}) = -2\pi\langle \mathbf{u}\ \frac{\boldsymbol{\Delta}}{W}\rangle + \theta \cdot \arg(\mathbf{u})$，编码了组合相位偏移；$\arg(\mathbf{u}) = \operatorname{atan2}(u_y, u_x)$ 表示频率向量 $\mathbf{u}$ 的角度。尽管几何变换能够保留幅度谱，却会系统性地旋转复平面中的每个傅里叶系数。关键的是，式（1）中的检测度量计算复数值之间的 L_1距离，而该距离会随着相位不一致而增大：$|k^*_{\mathbf{u}} - Z'_{\mathbf{u}}| \approx 2|k^*_{\mathbf{u}}||\sin(\Phi_{\mathbf{u}}/2)|$。随着位移增大，相位斜坡 $\Phi_{\mathbf{u}}$ 会使潜在向量偏离 k^*的实轴流形，从而增大L_1 误差。同时，旋转角度为 θ 时会导致频谱旋转，使恢复的系数与嵌入的密钥 $k^*_{\mathbf{u}}$错位。由此产生的相消干涉会导致



(a) 干净图像　(b) 带水印　(c) 变换后

图 2. **几何变换所致相位扰动的频域分析**。从相同的初始潜变量出发，我们展示了 (a) 标准生成、(b) 带潜变量水印的生成，以及 (c) 对 (b) 进行几何变换后的结果。每列均展示 RGB图像及其潜变量幅度和相位谱。几何变换会破坏水印引起的相位波纹，同时保留幅度结构，这表明水印失效源于相位调制，而非内容改变。TPR@1%FPR 是基于 1000 张图像计算得出的。（放大查看细节。）

检测得分 d 从 0 逐渐趋近均值幅度 $A \gg \tau$，进而导致检测失败（详见附录 C），图 2中的实证结果验证了这一点：带水印图像与变换后图像的幅度谱几乎完全相同，证实视觉保真度得以保留；而经过 7px 平移和 5°旋转后，相位模式则变得杂乱。在所示的代表性示例中，一种变换增大了检测距离。在 1000 张生成图像上进行评估后发现，这种极小的扰动使 TPR@1% FPR=100% 从原值降至 60%，表明微几何变换能够在保持感知质量的同时，有效地使 TreeRing 水印检测失效。

我们的分析表明，语义水印依赖于图像空间与潜在空间之间稳定的空间对应关系。这暴露出一个关键漏洞：通过微几何扰动，可以在保持感知保真度的同时使其可检测性失效，这表明空间锁定水印机制存在根本性局限。更重要的是，这种敏感性并非语义水印所独有。空间域和频域中的传统方案长期以来一直容易受到旋转、缩放和平移攻击的影响，从而破坏嵌入信号（Lin 等人，2024年）。尽管相关研究已持续数十年，近期基准测试仍证实，几何攻击在现代水印范式中依然是一个持续存在的挑战（Ding 等人，2024年），这凸显出空间对齐是一项根本性限制，而非特定方案所独有的局限。受这一观察启发，我们提出 MarkCleaner——一个利用这种依赖关系，通过受控几何变换实现高效、统一水印去除且不产生语义漂移的框架。

4. 方法

受第3节分析的启发，我们提出 MarkCleaner，一种通过受控微几何扰动实现统一水印去除的框架。如图 3所示，MarkCleaner 将输入图像表示为一组二维高斯基元，并通过可微高斯渲染学习不可感知的几何位移，从而实现水印去除。该框架采用基于 UNet 的编码器-解码器架构。给定一幅输入图像，我们分别在空域和频域中施加双域掩码攻击，以破坏潜在的水印模式（第4.1节）。随后，视觉编码器提取多尺度特征，并将其解码为逐图像块的二维高斯参数，再通过可微光栅化进行渲染（第4.2节）。在训练过程中，模型并不以重建输入为监督目标。相反，渲染输出会与输入图像经过微几何扰动后的版本进行匹配，使模型能够在保留视觉内容的同时学习几何位移（第4.3节）。为确保空间位移下的语义一致性，我们进一步引入特征级对齐和感知约束（第4.4节）。下面将详细介绍各个组成部分。

4.1. 掩码引导编码

为了破坏嵌入在空域或频域中的水印，我们在编码前应用双域掩码。通过对输入的部分区域进行随机遮罩，网络被促使基于全局语义结构重建底层的干净内容。

频带掩码（FBM）。我们首先通过二维快速傅里叶变换（FFT）将输入图像 I 变换到频域。以频谱原点为中心、半径为 r 的圆形掩码用于分离低频与高频分量。在训练过程中，我们以相等概率随机应用低通或高通滤波。这种随机频域掩码会破坏频谱水印特征（例如 TreeRing 中的环形图案），而与其嵌入位置无关，同时保留足够的结构信息以供重建。

空间随机掩码（SRM）。随后，我们施加空间掩码以破坏空间连续性，从而实现干净的语义修复。具体而言，我们随机掩码 r% 的像素，其中掩码比例 r、空间位置和区域尺寸均在训练过程中随机采样。这种不规则的空间掩码训练编码器从不完整的空间信息中提取连贯的语义特征，这对于在解码期间施加几何扰动时保持内容一致性至关重要，同时可防止对完整图像结构的过拟合。作为额外益处，用户可以选择性地提供可见水印区域的二值掩码，以引导局部移除。

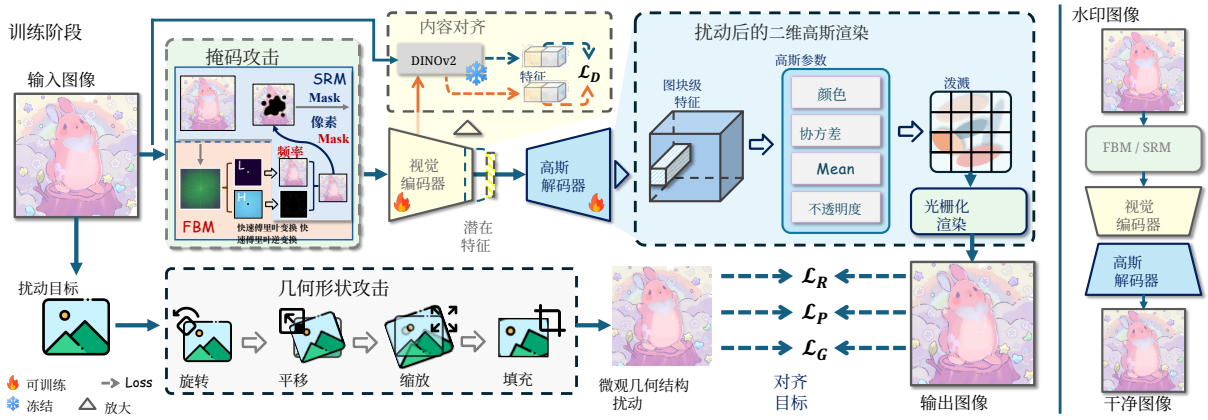


图 *3.* **MarkCleaner 概览**。 训练阶段：输入图像首先经过双域掩码处理，包括频带掩码（FBM）和空间随机掩码（SRM），随后被编码并解码为每个图像块的二维高斯参数（均值、协方差、颜色、不透明度），以进行可微光栅化。模型使用几何扰动目标而非原始输入进行监督，采用损失 $\mathcal{L}_R$、 $\mathcal{L}_P$、 $\mathcal{L}_G$，以及基于 DINOv2 的对齐损失 $\mathcal{L}_D$。推理阶段：带水印图像经过相同的处理流程，并可选择使用用户提供的掩码标注可见水印区域，从而生成语义一致但经过几何扰动的输出。

在训练过程中，FBM 与 SRM 被组合为随机掩码操作。当采样到频域掩码时，它会在空间掩码之前施加。掩码图像在编码前仍会被变换回空间域，以确保训练与推理之间输入表征的一致性（Xie 等人，2023）。

4.2. 扰动二维高斯渲染

通用的逐像素解码器会在固定的空间网格上生成输出，其中像素值与预先确定的位置紧密耦合，限制了空间重组的灵活性（Chen等人，2021年；Zhang等人，2024年）。由于这种僵化，即使是轻微的亚像素偏移，也会导致整个网格出现显著的语义错位，迫使解码器重新合成全局像素值来模拟位移。相比之下，2DGS将图像表示为由重叠高斯基元构成的连续场，其中每个基元都具有可学习的显式参数。尤为重要的是，这一特性将几何属性（位置、协方差）与外观（颜色、不透明度）解耦。显式表征对于水印去除尤为有益，因为评估时水印失效主要源于空间错位，而非语义退化。我们可以学习一个作用于高斯中心的微偏移场，而不必强迫网络回归复杂的像素级变化来隐藏水印。因此，我们采用2DGS作为渲染解码器。

高斯表示。 每个二维高斯基元由均值 $\mu_i \in \mathbb{R}^2$ （空域位置）、协方差 $\Sigma_i \in \mathbb{R}^{2\times2}$ （空域范围）、颜色 $c_i \in \mathbb{R}^3$和不透明度$\alpha_i \in \mathbb{R}$定义。协方差参数化为 $\Sigma_i = L_i L_i^\top$，并采用下三角形式 L_i 以确保半正定性。位置 p 处的像素值通过对高斯贡献求和计算：

$$I_p = \sum_i c_i \alpha_i \exp\left(-\frac{1}{2}(p-\mu_i)^\top \Sigma_i^{-1}(p-\mu_i)\right). \quad (3)$$

块级高斯解码。 首先，使用基于UNet的编码器将掩码图像编码为多尺度潜在特征。随后，高斯解码器根据这些特征预测每个块的高斯参数。为高效处理高分辨率图像，我们将图像划分为非重叠块，并为每个块分配一组专用高斯基元，以此进行块级光栅化（Li et al.，2025）。相邻块之间设有重叠边界，并通过混合处理确保平滑过渡。最终输出通过可微光栅化获得，从而实现端到端训练。更多详情见附录D。

4.3. 几何偏差策略

我们施加微几何扰动以生成一个发生位移的监督目标，使模型能够在训练期间学习度量位移。

扰动流程。 给定输入图像 I，我们依次应用四种几何变换，得到扰动目标 I_{pert}：

- **旋转**：围绕图像中心进行轻微旋转。
- **平移**：沿两个坐标轴进行微小的空间偏移。
- **缩放**：进行轻微的尺度调整，然后调整至原始分辨率。
- **填充**：上述变换造成的缺失区域将使用原始图像中的相应内容填充。

