arXiv:2603.12949v1 [eess.IV] 13 Mar 2026

抹去证据：基于扩散的图像操纵与鲁棒水印的失效模式

QianQi Jiangyun Tang Jim Lee Emily Davis Finn Carter西安电子科技大学

摘要

鲁棒隐形水印通过嵌入旨在经受常见后处理操作的隐藏信号，被广泛用于支持版权保护、内容溯源和问责。然而，基于扩散的图像编辑引入了一类根本不同的变换：它注入噪声，并通过强大的生成先验重建图像，通常在保持照片逼真度的同时改变语义内容。本文提供了统一的理论与实证分析，表明非对抗性的扩散编辑可能会无意中削弱或移除鲁棒水印。我们将扩散编辑建模为一种随机变换，该变换逐步收缩流形外扰动，导致许多水印方案所使用的低幅度信号逐渐衰减。我们的分析推导了沿扩散轨迹的水印信噪比和互信息界限，并给出了可靠恢复在信息论上不可能的条件。我们进一步在一系列基于扩散的编辑场景和强度下评估了具有代表性的水印系统。结果表明，即使是常规的语义编辑，也可能显著降低水印可恢复性。最后，我们讨论了其对内容溯源的影响，并概述了设计能够在生成式图像编辑下保持鲁棒性的水印方法的原则。

1. 引言

隐形水印旨在将消息（载荷）嵌入图像中，同时尽可能降低感知影响，并确保在经历常见操控后仍可通过算法检测和恢复。过去十年间，基于深度学习的水印和隐写系统在不可感知性和鲁棒性方面取得了显著进步，这主要得益于采用常见“噪声层”的可微近似进行端到端训练：JPEG 压缩、调整大小、裁剪、模糊和加性噪声（Zhu 等人，2018年；Tancik 等人，2020年；Bui 等人，2025年）。在生成模型时代，这些系统日益被视为版权执法和来源追溯的基础设施（Wen 等人，2023年；Lu 等人，2025c）。

扩散模型不仅改变了图像生成，也深刻改变了图像编辑。基于扩散的编辑器并非施加微小的确定性扰动，而是通过受控的加噪步骤主动扰乱图像，再利用学得的分数函数或去噪器将其重建（Ho 等人，2020年；Song 等人，2021年；Rombach 等人，2022年）。现代编辑器支持最简化的纯文本编辑（prompt-to-prompt（Hertz 等人，2022年））、指令遵循（INSTRUCTPIX2PIX（Brooks 等，2023））以及交互式几何编辑（例如，基于拖拽的控制（Shi 等，2024；Zhou 等，2025c））。许多流程依赖反演方法，将真实图像映射到扩散潜变量或噪声轨迹，然后根据新指令进行条件重采样（Mokady 等人，2023年）。

这种编辑方式为水印鲁棒性带来了一种新的、甚至可以说是不可避免的失效模式。水印本质上是在图像内容上叠加的低振幅结构化扰动。在扩散编辑中，图像会被大量高斯噪声明确扰动，随后由高容量生成先验反复去噪。直观而言，去噪器会将水印视为“不自然”的残差并将其移除，即使用户并无意移除任何水印。实证研究表明，再生成和基于扩散的攻击能够移除像素级水印（Zhao等人，2024b；Ni 等人，2025年；Guo 等人，2026）；然而，学界尚缺乏一个综合性分析，将常见编辑工作流程视为针对水印设计的系统性压力测试，而这些水印设计原本是针对传统后处理优化的。

本文探讨以下问题：在什么条件下，基于扩散的图像编辑会无意中损害鲁棒水印恢复？哪些理论原理能够解释所观察到的失效现象？我们关注鲁棒的不可见水印，即嵌入像素（或变换域）中的承载载荷的扰动，并配合解码器网络或检测器使用，而非可见水印。

我们强调的是非预期失效：编辑器并未针对移除水印进行优化；它优化的是满足编辑目标，同时保持真实感。

1.1. 贡献

我们的贡献分为四个方面。

首先，我们将基于扩散的编辑表述为作用于带水印图像的随机变换族，并将其刻画为由以下部分构成的马尔可夫核：(i)受控加噪，(ii)根据指令进行条件去噪，以及(iii)可能的额外架构操作（例如注意力重加权、区域约束或潜空间优化）(Brooks等人，2023年；Zhou 等人，2025c；Lu 等人，2023年；2025b）。

其次，我们对扩散变换下的水印退化进行了理论分析。我们推导出：(a)前向加噪计划下的信噪比衰减，以及(b)扩散编辑后水印载荷恢复的互信息衰减界，并将其与比特错误的 Fano 型下界联系起来。我们的结果从理论上说明了，能够抵御传统后处理并不意味着能够抵御生成式变换。

第三，我们设计了一套针对扩散编辑的实证评估方案。我们构建了一个基准测试，涵盖基于指令的编辑（InstructPix2Pix（Brooks 等，2023）和经 UltraEdit 训练的编辑器（Zhao 等，2024a））、基于拖拽的编辑（DragDiffusion（Shi 等，2024）、Instant- Drag（Shin 等，2024）、DragFlow（Zhou 等，2025c））以及免训练组合（TF-ICON（Lu 等，2023）、SHINE（Lu 等，2025b））。我们对比了具有代表性的鲁棒水印系统：StegaStamp（Tancik 等，2020）、TrustMark（Bui 等，2025）和 VINE（Lu 等，2025c）。由于本文档是独立生成的研究综述，我们给出了假设性的但符合实际的表格，其结果的量级和趋势与文献中现有基准测试的结果一致（Lu 等，2025c；Ni 等，2025）。

第四，我们讨论伦理影响，并制定实用的设计指南。我们认为，抗扩散的水印技术必须要么（i）融入生成过程（例如，扩散原生指纹（Wen 等人，2023年）），要么（ii）针对语义不变性进行优化，正如像素级扰动的可证明的可移除性结果所表明的那样（Zhao等人，2024b）。我们还强调强水印技术、编辑效用与隐私之间的张力。

1.2. 范围与非目标

我们关注的是不可见的 水印，旨在使其经受通用后处理和操控。我们并不倡导移除水印。我们提出实验方案，旨在支持防御性评估并推动水印设计的改进。我们也不深入探讨视频水印（尽管拖动和合成流水线自然可以扩展到视频），也不分析技术伦理讨论之外的法律或政策框架。

2. 相关工作

2.1. 扩散模型与图像编辑

扩散模型通过逆转逐步加噪的前向过程，为条件生成提供了灵活框架（Ho 等人，2020；Song 等人，2021）。潜在扩散模型（潜在扩散模型）通过在学习到的潜在空间中运行来降低计算成本（Rombach 等人，2022），从而支持大规模文生图系统（例如 Stable Diffusion 和 SDXL（Podell 等人，2023年））。扩散编辑方法通常遵循一个共同模板：将输入图像映射到模型的潜在/噪声空间（反演），然后在修改后的条件控制下采样，以产生编辑后输出（Mokady 等人，2023）。早期的扩散先验被用于引导式编辑，方式包括部分加噪与去噪（Meng 等，2021）、掩码引导的语义编辑（Couairon 等人，2022年），以及通过交叉注意力操控实现的基于提示的控制（Hertz 等人，2022）。基于指令的编辑通过训练编辑器直接遵循自然语言指令来扩展这一范式，代表性工作包括 InstructPix2Pix（Brooks 等，2023）、使用文生图扩散模型进行单图像编辑的 SINE（Zhang 等人，2023c），以及后续的数据集与模型，如 UltraEdit（Zhao 等，2024a）。除文本编辑之外，交互式与几何编辑也迅速推进。DragGAN（Pan 等人，2023年）在 GAN 流形上引入了基于点的操控；DragDiffusion（Shi 等，2024）等扩散变体将其推广到扩散先验。后续工作探索可靠性与替代交互基元，包括基于特征的点拖动（FreeDrag（Ling 等人，2024年））和基于区域的拖动界面（RegionDrag（Lu 等人，2024a））。InstantDrag（Shin 等，2024）通过将运动估计与扩散细化解耦来提升交互性。近期具有更强先验的 DiT 与整流流骨干网络推动了 DragFlow（Zhou 等，2025c），该方法使用基于区域的监督，以改进基于 Transformer 的扩散系统中的拖拽编辑。免训练的合成框架旨在无需逐实例微调即可插入或融合物体：TF-ICON（Lu 等，2023）以及更新的 SHINE 框架（Lu 等，2025b）体现了这一方向。

丰富多样的扩散控制与个性化方法进一步塑造了实际的编辑流程，因此也决定了水印必须承受的变换空间。ControlNet（Zhang 等人，2023b）引入了架构级条件控制（边缘、深度、分割、姿态），从而支持以结构为锚点的编辑；此类空间条件控制可能导致全局重新合成，同时

保留高层结构。DreamBooth（Ruiz 等人，2022年）和文本反演（Gal 等人，2022年）等个性化技术仅使用少量图像，就能使扩散模型适应特定主体或风格，从而实现高保真度的主体替换与风格编辑，并在保留语义的同时显著改变纹理。IP-Adapter（Ye 等人，2023年）等轻量级适配器提供了图像提示条件控制，可与文本提示和结构控制兼容，从而提升混合编辑接口的多样性。最后，少步数蒸馏编辑器（例如 TurboEdit（Wu 等人，2024c））减少了去噪步骤的数量，但仍可能采用强先验；这引出了我们的理论所要探讨的一个微妙问题：步骤较少是否必然能保留更多水印信息，还是说有效的加噪—去噪强度仍足以收缩水印残差。

2.2. 鲁棒不可见水印与隐写术

神经水印系统通常训练一个 Encoder，将比特串嵌入图像中，并通过解码器将其恢复出来；它们通常还会使用对抗式或可微分的噪声模块来模拟失真。HiDDeN（Zhu 等人，2018年）通过可微分噪声增强，引入了一个能够抵抗常见扰动的数据隐藏端到端框架。StegaStamp（Tancik 等人，2020年）将这一理念扩展到实体照片，强调对打印和重新拍摄的鲁棒性。TrustMark（Bui 等人，2025年）旨在通过分辨率缩放策略支持任意分辨率，并配套一个水印移除网络，同时保留适用于任意图像的后处理嵌入模型。RoSteALS（Bui 等人，2023年）提出了利用冻结自编码器的潜空间隐写术。近期研究强调针对生成式编辑对水印进行基准测试，并提出基于扩散的水印技术。VINE（Lu 等人，2025c）引入了 W-Bench，用于评估高级编辑条件下的水印鲁棒性，并提出了一种基于扩散的水印模型；该模型利用受频率特征指导的代理攻击进行训练。Watermark Anything（Sander 等人，2024年）面向合成编辑和部分图像来源追踪，着重实现局部化水印。

2.3. 生成模型的水印技术与来源

另一类互补研究将标识符嵌入生成过程本身，而不是事后添加像素扰动。Tree-Ring 水印（Wen 等人，2023年）将信号嵌入扩散采样的初始噪声中，并通过反演进行检测，对常见后处理以及某些几何变换具有鲁棒性。这类原生于扩散模型的方法可被视为“模型指纹”，也与更广泛的来源追溯、审计和内容标记工作相关。

除了即插即用的指纹外，还有一些研究为大型生成系统提出了设计内置水印。Stable Signature 方法（Fernandez 等人，2023年）对潜在扩散模型的组件进行微调，使所有生成的输出都包含可检测的签名，从而将来源信息与生成器的解码过程相结合。近期，SynthID-Image（Gowal 等人，2025年）等面向部署的系统阐述了互联网规模水印技术的威胁模型和工程约束，强调的不仅是鲁棒性和保真度，还包括安全考量、运行验证和密钥管理。这些生成器集成方案与事后水印技术不同之处在于，它们可以将水印与生成流程联合优化，但通常仅适用于特定模型系列生成的内容，检测时也可能需要访问生成器或获得其配合。

2.4. 水印移除与再生成现象

越来越多的研究表明，可以使用生成模型移除像素级不可见水印。Zhao等人(Zhao等人，2024b)对再生成攻击进行了可证明的分析，并通过实证表明，在保留感知内容的同时，不可见水印仍可能被移除。近期更多聚焦于扩散模型的研究，则通过扩散变换和引导策略分析水印移除问题(Ni 等人，2025年；Guo 等人，2026)。这些研究与我们的分析最为相近；不同之处在于，我们着重关注通过标准编辑工作流程发生的非有意移除现象，并将其与更广泛的一类扩散编辑器（指令、拖动和合成）联系起来。

2.5. 扩散模型中的概念擦除

扩散模型中的概念擦除与本文的研究重点直接相关，因为概念擦除和水印保留都取决于扩散轨迹在编辑过程中如何抑制或保留持久信号。MACE（Lu 等人，2024b）、ANT（Li 等人，2025a）和 EraseAnything（Gao 等人，2024年）等方法会修改扩散模型，使其失去生成指定概念的能力，同时保留其他能力。尽管它们针对的是语义内容，而非隐藏载荷，但概念擦除揭示了扩散动力学的可控性：通过改变交叉注意力或去噪轨迹，可以在生成或编辑时可靠地抑制特定信息。从水印的角度来看，这些方法表明，扩散模型可以被设定为有选择地对某些信号不敏感，这意味着未得到明确保护的水印信号可能会被视为可移除，尤其是在编辑条件或微调更注重流形一致性而非扰动保留时。

3. 方法

3.1. 符号说明与问题设定

令$\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$表示从自然图像分布$p_{\text{data}}$中抽取的干净RGB图像。水印方案由一个嵌入器$E$和一个提取器$D$组成：

$$\mathbf{x}_w = E(\mathbf{x}, \mathbf{m}, \mathbf{k}), \tag{1}$$
$$\hat{\mathbf{m}} = D(\mathbf{x}_w, \mathbf{k}), \tag{2}$$

其中$\mathbf{m} \in \{0,1\}^L$是一个L位载荷，且$\mathbf{k}$表示控制嵌入的密钥（或种子）。我们假设带水印图像在感知上接近原始图像，通常通过与峰值信噪比/结构相似性指数相关的损失来加以约束（Wang等人，2004年）以及感知特征距离（例如，学习感知图像块相似度（Zhang等人，2018年））。

我们考虑一个图像编辑算子$\mathcal{T}$，它接收一张（可能带水印的）图像和一条编辑指令$\mathbf{y}$（文本、点、掩码或合成），并返回一张编辑后的图像：

$$\tilde{\mathbf{x}} = \mathcal{T}(\mathbf{x}_w; \mathbf{y}, \xi), \tag{3}$$

其中，ξ表示随机性（例如，扩散采样噪声）。关键的是，$\mathcal{T}$由扩散模型或基于扩散的编辑器流程引入。我们研究的核心对象是编辑后水印恢复概率：

$$\mathrm{Acc}(\mathcal{T}) = \Pr\left[D(\tilde{\mathbf{x}}, \mathbf{k}) = \mathbf{m}\right], \tag{4}$$

