arXiv:2602.06101v1 [eess.IV] 5 Feb 2026

ALIEN：面向可控生成的解析潜在水印 n

Liangqi Lei [1] Keke Gai [1] Jing Yu [2] Liehuang Zhu [1] Qi Wu [3]

摘要

水印是一种保护知识产权、减少滥用的技术替代方案。现有方法侧重于优化带水印的潜变量，以平衡水印鲁棒性与保真度，因为潜在扩散模型（LDMs）被视为生成任务的强大工具。然而，依赖计算密集型启发式优化进行迭代信号细化，会导致训练开销高和陷入局部最优。为解决这些问题，我们提出了一种面向可控生成的解析水印框架（ALIEN）。我们首次解析推导出随时间变化的调制系数，用于引导水印残差的扩散，从而实现可控的水印嵌入模式。实验结果表明，ALIEN-Q 在 5 项质量指标上超越最先进方法 33.1%；与最先进方法相比，ALIEN-R 在 15 种不同条件下对生成式变体和稳定性威胁的鲁棒性提高了 14.0%。代码见 https://anonymous.4open.science/r/ALIEN/。

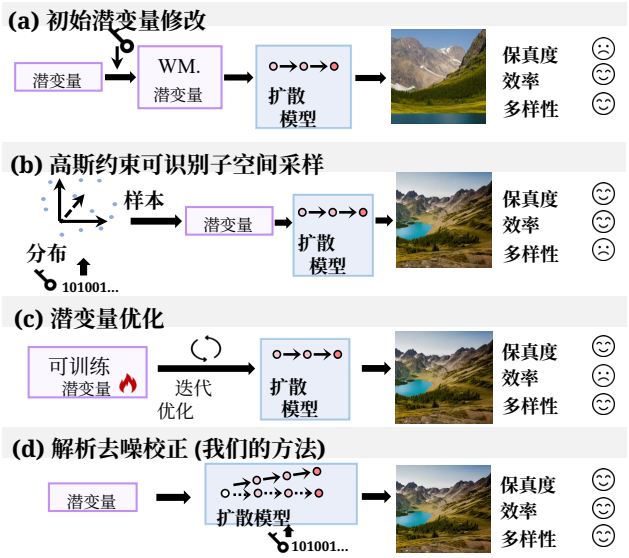


图 *1.* (a) 潜变量修改，(b) 约束采样，(c) 迭代优化，(d) 我们的 ALIEN，采用原理性嵌入。

1. 引言

文生图扩散模型，例如 Stable Diffusion（Rombach et al.，2022）和 DALL·E（Rameshet al.，2022），展现出了生成高质量图像的卓越能力。为了保护生成模型的知识产权（Gowal& Kohli，2023年）并促进滥用追踪（Barrett，2023年），各国政府日益呼吁制定法规（欧盟，2024年；拜登，2023年），以强制采用水印。对生成图像进行稳健且不可感知的水印处理，已成为关键而紧迫的研究重点。后处理水印（Cox et al.，2007）应用于扩散模型生成的内容，但其鲁棒性不足以在现实应用中实现可靠检测（Ren et al.，2024年）。一些研究考虑了模型分发场景（Rezaeiet al.，2024年；Ciet al.，2024年；Feng et al.，2024年；Wang 等人，2025 年）。研究人员对扩散模型进行微调，将水印嵌入模型参数中，这不可避免地限制了效率和可扩展性。

近期研究主要集中于语义水印（Lee & Cho，2025 年），旨在将水印信号嵌入语义或潜在特征中，以更好地抵御基于图像处理的攻击（Zhao 等人，2024年）。基于初始潜变量修改的水印方法（Wen et al.，2024年；Ci et al.，2025 年）通过直接修改初始潜变量或向其中添加扰动来嵌入水印（图 1（a））。尽管这种机制直观易懂，但初始潜空间与最终图像之间的映射高度复杂且非线性，因此直接修改容易导致语义漂移。此外，为了保持不可见性，对初始潜变量的修改幅度受到严格限制，因而难以提高水印容量。为确保高容量和无损水印嵌入，基于高斯约束可识别子空间采样的水印（Yang 等人，2024c；Gunn et al.，2024）