扰动幅度被刻意保持在极小范围内（例如旋转仅几度、平移仅几个像素），

确保由此产生的位移对人类观察者而言难以察觉，同时足以破坏频域和语义水印检测器所利用的精确空域与相位对齐。通过监督模型重构 I_{pert} 而非 I，网络学会将几何位移融入生成过程。在推理时，这种习得的行为会自然地产生语义上与输入一致、但几何位置有所偏移的输出，从而使水印解码失效，同时不改变视觉内容。

微几何扰动监督。 与传统图像重建通过最小化输出与原始输入之间的距离不同，我们的训练以扰动目标 I_{pert}为参照，对渲染输出 I_{out} 进行监督。这种设计促使模型将空间位移内化为所学习表征的一部分，而不是将其视为需要纠正的误差。在推理时，这种习得的行为会自然地产生与输入在语义上保持一致、但在几何上发生位移的输出，从而使水印解码失效，同时不改变视觉内容。

4.4. 内容对齐。

自监督特征对齐。 为在发生几何位移时保留内容和语义一致性，我们采用冻结的 DINOv2 编码器（Oquab 等人，2024年）提取的自监督视觉特征作为内容锚点。DINOv2 特征能够捕捉高层场景结构，并且对外观变化和局部扰动具有很强的不变性，因此非常适合我们的任务。具体而言，我们使用冻结的 DINOv2 编码器从掩码输入图像中提取特征。为适应不同的掩码比例和空间损坏，我们采用轻量级多层感知机在将特征用于条件化高斯解码器之前对其进行调整，该条件化通过自适应层归一化（AdaLN）（Karras 等人，2019）实现。这种条件化促使解码器在允许局部几何位移的同时保留全局语义结构。此外，我们采用基于 DINO 的特征对齐损失，通过计算渲染输出与扰动目标之间的余弦相似度，进一步确保空间偏移下的语义一致性。详情见附录E。

损失函数。 总训练目标结合了多种互补损失：

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_P + \lambda_3 \mathcal{L}_G + \lambda_4 \mathcal{L}_D, \quad (4)$$

其中 $\mathcal{L}_{rec}$ 是重建损失，结合了 L_1 和频域项，以确保像素级和频谱保真度； $\mathcal{L}_P$是用于提升高层视觉质量的 LPIPS 感知损失； $\mathcal{L}_G$ 是用于鼓励生成逼真输出的对抗损失； $\mathcal{L}_D$ 是用于保证语义一致性的内容特征对齐损失。损失函数的具体形式详见附录F。

5. 实验

5.1. 实验设置

数据集与模型设置。我们使用来自 MS-COCO-2017 的 5,000 张图像训练 MarkCleaner（Lin 等人，2014），并使用来自 DiffusionDB-2M 的 1,000 对带水印和干净图像进行评估（Wang 等人，2023）。所有图像均由 Stable Diffusion v2-1 生成（Rombach 等人，2022年），分辨率为 512×512，采用 7.5 的无分类器引导和 50 个去噪扩散隐式模型（DDIM）步骤（Song 等人，2021年）。

水印嵌入方法。MarkCleaner 在十五种公开发布的水印上进行了评估，其中包括后处理方法（DwtDct（Cox 等，2007）、SSL（Fernandez 等，2022）、StegaStamp（Tancik 等，2020）、VINE（Lu 等，2025）、WOFA（Liu 等，2025a））和处理中方法（Stable Sig.（Fernandez 等，2023）、TreeRing（Wen 等，2023）、RingID（Ci 等，2025）、HSTR（Lee &Cho，2025）、HSQR（Lee 和 Cho，2025）、Gaussian Shading（Yang 等，2024）、T2SMark（Yang 等，2025））。每个基线方法都显式地以鲁棒性为目标，为移除性能提供了严格的基准。

水印去除方法。 我们将 Mark-Cleaner 与 14 种具有代表性的水印去除方法进行比较，涵盖三类：基于失真的攻击、利用生成先验的基于再生成的攻击（Cheng等人，2020；Saberi等人，2024年；Liu等人，2025b），以及基于优化的攻击（Kassis和Hengartner，2025）。所有评估均在严格的黑盒设置下进行。为确保比较公平，我们采用等质量协议，使各方法的输出视觉保真度保持一致，从而单独评估它们在水印擦除方面的相对效果。详细配置见附录 G.1。

实现与评估指标。 所有模型均使用 Adam 训练，学习率为 2×10^{-4}，批量大小为 16。图块大小为 16×16，每个高斯均以 12 维嵌入进行参数化。实验使用四张 NVIDIA L40 GPU 进行。水印鲁棒性通过 TPR@1%FPR（越低越好）和比特准确率（ACC，0.5 表示随机化）进行评估，遵循先前工作（Wen等人，2023；Liu等人，2025b）。图像质量使用 FID（Heusel等人，2017）、CLIP 分数（Hessel等人，2021年）和 LPIPS（Zhang等人，2018）进行评估。

5.2. 实验结果

我们采用各方法均经过校准、以生成保真度相当的图像的统一质量评估方案，评估水印去除效果、视觉质量和推理速度。

水印擦除性能。。表1报告了在 DiffusionDB- 2M Prompts（Santana，2022年）上 14 种攻击方法的水印移除结果。MarkCleaner 始终实现了最低的 mTPR 和 mΔA，在所有评估的

表 *1.* 针对各种移除攻击的鲁棒性定量比较。结果以 TPR@1%FPR / 准确率表示。最后一列通过 mTPR / mΔA（平均 TPR 和平均 |准确率 − 0.5|）概述总体表现。数值越低，表示移除效果越好，检测结果越接近随机猜测。我们的方法在所评估的移除攻击中总体表现最佳。

攻击类型	DwtDct	SSL	Stega 印章	Stable 标记	VINE	WOFA	高斯阴影	T2S Mark	Tree Ring	RingID	HSTR	HSQR	mTPR(↓) mΔA(↓)
None	.800/.888	1.0/1.0	1.0/.999	1.0/.993	1.0/.999	.977/.832	1.0/1.0	1.0/1.0	.943/.970	1.0/1.0	1.0/1.0	1.0/1.0	.977 / .473
JPEG 压缩	.003/.490	.243/.680	1.0/.997	.782/.695	1.0/.989	.040/.458	1.0/1.0	1.0/.998	.083/.945	1.0/1.0	.980/.988	1.0/1.0	.678 / .353
裁剪与缩放	.007/.521	.927/.875	.033/.554	.999/.978	.013/.505	.800/.631	.003/.501	.000/.501	.000/.842	.010/.582	.070/.597	.153/.702	.251 / .154
高斯模糊	.357/.682	.987/.981	1.0/1.0	.818/.817	1.0/.998	.693/.707	1.0/1.0	1.0/1.0	.593/.965	1.0/1.0	1.0/1.0	1.0/1.0	.871 / .429
高斯噪声	.000/.505	.010/.508	.987/.872	.000/.543	1.0/.901	.069/.439	1.0/.991	.987/.951	.000/.925	.993/1.0	.103/.578	.987/.990	.511 / .267
颜色抖动	.061/.523	.517/.769	.927/.868	.895/.848	.953/.884	.278/.470	.990/.993	.953/.973	.247/.962	.960/.998	.927/.967	1.0/1.0	233 / .138
旋转	.000/.522	.973/.918	.000/.510	.998/.813	.010/.497	.987/.845	.007/.539	.000/.499	.000/.850	1.0/1.0	.103/.578	.060/.613	.251 / .154
平移	.003/.501	.980/.923	.255/.587	1.0/.991	.013/.504	.822/.963	.027/.568	.000/.501	.000/.888	.013/.484	.087/.585	.057/.563	.271 / .172
VAE-B（ICLR，2018）	.000/.499	.760/.818	1.0/.999	.730/.680	1.0/.990	.157/.439	1.0/1.0	1.0/1.0	.190/.957	1.0/1.0	1.0/1.0	1.0/1.0	.686 / .365
VAE-C（CVPR，2020）	.000/.496	.410/.720	1.0/.998	.582/.652	.997/.959	.173/.440	1.0/1.0	1.0/1.0	.123/.947	1.0/1.0	.990/.993	1.0/1.0	.606 / .351
DA（ICLR，2024年）	.000/.494	.003/.510	.043/.531	.625/.463	.450/.609	.477/.479	1.0/.987	.993/.939	.000/.880	.927/.996	.800/.828	1.0/.998	.526 / .226
CtrlRegen+（ICLR，2025b）	.003/.494	.030/.557	.197/.603	.023/.479	.833/.671	.110/.416	1.0/.999	1.0/.987	.001/.885	.980/.999	.437/.890	1.0/.998	.468 / .248
UnMarker (S&P, 2025)	.010/.538	.973/.918	.990/.946	.999/.981	.007/.502	.997/.917	1.0/1.0	.017/.502	.010/.948	.460/.939	.097/.693	.033/.575	.466 / .288
IRA（CVPR，2025）	.000/.487	.680/.759	.990/.905	.625/.508	1.0/.924	.100/.418	1.0/1.0	1.0/1.0	.800/.975	1.0/1.0	.990/.995	1.0/1.0	.682 / .331
NFPA（神经信息处理系统大会，2025）	.000/.500	.030/.565	.013/.481	.043/.494	.013/.502	.087/.387	.003/.514	.000/.499	.003/.913	.030/.672	.153/.645	.367/.795	.061 / .154
MarkCleaner（我们的方法）	**.003/.500**	**.000/.504**	**.010/.488**	**.000/.446**	**.063/.554**	**.002/.494**	**.003/.554**	**.003/.500**	**.001/.795**	**.003/.802**	**.051/.630**	**.022/.766**	**.014 / .094**