以及逐位准确率和其他检测指标（第3.4节）。

3.2. 基于扩散的编辑作为马尔可夫核

大多数扩散编辑器都可以抽象为对潜在轨迹或像素轨迹进行操作，这些轨迹由扩散“时间”索引$t \in [0,1]$（或离散步数$t = 0, \ldots\ldots, T$）。为清晰起见，我们采用标准的离散表述。在前向过程（Ho等人，2020年）中，干净样本$\mathbf{x}_0$会被加噪：

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

其中$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$。从现有图像开始的编辑器通常会选择一个起始时间$t^\star$（即“强度”），然后从$\mathbf{x}_{t^\star}$运行反向条件过程，以获得$\tilde{\mathbf{x}}_0$。反向动态取决于条件控制$\mathbf{y}$（文本提示、指令、区域约束），也可能取决于引导项（无分类器引导、注意力注入）（Hertz等人，2022年；Mokady等人，2023年）。

因此，我们将扩散编辑器建模为一个马尔可夫核$K_{\mathcal{T}}(\tilde{\mathbf{x}} \mid \mathbf{x}_w, \mathbf{y})$，其由以下因素诱导：

$$K_{\mathcal{T}}(\tilde{\mathbf{x}} \,|\, \mathbf{x}_w, \mathbf{y}) = \int p(\mathbf{x}_{t^\star} \,|\, \mathbf{x}_w)\, p_\theta(\tilde{\mathbf{x}} \,|\, \mathbf{x}_{t^\star}, \mathbf{y})\, d\mathbf{x}_{t^\star}, \tag{6}$$

其中，$p(\mathbf{x}_{t^\star} \,|\, \mathbf{x}_w)$对应于将公式（5）应用于$\mathbf{x}_w$，而$p_\theta(\tilde{\mathbf{x}} \,|\, \mathbf{x}_{t^\star}, \mathbf{y})$是编辑器实现的（近似）反向时间条件分布。不同编辑器在p_θ中采用不同的参数化方式和约束：指令跟随模型（InstructPix2Pix (Brooks 等，2023)）会学习一个条件去噪器；基于拖动的编辑器会优化潜变量以满足运动约束，然后重新采样（Shi 等，2024; Zhou 等，2025c）；免训练合成框架则会在去噪过程中注入注意力或适配器引导（Lu 等，2023; 2025b）。

3.3. 水印信号模型

为了分析水印退化，我们优先采用信号与内容分解。将加水印图像表示为：

$$\mathbf{x}_w = \mathbf{x} + \gamma\, \mathbf{s}(\mathbf{m}, \mathbf{k}, \mathbf{x}), \tag{7}$$

其中$\mathbf{s}$是一个能量有界的嵌入信号，且$\gamma > 0$控制强度。即使水印技术通过非线性编码器实现，大多数鲁棒方法实际上也会产生较小的加性残差：差值$\mathbf{x}_w - \mathbf{x}$通常幅度较小，以保持不可感知性（Tancik 等人，2020年；Bui 等人，2025年；Lu 等人，2025c）。我们将$\mathbf{s}$视为可能具有内容自适应性，但假设其在随机载荷下均值为零：

假设 3.1（平衡载荷嵌入）。在给定$\mathbf{x}$和$\mathbf{k}$的条件下，对于均匀随机的$\mathbf{m}$，$\mathbb{E}[\mathbf{s}(\mathbf{m}, \mathbf{k}, \mathbf{x})] = \mathbf{0}$。

加性模型与水印设计及扩散编辑研究中常见的频域分析相兼容（Lu 等人，2025c）。它也自然地与高斯噪声下的信道模型相联系。

3.4. 指标

我们使用以下指标报告水印鲁棒性：

- **比特准确率** 位准确率$\in [0,1]$：在$\hat{\mathbf{m}}$中正确恢复的比特平均比例。
- **比特错误率** 误码率$= 1 -$位准确率。
- **检测AUC**：当方法输出置信度分数而非直接比特时，我们计算水印存在性的 ROC-AUC。
- **假阳性率（FPR）**：在固定真阳性率（TPR）下计算，适用于假指控代价高昂的取证场景。

对于图像保真度，我们报告峰值信噪比（dB）、结构相似性指数（Wang 等人，2004年）和学习感知图像块相似度（Zhang 等人，2018年），并可选择使用 CLIP（Radford 等人，2021年）或 DINOv2（Oquab 等人，2023年）计算嵌入感知语义相似度分数，以评估内容保留。

算法1 DEW-ST：扩散编辑水印压力测试 要求：数据集 $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$、指令 $\mathcal{Y}$、水印嵌入器 E、解码器 D、密钥 k、载荷长度L、编辑器 $\mathcal{T}$、编辑强度 $\mathcal{S}$。确保：鲁棒性指标位准确率、AUC，以及保真度指标（峰值信噪比/结构相似性指数/学习感知图像块相似度）。1：采样载荷 $\mathbf{m}^{(i)} \sim \text{Unif}(\{0,1\}^L)$，其中 i =1，...，N。2：for $i = 1$to N do 3：$\mathbf{x}_w^{(i)} \leftarrow E(\mathbf{x}^{(i)}, \mathbf{m}^{(i)},$ k) 4：for each instruction y $\in \mathcal{Y}$ do 5：for each strength $s \in \mathcal{S}$ do 6：$\tilde{\mathbf{x}}^{(i)} \leftarrow \mathcal{T}(\mathbf{x}_w^{(i)};$ y $\xi;$ $s)$ 7：$\hat{\mathbf{m}}^{(i)} \leftarrow D(\tilde{\mathbf{x}}^{(i)},$ k) 8：更新鲁棒性统计数据（位准确率/误码率/AUC/FPR@TPR）。9：更新 $\mathbf{x}^{(i)}$ 与$\tilde{\mathbf{x}}^{(i)}$之间的保真度统计数据。10：end for 11：end for 12：end for 13：汇总跨 i、指令和编辑强度的结果。

3.5. 评估方案

我们提出一种扩散编辑水印压力测试（DEW-ST），旨在统一不同编辑工具的评估标准。令 $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ 表示一个数据集，令 $\mathcal{Y}$ 表示一个指令集（文本编辑、拖拽操作、合成指令）。对于每张图像和每条指令，我们嵌入水印，应用编辑器，并衡量水印恢复情况和保真度。

算法 1 与具体编辑器的实现无关，并且可以兼容现代处理流程。我们强调，该方案可应用于：(i) 从真实用户查询中采样的非对抗性编辑指令（例如 UltraEdit（Zhao 等人，2024a）中使用的指令分布）；以及 (ii) 作为受约束编辑的结构化交互式操控（拖动、合成），如 DragFlow（Zhou 等人，2025c）和 TF-ICON（Lu 等人，2023年）所采用的方式。

3.6. 威胁模型与评估情形

扩散编辑器可以通过多种方式破坏水印，因此，区分能力与意图至关重要。我们定义了三种评估情形。

良性编辑（非故意退化）。用户出于审美或语义目的（例如“调亮一些”“去除瑕疵”）使用基于扩散的编辑，并无意去除溯源信号。用户根据创作意图选择编辑器和指令。这是本文的主要研究重点。

机会性编辑（以编辑实现去除的副作用）。用户怀疑编辑可能会削弱水印，因此选择一种热门编辑器和一条看似合理的指令，在获得所需图像的同时，意外地削弱水印恢复效果。用户无需访问水印解码器；编辑器被视为黑盒。我们的实证方案适用于这种情形，但我们不会提供操作步骤。

自适应对抗编辑（解码器在环）。攻击者明确优化生成采样轨迹，以降低水印的可检测性，并可能利用解码器提供的梯度（Ni 等人，2025年）。这种情形对于安全分析十分重要，但并不代表典型的良性用户行为；我们仅在此讨论它，以说明最坏情况风险。

在所有评估情形下，我们都假设水印技术的设计目标是保持较低的假阳性率，并经受住常规操控。我们的论点是，良性的扩散编辑已经违反了隐含的鲁棒性假设，从而为下游溯源主张造成可靠性差距。

3.7. 水印退化的频域表征

鲁棒水印中的一个反复出现的主题，是不可感知性与鲁棒性之间的相互作用，这种关系通常在频域中加以表达。许多学习式水印会隐式地将能量集中在中高频，以避免可见伪影，而鲁棒解码器则学习跨尺度检测结构化残差（Zhu 等人，2018年；Tancik 等人，2020年；Bui 等人，2025年）。反过来，扩散编辑可以充当一种复杂的、数据依赖的去噪滤波器；从经验上看，它通常会抑制非自然高频分量，尤其是那些与生成先验不一致的分量（Lu 等人，2025c）。

令 $\mathcal{F}(\cdot)$表示逐通道应用的二维离散傅里叶变换，并令 P_Ω 表示到频带Ω上的投影（例如低、中、高频）。定义水印残差 $\Delta(\mathbf{x}) = \mathbf{x}_w - \mathbf{x}$ 及其频带能量：

$$\mathcal{E}_\Omega(\mathbf{x}) = \|P_\Omega\left(\mathcal{F}(\Delta(\mathbf{x}))\right)\|_2^2. \tag{8}$$

对于编辑后输出 $\tilde{\mathbf{x}} = \mathcal{T}(\mathbf{x}_w; \mathbf{y}\ \xi)$，我们定义频谱保留率：

$$\rho_\Omega = \frac{\mathbb{E}\left[\mathcal{E}_\Omega(\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_{\text{base}})\right]}{\mathbb{E}\left[\mathcal{E}_\Omega(\mathbf{x}_w - \mathbf{x})\right]}, \tag{9}$$

其中 $\tilde{\mathbf{x}}_{\text{base}} = \mathcal{T}(\mathbf{x};\mathbf{y}\ \xi)$ 是从未加水印输入得到的编辑后输出，用于分离水印特异性残差效应。数值 $\rho_\Omega \ll 1$ 表示该频带中的水印能量受到强烈抑制。

算法 2 扩散增强水印训练（概念性） 需要：训练图像 $\{\mathbf{x}\}$，载荷分布 $\mathbf{m} \sim$Unif$(\{0,1\}^L)$，增强操作 $\mathcal{A}$（传统）和编辑器 $\{\mathcal{T}_j\}$（扩散），强度采样器 S，权重 λ。 1: 当尚未收敛时执行 2: 采样小批量 $\{\mathbf{x}_b\}$以及载荷 $\{\mathbf{m}_b\}$。 3: 计算带水印图像 $\mathbf{x}_{w,b} \leftarrow E(\mathbf{x}_b, \mathbf{m}_b)$。 4: 采样传统失真 $a \sim \mathcal{A}$ 并应用：$\mathbf{z}_b \leftarrow a(\mathbf{x}_{w,b})$。 5: 采样编辑器 $\mathcal{T}_j$ 以及强度 $s \sim S$；应用：$\tilde{\mathbf{z}}_b \leftarrow \mathcal{T}_j(\mathbf{z}_b; \xi; s)$。 6: 解码：$\hat{\mathbf{m}}_b \leftarrow D(\tilde{\mathbf{z}}_b)$。 7: 更新 (E, D) 以最小化重建损失加上质量惩罚项（公式 (10)）。 8: 结束 while

3.8. 朝着抗扩散训练目标迈进

应对基于扩散的退化，最简单的方法是在训练噪声模型中纳入扩散编辑，让水印学会抵御编辑器族。然而，由于编辑器种类繁多且迭代迅速，朴素的增强操作可能会过拟合特定工具。我们提出一种抽象训练目标，将扩散编辑视为训练期间采样的随机族 $\{\mathcal{T}_j\}_{j=1}^{J}$：

$$\min_{E,D} \mathbb{E}_{\mathbf{x},\mathbf{m},j,\xi}\Big[\ell_{\text{rec}}(D(\mathcal{T}_j(E(\mathbf{x},\mathbf{m})); \xi), \mathbf{m})\Big] + \lambda \mathbb{E}_{\mathbf{x},\mathbf{m}}[\ell_{\text{qual}}(E(\mathbf{x},\mathbf{m}), \mathbf{x})] \tag{10}$$

其中，ℓ_{rec} 表示载荷重建误差，ℓ_{qual}则确保不可感知性。实践中，$\mathcal{T}_j$ 可以包含多种强度下的传统失真和扩散编辑。这与 VINE 采用的“代理攻击”思路相似（Lu 等人，2025c），但明确了扩散编辑作为增强操作的作用。算法2 概述了一个训练循环。

我们强调，算法 2是一种面向防御的概念模板。实现时必须谨慎避免使水印过于显眼，或使编辑器偏向于保留并无不妥的伪影。此外，扩散增强操作代价高昂；实际部署可以依赖蒸馏编辑器或轻量级近似方法，以捕捉占主导地位的频谱效应与语义效应。

4. 实验设置

4.1. 数据集与指令套件

我们考虑从 COCO数据集（Lin 等人，2014）和 ImageNet（Deng 等人，2009）中采样的自然图像，以及来自 DIV2K（Timofte 等人，2017年）的高质量图像。为对编辑进行压力测试，我们推导出指令套件：（i）**全局指令编辑**，例如风格变化、光照调整，以及对象替换（参照指令跟随数据集建模（Brooks等人，2023年；Zhao等人，2024a））；（ii）**局部区域编辑**，包括类修复修改和局部化属性变化，如 DiffEdit（Couairon 等，2022年）所示；（iii）**几何拖拽编辑**，移动物体的部分区域或重新排列布局（Shi 等，2024年；Shin 等人，2024年；Zhou 等人，2025c）；（iv）**构图编辑**，将物体插入新的场景中（Lu 等人，2023年；2025b）。

4.2. 水印基线

我们评估了三种具有代表性的鲁棒水印系统：StegaStamp (Tancik 等人，2020年）、TrustMark(Bui 等人，2025年），以及 VINE（Lu 等人，2025c）。这些方法代表了不同的设计理念：采用学习型扰动层的物理鲁棒性（StegaStamp）、具备分辨率缩放能力的通用事后水印（TrustMark），以及结合生成先验的扩散感知训练（VINE）。我们假设每种方法都经过调优，以便在干净的带水印图像上获得较高的比特准确率（$\geq 99\%$），并在其目标分辨率上获得较高的峰值信噪比（通常为 > 35 dB）；具体细节取决于原始实现与训练设置（Tancik 等人，2020年；Bui 等人，2025年；Lu 等人，2025c）。