[1]北京理工大学，中国北京 [2]中央民族大学信息工程学院，中国北京 [3]阿德莱德大学计算机科学学院，阿德莱德，澳大利亚。通讯作者：Liangqi Lei<3120245873@bit.edu.cn>，KekeGai <gaikeke@bit.edu.cn>。

预印本。*2026*年*2*月*9*日。

结合密码学，将初始潜空间划分为互不重叠的可识别子空间，迫使用户从与特定水印信息关联的子空间中进行采样（图 1（b））。这种对初始潜空间的严格约束限制了生成多样性。基于优化的水印方法（Huang et al.，2024年；Zhang et al.，2024年）在潜空间中进行水印优化，以更好地平衡鲁棒性与保真度（图1（c））。然而，这些方法依赖计算密集型启发式优化，通过迭代寻找最优水印，导致训练开销巨大，且容易陷入局部最优。此外，语义水印的检测通常依赖扩散反演，这使其仅适用于可逆采样器（Lu etal.，2022）；而使用不可逆采样器生成图像时，水印将无法检测（Karras et al.，2022）。攻击者可利用这一漏洞，通过使用不可逆调度器重新生成图像来移除水印。（An 等人，2024年）。

尽管前述语义水印方案在鲁棒嵌入方面已取得进展，但它们回避了一个根本问题：无法从扩散模型的生成原理中推导出精确且高效的水印嵌入机制。当前的优化或基于约束的方法本质上回避了这一问题，转而依赖计算密集型优化或采样约束。这些方法固有地限制了保真度、效率和多样性。

为解决这些问题，我们提出了可控生成解析水印框架（ALIEN）。如图1(d)所示，与依赖计算密集的启发式潜变量优化或对初始潜变量采样施加约束的现有方法不同，我们从扩散模型随机微分方程的反向过程出发，首次给出了水印残差传播机制的解析推导。具体而言，给定目标水印残差，我们解析推导出一个时变调制系数，将该目标转化为噪声预测的精确校正项。通过在每个去噪步骤注入该校正项，我们有效地改变了扩散模型底层的概率流。这施加了一种确定性力，无论采样路径如何，都能无缝引导生成轨迹趋向带水印状态，从而消除对迭代优化的需求，并确保与各类采样器兼容。ALIEN实现了一种原理性的水印嵌入模式，而非依赖启发式方法。无需迭代优化，我们在去噪过程的每一步将精确调制后的水印注入噪声预测目标。

我们的贡献主要包括以下三点：(1) 首次对水印传播进行了解析推导，实现了无缝、低推理成本的水印嵌入，无需迭代优化。(2) 由于水印是基于扩散模型进行精确补偿的，ALIEN 通过对水印信号进行精确的理论补偿，保持了语义一致性和视觉保真度。(3) ALIEN 的水印嵌入机制不依赖特定的初始潜变量和采样器类型（包括不可逆采样器），解决了语义水印对可逆采样器高度依赖的问题。

2. 相关工作

扩散模型在图像生成方面表现出卓越性能（Dhariwal & Nichol, 2021），这得益于相关方法（Ho et al., 2020; Song et al., 2020b）和采样技术（Song et al., 2020a; Song & Ermon, 2020）。潜在扩散模型（LDMs）在预训练变分自编码器（VAE）的潜在空间中优化图像生成，从而进一步加速扩散模型的实际应用。在推理阶段，LDM 首先从标准高斯分布 $\mathcal{N}(0, \mathrm{I})$ 中采样得到初始潜变量 $\mathbf{z}_T \in \mathbb{R}^{c\times w\times h}$，其中 T表示扩散模型的总时间步数。经过迭代去噪后，潜在向量逐步演变为无噪声表示 $\mathbf{z}_0$。VAE解码器根据 $\mathbf{z}_0$ 重建出最终图像 $\mathbf{x}_0$。保护生成内容的知识产权并防止其遭到滥用，已成为重要的研究重点。本文聚焦于将水印融入扩散过程的现有方法在保真度、效率和可控性方面的关键问题。