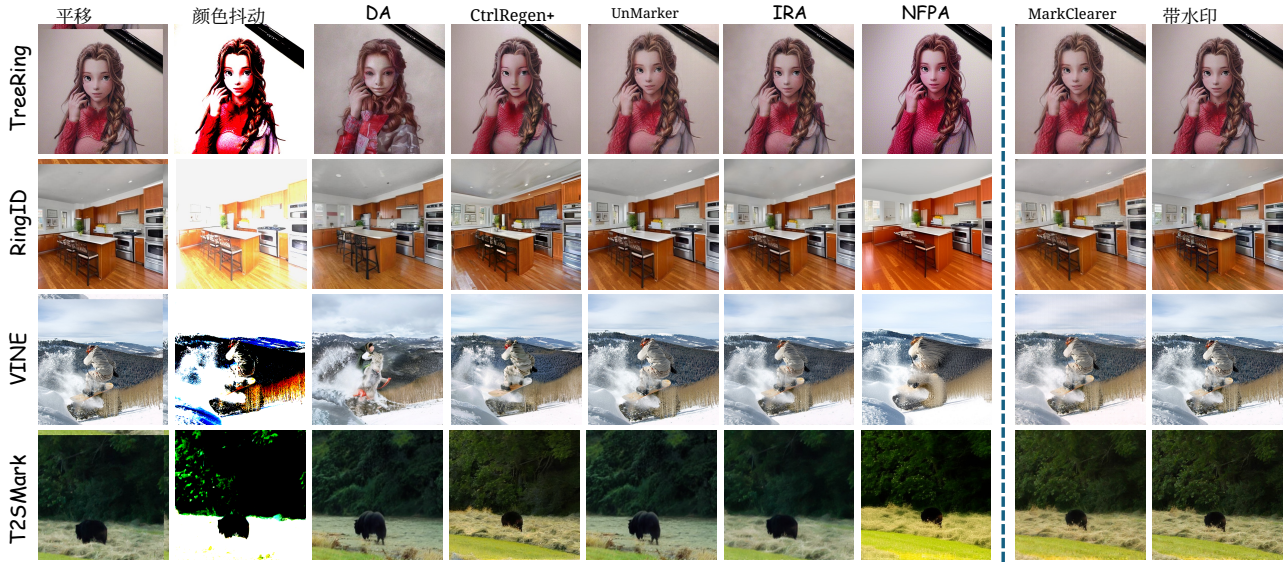


图*4*。 **水印去除方法的定性比较**。 传统的像素空间失真会严重降低视觉质量，而基于生成的方法往往以语义漂移为代价去除水印。我们的 MarkCleaner 能更有效地抑制水印，同时保持更好的视觉保真度。更多结果见附录 G.2。

类别中均表现出卓越且稳定的擦除性能。值得注意的是，我们的方法能够有效抑制语义水印和非语义方案的检测，验证了其对多种水印范式的普适性。我们将这种广泛的泛化能力归因于 2DGS 渲染的位移机制。传统检测器依赖严格的像素网格对齐来提取信号。然而，2DGS 中的连续光栅化会引入局部亚像素偏移。这些微小的几何扰动有效打破了水印验证所需的空域同步，使图案无法读取，且无需针对特定结构。相比之下，基线方法存在显著局限。基于失真的攻击可以成功攻击脆弱的水印，但

在鲁棒语义方案上失效，且往往仍保留近乎完美的检测准确率。基于再生成的方法提高了鲁棒性，但表现仍不稳定，检测率与随机水平有显著偏差。基于优化的方法展现出更强的泛化能力，但仍存在明显的残余可检测性。MarkCleaner 在所有水印类型上近乎均匀地抑制真正率和准确率，表明其有效地将伪影与语义解耦，实现了可靠的通用擦除

移除保真度。图4 展示了不同水印移除策略的定性比较。传统的图像空间失真可以使语义水印失效，但需要施加强烈的扰动，从而严重损害

表 2. 在单块 L40 GPU 上进行推理。TreeRing 和 HSQR 表示各自水印方案的真正率。时间和帧每秒用于衡量推理速度。粗体表示最佳性能。

指标	IRA	UnMarker	CtrlRegen+	NFPA	Ours
TreeRing ↓	0.800	0.010	0.001	0.003	**0.001**
HSQR ↓	1.0	0.033	1.0	0.367	**0.022**
时间（秒）↓	72.155	6.527	2.732	3.887	**0.179**
帧每秒 ↑	0.01	0.15	0.37	0.26	**5.58**

视觉质量。基于生成的方法可以使水印检测失效，但会引入明显的语义漂移，例如改变面部属性（第 1 行），或使纹理变得模糊且过度平滑（第 3 行）。此外，这些方法还会引入明显的色偏和意外的几何偏移，反映出再生成过程中缺乏显式的空间控制。相比之下，MarkCleaner 可去除可见水印和不可见水印，同时保留精细的图像结构和语义内容。其输出在视觉上与原始图像难以区分，且不会引入色彩失真或可感知的几何伪影。更多结果见附录G.2。

推理效率。 表2 比较了在单张 NVIDIA L40 GPU 上评估的推理效率和攻击效果。MarkCleaner 实现了高达 5.58 帧每秒的推理速度（每张图像耗时 0.179 秒），比基于优化的方法（例如 400×比 IRA 快， 36× 比 UnMarker 快）快了几个数量级。值得注意的是，效率的提升并未降低攻击性能。快速基线方法往往无法对抗最先进的语义水印，而我们的方法始终能实现接近于零的检测率。我们的方法打破了速度与攻击强度之间的权衡，为击败最先进的语义水印方案提供了实时解决方案。

5.3. 消融研究

为分析 MarkCleaner 中各组件的作用，我们通过移除关键模块进行了全面的消融研究： **掩码引导编码器**（ME）、**高斯渲染**（GR）（替换为 UNet 解码器）、**内容对齐**（CA），以及 **几何攻击监督**（GA）。结果汇总于图5和表 3。

结果表明，GA 是实现语义水印去除的决定性因素。移除 GA 虽然能保持较高的视觉保真度，但几乎完全无法去除语义水-

水印，证实几何偏差对于打破语义水印所依赖的相位对齐至关重要。移除 ME 会导致水印移除性能和视觉质量均显著下降，在可见水印去除上的结果尤其较差。这表明显式空间掩码对于准确定位并抑制可见水印区域至关重要。将 GR 替换为 UNet 解码器会导致降低的几何

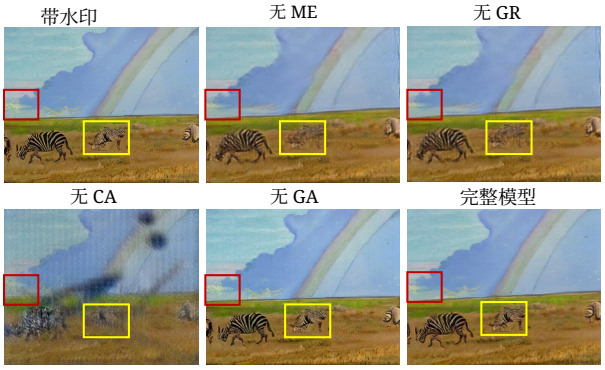


图 5。 **消融研究可视化**。 红色方框突出显示细微几何偏移，黄色方框突出显示结构保留。完整模型通过引入肉眼难以察觉的微小偏移来擦除水印，同时保持视觉连贯性。

表 3。 MarkCleaner 关键组件的消融研究

模块	真正率 ↓	FID ↓	CLIP ↑	LPIPS ↓
去掉 ME	0.2467	145.145	0.3154	0.6846
去掉 GR	0.1367	146.604	0.3163	0.6837
去掉 CA	0.1567	149.416	0.2528	0.7064
去掉几何增强	1.0000	131.487	0.3324	0.6676
完整模型	**0.0001**	**133.196**	**0.3705**	**0.6295**

度量位移以及错位监督下的纹理模糊，反映出 GR 更适合建模细粒度、不可感知的几何扰动，同时保留局部细节。移除 CA 会导致语义一致性严重下降，即便仍然施加了几何扰动。这证实语义引导对于在几何扰动过程中维持高层内容结构不可或缺。总体而言，完整的 MarkCleaner 框架实现了近乎完美的擦除（真正率 ≈ 0），与无扰动基线相比，FID 仅有可忽略的权衡，展现了擦除与保真度之间的最优平衡。

6. 结论

我们发现，当前的语义水印因其依赖相干相位对齐，本质上对几何扰动敏感，从而使不可感知的空间位移能够在不降低图像内容质量的情况下使验证失效。MarkCleaner 并非进行激进的内容再生成，而是将用于模式级修复的掩码建模与通过 2DGS 渲染实现的随机微几何位移相结合，从而实现水印去除。大量实验表明，MarkCleaner 在水印去除效果、视觉保真度和推理效率方面均优于最先进的生成式去除方法。这一视角表明，在未来的水印设计中，应将几何鲁棒性作为首要考量。

影响声明

本文研究数字水印嵌入方法的鲁棒性。我们揭示了水印在微几何扰动下的一个关键脆弱性。我们承认，所提出的方法 MarkCleaner 可能被滥用于未经授权的水印去除，从而导致版权侵权，或规避人工智能生成内容检测。然而，我们认为，识别并公开这些脆弱性是迈向更鲁棒、更可靠水印的必要一步。通过揭示当前空间对齐机制的脆弱性，本工作推动了能够抵御几何畸变的水印嵌入方法的发展。我们敦促社区在未来的图像水印系统中，将几何鲁棒性视为一项重要约束。

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等。Gpt-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。Alam, I., Islam, M. T., 和 Woo, S. S. 面向有效隐形水印去除的显著性感知扩散重建。发表于 *ACMWeb Conf.Companion*，2025。Ball´e, J., Minnen, D., Singh, S., Hwang, S. J., 和 Johnston, N. 采用尺度超先验的变分图像压缩。发表于国际学习表征会议，2018。Bhat, G., Danelljan, M., Van Gool, L., 和 Timofte, R. 深度突发超分辨率。发表于 *IEEE*计算机视觉与模式识别会议，第 9209–9218 页，2021年6月。B´egaint, J., Racap´e, F., Feltman, S., 和 Pushparaja, A. Compressai：用于端到端压缩研究的 pytorch 库和评估平台。arXiv 预印本 arXiv:2011.03029，2020。Chen, D., Chen, L., Zhang, Z., 和 Zhang, L. 面向任意尺度超分辨率的通用高效二维高斯泼溅。发表于国际计算机视觉会议，2025年6月。Chen, Y., Liu, S., 和 Wang, X. 使用局部隐式图像函数学习连续图像表示。发表于*IEEE*计算机视觉与模式识别会议，第 8628–8638 页，2021年6月。Cheng, Z., Sun, H., Takeuchi, M., 和 Katto, J. 采用离散化高斯混合似然和注意力模块的学习式图像压缩。发表于 *IEEE*计算机视觉与模式识别会议，第 7939–7948 页，2020。Ci, H., Yang, P., Song, Y., 和 Shou, M. Z. Ringid：重新思考树轮水印嵌入，以增强多密钥识别。在欧洲计算机视觉会议，第338–354页，2025年。Cox, I.、Miller, M.、Bloom, J.、Fridrich, J.和Kalker, T. 数字水印与隐写术。Morgan Kaufmann，伯灵顿，马萨诸塞州，2007年。Dabov, K.、Foi, A.、Katkovnik, V.和Egiazarian, K. 稀疏三维变换域协同滤波进行图像去噪。*IEEE*图像处理汇刊，16(8):2080–2095，2007年。Ding, M.、Rabbani, T.、An, B.、Agrawal, A.、Xu, Y.、Deng, C.、Zhu, S.、Mohamed, A.、Wen, Y.、Goldstein, T.等。Waves：图像水印鲁棒性的基准测试。载于*ICLR*负责任基础模型研讨会，2024年。Fernandez, P.、Sablayrolles, A.、Furon, T.、J´egou, H.和Douze, M. 在自监督潜在空间中嵌入图像水印。载于*ICASSP*，第3054–3058页，2022年5月。Fernandez, P.、Couairon, G.、J´egou, H.、Douze, M.和Furon, T. 稳定签名：将水印植入潜在扩散模型。载于国际计算机视觉会议，第22466–22477页，2023年10月。Gowal, S.、Bunel, R.、Stimberg, F.、Stutz, D.、Ortiz-Jimenez, G.、Kouridi, C.、Vecerik, M.、Hayes, J.、Rebuffi, S.-A.、Bernard, P.等。Synthid-image：互联网规模的图像水印嵌入。*arXiv*预印本 *arXiv:2510.09263*，2025年。Hessel, J.、Holtzman, A.、Forbes, M.、Le Bras, R.和Choi, Y. CLIPScore：一种无需参考的图像描述评估指标。载于自然语言处理实证方法，第7514–7528页，2021年11月。Heusel, M.、Ramsauer, H.、Unterthiner, T.、Nessler, B.和Hochreiter, S. 通过双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡。载于神经信息处理系统进展，第30卷，2017年。Hu, J.、Xia, B.、Chen, B.、Yang, W.和Zhang, L. Gaussiansr：用于任意缩放图像超分辨率的高保真二维高斯泼溅。载于*AAAI*，第4期，第3554–3562页，2025年。Karras, T.、Laine, S.和Aila, T. 面向生成对抗网络的基于风格的生成器架构。载于*IEEE*计算机视觉与模式识别会议，2019年6月。Kassis, A.和Hengartner, U. Unmarker：针对防御性图像水印的通用攻击。载于*IEEE*安全与隐私研讨会，第2602–2620页，2025年5月。Kerbl, B.、Kopanas, G.、Leimk¨uhler, T.和Drettakis, G. 用于实时辐射场渲染的3D高斯泼溅。*ACM*图形学汇刊，42(4):1–14，2023年7月。