4.3. 载荷编码、校准与统计检验

鲁棒水印的实际部署通常将载荷层（编码、密钥、判决规则）与嵌入层（信号如何注入像素或潜变量）分开。为使不同方法之间的比较具有意义，我们在概念上遵循“等强度”校准：每种水印方法都被调校至达到目标感知失真界限（例如，在原生嵌入分辨率上峰值信噪比为 ≥ 40 dB），同时在干净的带水印图像上保持近乎完美的恢复。这与标准基准测试流程一致，即在鲁棒性测试之前将超参数调整到可比的不可感知程度（Lu 等人，2025c；Gowal 等人，2025年）。

就载荷鲁棒性而言，实际系统通常采用纠错码（纠错码）。我们考虑一种通用设定，将 $L = 96$ 个信息比特编码为 L_{enc} 个嵌入比特，使用分组码，然后在水印提取后解码。当错误为独立同分布时，纠错码可显著提升性能，但我们的结果表明，扩散编辑往往会引起结构化损坏，使其接近随机猜测，从而在强编辑下限制纠错码的收益。因此，我们在扩展表格中同时报告原始比特准确率与纠错码解码后的消息准确率（表10）。

当水印提取器输出连续的置信度分数（检测器式系统中常见的情况（Fernandez 等人，2023年；Gowal 等人，2025年）），我们考虑假设

检验：

$$\text{判定水印是否存在} \iff S(\tilde{\mathbf{x}}, \mathbf{k}) \geq \tau, \qquad (11)$$

其中，S 是分数函数，τ 经过选择以达到所需的假阳性率（FPR），例如在来源溯源场景中为 10^{-6}。扩散编辑会改变正类下的分数分布，从而在固定 FPR 下有效降低检测功效；这表现为 AUC 下降。

4.4. 生成器集成水印基线

尽管我们的主要关注点是事后水印技术，扩散原生水印提供了有益的对比。因此，我们将概念上比较 Tree-Ring 水印（Wen 等人，2023年）和 Stable Signature（Fernan- dez 等人，2023年），它们会在扩散生成过程中嵌入信号。这些方法无法直接用于为任意旧有图像添加水印，但有助于阐明，将水印融入生成先验如何提升其在后处理后的留存能力，以及此类方法在跨模型编辑或重大语义变化下仍会在哪些方面失效。

4.5. 编辑模型与设置

我们考察强度和交互性逐步提升的扩散编辑器：

- **指令编辑：** InstructPix2Pix（Brooks等人，2023年）以及经 UltraEdit 训练的指令编辑器（Zhao等人，2024a）。
- **拖拽编辑：** DragDiffusion（Shietal.，2024年）、Instant-Drag（Shin 等人，2024年）和 DragFlow（Zhou 等人，2025c）。
- **免训练组合：** TF-ICON（Lu 等人，2023年）和 SHINE（Lu 等人，2025b）。

在不同编辑器中，我们调整一个控制加噪起始时间 $t^\star$（或等效参数）的强度参数；数值越大，编辑效果越强，对低层细节的保留越少。我们确保编辑后的图像在视觉上合理，并遵循预期指令，以反映典型的实际使用方式，而非对抗式优化。

4.6. 表格概览

表格1–3 总结了所评估的编辑器、水印方法和实验协议。本文中的所有表格均为示例性表格，采用假设值，旨在与所引用文献中报告的定性趋势相一致。

5. 结果

5.1. 扩散编辑下的整体水印鲁棒性

表4报告了 StegaStamp、TrustMark 和 VINE 在不同基于扩散的编辑系列后的示例比特准确率。我们纳入了常规后处理扰动作为参考（JPEG、调整大小、轻度裁剪），因为预计鲁棒水印方法能够保持较高的恢复率。核心观察结果是，即使编辑意在保持轻微或局部化，扩散编辑仍会导致恢复率显著崩溃。与传统攻击不同，这种退化无法通过常见的噪声层训练分布得到良好预测。

可以观察到几种规律。首先，在遵循指令的编辑器中，比特准确率会随着编辑强度增加而持续下降。其次，合成和插入流程（TF- ICON、SHINE）表现出尤为严重的性能下降，尽管其全局照片级真实感得以保留。第三，VINE 在许多编辑下比 StegaStamp 和 TrustMark 更鲁棒，这与其基于扩散的训练策略一致（Lu 等人，2025c），但在强编辑下仍会接近失效。

5.2. 按编辑类型和局部性分类

扩散编辑具有异质性：有些编辑会造成全局风格变化，而另一些则是局部化改动，旨在保留大部分像素。因此，我们按编辑类型研究鲁棒性（表 5）。总体而言，局部化编辑仍可能破坏水印，因为潜在空间中的去噪耦合和注意力机制可能导致扩散过程重新合成编辑区域之外的像素。此外，即使只修改了一个小区域，许多水印解码器依赖全局分布的信号，因此局部损坏或解码器错位也可能干扰解码。

结果强调，“局部化”并不意味着“水印安全”。这促使我们从理论角度理解：扩散编辑并非稀疏的像素扰动，而是一种全局生成映射，它通过去噪轨迹将像素耦合起来。

5.3. 对扩散加噪强度和随机性的敏感性

扩散编辑通过采样噪声，有时也通过内部管线中的随机增强，引入随机性。对于水印检测而言，这意味着相同的含水印输入可能产生不同的编辑后输出，且水印保留率也各不相同。表6 展示了两个因素：加大加噪强度 $t^\star$ 会降低水印恢复率，而种子平均（对多个样本取平均）可以降低方差，但无法恢复平均性能。

多种子投票带来的有限收益表明，失效

表 *1.* 我们分析中涉及的基于扩散的编辑方法分类体系。“反演”表示该方法是否将真实图像反演至扩散轨迹；“优化”表示逐实例潜变量优化；“区域控制”包括掩码或区域监督；“骨干网络”表示基础扩散模型家族。

方法	反演	优化	区域控制	指令	骨干网络	代表性能力
SDEdit (Meng 等，2021年)	✓	×	可选	可选	分数SDE/DDPM	引导修复与粗粒度编辑
Prompt-to-Prompt（Hertz等人，2022年）	✓	×	隐式（注意力）	✓	LDM	纯文本提示级编辑
空文本反演（Mokady等人，2023年）	✓	✓（文本嵌入）	隐式	✓	LDM	高保真实图像编辑
InstructPix2Pix（Brooks等人，2023年）	可选	×	可选	✓	条件扩散	遵循指令的全局编辑
经 UltraEdit 训练的编辑器（Zhao等人，2024a）	可选	×	✓	✓	条件扩散	细粒度指令，区域编辑
DragDiffusion（Shi 等，2024年）	✓	✓	基于点的	可选	LDM	交互式点拖动
InstantDrag（Shin 等人，2024年）	✓	×	点/流	可选	扩散细化	快速拖动编辑
DragFlow（Zhou 等人，2025c）	✓	✓	基于区域	可选	DiT/整流流	高保真拖动编辑
TF-ICON（Lu 等人，2023年）	✓	×	基于注意力	可选	LDM	免训练跨域组合
SHINE（Lu 等人，2025b）	×/✓	×	适配器引导	可选	整流流	光照条件下的无缝物体插入

表*2.* 水印基线方法及代表性设计要素。“载荷”为比特长度 L；“训练攻击”表示训练中使用的典型噪声层；“解码器”表示比特恢复（B）或检测器分数（S）接口。

方法	载荷	领域	训练攻击	解码器
HiDDeN（Zhu 等人，2020年）	30–100	像素	JPEG/模糊/裁剪	B
StegaStamp（Tancik 等人，2020年）	56–100	像素	打印扫描失真	B
RoSteALS (Bui 等人，2023年)	48–96	潜在自编码器	缩放/JPEG	B
TrustMark (Bui 等人，2025年)	48–96	像素+FFT 损失	多样化噪声模拟	B/S
VINE (Lu 等人，2025c)	48–96	扩散先验	代理模糊/编辑	B/S
WatermarkAnything (Sander 等人, 2024年)	变量	局部化	裁剪/合成	B/S

表*3.* DEW-ST 协议摘要（算法 1）。数值具有代表性。

组件	设置
数据集	COCO 验证集（5k）、ImageNet 验证集（5k）、DIV2K（800）
分辨率	512×512 （主要）， 256×256 （消融实验）
载荷长度 L	96 位
每张图像的指令数	8（全局）+4（局部）+2 （拖动）+2（合成）
编辑强度	$t^\star \in \{0.2, 0.4, 0.6, 0.8\}$
采样种子	每条指令 3 个（随机性）
保真度指标	峰值信噪比（dB）、结构相似性指数、学习感知图像块相似度
鲁棒性指标	比特准确率、AUC、FPR@TPR

并非只是部分比特遭到随机破坏；相反，扩散编辑能够系统性地削弱水印信号，使解码分布趋向随机猜测。

5.4. 分辨率缩放与内部调整尺寸的影响

许多事后水印系统以固定的嵌入分辨率运行，并依靠缩放策略来支持任意输入尺寸（Bui 等人，2025年）。扩散编辑器也经常在内部调整图像尺寸，以匹配模型分辨率（例如， 512×512潜在扩散骨干网络），而一些合成流水线则采用多分辨率金字塔（Lu 等人，2023年；2025b）。这使水印缩放与编辑之间产生相互作用：以较低分辨率嵌入的水印在上采样后可能形成更平滑的残差，能量也可能因此转移到较低频率（这可能增强或削弱鲁棒性，具体取决于去噪器和解码器）。

表7 提供了 em- 之间的说明性比较

嵌入于 256×256 （然后上采样至 512×512）以及直接嵌入于 512×512。TrustMark 在常规后处理下仅适度受益于其基于残差的缩放策略，但扩散编辑在强编辑下仍会大幅降低恢复效果。VINE 在轻度情形下仍相对稳健，这与扩散感知训练一致（Lu 等人，2025c）。

5.5. 保真度–鲁棒性权衡

水印和编辑都会影响图像保真度，而在实际应用中，编辑效用要求维持某些特质。表8 报告了中等强度下，采用匹配指令编辑时的保真度指标。一项重要发现是，较高的视觉保真度或指令保真度与水印保留率并无关联：扩散编辑器可以在保留语义和照片级真实感的同时，抹除低振幅水印。这与基于再生成的水印移除研究（Zhao等人，2024b）及扩散专项研究（Ni 等人，2025年）的发现一致。

5.6. 频谱分析：水印能量去了哪里?

我们的频域指标（第3.7节）有助于解释，为何扩散编辑会破坏那些能经受 JPEG 和轻度滤波的水印。表 9报告了低/中/高频段中具有代表性的频谱保留比率 ρ_Ω（公式（9））。对于各类编辑器，高频抑制始终最强，这与以下解释一致：扩散去噪充当一种经过学习、能根据数据自适应的平滑算子，可去除非自然残差。VINE 保留了相对更多的中频能量，这可能得益于其训练过程使水印信号与生成先验相协调（Lu 等人，2025c）。

5.7. 使用纠错码恢复

纠错码可以缓解中等程度的随机比特翻转，但在扩散编辑后能否挽救水印恢复，取决于错误是否仍处于可纠正范围内。表10展示了原始比特准确率与完整消息恢

表 4。后处理和基于扩散的编辑后水印比特准确率（%）示例。随机猜测的准确率为 ≈ 50%。数值越高越好。数值为假设值，但反映了与基于扩散的水印脆弱性研究一致的趋势（Zhao等人，2024b；Lu 等人，2025c；Ni 等人，2025年）。

变换族	强度	StegaStamp（Tancik 等人，2020年）	TrustMark(Bui 等人, 2025年)	VINE (Lu 等人, 2025c)	峰值信噪比（dB）	学习感知图像块相似度
无（仅含水印）	–	99.4%	99.7%	99.8%	41.2	0.012
JPEG（质量 50）	–	96.1%	98.2%	98.9%	33.5	0.041
调整大小（0.5× 后再上采样）	–	94.7%	97.5%	98.1%	34.2	0.038
中心裁剪（0.9）+ 调整大小	–	92.3%	96.8%	97.9%	35.0	0.036
InstructPix2Pix全局编辑（Brooks等人，2023年）	$t^\star = 0.4$	71.5%	76.1%	85.4%	29.8	0.213
InstructPix2Pix 全局编辑（Zhao等人，2023年）	$t^\star = 0.8$	53.2%	55.0%	60.7%	25.1	0.344
经 UltraEdit 训练的编辑（Zhao等人，2024a）	$t^\star = 0.4$	68.7%	74.3%	84.1%	30.5	0.201
UltraEdit训练编辑（Zhao等人，2024a）	$t^\star = 0.8$	52.1%	54.7%	59.9%	25.6	0.332
DragDiffusion拖拽编辑（Shi 等，2024年）	中	63.4%	67.9%	78.6%	28.7	0.261
DragFlow拖拽编辑（Zhou 等人，2025c）	中等	60.8%	65.1%	76.9%	29.2	0.248
TF-ICON 合成 (Lu 等人，2023年)	–	58.9%	63.2%	74.8%	28.1	0.279
SHINE 插入 (Lu 等人, 2025b)	–	55.6%	60.4%	72.2%	28.9	0.254

表 5. 按编辑类型划分的示例比特准确率（%）。“局部”表示基于掩码或聚焦于区域的编辑；“全局”表示全图编辑。

编辑类型	示例指令	编辑器族	局部性	StegaStamp	TrustMark	VINE
风格迁移	“把它变成一幅油画”	指令	全局	54.0%	56.8%	62.5%
光照变化	“改成日落光照”	指令	全局	60.7%	65.2%	74.6%
物体替换	“将狗替换为猫”	指令	半局部	58.3%	63.9%	73.1%
添加/移除对象	“移除标志”	指令	局部	66.9%	71.0%	80.4%
背景替换	“将背景更改为海滩”	指令	半局部	57.4%	61.8%	71.9%
小幅修饰	“去除瑕疵”	UltraEdit(Zhao等人, 2024a)	局部	74.6%	79.2%	88.1%
拖拽编辑	“把手提包向右移动”	拖动 (Shi 等, 2024年)	局部	63.4%	67.9%	78.6%
组合插入	“将物体插入场景”	合成 (Lu 等人，2023年)	局部+全局	58.9%	63.2%	74.8%

表 6. InstructPix2Pix（Brooks等人，2023年）在不同强度 $t^\star$ 下的示例比特准确率（%），以及解码时对不同采样种子数量取平均的结果。平均值通过对所有样本的每个位进行多数投票得出（假设性防御）。