Labs, B. F.、Batifol, S.、Blattmann, A.、Boesel, F.、Con-sul, S.、Diagne, C.、Dockhorn, T.、English, J.、English, Z.、Esser, P.等。Flux. 1 kontext：潜在空间中的流匹配，用于上下文内图像生成与编辑。*arXiv*预印本 *arXiv:2506.15742*，2025年。

Lee, B.、Lee, H.、Sun, X.、Ali, U.与Park, E. 3D高斯泼溅去模糊。载于欧洲计算机视觉会议，2024年6月。

Lee, S. J.与Cho, N. I. 语义水印重塑：借助傅里叶完整性提升鲁棒性与生成质量。载于国际计算机视觉会议，第18759–18769页，2025。

Li, H.、Chen, C.、Li, X.与Lu, G. 面向图像修复的语义对齐二维高斯泼溅。arXiv预印本arXiv:2509.01964，2025。

Lin, C.-Y.、Wu, M.、Bloom, J.、Cox, I.、Miller, M.与Lui, Y. 面向图像的抗旋转、缩放和平移水印。*IEEE*图像处理汇刊，10(5):767–782，2001。数字对象标识符：10.1109/83.918569。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P. 和 Zitnick, C. L. 微软COCO：上下文中的常见物体。载于 欧洲计算机视觉会议，第740–755页，2014年。

Liu, G., Cao, S., Qian, Z., Zhang, X., Li, S. 和 Peng, W. 一种适用于所有场景的水印：一种抵御部分图像盗用的鲁棒水印方案。载于*IEEE*计算机视觉与模式识别会议，第8225–8234页，2025年6月a。

Liu, Y., Song, Y., Ci, H., Zhang, Y., Wang, H., Shou, M. Z. 和 Bu, Y. 图像水印可通过利用干净噪声进行可控再生成来移除。载于 国际学习表征会议，2025b。

Lu, S.、Zhou, Z.、Lu, J.、Zhu, Y. 和 Kong, A. W.-K. 利用生成先验实现抵御图像编辑的鲁棒水印嵌入：从基准测试到发展。在 国际学习表征会议，2025年。

Lukas, N.、Diaa, A.、Fenaux, L. 和 Kerschbaum, F. 利用优化实现针对图像水印的自适应攻击。在国际学习表征会议，2024年。

M¨uller, A.、Lukovnikov, D.、Thietke, J.、Fischer, A. 和 Quiring, E. 针对扩散模型语义水印的黑盒伪造攻击。在 *IEEE* 计算机视觉与模式识别会议，第20937–20946页，2025年6月。

Navas, K. A., Ajay, M. C., Lekshmi, M., Archana, T. S., 和 Sasikumar, M. 基于 DWT-DCT-SVD 的水印嵌入。载于国际通信系统与软件中间件会议论文集，第 271–274 页，2008年。

Oppenheim, A. 和 Lim, J. 相位在信号中的重要性。*IEEE*会刊，69(5):529–541，1981年。数字对象标识符：10.1109/PROC.1981.12022。Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA,D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rab- bat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., 和 Bojanowski, P. DINOv2：无需监督地学习鲁棒视觉特征。机器学习研究汇刊，2024年。国际标准连续出版物号 2835-8856。特色认证。

Peng, L.、Wu, A.、Li, W.、Xia, P.、Dai, X.、Zhang, X.、Di, X.、Sun, H.、Pei, R.、Wang, Y.、Cao, Y. 和 Zha, Z.-J. 从像素到高斯：利用二维高斯建模实现超快速连续超分辨率。arXiv预印本 arXiv:2503.06617，2025。

Qiu, H.、Wang, Z.、Zhang, M.、Zhang, X.、You, X. 和 Yang, M.《未来无标记：通过下一帧预测去除人工智能生成图像中的水印》。载于神经信息处理系统进展，2025。

Ren, K.、Yang, Z.、Lu, L.、Liu, J.、Li, Y.、Wan, J.、Zhao, X.、Feng, X. 和 Shao, S.《综述：生成式人工智能时代 AIGC水印的作用与未来》。arXiv 预印本 arXiv:2411.11478，2024年。

Rombach, R.、Blattmann, A.、Lorenz, D.、Esser, P. 和 Ommer, B.《利用潜在扩散模型进行高分辨率图像合成》。载于 *IEEE*计算机视觉与模式识别会议，第 10684–10695 页，2022年6月。

Saberi, M.、Sadasivan, V. S.、Rezaei, K.、Kumar, A.、Chegini, A.、Wang, W. 和 Feizi, S.《AI 图像检测器的鲁棒性：基本极限与实际攻击》。载于国际学习表征会议，2024。

Santana, G. Gustavosta：稳定扩散提示词。https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts，2022年。数据集。

Sharma, S.、Zou, J. J.、Fang, G.、Shukla, P. 和 Cai, W. 图像水印嵌入在身份保护与验证中的综述。《多媒体工具与应用》，83(11)：31829–31891，2024年。

Song, J.、Meng, C. 和 Ermon, S. 去噪扩散隐式模型。载于国际学习表征会议，2021年。

Tancik, M., Mildenhall, B., and Ng, R. StegaStamp：实体照片中的不可见超链接。发表于*IEEE*计算机视觉与模式识别会议，2020年6月。

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. DiffusionDB：面向文生图生成模型的大规模提示词图库数据集。发表于计算语言学协会年会，第893–911页，2023年。Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-rings水印：扩散图像的不可见指纹。发表于 神经信息处理系统进展，第36卷，第58047–58063页，2023年。Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., S han, Y., Qie, X., and Shou, M. Z. Tune-A-Video：通过单次调优图像扩散模型实现文生视频生成。发表于*IEEE*计算机视觉与模式识别会议，第7623–7633页，2023年。Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y. S., and Loy, C. C. 用于自监督视觉预训练的掩码频率建模。发表于 国际学习表征会议，2023年。Xu, Q., Z hang, R., Zhang, Y., Wang, Y., and Tian, Q. 一种基于傅里叶的域泛化框架。发表于 *IEEE*计算机视觉与模式识别会议，第14383–14392页，2021年。Yang, J., Fang, H., Zhang, W., Yu, N., and Chen, K. T2SMark：在扩散模型的噪声水印中平衡鲁棒性与多样性。发表于神经信息处理系统进展，2025年。Yang, Y. and Soatto, S. FDA：用于语义分割的傅里叶域自适应。发表于*IEEE*计算机视觉与模式识别会议，第4085–4095页，2020年。Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., and Yu, N. 高斯着色：扩散模型的可证明无性能损失图像水印。发表于*I EEE*计算机视觉与模式识别会议，第12162–12171页，2024年6月。Zhang, K. A., Xu, L., Cuesta-Infante, A., and Veeramach- aneni, K. 通过注意力机制实现鲁棒的不可见视频水印。arXiv预印本 arXiv:1909.01285，2019年。Zhang, L., Rao, A., and Agrawala, M. 为文生图扩散模型添加条件控制。发表于 国际计算机视觉会议，2023年。Zhang, R. 让卷积网络重新具备平移不变性。发表于 国际机器学习会议，第97卷，第7324–7334页。PMLR，2019年6月9日至15日。Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. 深层特征作为感知度量的惊人效力。发表于*IEEE*计算机视觉与模式识别会议，2018年。Zhang, X., Ge, X., Xu, T., He, D., Wang, Y., Qin, H., Lu, G., Geng, J., and Zhang, J. Gaussianimage：通过二维高斯泼溅实现 1000帧/秒的图像表示与压缩。载于欧洲计算机视觉会议，2024年6月。Zhao, W., Bai, L., Rao, Y., Zhou, J., and Lu, J. Unipc：一种用于扩散模型快速采样的统一预测-校正框架。神经信息处理系统进展，2023。Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., Vigna, G., Wang, Y.-X., and Li, L. 隐形图像水印可被生成式AI 确证移除。神经信息处理系统进展，第37卷：8643–8672页，2024年。Zhou, Y., Xiao, Z., Yang, S., and Pan, X. 无混叠潜扩散模型：提升扩散潜空间的分数平移等变性。载于*IEEE*计算机视觉与模式识别会议，第34–44页，2025年6月。

A. 附录概述。

附录提供了全面的理论分析、架构细节和扩展实验结果，以支持正文。我们首先定义威胁模型和攻击方法的适用范围。随后，我们正式推导微几何扰动如何破坏潜在水印相位相干性，并详细阐述二维高斯泼溅解码器和内容对齐模块的实现细节。我们针对多种鲁棒水印方案进行了广泛的定性和定量比较。我们还展示了如何将本方法扩展至借助用户提供的掩码进行交互式可见水印去除。最后，我们批判性地分析了 MarkCleaner 当前的局限性，并探讨未来研究的有前景方向。

B. 威胁模型

我们在一个现实的威胁模型中，对水印去除问题进行建模，其中涉及两个实体：运营受水印保护的图像服务的提供方，以及试图在保留语义和视觉效用的同时去除水印的攻击者。

提供方的目标与能力：提供方部署生成模型 $\mathcal{G}$ 及水印机制 $\mathcal{W}$，以确立内容来源并防止生成内容被滥用。对于给定的生成图像 I，水印可以在生成后嵌入，也可以在生成过程中嵌入。提供方同时控制生成模型、水印编码器 $\mathcal{E}$和水印解码器 $\mathcal{D}$，其中 $\mathcal{D}(x_w) \rightarrow \{0,1\}$ 表示水印是否存在。检测只需一张图像，并假设提供方拥有足够的计算资源来训练稳健的水印系统。

攻击者的目标与能力：攻击者旨在生成 $I' = \mathcal{T}(I)$，以去除水印（$\mathcal{D}(I') \rightarrow 0$），同时保留语义和感知内容。水印检测置信度降至随机水平以下（准确率 $= 0.5$或 TPR@1%FPR $= 0$）。攻击者面临严格的约束：

- **访问权限：** 不了解模型参数、水印方案或检测器内部机制；无法访问检测 API 进行迭代查询。
- **数据：** 没有成对的干净与带水印样本；没有专有数据集；只能访问一张可疑图像。
- **计算资源：** 资源有限，无法进行多 GPU 训练；只能在消费级硬件上进行轻量级推理。
- **时间：** 每张图像的处理必须在几秒内完成，不允许进行耗时数分钟的优化。

攻击者不了解水印嵌入算法、检测机制或生成模型参数。这反映了现实中的黑盒场景：水印去除必须仅凭带水印图像本身成功，无法访问水印系统或模型内部机制。

C. 语义水印的几何脆弱性分析

本节将正式证明，像素域平移如何经由 VAE 编码器传播并引发潜域相位斜坡，从而破坏嵌入在潜变量傅里叶频谱实部中的水印。

C.1. 实部潜变量水印的形式化描述

考虑一个潜扩散模型，其中 VAE 编码器 $\mathcal{E}$ 将图像 $I \in \mathbb{R}^{W\times H\times 3}$ 映射为潜变量 $z \in \mathbb{R}^{w\times h\times c}$，下采样因子为 $d = 8$。TreeRing 水印（Wen 等人, 2023）会修改初始噪声 z_T，该噪声会传播至生成的潜变量 z_0。令 $Z_{u,v} = \mathcal{F}(z)[u,v] = A_{u,v}e^{i\phi_{u,v}}$ 表示频率 (u,v) 处的傅里叶系数。水印通过约束频率掩码 M 内 $Z_{u,v}$ 的**实部**趋向目标值$\eta_{u,v} \in \mathbb{R}$来嵌入：

$$\mathrm{Re}(Z_{u,v}) = A_{u,v}\cos(\phi_{u,v}) = \eta_{u,v}, \quad \forall (u,v) \in M. \tag{5}$$

检测依赖于 L1 距离： $d = \frac{1}{|M|}\sum_M |\mathrm{Re}(\hat{Z}_{u,v}) - \eta_{u,v}| < \tau$，其中 $\hat{Z}$ 通过 DDIM反演恢复。

C.2. 通过像素平移和旋转进行相位斜坡诱导

我们严格推导了复合几何变换$\mathcal{T}_{\boldsymbol{\Delta},\,\theta}$（平移 $\boldsymbol{\Delta}$ 与旋转 θ）如何经由 VAE 反演流程传播，从而使潜在水印失效。

基于分数潜变量偏移的相位调制。 设 $\mathcal{T}_{\mathbf{\Delta}}$ 为平移量 $\mathbf{\Delta} = (\Delta_x, \Delta_y)$ 下的像素域平移。VAE 编码器 $\mathcal{E}$ 可建模为以核 ψ 进行的连续卷积，随后接步幅采样算子 $\mathcal{S}_d$（Zhou 等人，2025）。离散索引 $\mathbf{n}$ 处平移后图像 $I' = \mathcal{T}_{\mathbf{\Delta}}(I)$ 的潜在值为：

$$z'[\mathbf{n}] = \int_{\mathbb{R}^2} I(\mathbf{x} - \mathbf{\Delta})\psi(d\mathbf{n} - \mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^2} I(\mathbf{y})\psi(d(\mathbf{n} - \frac{\mathbf{\Delta}}{d}) - \mathbf{y})d\mathbf{y}. \tag{6}$$

若 $\mathbf{\Delta}$ 不是 d的整数倍，则 $\mathbf{\Delta}/d$ 会在潜变量网格上形成分数偏移。根据傅里叶移位定理，这种在 z 中的空间偏移会在潜变量频谱中引入频率相关的相位斜坡：

$$\hat{Z}'_{u,v} = \mathcal{F}(z')_{u,v} = Z_{u,v} \cdot \exp\left(-i2\pi\left(\frac{u\Delta_x}{dw} + \frac{v\Delta_y}{dh}\right)\right) + \epsilon_{\text{alias}}, \tag{7}$$

其中 $\Delta\phi_{u,\,v} = -2\pi(\frac{u\Delta_x}{dw} + \frac{v\Delta_y}{dh})$ 是诱导产生的相位偏移，而 ϵ_{alias} 是变分自编码器混叠噪声。

坐标漂移与核非等变性。尽管 Tree-ring 使用圆形掩码 M 来缓解旋转的影响，但其检测是**逐点**进行的，因此容易受到坐标漂移的影响。像素域旋转 θ 会导致潜变量频率坐标发生旋转：

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{8}$$

由于变分自编码器的核 ψ 固定在矩形网格上，因此会表现出**旋转非等变性**(Zhang，2019）。旋转坐标$(u',\ v')$ 处的编码特征并非原始特征的简单旋转版本，而是受到核的空间偏置的扭曲，该偏置记为 ϵ_{rot}。这种扭曲与平移产生的相位斜坡相结合，形成复合潜变量频谱：

$$\hat{Z}'_{u,v} = A_{u',v'}e^{i(\phi_{u',v'} + \Delta\phi_{u,v})} + \epsilon_{\text{alias}} + \epsilon_{\text{rot}} \tag{9}$$

C.3. 实部特征破坏机制

水印通常嵌入实部：$\text{Re}(Z_{u,v}) = A_{u,v}\cos(\phi_{u,v}) = \eta_{u,v}$。在复合相位扰动 $\Delta\Phi'$作用下，恢复的实部 $\text{Re}(\hat{Z}'_{u,v})$ 为：

$$\text{Re}(\hat{Z}'_{u,v}) = \underbrace{A_{u,v}\cos(\phi_{u,v})}_{\text{Signal } \eta_{u,v}}\cos(\Delta\Phi') - \underbrace{A_{u,v}\sin(\phi_{u,v})}_{\text{Imaginary Leakage}}\sin(\Delta\Phi') \tag{10}$$

这种**相位到实部泄漏**会导致两种失败模式：

1. **衰减**：原始信号 $\eta_{u,v}$ 会被 $\cos(\Delta\Phi')$ 抑制。当 $\Delta\Phi' \to \pi/2$ 时，信号消失。
2. **干扰**：高熵虚部 $\text{Im}(Z)$ 通过 $\sin(\Delta\Phi')$ 投影到实轴上，形成非相干噪声。

C.4. 统计失效与检测失败

在由平移 $\mathbf{\Delta}$ 和旋转 θ组成的复合几何变换 $\mathcal{T}_{\mathbf{\Delta},\theta}$ 下，检测得分 d' 因相位扰乱与坐标漂移的协同效应而偏离原始密钥 η：

$$d' = \frac{1}{|M|}\sum_{(u,v)\in M} |A_{u_\theta,v_\theta}\cos(\phi_{u_\theta,v_\theta} + \Delta\phi_{u,v}) - \eta_{u,v}|, \tag{11}$$

其中 $(u_\theta, v_\theta) = \mathbf{R}_{-\theta}(u, v)$ 表示旋转后的频率坐标。

这种失效由两种截然不同但相互耦合的机制驱动：

1. 通过平移产生相位斜坡： 正如傅里叶移位定理所确立的那样，像素域平移 $\mathbf{\Delta}$ 由于变分自编码器的步幅 $d = 8$，会在潜变量网格中诱导出分数偏移。这表现为频率相关的相位斜坡$\Delta\phi_{u,\,v}$。在半径为 $r \in [r_{\min}, r_{\max}]$ 的环形掩码 M 内，相位变化范围为：

$$\Delta\phi_{\text{range}} \approx 4\pi \cdot \frac{r_{\max} \cdot \|\mathbf{\Delta}\|_2}{d \cdot w}. \tag{12}$$

2. 通过旋转产生坐标漂移： 尽管掩码 M 具有径向对称性，树环检测却是**逐点进行的**。旋转会将原本位于（u, v）处的信号重新映射到新坐标（u', v'）。对于半径为 r 的频率索引，离散傅里叶网格中的空域漂移 s 为：

$$s \approx r \cdot \theta \quad (\text{in grid units}). \tag{13}$$

在复平面中，这些机制表现为**复合相位-坐标坍缩**。平移会使水印向量偏离实轴密钥 η，而旋转会使信号漂移至错误的频率箱，实际上以不相关的背景系数取代水印系数。

统计收敛： 随着变换幅度增大，复合相位 $\Phi = \phi_{u',\,v'} + \Delta\phi_{u,\,v}$ 在 $[0, 2\pi)$ 上的分布会变得越来越均匀。因此，恢复后的实部期望值呈现类 sinc 衰减：

$$\left|\mathbb{E}[\mathrm{Re}(\hat{Z}'_{u,v})]\right| = \left|\eta_{u,v} \cdot \frac{\sin\alpha}{\alpha}\right| \cdot \mathcal{C}(\theta), \tag{14}$$

其中，$\alpha = \Delta\phi_{\text{range}}/2$ 和 $\mathcal{C}(\theta) \in [0,1]$ 是表示坐标对齐的相干因子。对于 Stable Diffusion（$d = 8$, $w = 64$, $r_{\max} = 16$），我们确定了实现可靠检测失败（$P > 0.95$）所需的以下临界阈值：

- **平移阈值：** $\|\boldsymbol{\Delta}\|_2 \geq 8$ 像素。达到此幅度时，$\Delta\phi_{\text{范围}} \geq \pi$，从而导致 $\alpha \geq \pi/2$ 并造成显著的信号衰减（$\sin(\alpha)/\alpha \leq 0.637$）。
- **旋转阈值：** $\theta \geq 3.6°$。在此角度下，漂移 $s \approx 16 \cdot (3.6 \cdot \pi/180) \approx 1.0$，导致信号跳至相邻的离散频率箱，从而与逐点密钥 $\eta_{u,\,v}$完全不匹配。
- **复合效应：** 将**8 个像素**与 $5°$ 组合变换，可确保 $\mathbb{E}[\mathrm{Re}(\hat{Z}')] \to 0$，从而迫使 d' 趋向于潜变量频谱 $A \approx 0.886$ 的均值幅度（对于标准高斯噪声 z_T）。