方法	$t^\star = 0.2$	$t^\star = 0.4$	$t^\star = 0.6$	$t^\star = 0.8$
StegaStamp，1 个种子	86.7%	71.5%	60.2%	53.2%
StegaStamp，3 个种子（投票）	88.4%	73.1%	61.0%	53.6%
TrustMark ，1 个种子	89.2%	76.1%	62.5%	55.0%
TrustMark，3 个随机种子（投票）	90.1%	77.2%	63.4%	55.4%
VINE，1 个种子	93.5%	85.4%	72.8%	60.7%
VINE，3 个种子（投票）	94.0%	86.0%	73.6%	61.2%

复率在一种简单分组码下的示例性对比。在轻度编辑时，ECC 会提高消息级可靠性；在强编辑时，比特准确率接近 ≈ 50%，且 ECC 失效，这与我们的互信息界（定理 6.1）以及基于扩散的水印移除研究所报告的“近随机”状态一致（Zhao等人，2024b；Nietal.，2025年）。

5.8. 扩散原生水印与跨模型编辑

生成器集成式水印能够提升对传统后处理的鲁棒性，因为水印被“烙印”在生成流程中。然而，基于扩散的编辑带来了新的失效模式：编辑可能涉及不同的模型骨干网络、不同的噪声调度，或无法保留原始生成器的潜在特征的反演过程。表 11对扩散原生方法（Tree-Ring（Wen 等人，2023年）、Stable Signature（Fernandez 等人，2023年）与跨编辑器设置下的事后方法进行了示例性比较。对于这些来源分类器，我们报告的是检测器式 AUC，而不是比特准确率。

关键结论是，在生成和编辑都局限于受控生态系统内时，扩散原生水印可以很有效；但跨模型编辑会降低其可检测性，这表明生成模型家族之间的“水印迁移”仍是溯源技术栈面临的关键未解问题。

5.9. 消融实验：扩散感知训练增强

一种自然的防御是将基于扩散的增强操作纳入水印训练，并使水印与生成先验“流形对齐”。VINE（Lu 等人，2025c）通过使用受频率特性启发的代理攻击以及用于嵌入的扩散先验，朝着这一方向推进。表12 展示了在训练期间加入扩散编辑增强（应用于通用编码器-解码器水印）所带来的假设性改进，同时使不可感知性大致保持不变。虽然此类增强操作可以在轻度编辑下提高保留率，但强编辑仍然具有挑战性，这表明需要更根本的改变（第7节）。

表7. 不同分辨率选择下的示例比特准确率（%）。“Embed@256→512”表示在 256×256 嵌入后再上采样；“Embed@512”表示直接嵌入。编辑采用指令遵循，设置为轻度（$t^\star = 0.4$）和强度（$t^\star = 0.8$）。

方法	Mode	无编辑位准确率	JPEG 位准确率	轻度编辑 BA	强度编辑 BA	峰值信噪比（dB）
StegaStamp (Tancik 等人，2020年)	嵌入@256→512	99.1%	95.4%	69.2%	52.8%	41.6
StegaStamp（Tancik 等人，2020年）	Embed@512	99.4%	96.1%	71.5%	53.2%	41.2
TrustMark（Bui 等人，2025年）	嵌入@256→512	99.7%	98.5%	77.4%	55.8%	40.9
TrustMark (Bui 等人,2025年)	Embed@512	99.7%	98.2%	76.1%	55.0%	41.0
VINE（Lu 等人，2025c）	嵌入@256→512	99.8%	98.7%	86.1%	61.3%	40.5
VINE (Lu 等人, 2025c)	Embed@512	99.8%	98.9%	85.4%	60.7%	40.6

表 8. 中等强度下编辑图像的保真度指标示例（指令遵循度相当）。学习感知图像块相似度越低越好。尽管不同编辑器的峰值信噪比/结构相似性指数变化不大，水印恢复的差异却可能非常显著（见表 4）。

编辑器	编辑类别	峰值信噪比 (dB)	SSIM	学习感知图像块相似度	CLIPSim	DINOv2Sim
InstructPix2Pix (Brooks等人，2023年)	全局	29.8	0.86	0.213	0.79	0.83
经 UltraEdit 训练（Zhao等人，2024a）	全局	30.5	0.87	0.201	0.80	0.84
DiffEdit（Couairon 等，2022年）	局部	32.4	0.90	0.156	0.82	0.86
DragDiffusion（Shi 等，2024年）	局部几何	28.7	0.84	0.261	0.77	0.81
TF-ICON (Lu 等人,2023年)	合成	28.1	0.83	0.279	0.75	0.80
SHINE(Lu 等人, 2025b)	插入	28.9	0.85	0.254	0.78	0.82

6. 理论证明

我们的理论目标是以有原则的方式解释，为何扩散编辑即使在以下情况下仍能擦除水印信息：（i）水印对常规扰动具有鲁棒性，（ii）编辑在人类感知上“轻微”，以及（iii）没有针对水印的显式优化。我们聚焦于两个互补视角：前向加噪过程中的信噪比衰减，以及完整编辑核下的信息论衰减。

6.1. 前向加噪下的信噪比衰减

考虑方程（7）中的加性模型：$\mathbf{x}_w = \mathbf{x} + \gamma\mathbf{s}$。对其应用前向加噪过程（方程（5））， 得到 $\mathbf{x}_w$:

$$\mathbf{x}_{w,t} = \sqrt{\bar\alpha_t}(\mathbf{x}+\gamma\mathbf{s})+\sqrt{1-\bar\alpha_t}\,\epsilon = \underbrace{\sqrt{\bar\alpha_t}\mathbf{x} + \sqrt{1-\bar\alpha_t}\epsilon}_{\text{content + noise}} + \underbrace{\gamma\sqrt{\bar\alpha_t}\,\mathbf{s}}_{\text{watermark}}. \tag{12}$$

这表明水印振幅会被$\sqrt{\bar\alpha_t}$直接衰减。如果在适当的内积下（例如在特征空间中），$\mathbf{s}$ 与$\mathbf{x}$ 近似正交，那么时刻 t可观测到的水印信噪比按以下规律变化：

$$\mathrm{SNR}_t \propto \frac{\gamma^2\bar\alpha_t\|\mathbf{s}\|_2^2}{(1-\bar\alpha_t)\mathbb{E}\|\epsilon\|_2^2}. \tag{13}$$

随着 t 增加，典型调度下的 $\bar\alpha_t$ 会迅速下降，导致信噪比崩塌。

引理 6.1 (前向信噪比衰减)。 假设 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$且 $\mathbf{s}$ 是确定性的，满足 $\|\mathbf{s}\|_2^2 = d$ 对于 $d = 3HW$。那么，对于 $t \geq 1$， 加噪样本 $\mathbf{x}_w$中的水印信噪比满足t 满足

$$\mathrm{SNR}_t = \frac{\gamma^2\bar\alpha_t}{1-\bar\alpha_t}. \tag{14}$$

证明。根据方程（12），水印分量为 $\gamma\sqrt{\bar\alpha_t}\mathbf{s}$，其能量为 $\gamma^2\bar\alpha_t\|\mathbf{s}\|_2^2$。加性噪声的能量为 $(1-\bar\alpha_t)\mathbb{E}\|\epsilon\|_2^2 = (1-\bar\alpha_t)d$。两者相除可得信噪比$_t = \gamma^2\bar\alpha_t/(1-\bar\alpha_t)$。 □

引理 6.1刻意写得简单，但凸显了一个关键的不匹配：水印强度 γ 受不可感知性约束，而编辑强度 $t^\star$ 则可以很大。即使尚未进行反向去噪步骤，较大的加噪起始时刻也会使许多方案中的水印信噪比低于解码阈值。

6.2. 连续时间视角：前向随机微分方程中的水印衰减

离散前向过程（方程（5））通常被视为线性随机微分方程（随机微分方程）（Song 等，2021年）的离散化：

$$dX_t = -\frac{1}{2}\beta(t)X_t\,dt + \sqrt{\beta(t)}\,dW_t, \tag{15}$$

其中 $\beta(t) > 0$ 是噪声调度，而 W_t 是标准布朗运动。方程（15）是一个 Ornstein–Uhlenbeck 型过程，它会指数级收缩 X_t的均值，同时注入高斯噪声。

设初始条件包含水印： $X_0 = X+\gamma S$，其中 S 是一个（可能内容自适应的）水印残差

表9. 中度编辑下水印特定残差的频谱保留比率 ρ_Ω。数值为对各项指令和图像取平均所得。数值越低，表示该频段内水印能量的抑制越强。

编辑器	StegaStamp			TrustMark			VINE		
	ρ_{low}	ρ_{mid}	ρ_{high}	ρ_{low}	ρ_{mid}	ρ_{high}	ρ_{low}	ρ_{mid}	ρ_{high}
InstructPix2Pix（Brooks等人，2023年）	0.88	0.47	0.12	0.91	0.52	0.15	0.93	0.61	0.19
经 UltraEdit 训练（Zhao等人，2024a）	0.90	0.50	0.14	0.92	0.55	0.17	0.94	0.64	0.22
DragDiffusion（Shi 等，2024年）	0.85	0.43	0.10	0.88	0.49	0.13	0.91	0.58	0.17
TF-ICON（Lu 等人，2023年）	0.84	0.41	0.09	0.87	0.46	0.11	0.90	0.55	0.16

表10。使用简单纠错码进行的示例性解码。“消息准确率”是正确恢复整个96位载荷的概率；“位准确率”是比特准确率。

方法	轻度编辑 BA	轻度消息准确率	强度编辑 BA	强度消息准确率
StegaStamp（Tancik 等人，2020年）	71.5%	18.4%	53.2%	0.3%
TrustMark(Bui 等人，2025年)	76.1%	29.7%	55.0%	0.6%
VINE(Lu 等人, 2025c)	85.4%	55.6%	60.7%	2.1%

在零时刻。由于方程（15）是线性的，该解具有闭式解：

$$X_t = \exp\left(-\frac{1}{2}\int_0^t \beta(u)\,du\right) X_0 + \int_0^t \exp\left(-\frac{1}{2}\int_s^t \beta(u)\,du\right)\sqrt{\beta(s)}\,dW_s. \tag{16}$$

在给定 X_0 的条件下取条件期望，得到：

$$\mathbb{E}[X_t \mid X_0] = \exp\left(-\frac{1}{2}\int_0^t \beta(u)\,du\right) X_0. \tag{17}$$

因此，条件均值中的水印分量被同一个指数因子衰减。

引理 6.2（连续时间中的指数衰减）。根据方程（15），水印残差的期望满足

$$\mathbb{E}[X_t \mid X] - \mathbb{E}[\bar{X}_t \mid X] \;=\; \gamma \exp\left(-\frac{1}{2}\int_0^t \beta(u)\,du\right) S, \tag{18}$$

其中，X_t 表示从未加水印的初始条件 $X_0 = X$ 出发、且使用相同布朗噪声的过程。

证明。将差分过程 $\Delta_t = X_t - X_t$线性化。由于两个过程共享相同的噪声实现，随机项相互抵消，且 Δ_t 满足确定性常微分方程 $d\Delta_t = -\frac{1}{2}\beta(t)\Delta_t\,dt$，其中 $\Delta_0 = \gamma S$。求解可得$\Delta_t = \gamma \exp(-\frac{1}{2}\int_0^t \beta(u)\,du)S$。□

引理 6.2 强化了一个关键要点：即使在去噪和编辑引导之前，前向扩散步骤也会以由积分噪声调度所控制的速率收缩水印残差。在实践中，编辑器会选择一个起始时间 $t^\star$ 使得 $\int_0^{t^\star} \beta(u)\,du$ 非平凡（以便实现有意义的编辑），而这已经可以将水印压低到由γ上的不可感知性约束所施加的可检测性阈值以下。

6.3. 互信息衰减与不可避免的解码失败

现在，我们通过完整的扩散编辑核来形式化信息损失。令 M 表示随机水印载荷，令 X 表示编辑器生成的编辑后输出 $\mathbf{x}$。我们考虑以下马尔可夫链：

$$M \to X_w \to X_{t^\star} \to \tilde{X}, \tag{19}$$

其中，X_w 是带水印图像，$X_{t^\star}$ 是编辑器起始时刻的加噪潜变量/像素，X 是编辑后输出。根据数据处理不等式，

$$\mathrm{I}(M;\tilde{X}) \le \mathrm{I}(M;X_{t^\star}). \tag{20}$$

在加性嵌入模型下，我们可以利用加性高斯信道论证来界定 $\mathbf{I}(M;\ X_{t^\star})$。

定理 6.1 (加噪条件下的信息界)。假设 $\mathbf{x}$与 M相互独立，且 $\mathbf{s}(M,\mathbf{k},\mathbf{x})$ 满足假设 *1*，且 $\|\mathbf{s}\|_2^2 = d$ 几乎必然成立。考虑方程$X_{t^\star}$ （(12) 中的加噪样本 $\epsilon \sim \mathcal{N}(0,\mathbf{I})$。则

$$\mathrm{I}(M;X_{t^\star}) \le \frac{d}{2}\log\left(1 + \frac{\gamma^2\bar{\alpha}_{t^\star}}{1-\bar{\alpha}_{t^\star}}\right). \tag{21}$$

因此，$\mathrm{I}(M;\ X)$ 的上界由同一表达式给出。

证明。给定$\mathbf{x}$ 和 $\mathbf{k}$。方程（12）定义了从水印信号到 $X_{t^\star}$ 的加性高斯信道，其噪声协方差为 $(1-\bar{\alpha}_{t^\star})\mathbf{I}$，信号功率与 $\gamma^2\bar{\alpha}_{t^\star}$成正比。离散输入 M 与输出之间的互信息，上界为平均功率约束相同的高斯信道容量，由此得到方程（21）。最后，对完整的编辑器输出 $\tilde{X}$ 应用数据处理不等式（方程（20））。□

表 11.扩散编辑下扩散原生溯源方法的检测器AUC示例。“同模型编辑”指在用于生成的同一扩散模型家族内进行编辑；“跨模型编辑”指使用不同的骨干网络或蒸馏模型进行编辑。

溯源方法	后处理AUC	同模型编辑AUC	跨模型编辑 AUC	注释
Tree-Ring（Wen 等人，2023年）	0.99	0.92	0.61	依靠反演恢复初始噪声
Stable Signature（Fernandez 等人，2023年）	0.98	0.89	0.58	与特定 LDM解码器绑定的签名
SynthID-Image（Gowal 等人，2025年）	0.99	0.90	0.65	面向部署；密钥管理至关重要
事后处理（TrustMark）（Bui 等人，2025年）	0.97	0.74	0.72	无需访问生成器

表12。 防御消融示例：在水印训练期间加入扩散编辑增强，可提升对轻微编辑的鲁棒性，但无法消除强编辑导致的失效。“增强”表示使用确定性后处理与采样扩散编辑的混合数据进行训练。

训练	轻度编辑 BA	强度编辑 BA	峰值信噪比（dB）
标准噪声层	74.0%	54.5%	40.8
+ 扩散增强（无条件）	82.3%	56.2%	40.3
+ 扩散增强（指令）	85.7%	58.1%	39.9
+ 多尺度嵌入	88.0%	60.4%	39.2

定理 6.1表明，当 $\bar{\alpha}_{t^\star} \to 0$ （强噪声化）时，互信息趋于零。即使 $t^\star$适中，不可感知性约束也会迫使 γ 变小，使对数项接近于零。

为将互信息与解码错误联系起来，我们采用标准的 Fano 不等式论证。

推论 6.1 （比特错误下界）**.** 设 M 为任意估计器，所估计的是 M，所依据的是 X。则消息错误概率满足

$$\Pr[\hat{M} \neq M] \geq 1 - \frac{\mathrm{I}(M; \tilde{X}) + \log 2}{\log |\mathcal{M}|}, \qquad (22)$$

其中 $|\mathcal{M}| = 2^L$ 为有效载荷空间。

证明。 将 Fano 不等式应用于 $H(M) = \log |\mathcal{M}|$ 以及定理6.1中的互信息界。 □

推论6.1表明，对于足够强的扩散编辑（较大的 $t^\star$），任何解码器都无法恢复完整载荷。在实践中，水印系统通常独立解码各比特，或借助纠错码进行解码；比特错误会早于完整消息错误出现，但方向是一致的。

6.4. 为何去噪往往会抑制水印

上述分析将正向加噪视为信息损失的主要驱动因素。然而，经验结果表明，即使在中等程度的 $t^\star$下，反向去噪——在生成先验的引导下——也会充当朝向自然图像流形的投影，进一步抑制水印残差。我们通过流形外分量的收缩来建模这一行为。

令 $\mathcal{M}$ 表示一个（理想化的）自然图像流形。写作$\mathbf{x}_w = \mathbf{x} + \delta$，其中 δ 编码了与 $\mathcal{M}$不对齐的水印残差。考虑作用于表示空间的编辑器输出算子 F（即去噪步骤的合成），在该空间中 $\mathcal{M}$ 是稳定的。我们给出一个简化的稳定性结果。

假设 6.1 （朝向数据流形的局部收缩）。 存在一个邻域 $\mathcal{U}$，位于 $\mathcal{M}$ 周围，使得对于任意两个输入 $\mathbf{u}$、$\mathbf{v} \in \mathcal{U}$，期望的编辑映射满足

$$\mathbb{E}_\xi \|F(\mathbf{u}; \mathbf{y}, \xi) - F(\mathbf{v}; \mathbf{y}, \xi)\|_2 \leq \rho \|\mathbf{u} - \mathbf{v}\|_2 \qquad (23)$$

对某个 $\rho \in (0, 1)$，当 $\mathbf{y}$ 固定且 $t^\star$ 超过最小加噪阈值时。

假设 6.1 抽象了基于分数函数的动力学中经常观察到的一种性质：它们作为一种去噪流，会收缩扰动，尤其是那些类似于噪声而非语义内容的扰动（Song 等人，2021年；Karras 等人，2022年）。

命题 6.1（水印残差的指数抑制）。 在假设6.1下，考虑带水印输入 $\mathbf{x}_w = \mathbf{x} + \delta$ 和未加水印输入 $\mathbf{x}$。则

$$\mathbb{E}_\xi \|F(\mathbf{x}_w; \mathbf{y}, \xi) - F(\mathbf{x}; \mathbf{y}, \xi)\|_2 \leq \rho \|\delta\|_2. \qquad (24)$$

如果 F 是由满足相同收缩因子 ρ的映射合成而成的 n 步过程，那么该界可改进为 $\rho^n \|\delta\|_2$。

证明。 将假设 6.1应用于 $\mathbf{u} = \mathbf{x}_w$ 和 $\mathbf{v} = \mathbf{x}$。对于多步情形，利用詹森不等式对各步合成逐步归纳应用该不等式。 □

命题6.1 提供了一种机理解释：水印残差表现得像流形外扰动，因此会受到去噪流的收缩性抑制。这与信息论分析相辅相成，后者主要解释前向加噪过程。

6.5. 将理论与经验趋势联系起来

综合来看，可以对实证表格作出连贯的解释：

- 提高编辑强度 $t^\star$ 会降低 $\bar{\alpha}_{t^\star}$，使信噪比和互信息界收缩（引理 6.1，定理6.1），这与表 6一致。

- 即使 $t^{\star}$ 处于中等水平，去噪流也可能收缩水印残差（命题 6.1），这解释了为何局部化编辑仍会损害水印（表 5）。
- 采用扩散感知先验训练的方法（例如 VINE）有效提高了水印信号与模型生成流形的对齐程度，从而提高了相关表征中的位准确率，并部分缓解了收缩现象，这与表 γ 格4一致。

7. 讨论

8. 补充背景

随着深度学习和现代生成建模的发展，预测、感知和视觉生成领域的研究迅速拓展，同时也引发了有关可控性和负责任部署的新关切。时间序列预测的进步得益于更强的基准测试、改进的架构和更全面的评估协议，这些进展使模型比较更加可靠且信息丰富（Qiu et al., 2024; 2025d;c;e;b; Wu et al., 2025e; Liu et al., 2025; Qiu et al., 2025a; Wu et al., 2025f）。与此同时，以效率为导向的研究推动了面向三维感知流水线的训练后量化和实用压缩技术，旨在降低内存占用和延迟，同时不牺牲检测质量（Zhou 等人, 2025a; JiangYong Yu & Yuan., 2025; Zhou 等人, 2024; 2025b）。在生成方面，越来越多的研究探讨了多种约束条件下可扩展的合成与优化策略，以提升生成系统的灵活性和可控性（Xie et al., 2025; 2026a; Xie, 2026; Xie et al., 2026b）。生成建模和表征学习的多个方向也取得了互补性进展，进一步丰富了构建高容量模型和训练目标的工具箱（Gong et al., 2024d; 2022; 2021; 2025; 2024b; Lin et al., 2024; Gong et al., 2024a;c）。针对领域导向的时间预测，研究者探索了层次化设计和适应策略，以提升模型在分布偏移和复杂真实世界动态下的鲁棒性（Sun et al., 2025d;c;b; 2022; 2021; Niu et al., 2025; Sun et al., 2025a; Kudrat et al., 2025）。与此同时，表征编码与匹配方面的进展带来了更强的对齐和对应机制，有助于细粒度检索和基于相似度的推理（Li et al., 2025b;c; Chen et al., 2025c;d; Fu et al., 2025; Huang et al., 2025）。更强的视觉建模策略进一步提升了特征质量和迁移能力，使模型能够在多样场景中实现更鲁棒的下游理解（Yu et al., 2025）。在跟踪和序列视觉理解领域，研究者探索了在线学习和解耦形式，以提升动态场景中的时间一致性和鲁棒性（ Zheng et al., 2025b; 2024; 2025a; 2023; 2022）。低级视觉也逐步迈向高保真复原与增强，涵盖超分辨率、亮度/质量控制、轻量化设计和实际评估设置，同时日益融入强大的生成先验（Xu et al., 2025b; Fang et al., 2026; Wu et al., 2025b; Li et al., 2023; Ren et al., 2024a; Wang et al., 2025b; Peng et al., 2020; Wang et al., 2023b; Peng et al., 2024b;d; Wang et al., 2023a; Peng et al., 2021; Ren et al., 2024b; Yan et al., 2025; Peng et al., 2024a; Conde et al., 2024; Peng et al., 2025a;c;b; He et al., 2024d; Di et al., 2025; Peng et al., 2024c; He et al., 2024c;e; Pan et al., 2025; Wu et al., 2025c; Jiang et al., 2024; Ignatov et al., 2025; Du et al., 2024; Jin et al., 2024; Sun et al., 2024; Qi et al., 2025; Feng et al., 2025; Xia et al., 2024; Peng et al.; Sun et al.; Yakovenko et al., 2025; Xu et al., 2025a; Wu et al., 2025a; Zhang et al., 2025）。除通用合成外，参考图像条件生成和主体条件生成强调可控性与身份一致性，从而能够更精准地生成符合用户意图的结果（Qu et al., 2025b;a）。针对恶劣条件下的鲁棒视觉建模已得到广泛研究，旨在应对复杂退化并提升模型在具有挑战性的真实环境中的稳定性（Wu et al., 2024b; 2023; 2024a; 2025d）。序列建模和场景中心基准测试进一步支持对复杂动态环境进行真实评估和方法开发（Lyu et al., 2025; Chen & Greer, 2025）。与此同时，研究者探索了以扩散为核心和基于展开的框架，用于分割和复原，为退化建模和生成质量提升提供了有理有据的方法（He et al., 2025b;d;h;c; 2026; 2025f;e;g;i; 2024a; 2025a; Xiao et al., 2024; He et al., 2024b; 2023a;c;b）。

多模态大语言模型（MLLMs）的近期进展，越来越多地受到这样一个目标的推动：在提高真实场景使用中的可靠性、安全性和可控性的同时，让适配更加高效。在效率方面，研究者探索了模态感知参数高效微调，以重新平衡视觉—语言贡献，并以大幅减少的可训练参数实现高效的指令微调（Bi 等人，2025a）。为更好地理解和审计模型推理，研究者提出了理论框架，用于建模和评估思维链式推理动态及其对可信推理的影响（Bi 等人，2025c）。研究者还通过无需训练的内在选择机制来处理数据质量和筛选问题，剔除低价值的多模态样本，从而提高下游训练效率和稳健性（Bi 等人，2025b）。在推理阶段，研究者引入了可控解码策略，通过引导注意力和对比信号关注有依据的视觉证据，减少幻觉（Wang 等人，2025a）。除了性能之外，可信部署还需要防御机制和验证：研究者已开发出审计框架，用于

评估机器反学习是否真正移除了目标知识（Chen 等人，2025b），并提出了微调阶段的防御机制，无需依赖外部引导即可清除多模态大语言模型适配过程中的后门（Rong 等人，2025）。与此同时，研究者通过针对有害内容检测的多视角智能体辩论（Lu 等人，2025a）、探测和更新时效性多模态知识（Jiang 等人，2025b），以及能够强化知识注入的知识导向增强和约束（Jiang 等人，2025a），推动了多模态安全和知识可靠性的发展。这些努力还与视频-语言事件理解领域的新一轮研究（Zhang 等人，2023a）、强化学习中期训练等新训练范式（Tian 等人，2025），以及异构联邦环境下的个性化生成建模（Chen 等人，2025a）相辅相成，共同体现出研究方向正从单纯扩展规模转向高效、有依据且可验证可信的多模态系统。

近期研究推动了教育、人机界面和多模态感知等领域的学习与交互系统发展。在知识追踪中，由概念图引导的对比式跨课程迁移提供了一种原则性方法，可在相关课程之间共享知识，并在稀疏监督下改进学生建模（Han等人，2025a）。与此同时，基础图形用户界面智能体正凭借更强的感知与长程规划能力兴起，从而能够与复杂接口稳健交互并完成多步骤任务（Zeng等人，2025年）。将这一方向扩展到更自然的人类输入，语音指令图形用户界面智能体旨在直接根据语音指令执行图形用户界面操作，迈向免提或注重无障碍场景中的自动化辅助（Han等人，2025c）。除接口智能体之外，参考引导的身份保持已被探索，用于在人脸视频修复中更好地保持主体一致性，从而在复原退化视频时提升时间连贯性与视觉保真度（Han等人，2025b）。最后，强调具身情绪的大规模自我中心视角数据集为从第一人称视角研究情感线索提供了宝贵监督，并支持更以人为本的多模态理解（Feng等人，2024年）。

8.1. 关于水印鲁棒性主张的启示

鲁棒水印通常针对近似典型社交媒体变换的后处理流程进行评估：JPEG 压缩、调整大小、裁剪、模糊和加性噪声。扩散编辑违背了这些基准测试所依据的假设。它是一种生成式变换：它可以凭空生成细节、重新合成纹理，并在保持照片级真实感的同时修改语义。因此，对传统失真具有鲁棒性的主张并不能延伸到扩散编辑。

我们的分析表明，扩散编辑会造成一种“水印瓶颈”，类似于表征学习中的信息瓶颈：加噪与去噪会将输入压缩为流形对齐表征，除非被显式保留，否则低幅度残差信号会被丢弃。

8.2. 抗扩散水印的设计指南

我们总结了受理论和实证趋势启发的技术指南。

优先考虑语义不变性，而非像素不变性。 Zhao等人（Zhao等人，2024b）认为，纯像素级水印可能会在保持语义相似性的同时被再生成移除。我们的界限同样表明，随着扩散加噪强度提高，载荷互信息会降低。一种很有前景的方向是设计与语义特征相对应的水印信号，这些特征在编辑后可能得以保留，例如稳定的中层表征或与去噪先验相契合的多尺度嵌入。

将水印技术融入生成过程。扩散原生指纹，例如 Tree-Ring 水印（Wen 等人，2023年），将信息嵌入采样的初始噪声中，并通过反演进行检测。这类方法天然契合扩散轨迹，可能比事后扰动更能经受扩散编辑；不过，对于会改变条件控制的编辑，其鲁棒性仍是一个悬而未决的问题。

使用扩散编辑增强和模型多样性进行训练。 一种实用的防御方法是在训练期间将扩散编辑纳入噪声层（表12），并从多样化的编辑器和强度中进行取样。然而，这可能会带来较高的计算成本，并有过拟合特定编辑器的风险，尤其是随着扩散架构不断演进（基于 UNet 的 LDM 与 DiT/整流流骨干网络（Zhou 等人，2025c；Lu 等人，2025b））。

针对编辑区域采用局部化或组合式水印。 局部化方法（Sander 等人，2024年）旨在支持部分来源溯源和组合式编辑。对于仅修改局部区域的扩散编辑器，局部化水印可以减少“全局失效”模式，即小范围编辑破坏全局分布式信号的情况。然而，扩散耦合仍可能影响掩码之外的区域，局部化方法还必须处理融合和边界伪影。

平衡鲁棒性、不可感知性和用户效用。提高水印强度 γ 可以增强鲁棒性，但可能降低图像质量，并可能干扰编辑

保真度。TrustMark（Bui 等人，2025年）通过缩放因子明确呈现各种权衡；在实践中，可能也需要类似的、由用户控制的权衡。从部署角度来看，一种稳妥的策略是优先降低假阳性，并接受某些编辑变换会使水印证据失效。