由于嵌入阈值满足 $\tau \ll A$，失去同步的向量会发生相消干涉，从而导致$d' > \tau$。虽然某些水印变体（例如恒定环）声称具有部分旋转不变性，但我们的分析强调，坐标漂移与 VAE引起的采样相位不一致相互耦合，确保了即使是各向同性图案也会遭遇相干性坍缩。此外，我们观察到，相位域干涉发生时所需的扰动尺度远小于引发可感知幅度损失所需的尺度，这证实锁相同步才是潜空间水印嵌入的关键。

证明结论。 我们已经证明，复合几何变换 $\mathcal{T}_{\boldsymbol{\Delta},\,\theta}$ 利用了变分自编码器的采样相位不一致。平移通过线性斜坡扰乱相位相干性，而旋转则引入坐标错位和非线性结构噪声。二者共同触发了检测所需的结构化锁相标识的崩溃，使水印在统计上与随机噪声无法区分，同时保留基于幅度的视觉内容。

□

D. 高斯渲染的更多细节

D.1. 3D高斯泼溅

3D高斯泼溅（3DGS）（Kerbl等人，2023）已成为一种强大的实时新视角合成技术。它使用显式、可学习的高斯基元表示三维场景，而不是使用隐式神经表示。3DGS通过将大量三维高斯基元与可微光栅化相结合来建模场景，从而实现实时渲染和细粒度的可编辑性。

在3DGS中，每个高斯基元由以下参数化：

- **位置** $\mu \in \mathbb{R}^3$：高斯的三维中心，
- **协方差** $\Sigma \in \mathbb{R}^{3\times3}$：描述形状和方向的三维协方差矩阵，
- **不透明度** $\alpha \in \mathbb{R}$：控制透明度，
- **颜色**：通过球谐函数（SH）系数表示视点相关外观。

三维　协方差矩阵 Σ 通常分解为旋转矩阵 R 和缩放矩阵 S:

$$\Sigma = RSS^\top R^\top \tag{15}$$

该参数化在确保半正定性的同时，提供了对高斯形状的直观控制。

在渲染时，三维高斯被投影到二维图像平面。给定视图变换 W 以及投影变换的仿射近似的雅可比矩阵 J，投影后的二维协方差为:

$$\Sigma' = JW\Sigma W^\top J^\top \tag{16}$$

最终像素颜色通过对按深度排序的重叠高斯进行阿尔法混合来计算:

$$C = \sum_{i\in\mathcal{N}} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j) \tag{17}$$

w其中 c_i 为颜色， α_i 为不透明度乘以该高斯在该像素处的贡献　　.

D.2. 二维高斯泼溅

在 3DGS 的基础上，二维高斯泼溅将类似的原理应用于二维图像处理。图像由大量二维高斯分布表示，每个分布都由均值 $\mu_i \in \mathbb{R}^2$、协方差矩阵 $\Sigma_i \in \mathbb{R}^{2\times2}$、不透明度 $\alpha_i \in \mathbb{R}$ 和颜色 $c_i \in \mathbb{R}^3$ 表征。

为确保半正定性，协方差被参数化为 $\Sigma_i = L_iL_i^\top$，其中 $L_i \in \mathbb{R}^{2\times2}$ 是一个下三角矩阵:

$$L_i = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \tag{18}$$

这将协方差表征所需的参数从 4 个减少到 3 个。我们将完整的参数集记为高斯空间 $\Theta = \{\mu, L, c, \alpha\}$。

位置 $\mathbf{p} = [x, y]^\top$ 处的像素值计算如下:

$$I_{\mathbf{p}} = \sum_i c_i\alpha_i \exp\left(-\frac{1}{2}(\mathbf{p}-\mu_i)^\top\Sigma_i^{-1}(\mathbf{p}-\mu_i)\right) \tag{19}$$

该公式利用了高斯混合模型良好的数学性质，即可以通过多个高斯核构建复杂分布。

D.3. 高斯建模的更多细节

给定掩码输入图像 $I_{\text{mask}} \in \mathbb{R}^{3\times H\times W}$，我们使用 UNet 编码器提取高斯特征。编码器输出稠密特征图 $F_g \in \mathbb{R}^{C'\times H\times W}$，再通过步幅卷积下采样为高斯级特征$F_g' \in \mathbb{R}^{N\times C'}$，其中 N 表示高斯核的数量。

编码器的层次结构及其跳跃连接能够捕获丰富的上下文，同时提高训练稳定性。高斯均值 μ在训练过程中经验证较为敏感，因此初始化为均匀分布于二维平面上的 μ_{fix}，且仅学习偏移量 μ_{bias}。

具体而言，高斯参数通过轻量级多层感知机从 F_g' 解码得到:

$$\mu_{bias} = E_\mu(F_g'), \quad c = E_c(F_g'), \quad \sigma = E_\sigma(F_g'), \quad \alpha = E_\alpha(F_g') \tag{20}$$

其中， $\mu_{bias} \in \mathbb{R}^{N\times2}$ 表示可学习的位置偏移， $c \in \mathbb{R}^{N\times3}$ 表示颜色， $\sigma \in \mathbb{R}^{N\times3}$ 表示协方差参数， $\alpha \in \mathbb{R}^{N\times1}$ 是不透明度。

为了约束 μ_{bias} 的空域范围，我们应用双曲正切激活，将取值限制在 $[-1, 1]$。最终位置通过以下方式获得:

$\mu = \mu_{bias} + \mu_{fix}$。

随后，这些参数会被栅格化到图像平面上，其软重叠特性可确保跨像素的颜色和强度平滑过渡。

D.4. 块级光栅化

高保真度图像表示需要大量高斯基元，导致图形处理器内存消耗显著增加。遵循 3DGS（Kerbl等人，2023）的基于瓦片的光栅化方法，我们采用块级光栅化来应对这一挑战。

将图像 $I \in \mathbb{R}^{H\times W}$ 划分为 $N_p = \frac{H}{p} \times \frac{W}{p}$ 个大小为 $p \times p$的非重叠块。每个块（i， j）都维护一组专用的 N_{patch} 高斯基元。参数空间变换为：

$$\mu \in \mathbb{R}^{N_p\times N_{patch}\times 2}, \quad c \in \mathbb{R}^{N_p\times N_{patch}\times 3}, \quad l \in \mathbb{R}^{N_p\times N_{patch}\times 3} \tag{21}$$

虽然这可能不会减少高斯基元的总数（$N_p \times N_{patch}$），但可以降低光栅化过程中的并发内存需求，并支持跨块并行渲染。

重叠混合。 为避免边界不连续，每个块都会在扩展后的（$p + 2a$） $\times$ （$p + 2a$）区域内进行渲染，该区域四周均包含 a 像素的边界。最终图像保留中心（$p - 2a$） $\times$（$p - 2a$）区域，并使用距离加权平均对重叠边界进行混合，从而确保平滑过渡。

E. 内容对齐详情

为确保重建过程中的全局语义一致性，我们引入预训练的 DINOv2 特征作为高层引导。这在采用补丁级处理时尤为重要，因为它有助于跨补丁边界保持连贯的语义。

E.1. DINOv2特征提取

DINOv2 是一种自监督视觉变换器，可在无标签条件下学习鲁棒视觉特征。其特征能够捕捉高层语义结构，并对外观变化表现出鲁棒性（Oquab 等人，2024年）。

我们从预训练的 DINOv2 视觉变换器模型中提取特征（训练期间保持冻结）。然而，当输入图像被重度掩码时，DINOv2 特征的可靠性会下降。受（Li 等人，2025）启发，从轻度掩码输入中提取的特征仍能保留丰富的语义信息，而掩码比例的增加则会导致表征逐渐退化。因此，我们采用轻量级多层感知机，将掩码特征转换为估计的干净特征：

$$f_{pred} = \text{MLP}(f_{masked}) \tag{22}$$

其中 f_{masked} 表示从掩码输入图像中提取的 DINOv2 特征， f_{pred} 为适配后的语义表征。

E.2. 自适应层归一化

我们使用自适应层归一化（AdaLN）（Karras 等人，2019）将语义特征注入解码器，如图 6所示。给定预测的语义特征向量 $f_{pred} \in \mathbb{R}^C$ 和隐藏特征图 $h \in \mathbb{R}^{C\times H\times W}$，AdaLN 操作定义如下：

$$\text{AdaLN}(h, f_{pred}) = B(\text{LN}(h) \cdot \alpha + \beta) \cdot \gamma \tag{23}$$

其中，LN(·) 表示层归一化， $B(\cdot)$ 表示主卷积块，而 α、 β、 $\gamma \in \mathbb{R}^C$ 分别是缩放、偏移和输出调制参数，由轻量级多层感知机根据 f_{pred} 预测。这种调制应用于多个解码器层，使语义信息能够在不同尺度上引导重建。

F. 损失函数详述

其中， I_{out} 表示由高斯解码器生成的渲染输出图像， I_{pert} 表示用于监督的几何扰动目标图像。训练目标由四个损失项加权组合而成：

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_P + \lambda_3\mathcal{L}_G + \lambda_4\mathcal{L}_D, \tag{24}$$

其中权重根据经验设置为 $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4 = 1 : 0.3 : 3 : 1$。

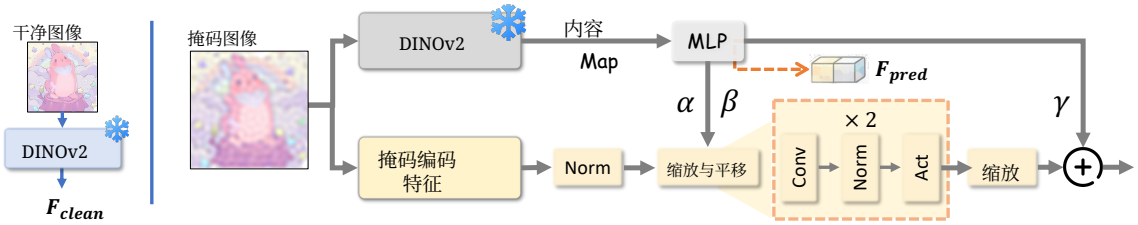


图 6. **内容对齐模块**。 DINO 特征经多层感知机处理，以预测自适应层归一化的缩放（α，β）和偏移（γ）参数，从而以语义上下文为解码器提供条件信息。

重建损失。 重建损失衡量渲染输出 I_{out} 与扰动目标 I_{pert} 在空域和频域中的差异:

$$\mathcal{L}_{rec} = \mathcal{L}_1 + \lambda_{freq}\mathcal{L}_{freq}, \tag{25}$$