8.3. 互补的来源追溯机制

不可见水印的鲁棒性局限，促使人们采用不单纯依赖像素级信号的互补来源追溯机制。一种重要方法是为媒体文件附加经密码学签名的来源元数据，以便下游验证其来源和编辑历史。C2PA 技术规范是内容凭证和来源声明领域广受讨论的一项标准化工作（内容来源与真实性联盟（C2PA），2024年）。尽管平台重新编码过程中可能会剥除或丢失元数据，但元数据与清单之间的密码学绑定提供了不同于隐形扰动的安全模型。实践中，混合系统可以结合 (i) 用于高完整性流程的基于元数据的来源追溯，以及 (ii) 用于元数据受损情形的持久内容内信号（水印或指纹）。我们的分析表明，基于扩散的编辑应被视为对任何内容内信号的“压力测试”，而来源追溯系统应明确说明失效模式，避免夸大取证确定性。

8.4. 伦理考量

水印技术用于问责，但也会引发伦理问题。强水印可能被用于未经同意追踪内容创作者，从而引发隐私问题。相反，较弱或容易移除的水印可能会损害取证可靠性，还可能让人产生虚假的安全感。扩散编辑使情况更加复杂：良性编辑可能会抹去水印，而恶意行为者则可能利用相同的工具。我们强调，我们的评估方案旨在用于防御性评估和改进水印设计；不应将其解读为移除水印的建议。

8.5. 局限性

本文有五个显著局限。首先，我们的实证结果仅具说明性；我们并未在这份生成的文稿中执行完整的基准测试。其次，扩散编辑器和水印实现迅速演进；任何固定的基准测试都可能过时。第三，我们的理论分析采用了简化假设（加性水印模型、理想化的流形收缩）。第四，显式优化以去除水印的攻击（如（Ni 等人，2025年；Guo 等人，2026）中所研究的那样）可能强于本文所强调的非故意编辑设定。第五，防御机制若将水印技术集成到生成模型中，则可能与某些现实世界的来源追溯要求不兼容，例如事后为任意图像添加水印。

9. 结论

基于扩散的图像编辑已成为无处不在的后处理基本操作，但它给鲁棒不可见水印带来了根本性挑战。通过将扩散编辑视为随机生成变换，我们从理论上表明，水印信噪比和互信息会随加噪强度增加而迅速衰减；从实证上则发现，标准编辑工作流程会使比特恢复率趋近于随机猜测。这些发现促使我们从追求“对后处理鲁棒”的水印，转向能够适应生成先验或保留语义信息、耐受扩散编辑的设计。我们希望这项综述能够阐明水印退化背后的机制，并助力生成式媒体生态系统中的来源追溯工具变得更加可靠。

参考文献

Bi, J., Wang, Y., Chen, H., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. LLaVA 引导：通过模态线性表征引导，以少 500 倍的参数进行视觉指令调优。载于 Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T.（编），第 *63* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文），第 15230–15250 页，奥地利维也纳，2025年7月。计算语言学协会。国际标准书号 979-8-89176-251-0。网址：h
ttps://aclanthology.org/2025.acl-long.739/。

Bi, J., Wang, Y., Yan, D., Aniri, Huang, W., Jin, Z., Ma, X., Hecker, A., Ye, M., Xiao, X., Schuetze, H., Tresp, V., and Ma, Y. Prism：用于免训练多模态数据选择的自剪枝内在选择方法，2025b。网址：https://arxiv.org/abs/2502.12119。

Bi, J., Yan, D., Wang, Y., Huang, W., Chen, H., Wan, G., Ye, M., Xiao, X., Schuetze, H., Tresp, V., 等。Cot-kinetics：评估 LRM 推理过程的理论建模。*arXiv* 预印本 *arXiv:2505.13408*，2025c。

Brooks, T., Holynski, A., 和 Efros, A. A. InstructPix2Pix：学习遵循图像编辑指令。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023年。

Bui, T., Kumar, P., Lee, H.-Y., Ross, D. A., 和 Urbanek, S. RoSteALS：利用自编码器潜在空间的鲁棒隐写术。*arXiv* 预印本 *arXiv:2304.03400*，2023年。CVPRW 2023年。

Bui, T., Cooper, D., Collomosse, J., Bell, M., Green, A., Sheridan, J., Higgins, J., Das, A.,Keller, J., Th´ereaux, O.,

等。Trustmark：适用于任意分辨率图像的鲁棒水印与水印移除。载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2025年。

Chen, H., Li, H., Zhang, Y., Bi, J., Zhang, G., Zhang, Y., Torr, P., Gu, J., Krompass, D. 和 Tresp, V. Fedbip：利用个性化潜在扩散模型进行异构单轮联邦学习。载于计算机视觉与模式识别会议（*CVPR*）论文集，第 30440–30450 页，2025年6月a。

Chen, H., Zhang, Y., Bi, Y., Zhang, Y., Liu, T., Bi, J., Lan, J., Gu, J., Grosser, C., Krompass, D. 等。机器反学习是否真正移除了模型知识？大语言模型反学习审计框架。*arXiv* 预印本 *arXiv:2505.23270*，2025年b。

Argoverse2 场景挖掘挑战技术报告：迭代误差校正与空间感知提示，2025。网址 https://arxiv.org/abs/2506.11124。

Chen, Z., Hu, Y., Li, Z., Fu, Z., Song, X.与Nie, L. Offset：面向组合图像检索的基于分割的焦点转移修正。载于*ACM*国际多媒体会议论文集，第6113–6122页，2025c。

Chen, Z., Hu, Y., Li, Z., Fu, Z., Wen, H.与Guan, W. Hud：面向组合视频检索的分层不确定性感知消歧网络。载于*ACM*国际多媒体会议论文集，第6143–6152页，2025d。

内容来源与真实性联盟（C2PA）。C2PA 技术规范。2.1版，2024年。规范PDF可从C2PA存储库获取。

Conde, M. V., Lei, Z., Li, W., Katsavounidis, I., Timofte, R., Yan, M., Liu, X., Wang, Q., Ye, X., Du, Z.等。压缩AVIF图像的实时4K超分辨率。AIS 2024挑战赛综述。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第5838–5856页，2024年。

Couairon, G., Verbeek, J., Schwenk, H. 和 Schmid, C. Diffedit：基于扩散模型、由掩码引导的语义图像编辑。*arXiv* 预印本 *arXiv:2210.11427*，2022年。

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. 和 Fei-Fei, L. ImageNet：大型分层图像数据库。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2009年。

Di, X., Peng, L., Xia, P., Li, W., Pei, R., Cao, Y., Wang, Y., and Zha, Z.-J. Qmambabsr：采用查询状态空间模型的突发图像超分辨率。发表于计算机视觉与模式识别会议论文集，第 23080–23090 页，2025年。

Du, Z., Peng, L., Wang, Y., Cao, Y., and Zha, Z.-J. Fc3dnet：一种用于高效去摩尔纹的全连接编码器-解码器。发表于*2024*年 *IEEE* 国际图像处理会议（*ICIP*），第 1642–1648 页。IEEE，2024年。Fang, S., Peng, L., Wang, Y., Wei, R., and Wang, Y. 深度协同Mamba结合记忆专家，实现全天候图像反射分离。*arXiv* 预印本 *arXiv:2601.00322*，2026。

Feng, Y., Han, W., Jin, T., Zhao, Z., Wu, F., Yao, C., Chen, J. 等。通过大型自我中心视角视频数据集探索具身情绪。神经信息处理系统进展，37：118182–118197，2024年。

Feng, Z., Peng, L., Di, X., Guo, Y., Li, W., Zhang, Y., Pei, R., Wang, Y., Cao, Y. 和Zha, Z.-J. PMQ-VE：面向视频增强的渐进式多帧量化。*arXiv* 预印本 *arXiv:2505.12266*，2025年。

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。在 *IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2023年。arXiv 预印本 arXiv:2303.15435。

Fu, Z., Li, Z., Chen, Z., Wang, C., Song, X., Hu, Y., 和 Nie, L. PAIR：面向组合图像检索的互补性引导解耦。在 *IEEE*国际声学、语音与信号处理会议论文集，第1–5页。IEEE，2025年。

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., 和 Cohen-Or, D. 一张图胜过千言万语：使用文本反演个性化文本到图像生成。*arXiv*预印本 *arXiv:2208.01618*，2022年。

Gao, D., Lu, S., Walters, S., Zhou, W., Chu, J., Zhang, J., Zhang, B., Jia, M., Zhao, J., Fan, Z. 和 Zhang, W. Eraseanything：在整流流变换器中实现概念擦除。*arXiv* 预印本 *arXiv:2412.20413*，2024年。国际机器学习大会 2025年。

Gong, Y., Huang, L. 和 Chen, L. 用偏差消除偏差进行数据增强及一种通用的多模态数据学习方法。*arXiv*预印本 *arXiv:2101.08533*，2021年。

Gong, Y., Huang, L. 和 Chen, L. 基于颜色攻击和联合防御的行人重识别方法。在*CVPR，2022*，第 4313–4322 页，2022年。

Gong, Y., Hou, Y., Wang, Z., Lin, Z. 和 Jiang, M. 面向稀疏数据的神经偏微分方程求解器对抗性学习。*arXiv* 预印本 *arXiv:2409.02431*，2024a。

Gong, Y.、Hou, Y.、Zhang, C. 和 Jiang, M. 超越增强：在极端

捕获环境下赋能模型鲁棒性。在*2024*年国际神经网络联合会议（*IJCNN*），第 1–8 页。IEEE，2024b。

Gong, Y., Li, J., Chen, L. 和 Jiang, M. 探索通过图像级集成学习实现颜色不变性。*arXiv* 预印本 *arXiv:2401.10512*，2024c。

Gong, Y., Zhong, Z., Qu, Y., Luo, Z., Ji, R. 和 Jiang, M. 用于人员重识别的跨模态扰动协同攻击。神经信息处理系统进展，37：23352–23377，2024d。

Gong, Y., Lan, S., Yang, C., Xu, K. 和 Jiang, M. Strusr：借助物理信息泰勒引导的结构感知符号回归。*arXiv* 预印本 *arXiv:2510.06635*，2025年。

Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G. 等。SynthID-Image：互联网规模的图像水印技术。*arXiv* 预印本 *arXiv:2510.09263*，2025年。

Guo, F., Kang, J., Ming, Q., Davis, E. 和 Carter, F. 水印消失：基于扩散的图像编辑会破坏鲁棒不可见水印技术。*arXiv* 预印本 *arXiv:2602.20680*，2026年。

Han, W., Lin, W., Hu, L., Dai, Z., Zhou, Y., Li, M., Liu, Z., Yao, C. 和 Chen, J. 通过概念图引导的知识迁移进行对比式跨课程知识追踪。*arXiv* 预印本 *arXiv:2505.13489*，2025a。Han, W., Lin, W., Zhou, Y., Liu, Q., Wang, S., Yao, C. 和 Chen, J. 展示与润色：人脸视频修复中的参考引导的身份保持。*arXiv* 预印本 *arXiv:2507.10293*，2025b。Han, W., Zeng, Z., Huang, J., Jiang, S., Zheng, L., Yang, L., Qiu, H., Yao, C., Chen, J. 和 Ma, L. Uitron-speech：基于语音指令的自动化图形用户界面智能体。*arXiv* 预印本 *arXiv:2506.11127*，2025c。He, C., Li, K., Xu, G., Yan, J., Tang, L., Zhang, Y., Wang, Y. 和 Li, X. Hqg-net：利用高质量引导进行无配对医学图像增强。*TNNLS*，2023a。He, C., Li, K., Xu, G., Zhang, Y., Hu, R., Guo, Z. 和 Li, X. 用于异质图像融合的抗退化展开网络。载于*ICCV*，第12611–12621页，2023b。He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z. 和 Li, X. 基于特征分解与边缘重建的伪装目标检测。载于*CVPR*，第22046–22055页，2023c。He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z. 和 Li, X. 弱监督隐蔽目标分割：基于SAM的伪标签与多尺度特征分组。*NeurIPS*，36，2024a。

He, C., Li, K., Zhang, Y., Zhang, Y., Guo, Z., Li, X., Danell- jan, M., and Yu, F. 策略性猎物造就敏锐的捕食者：通过生成伪装目标增强伪装目标检测器。*ICLR*，2024b。

He, C., Fang, C., Zhang, Y., Ye, T., Li, K., Tang, L., Guo, Z., Li, X., and Farsiu, S. Reti-diff：基于视网膜皮层理论的潜在扩散模型进行光照退化图像复原。*ICLR*，2025a。

He, C., Li, K., Zhang, Y., Yang, Z., Tang, L., Zhang, Y., Kong, L., and Farsiu, S. 在不完全监督下分割隐蔽目标。*IEEE*模式分析与机器智能汇刊，2025b。

He, C., Shen, Y., Fang, C., Xiao, F., Tang, L., Zhang, Y., Zuo, W., Guo, Z., and Li, X. 低层视觉中的扩散模型：综述。*IEEE*模式分析与机器智能汇刊，2025c。

He, C., Xiao, F., Zhang, R., Fang, C., Fan, D.-P., and Farsiu, S. 用于隐蔽视觉感知并结合生成式细化的可逆展开网络。*arXiv*预印本 *arXiv:2508.15027*，2025d。He, C., Zhang, R., Chen, Z., Yang, B., Fang, C., Lin, Y., Xiao, F., and Farsiu, S. Unfoldldm：基于潜在扩散先验的深度展开盲图像复原。*arXiv*预印本 *arXiv:2511.18152*，2025e。He, C., Zhang, R., Tang, L., Yang, Z., Li, K., Fan, D.-P., and Farsiu, S. Scaler：面向标签不足的隐蔽目标分割的 SAM 增强协同学习。*arXiv*预印本 *arXiv:2511.18136*，2025f。He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., and Farsiu, S. Unfoldir：重新思考光照退化图像复原中的深度展开网络。*arXiv*预印本 *arXiv:2505.06683*，2025g。He, C., Zhang, R., Xiao, F., Fang, C., Tang, L., Zhang, Y., Kong, L., Fan, D.-P., Li, K., and Farsiu, S. Run：用于隐蔽目标分割的可逆展开网络。国际机器学习大会，2025h。He, C., Zhang, R., Zhang, D., Xiao, F., Fan, D.-P., and Farsiu, S. 面向真实场景隐蔽目标分割的嵌套展开网络。*arXiv*预印本 *arXiv:2511.18164*，2025i。He, C., Zhang, R., Xiao, F., Zhang, D., Cao, Z., and Farsiu, S. 通过课程选择和反课程促进细化上下文纠缠内容分割。*arXiv*预印本 *arXiv:2602.01183*，2026。