其中， $\mathcal{L}_1 = \|I_{out} - I_{pert}\|_1$ 为空间 L1 损失，且 $\lambda_{freq} = 1$。

对于频率损失，我们定义一种同时考虑振幅与相位的频率距离度量。将每个 ˉ频率值 $\mathcal{F}(I)(u,v)$ 视为二维欧几里得向量 f，其模长对应振幅，而角度对应相位。我们将频率距离 $\mathcal{D}(\cdot,\cdot)$ 定义为（Xie等人，2023）：

$$\begin{aligned}\mathcal{D}(\vec{f}_{out}, \vec{f}_{pert}) &= \|\vec{f}_{out} - \vec{f}_{pert}\|_2^{\gamma} = |\mathcal{F}(I_{out})(u,v) - \mathcal{F}(I_{pert})(u,v)|^{\gamma} \\ &= \left((\mathcal{R}_{out}(u,v) - \mathcal{R}_{pert}(u,v))^2 + (\mathcal{I}_{out}(u,v) - \mathcal{I}_{pert}(u,v))^2\right)^{\gamma/2}.\end{aligned} \tag{26}$$

其中 $\mathcal{R}$ 和 $\mathcal{I}$ 分别是傅里叶变换的实部与虚部， γ 控制距离函数的锐度（默认设为1）。频率损失是所有频谱位置上的平均距离:

$$\mathcal{L}_{freq} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} |\mathcal{F}(I_{out})(u,v) - \mathcal{F}(I_{pert})(u,v)|^{\gamma}. \tag{27}$$

感知损失。 我们使用LPIPS（Zhang等人，2018）来刻画感知相似性:

$$\mathcal{L}_P = \text{LPIPS}(I_{out}, I_{pert}). \tag{28}$$

对抗损失。 我们采用带铰链损失的判别器 D :

$$\mathcal{L}_G^D = \mathbb{E}[\max(0, 1 - D(I_{pert}))] + \mathbb{E}[\max(0, 1 + D(I_{out}))] \tag{29}$$

$$\mathcal{L}_G^G = -\mathbb{E}[D(I_{out})]. \tag{30}$$

内容对齐损失。 为了将预测特征与真实语义表征显式对齐，我们使用基于余弦相似度的特征对齐损失:

$$\mathcal{L}_D = 1 - \frac{f_{pred} \cdot f_{clean}}{\|f_{pred}\|_2 \cdot \|f_{clean}\|_2}. \tag{31}$$

其中 f_{pred} 是由多层感知机从掩码输入预测得到的语义特征（如第E节所述），并且 f_{clean} 是从原始输入图像中提取的DINOv2 特征。该损失促使预测特征与干净特征对齐，从而增强全局语义一致性。

G. 更多实验结果

G.1. 图像水印去除设置

我们评估 MarkCleaner 针对被划分为三类的多种水印去除方法。

失真攻击 包括先前水印去除研究中普遍采用的标准信号扰动，具体包括 JPEG 压缩（质量因子 20）、裁剪和缩放（缩放因子 0.8）、高斯模糊（核大小 15）、高斯噪声（标准差 0.3）、颜色抖动（亮度因子 4）、旋转（15 度）以及平移（20 像素）。每种扰动均单独施加，以评估对基本图像变换的鲁棒性。

再生成攻击利用生成先验重建图像。具体而言，我们考虑以下方法：Diffusion-Attack（DA）（Saberi等人，2024年），采用去噪扩散隐式模型反演，推理步数为50，指导尺度为7.5；VAE-Attack (VA-B) 基于采用尺度超先验的变分图像压缩（Ball´e等人，2018），通过学习到的率失真瓶颈在质量等级4下重建图像；VAE-Attack (VA-C)（Cheng等人，2020）基于 Stable Diffusion 变分自编码器编码器-解码器流程，质量等级为3；CtrlRegen+ （Liu等人，2025b）采用干净噪声初始化，使用 ControlNet（Zhang等人，2023），去噪步数为50，条件尺度为1.0，指导尺度为2.0；通过潜在扩散过程进行图像再生成，并采用 UniPC多步调度器（Zhao等人，2023）；NFPA（Qiu等人，2025）的噪声扰动强度为0.1，净化步数为25，并采用推理步数为50的去噪扩散隐式模型调度器。

基于优化的攻击由 UnMarker（Kassis和Hengartner，2025）代表，它是一个适用于异构水印的通用框架，先进行中心裁剪预处理（裁剪比例为 0.9），随后执行低频约束优化阶段；该阶段采用自适应 Adam优化器，初始学习率设为 0.01，进行 10 次优化迭代，并使用基于VGG的感知损失来保持视觉保真度。为确保严格的黑盒评估设置，我们排除依赖替代模型的方法 IRA（M¨uller 等人，2025），该方法进行了 100 步优化。

G.2. 视觉恢复的补充结果

为了进一步分析水印可检测性与视觉质量之间的平衡，我们提供了大量定性比较，以展示 MarkCleaner 在不同水印方案下的视觉质量。

补充窗性结果9 和10 分别展示了采用 TreeRing（Wen 等人，2023）、RingID（Ci 等人，2025）、HSTR（Lee& Cho，2025）和 HSQR（Lee & Cho，2025）的语义水印的详细视觉比较。这四种方法代表了具有挑战性的语义水印方法，对常规攻击具有很强的鲁棒性。这些结果表明，MarkCleaner 在各种语义水印方案中始终能够实现更优的水印去除效果，同时保持感知保真度，且不会引入基线方法中常见的语义漂移、色彩失真或结构伪影。

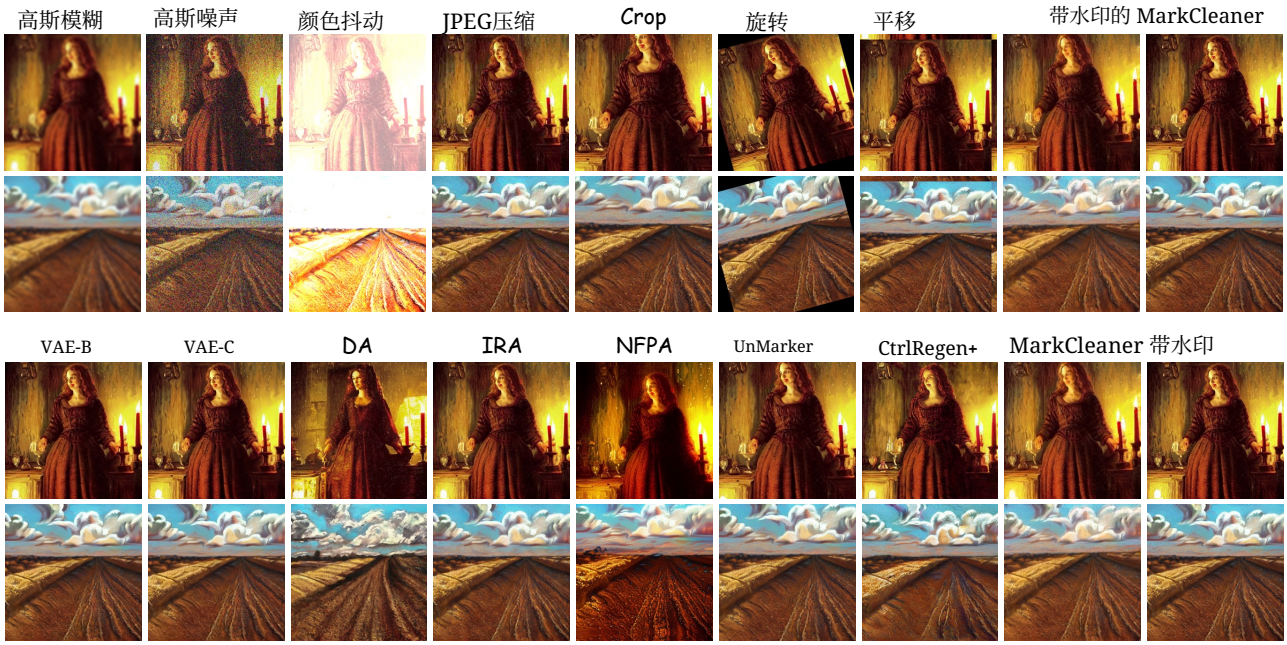


图 7. 在嵌入 Tree-Ring 水印的图像上，对水印去除方法进行更多定性比较（Wen 等，2023）。

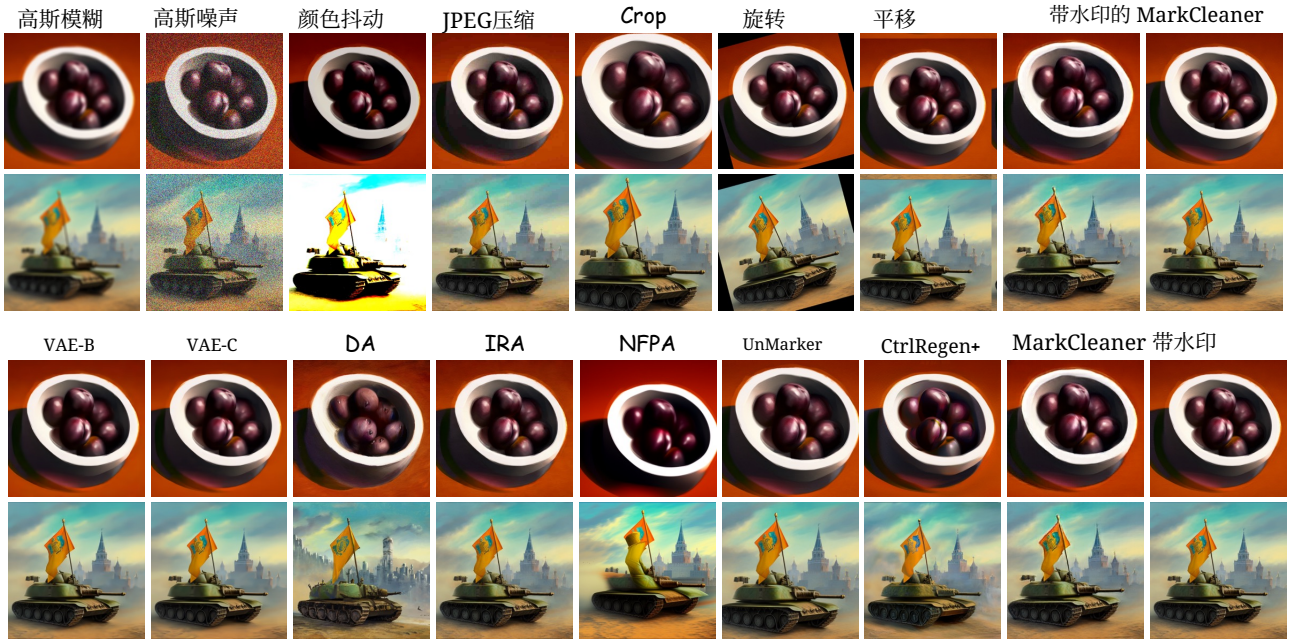


图 *8.* 更多定性 co mparisons：在嵌入 RingID 水印的图像上对水印去除方法的比较（Ci 等人，2025 ）.

像素度量与感知度量之间的差异。如表4所示，我们观察到 MarkCleaner 的评估结果存在显著差异。FID（133.6）和 LPIPS（0.627）等感知度量表明，其视觉保真度较高，与原始图像相当；而标准的逐像素度量（原始峰值信噪比：16.52，结构相似性指数：0.444）却低得出乎意料。这是我们的微几何扰动策略固有的特性。标准峰值信噪比假设空间对齐完美；然而，我们的方法会有意引入难以察觉的空间偏移，以破坏水印的相位相干性。仅 2–3 个像素的偏移就可能大幅拉低峰值信噪比，却不会降低实际视觉内容的质量。

为将几何位移与实际内容退化分离，我们计算**对齐峰值信噪比**和**对齐结构相似性指数**(Bhat 等人，2021年）。这些指标首先通过基于光流的对齐，将输出与输入进行配准，以补偿空间偏移，然后在对齐后的图像对上计算峰值信噪比/结构相似性指数。这样便可将内容保真度与几何错位区分开来。结果呈现出一个显著趋势：MarkCleaner 的峰值信噪比从 16.52（原始值）大幅提升至**25.30** （对齐后），结构相似性指数也从 0.444 跃升至**0.820**。这一显著恢复（$>$ 8dB 增益）证实，原始分数偏低主要是由空间错位造成，而非语义损坏。相比之下，DA 和 IRA 等生成式方法在对齐后的提升微乎其微，这表明它们保真度较低的根源是不可逆的内容幻觉，无法通过几何配准加以纠正。高对齐峰值信噪比与近乎完美的水印抑制（如表1所示）共同验证了我们的核心论点：MarkCleaner 能够有效隔离几何扰动，在保留语义完整性的同时干扰水印检测。

表 *4.* 不同水印去除方法的视觉质量指标定量比较

指标	带水印	VAE-B	VAE-C	DA	CtrlRegen+	UnMarker	NFPA	IRA	Ours
峰值信噪比 ↑	–	32.72	32.62	18.92	21.47	14.85	29.06	12.76	16.52
结构相似性指数 ↑	–	0.893	0.885	0.511	0.660	0.434	0.845	0.358	0.444
对齐峰值信噪比 ↑	–	30.54	30.47	18.39	28.52	28.96	21.43	19.17	25.30
对齐结构相似性 ↑	–	0.924	0.918	0.495	0.886	0.917	0.682	0.666	0.820
LPIPS ↓	0.620	0.614	0.615	0.644	0.618	0.614	0.609	0.641	0.627
FID ↓	131.5	142.5	145.7	133.8	134.0	130.9	142.2	135.04	133.6
CLIP ↑	0.380	0.386	0.385	0.356	0.382	0.386	0.361	0.3593	0.376



图 *9.* 更多定性　　嵌入 HSTR 水印图像上的水印去除方法的定性比较（Lee&Cho，2025）。

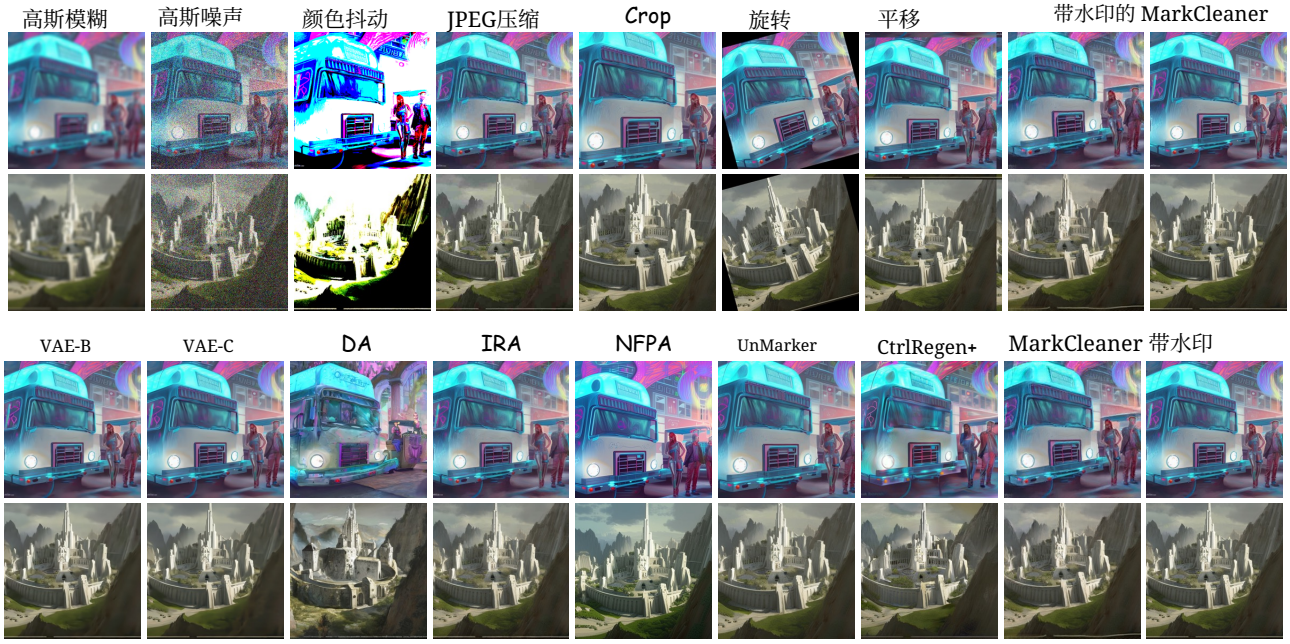


图 *10.* 嵌入 HSQR 水印图像上的水印去除方法的更多定性比较（Lee & Cho，2025）。

G.3. 潜在应用：去除可见水印

尽管 MarkCleaner 主要针对稳健的不可见水印而设计，其架构也天然支持可见水印去除。我们的框架采用掩码引导编码器，根据部分观测重建图像内容。从理论上讲，如果掩码专门针对被可见水印遮挡的区域，模型就应能够生成并恢复底层内容。

为验证这一能力，　　我们进行了一项探索性实验，手动掩蔽水印区域。　　t

推理时间。如图 11所示，MarkCleaner 能够有效去除可见水印，并在掩蔽区域重建合理的内容。此设置模拟了一种交互场景：用户显式选择要移除的可见水印区域。

然而，我们发现，在处理大型连续水印模式时，性能存在局限。由于 MarkCleaner 使用随机离散掩码策略（例如局部随机丢弃）进行训练，以学习鲁棒特征，因此它对图像修复的优化不足，而图像修复往往是去除可见水印所必需的，尤其是在处理大面积连续空洞时。训练阶段（离散掩码）与推理阶段（连续块状掩码）之间的领域差距，可能导致修复区域模糊或出现伪影。我们将此视为未来的研究方向：在训练过程中引入块级掩码策略，可进一步提升模型处理可见水印的性能。

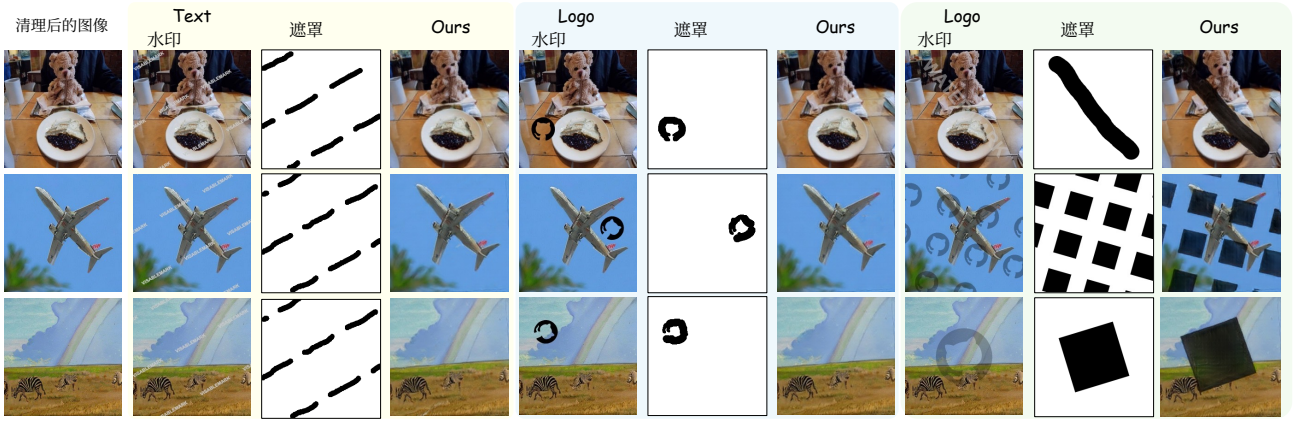


图*11*。 将 MarkCleaner 应用于交互式可见水印去除。

H. 讨论与未来工作

讨论。 虽然 MarkCleaner 能有效去除几乎不可见的水印，但我们发现它存在两项局限。首先，消融研究表明，几何增强（GA）对于有效擦除不可或缺；禁用 GA 虽能获得更高的重建保真度，却无法使水印检测失效。这种非此即彼的结果证实了我们的假设：去除水印的主要驱动力是相位位移，而非语义改变。值得注意的是，在不使用 GA 时观察到的高视觉质量，界定了这一范式的“保真度上限”。这表明，尽管当前实现已能有效去除水印，仍有很大空间进一步改进微扰动，以弥合这一保真度差距，并朝着相位扰乱完全不可感知的未来迈进。其次，在处理大型连续可见水印时，性能会下降。这源于训练与推理不匹配：我们的随机离散掩码策略针对分散扰动进行了优化，而非针对块级图像修复。未来的工作可通过在训练过程中引入连续掩码模式来解决这一问题。

未来工作。 这项工作的一个引人注目的拓展方向，是打通水印去除与鲁棒水印设计之间的闭环。未来研究应将可微几何算子融入对抗训练，以增强水印鲁棒性。通过对编码器进行正则化，使其抵御傅里叶移位定理所预测的相位扰乱，我们可以开发出几何不变的来源信息。这种方法旨在设计出即使在非等变步幅架构产生的采样伪影下，仍可检测到的签名。