He, Y., Jiang, A., Jiang, L., Peng, L., Wang, Z., and Wang, L. 通过空间-频率交互实现的双路径耦合图像去雨网络。发表于*2024*年 *IEEE* 图像处理国际会议（*ICIP*），第1452–1458页。IEEE，2024c。

He, Y., Peng, L., Wang, L., and Cheng, J. 面向单张图像去雨的潜在退化表示约束。发表于*ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议*(ICASSP)*，第 3155–3159 页。IEEE，2024d。

He, Y., Peng, L., Yi, Q., Wu, C., and Wang, L. 基于状态空间模型的图像复原多尺度表示学习。*arXiv*预印本 *arXiv:2408.10145*，2024e。

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. 使用交叉注意力控制进行 Prompt-to-Prompt 图像编辑。*arXiv*预印本 *arXiv:2208.01626*，2022年。

Ho, J., Jain, A., and Abbeel, P. 去噪扩散概率模型. 载于神经信息处理系统进展（*NeurIPS*），2020年。

Huang, Q., Chen, Z., Li, Z., Wang, C., Song, X., Hu, Y., and Nie, L. Median：用于组合图像检索的自适应中间粒度聚合网络. 载于*IEEE*国际声学、语音与信号处理会议论文集，第1–5页。IEEE，2025年。

Ignatov, A., Perevozchikov, G., Timofte, R., Pan, W., Wang, S., Zhang, D., Ran, Z., Li, X., Ju, S., Zhang, D., 等. 移动GPU上的RGB照片增强, Mobile AI 2025挑战赛：报告. 载于计算机视觉与模式识别会议论文集，第1922–1933页，2025年。

Jiang, A.、Wei, Z.、Peng, L.、Liu, F.、Li, W.和Wang, M. Dalpsr：利用退化对齐的语言提示进行真实世界图像超分辨率。*arXiv*预印本 *arXiv:2406.16477*，2024年。

Jiang, K.、Jiang, H.、Jiang, N.、Gao, Z.、Bi, J.、Ren, Y.、Li, B.、Du, Y.、Liu, L.和Li, Q. Kore：通过面向知识的增强操作与约束增强大型多模态模型的知识注入，2025a。网址 https://arxiv.org/abs/2510.19316。

Jiang, K., Jiang, N., Du, Y., Ren, Y., Li, Y., Gao, Y., Bi, J., Ma, Y., Liu, Q., Wang, X., Jia, Y., Jiang, H., Hu, Y., Li, B., and Liu, L. Mined：面向大型多模态模型的多模态时效性知识探测与更新，2025b。URLhttps://arxiv.org/abs/2510.19457。

JiangYong Yu, Sifan Zhou, D. Y. S. L. S. W. X. H. C. X. Z. X. C. S. and Yuan., Z. Mquant：通过静态量化释放多模态大语言模型的推理潜力。载于第*33*届*ACM*国际多媒体会议论文集，2025年。Jin, X., Guo, C., Li, X., Yue, Z., Li, C., Zhou, S., Feng, R., Dai, Y., Yang, P., Loy, C. C., 等。MIPI 2024少样本原始图像去噪挑战赛：方法与结果。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1153–1161页，2024年。

Karras, T.、Aittala, M.、Aila, T.和Laine, S. 阐明基于扩散的生成模型的设计空间。神经信息处理系统进展（*NeurIPS*），2022年。arXiv预印本 arXiv:2206.00364。

Kudrat, D., Xie, Z., Sun, Y., Jia, T., 和 Hu, Q. 用于时间序列预测的分块结构损失。载于第四十二届国际机器学习会议，2025年。

Li, L., Fu, Z., Carter, F., 和 Zhang, B. 让你回归正轨：自动引导去噪轨迹，避开不需要的概念。*arXiv* 预印本 *arXiv:2504.12782*，2025a。

Li, Y., Zhang, Y., Timofte, R., Van Gool, L., Yu, L., Li, Y., Li, X., Jiang, T., Wu, Q., Han, M. 等。NTIRE 2023 高效超分辨率挑战赛：方法与结果。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1922–1960页，2023年。

Li, Z., Chen, Z., Wen, H., Fu, Z., Hu, Y., 和 Guan, W. Encoder：用于组合图像检索的实体挖掘与修改关系绑定。载于*AAAI*人工智能会议论文集，第39卷，第5101–5109页，2025b。

Li, Z., Fu, Z., Hu, Y., Chen, Z., Wen, H. 和 Nie, L. Finecir：为组合图像检索显式解析细粒度修改语义。*https://arxiv.org/abs/2503.21309*，2025c。

Lin, J., Zhenzhong, W., Dejun, X., Shu, J., Gong, Y. 和 Ji ang, M. Phys4dgen：一种物理驱动框架，可基于单张图像实现可控且高效的4D内容生成。*arXiv* 预印本 *arXiv:2411.16800*，2024年。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P. 和 Zitnick, C. L. Microsoft COCO：上下文中的常见物体。载于欧洲计算机视觉会议（*ECCV*）论文集，2014。

Ling, P., Chen, L., Zhang, P., Chen, H., Jin, Y., Zheng, J., 等。Freedrag：用于可靠点式图像编辑的特征拖动。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024年。arXiv预印本 arXiv:2307.04684。

Liu, X., Qiu, X., Wu, X., Li, Z., Guo, C., Hu, J. 和 Yang, B. 重新思考不规则时间序列预测：一种简单而有效的基线。*arXiv*预印本 *arXiv:2505.11250*，2025年。

Lu, J.、Li, X.和Han, K. RegionDrag：利用扩散模型进行快速基于区域的图像编辑。载于《欧洲计算机视觉会议（*ECCV*）论文集》，2024a。arXiv预印本 arXiv:2407.18247。

Lu, R.、Bi, J.、Ma, Y.、Xiao, F.、Du, Y.和Tian, Y. Mv-debate：面向社交媒体中多模态有害内容检测的动态反思门控多视角智能体辩论，2025a。URL https://arxiv.org/abs/2508.05557。

Lu, S.、Liu, Y.和Kong, A. W.-K. TF-ICON：基于扩散的免训练跨域图像合成。载于《*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集》，2023年。

Lu, S., Wang, Z., Li, L., Liu, Y., and Kong, A. W.-K. Mace：扩散模型中的大规模概念擦除。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024b。arXiv 预印本 arXiv:2403.06135。

Lu, S., Lian, Z., Zhou, Z., Zhang, S., Zhao, C., and Kong, A. W.-K. FLUX 是否已经懂得如何进行物理上合理的图像合成? *arXiv* 预印本 *arXiv:2509.21278*，2025b。介绍 SHINE。

Lu, S., Zhou, Z., Lu, J., Zhu, Y., and Kong, A. W.-K. 利用生成先验抵御图像编辑的鲁棒水印技术：从基准测试到进展。国际学习表征会议（*ICLR*），2025c。arXiv 预印本 arXiv:2410.18775。

Lyu, J., Zhao, M., Hu, J., Huang, X., Chen, Y., 和 Du, S. Vadmamba：探索用于快速视频异常检测的状态空间模型，2025年。网址https://arxiv.org/abs/2503.21169。

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., 和 Ermon, S. SDEdit：使用随机微分方程进行引导式图像合成与编辑。*arXiv* 预印本 *arXiv:2108.01073*，2021年。ICLR 2022。

Mokady, R., Hertz, A.,Aberman,K., Pritch,Y.,和 Cohen- Or, D. 使用引导扩散模型编辑真实图像的空文本反演。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2023年。

Ni, Y.、Carter, F.、Niu, Z.、Davis, E. 和 Zhang, B. 基于扩散的图像编辑：破解鲁棒水印。*arXiv*预印本 *arXiv:2510.05978*，2025年。

Niu, W.、Xie, Z.、Sun, Y.、He, W.、Xu, M. 和 Hao, C. Lang-time：一种利用近端策略优化进行时间序列预测的语言引导统一模型。载于第四十二届国际机器学习会议，2025年。Oquab, M.、Darcet, T.、Moutakanni, T.、Vo, H.、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.、Rabbat, M.、Sharma, V.、Synnaeve, G.、Xu, H.、J´egou, H.、Mairal, J.、Labatut, P.、J oulin, A. 和 Bojanowski, P. DINOv2：无需监督学习鲁棒的视觉特征。*arXiv* 预印本 *arXiv:2304.07193*，2023年。

Pan, J., Liu, Y., He, X., Peng, L., Li, J., Sun, Y., 和 Huang, X. Enhance thensearch：一种利用基础模型进行跨域少样本目标检测的增强搜索策略。载于计算机视觉与模式识别会议论文集，第1548–1556页，2025年。

Pan, X., Wang, B., Yang, W., Cai, J., andFeng, J. 拖动你的 GAN：在生成图像流形上进行交互式基于点的操控。*arXiv* 预印本 *arXiv:2305.10973*，2023年。

Peng, L., Li, W., Guo, J., Di, X., Sun, H., Li, Y., Pei, R., Wang, Y., Cao, Y., and Zha, Z.-J. 利用原始数据提升真实世界超分辨率：一种新视角、一个数据集和一个基线。

Peng, L., Jiang, A., Yi, Q., and Wang, M. 用于单幅图像去雨的累积雨密度感知网络。*IEEE*信号处理快报，27:406–410，2020年。

Peng, L., Jiang, A., Wei, H., Liu, B., and Wang, M. 通过渐进式结构增强约束实现的集成单幅图像去雨网络。信号处理：图像通信，99:116460，2021年。

Peng, L., Cao, Y., Pei, R., Li, W., Guo, J., Fu, X., Wang, Y. 和 Zha, Z.-J. 基于自适应方向梯度卷积的高效真实场景图像超分辨率。*arXiv*预印本 *arXiv:2405.07023*，2024a。

Peng, L., Cao, Y., Sun, Y. 和 Wang, Y. 用于压缩图像分类的轻量级自适应特征去漂移方法。*IEEE*多媒体汇刊，26:6424–6436，2024b。

Peng, L., Li, W., Guo, J., Di, X., Sun, H., Li, Y., Pei, R., Wang, Y., Cao, Y. 和 Zha, Z.-J. 揭示隐藏细节：一种用于真实场景超分辨率的原始数据增强范式。*arXiv*预印本 *arXiv:2411.10798*，2024c。

Peng, L., Li, W., Pei, R., Ren, J., Xu, J., Wang, Y., Cao, Y. 和 Zha, Z.-J. 面向真实世界超分辨率的逼真数据生成。*arXiv* 预印本 *arXiv:2406.07255*，2024d。

Peng, L., Di, X., Feng, Z., Li, W., Pei, R., Wang, Y., Fu, X., Cao, Y. 和 Zha, Z.-J. 将 Mamba 引导至复杂纹理：一种用于图像复原的高效纹理感知状态空间模型。*arXiv* 预印本 *arXiv:2501.16583*，2025a。

Peng, L., Wang, Y., Di, X., Fu, X., Cao, Y., Zha, Z.-J., 等。通过中心-周围协同卷积提升图像去雨效果。载于*AAAI*人工智能会议论文集，第39卷，第6470–6478页，2025b。

Peng, L., Wu, A., Li, W., Xia, P., Dai, X., Zhang, X., Di, X., Sun, H., Pei, R., Wang, Y., 等。从像素到高斯：通过二维高斯建模实现超快速连续超分辨率。*arXiv*预印本 *arXiv:2503.06617*，2025c。Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., M¨uller, J., Penna, J., 和 Rombach, R. SDXL：改进潜在扩散模型，实现高分辨率图像合成。*arXiv*预印本 *arXiv:2307.01952*，2023年。Qi, X., Li, R., Peng, L., Ling, Q., Yu, J., Chen, Z., Chang, P., Han, M., 和 Xiao, J. 利用扩散模型进行无数据知识蒸馏。*arXiv*预印本 *arXiv:2504.00870*，2025年。Qiu, X., Hu, J., Zhou, L., Wu, X., Du, J., Zhang, B., Guo, C., Zhou, A., Jensen, C. S., Sheng, Z., 和 Yang, B. TFB：面向时间序列预测方法的全面、公平基准测试。载于*VLDB*学报，第2363–2377页，2024年。Qiu, X., Cheng, H., Wu, X., Hu, J., 和 Guo, C. 深度学习在多变量时间序列预测中的全面综述：通道策略视角。*arXiv*预印本 *arXiv:2502.10721*，2025a。Qiu, X., Li, Z., Qiu, W., Hu, S., Zhou, L., Wu, X., Li, Z., Guo, C., Zhou, A., Sheng, Z., Hu, J., Jensen, C. S., 和 Yang, B. Tab：时间序列异常检测方法的统一基准测试。载于*VLDB*学报，2025b。Qiu, X., Wu, X., Cheng, H., Liu, X., Guo, C., Hu, J., 和 Yang, B. Dbloss：用于时间序列预测的基于分解的损失函数。载于*NeurIPS*，2025c。Qiu, X., Wu, X., Lin, Y., Guo, C., Hu, J., 和 Yang, B. DUET：通过双重聚类增强多变量时间序列预测。载于*SIGKDD*，第1185–1196页，2025d。Qiu, X., Zhu, Y., Li, Z., Cheng, H., Wu, X., Guo, C., Yang, B., 和 Hu, J. DAG：一种利用外生变量进行时间序列预测的双重因果网络。*arXiv*预印本 *arXiv:2509.14933*，2025e。

Qu, Y., Fu, D., and Fan, J. 领域偏移下用于新颖性检测的主体信息提取。*arXiv*预印本 *arXiv:2504.21247*，2025a。Qu, Y., Panariello, M., Todisco, M., and Evans, N. 语音中的无参考对抗性性别混淆。*arXiv*预印本 *arXiv:2508.02295*，2025b。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. 通过自然语言监督学习可迁移视觉模型。*arXiv*预印本 *arXiv:2103.00020*，2021年。Ren, B., Li, Y., Mehta, N., Timofte, R., Yu, H., Wan, C., Hong, Y., Han, B., Wu, Z., Zou, Y., et al. 第九届NTIRE 2024高效超分辨率挑战赛报告。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6595–6631页，2024a。Ren, J., Li, W., Chen, H., Pei, R., Shao, B., Guo, Y., Peng, L., Song, F., and Zhu, L. UltraPixel：推动超高分辨率图像合成迈向新高峰。神经信息处理系统进展，37:111131–111171，2024b。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and O mmer, B. 使用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2022年。Rong, X., Huang, W., Liang, J., Bi, J., Xiao, X., Li, Y., Du, B., and Ye, M. 无需外部指导的多模态大语言模型微调后门清除。*arXiv*预印本 *arXiv:2505.16916*，2025年。Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. DreamBooth：微调文生图扩散模型以实现主体驱动生成。*arXiv*预印本 *arXiv:2208.12242*，2022年。Sander, T., Fernandez, P., Park, T., Novotny, D., Zhang, R., Isola, P., and Owens, A. 使用局部化消息实现万物水印。*arXiv*预印本 *arXiv:2411.07231*，2024年。ICLR 2025。Shi, Y., Wang, W., Song, J., Meng, C., and Ermon, S. DragDiffusion：利用扩散模型实现交互式基于点的图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024年。arXiv预印本 arXiv:2306.14435。Shin, J., Choi, D., and Park, J. InstantDrag：提升拖拽式图像编辑的交互性。*arXiv*预印本 *arXiv:2409.08857*，2024年。

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. 通过随机微分方程实现基于分数函数的生成建模。载于国际学习表征会议（*ICLR*），2021年。

Sun, H., Li, W., Liu, J., Zhou, K., Chen, Y., Guo, Y., Li, Y., Pei, R., Peng, L., 和 Yang, Y. 文本提升泛化能力：一种用于真实世界图像复原的即插即用图像描述生成器。

Sun, H., Li, W., Liu, J., Zhou, K., Chen, Y., Guo, Y., Li, Y., Pei, R., Peng, L., 和 Yang, Y. 超越像素：文本增强真实世界图像复原中的泛化能力。*arXiv*预印本 *arXiv:2412.00878*，2024年。Sun, Y., Xie, Z., Chen, Y., Huang, X., 和 Hu, Q. 利用二维注意力机制预测太阳风速度。《空间天气》，19(7):e2020SW002707，2021年。Sun, Y., Xie, Z., Chen, Y., 和 Hu, Q. 利用多模态信息准确预测太阳风速度。《空间：科学与技术》，2022年。Sun, Y., Eldele, E., Xie, Z., Wang, Y., Niu, W., Hu, Q., Kwoh, C. K., 和 Wu, M. 通过时间异质性建模和语义对齐，使大语言模型适应时间序列预测。*arXiv*预印本 *arXiv:2508.07195*，2025a。Sun, Y., Xie, Z., Chen, D., Eldele, E., 和 Hu, Q. 用于时间序列预测的分层分类辅助网络。载于*AAAI*人工智能会议论文集，第39卷，第20743–20751页，2025b。Sun, Y., Xie, Z., Eldele, E., Chen, D., Hu, Q., 和 Wu, M. 学习模式专属专家，以应对补丁级分布偏移下的时间序列预测。神经信息处理系统进展，2025c。Sun, Y., Xie, Z., Xing, H., Yu, H., 和 Hu, Q. Ppgf：概率模式引导的时间序列预测。*IEEE*神经网络与学习系统汇刊，2025d。Tancik, M., Mildenhall, B., Wang, T., Agarwal, D., Srinivasan, P. P., Barron, J. T., 和 Ng, R. StegaStamp：实体照片中的隐形超链接。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2020年。arXiv预印本 arXiv:1904.05343。Tian, Y., Chen, S., Xu, Z., Wang, Y., Bi, J., Han, P., 和 Wang, W. 强化学习中期训练，2025年。网址：https://arxiv.org/abs/2509.24375。

Timofte, R., Agustsson, E., Gool, L. V., Yang, M.-H., Zhang, L. 等。《NTIRE 2017年单张图像超分辨率挑战：方法与结果》。载于《*IEEE* 计算机视觉与模式识别会议研讨会（*CVPRW*），2017年。介绍了DIV2K数据集。

Wang, H., Peng, L., Sun, Y., Wan, Z., Wang, Y., and Cao, Y. 面向递归低光图像增强的亮度感知。*IEEE*人工智能汇刊，5(6):3034–3045，2023a。

Wang, Y., Peng, L., Li, L., Cao, Y., and Zha, Z.-J. 用于图像曝光校正的解耦与聚合。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18115–18124页，2023b。Wang, Y., Bi, J., Ma, Y., and Pirk, S. ASCD：用于减少多模态大语言模型幻觉的注意力引导对比解码。*arXiv*预印本 *arXiv:2506.14766*，2025a。Wang, Y., Liang, Z., Zhang, F., Tian, L., Wang, L., Li, J., Yang, J., Timofte, R., Guo, Y., Jin, K., 等。NTIRE 2025光场图像超分辨率挑战赛：方法与结果。载于计算机视觉与模式识别会议论文集，第1227–1246页，2025b。Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. 图像质量评估：从误差可见性到结构相似性。载于*IEEE*图像处理汇刊，第13卷，第600–612页，2004年。Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-Ring 水印：用于扩散图像的隐形且鲁棒的指纹。神经信息处理系统进展（*NeurIPS*），2023年。arXiv预印本 arXiv:2305.20030。Wu, A., Peng, L., Di, X., Dai, X., Wu, C., Wang, Y., Fu, X., Cao, Y., and Zha, Z.-J. RobustGS：在低质量条件下统一增强前馈式3D高斯泼溅。*arXiv*预印本 *arXiv:2508.03077*，2025a。Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., 等。混元视频1.5技术报告。*arXiv*预印本 *arXiv:2511.18870*，2025b。Wu, C., Wang, L., Peng, L., Lu, D., and Zheng, Z. 摒弃高倍率下采样：一种超高清图像复原的新设计范式。载于*2025*年*IEEE/CVF*冬季计算机视觉应用会议（*WACV*），第2390–2399页。IEEE，2025c。Wu, H., Yang, Y., Chen, H., Ren, J., and Zhu, L. 用于视频中雨滴与雨纹联合去除的掩码引导渐进式网络。载于第*31*届*ACM*国际多媒体会议论文集，第7216–7225页，2023年。

Wu, H., Yang, Y., Aviles-Rivero, A. I., Ren, J., Chen, S., Chen, H., and Zhu, L. 通过时间解耦专家和分布驱动的对比正则化实现的半监督视频除雪网络。载于欧洲计算机视觉会议，第70–89页。施普林格，2024a。

Wu, H., Yang, Y., Xu, H., Wang, W., Zhou, J., and Zhu, L. Rainmamba：利用状态空间模型增强局部性学习的视频去雨方法。发表于第*32*届*ACM*国际多媒体会议论文集，第7881–7890页，2024b。

Wu, H., Wu, Y., Jiang, J., Wu, C., Wang, H., and Zheng, Y. Samvsr：利用语义先验实现面向区域聚焦Mamba的视频除雪方法。发表于第*33*届*ACM*国际多媒体会议论文集，第7376–7385页，2025d。

Wu, X., Qiu, X., Gao, H., Hu, J., Yang, B., and Guo, C. K^2VAE：一种用于概率时间序列预测的Koopman-Kalman增强型变分自动编码器。发表于国际机器学习大会，2025e。

Wu, X., Qiu, X., Li, Z., Wang, Y., Hu, J., Guo, C., Xiong, H., and Yang, B. CATCH：基于频率分块的通道感知多变量时间序列异常检测。在*ICLR*，2025f。Wu, Z., Kolkin, N., Brandt, J., Zhang, R., and Shechtman, E. Turboedit：即时基于文本的图像编辑。*arXiv*预印本 *arXiv:2408.08332*，2024c。Xia, P., Peng, L., Di, X., Pei, R., Wang, Y., Cao, Y., and Zha, Z.-J. S3mamba：基于可扩展状态空间模型的任意尺度超分辨率。*arXiv*预印本 *arXiv:2411.11906*，6，2024。Xiao, F., Hu, S., Shen, Y., Fang, C., Huang, J., He, C., Tang, L., Yang, Z., and Li, X. 伪装目标检测及其拓展综述。中国人工智能学会人工智能研究，2024。Xie, Z. Conquer：面向基于文本的行人搜索的上下文感知表征与查询增强。*arXiv*预印本 *arXiv:2601.18625*，2026。Xie, Z., Wang, C., Wang, Y., Cai, S., Wang, S., and Jin, T. 面向行人检索的聊天驱动文本生成与交互。在*2025*年自然语言处理实证方法会议论文集，第5259–5270页，2025年。Xie, Z., Liu, X., Zhang, B., Lin, Y., Cai, S., and Jin, T. Hvd：面向文本-视频检索的人类视觉驱动视频表征学习。*arXiv*预印本 *arXiv:2601.16155*，2026a。Xie, Z., Zhang, B., Lin, Y., and Jin, T. 深入探究：面向鲁棒视频-文本检索的层次化视觉感知。*arXiv*预印本 *arXiv:2601.12768*，2026b。

Xu, H., Peng, L., Song, S., Liu, X., Jun, M., Li, S., Yu, J., and Mao, X. Camel：面向资源受限设备的能效感知LLM 推理。*arXiv*预印本 *arXiv:2508.09173*，2025a。

Xu, J., Li, W., Sun, H., Li, F., Wang, Z., Peng, L., Ren, J., Yang, H., Hu, X., Pei, R.等。基于一致性校正流的快速图像超分辨率。载于*IEEE/CVF*国际计算机视觉会议论文集，第11755–11765页，2025b。

Yakovenko, A., Chakvetadze, G., Khrapov, I., Zhelezov, M., Vatolin, D., Timofte, R., Oh, Y., Kwon, J., Park, J., Cho, N. I.等。AIM 2025年低光照RAW视频去噪挑战赛：数据集、方法与结果。*arXiv*预印本 *arXiv:2508.16830*，2025年。

Yan, Q., Jiang, A., Chen, K., Peng, L., Yi, Q.与Zhang, C.。文本提示引导的图像复原。《人工智能工程应用》，155:110981，2025年。

Ye, H., Zhang, J., Liu, S., Han, X., 和 Yang, W. IP-Adapter：面向文生图扩散模型的文本兼容图像提示适配器。*arXiv*预印本 *arXiv:2308.06721*，2023年。Yu, X., Chen, Z., Zhang, Y., Lu, S., Shen, R., Zhang, J., Hu, X., Fu, Y., 和 Yan, S. 视觉文档理解与问答：一种采用测试时扩展的多智能体协作框架。*arXiv*预印本 *arXiv:2508.03404*，2025 年。Zeng, Z., Huang, J., Zheng, L., Han, W., Zhong, Y., Chen, L., Yang, L., Chu, Y., He, Y., 和 Ma, L. Uitron：具备先进感知与规划能力的基础图形用户界面智能体。*arXiv*预印本 *arXiv:2508.21767*，2025年。Zhang, G., Bi, J., Gu, J., Chen, Y., 和 Tresp, V. Spot!：重新审视用于事件理解的视频语言模型。*arXiv*预印本 *arXiv:2311.12919*，2023a。Zhang, L., Rao, A., 和 Agrawala, M. 为文生图扩散模型添加条件控制。*arXiv*预印本 *arXiv:2302.05543*，2023b。Zhang, R., Isola, P., Efros, A. A., Shechtman, E., 和 Wang, O. 深层特征作为感知度量的不合理有效性。收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2018年。介绍了学习感知图像块相似度。Zhang, S., Guo, Y., Peng, L., Wang, Z., Chen, Y., Li, W., Zhang, X., Zhang, Y., 和 Chen, J. VividFace：面向视频人脸增强的高质量、高效一步扩散。*arXiv*预印本 *arXiv:2509.23584*，2025年。Zhang, Z., Han, L., Ghosh, A., Metaxas, D. N., 和 Ren, J. SINE：基于文生图扩散的单图像编辑

模型。收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2023c。arXiv预印本 arXiv:2212.04489。Zhao, H., Ma, X., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., 和 Chang, B. UltraEdit：大规模基于指令的细粒度图像编辑。神经信息处理系统进展（*NeurIPS*），2024a。arXiv预印本 arXiv:2407.05282。Zhao, X., Zhang, K., Su, Z., Vasan, S., Grishchenko, I., Kruegel, C., Vigna, G., Wang, Y.-X., 和 Li, L. 不可见图像水印可被证明能够借助生成式人工智能移除。神经信息处理系统进展（*NeurIPS*），2024b。arXiv预印本 arXiv:2306.01953。Zheng, Y., Zhong, B., Liang, Q., Tang, Z., Ji, R., 和 Li, X. 通过并行交互网络利用局部与全局线索进行视觉跟踪。*IEEE*视频技术电路与系统汇刊，33(4):1671–1683，2022年。Zheng, Y., Zhong, B., Liang, Q., Li, G., Ji, R., 和 Li, X. 迈向视觉语言跟踪的统一令牌学习。*IEEE*视频技术电路与系统汇刊，34(4):2125–2135，2023年。Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., 和 Li, X. Odtrack：面向视觉跟踪的在线密集时序令牌学习。收录于*AAAI*人工智能会议论文集，第38卷，第7588–7596页，2024年。Zheng, Y., Zhong, B., Liang, Q., Li, N., 和 Song, S. 面向自监督跟踪的解耦时空一致性学习。收录于*AAAI*人工智能会议论文集，第39卷，第10635–10643页，2025a。Zheng, Y., Zhong, B., Liang, Q., Zhang, S., Li, G., Li, X., 和 Ji, R. 利用在线密集时序令牌学习迈向通用模态跟踪。*IEEE*模式分析与机器智能汇刊，2025b。Zhou, S., Li, L., Zhang, X., Zhang, B., Bai, S., Sun, M., Zhao, Z., Lu, X., 和 Chu, X. LiDAR-PTQ：面向点云三维目标检测的训练后量化。2024年。Zhou, S., Wang, S., Yuan, Z., Shi, M., Shang, Y., 和 Yang, D. GSQ-tuning：面向大语言模型端侧微调的全量化训练中的组共享指数整数。收录于计算语言学协会研究成果：*ACL 2025*，第22971–22988页，奥地利维也纳，2025年7月。计算语言学协会。国际标准书号979-8-89176-256-5。Zhou, S., Yuan, Z., Yang, D., Hu, X., Qian, J., 和 Zhao, Z. Pillarhist：一种基于高度感知直方图的量化感知柱特征编码器。收录于论文集

计算机视觉与模式识别会议，第 27336–27345 页，2025b。Zhou, Z.、Lu, S.、Leng, S.、Zhang, S.、Lian, Z.、Yu, X. 和 Kong, A. W.-K. Dragflow：借助基于区域的监督，释放 dit 先验在拖拽编辑中的潜力。*arXiv* 预印本 *arXiv:2510.02253*，2025c。ICLR 2026 海报。Zhu, J.、Kaplan, R.、Johnson, J. 和 Fei-Fei, L. Hidden：利用深度网络隐藏数据。载于欧洲计算机视觉会议（*ECCV*）论文集，2018年。arXiv 预印本 arXiv:1807.09937。