潜在扩散模型中的水印旨在追踪内容来源并确保生成内容的可追责性。关于现有水印方法分类的全面背景介绍，请参阅附录D。我们重点关注在扩散过程中集成水印的既有工作。Tree-Ring（Wen etal.，2024年）和 RingID（Ci et al.，2025 年）等潜变量修改水印将水印嵌入初始潜变量的傅里叶空间，从而导致语义漂移。Gaussian Shading（Yang 等人，2024c）、GaussMarker（Liet al.，2025 年）和 PRC（Gunnet al.，2024年）等受限采样水印利用密码学原理修改采样模式，但严格的约束限制了采样范围，导致可控性降低。Zodiac（Zhang et al.，2024年）和 Robin（Huangetal.，2024年）等基于优化的方法通过优化潜变量来提升语义一致性，但依赖启发式优化会带来计算开销。上述方法的检测依赖扩散反演，因此仅适用于可逆采样器，无法有效用于不可逆采样器。我们的方法通过解析推导，实现了低开销且普遍适用的水印。

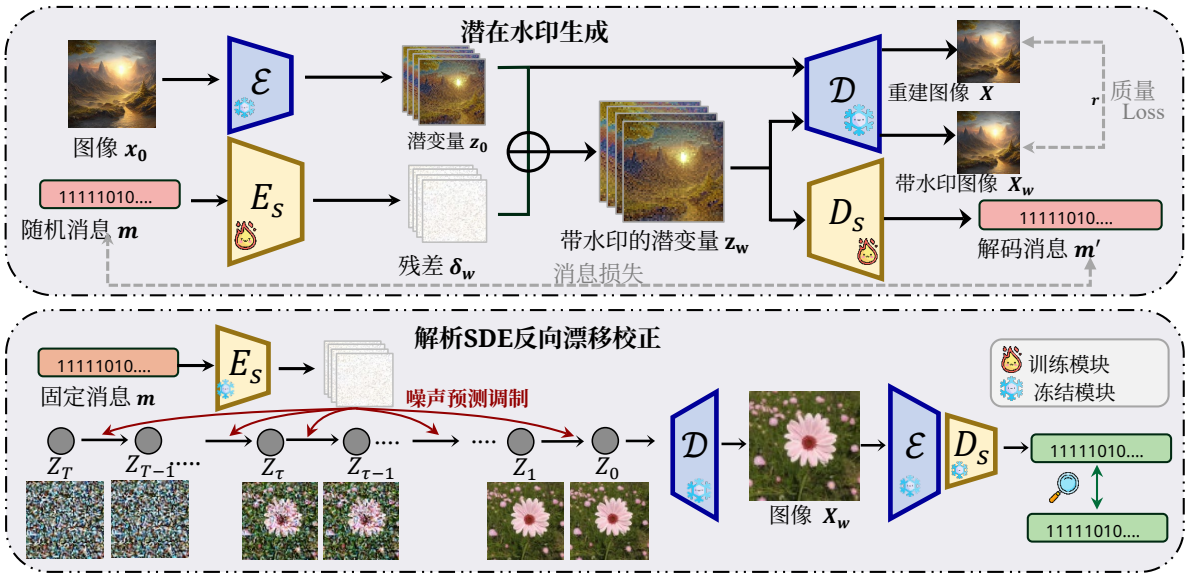


图 2. ALIEN 框架包含两个主要阶段：（上）不可感知潜在水印生成，其中训练秘密编码器 E_s 和解码器 D_s，将消息 m 嵌入到稳健的潜在残差 δ_w 中，同时保持图像质量。（下）解析SDE反向漂移校正，对噪声预测应用时变调制。这种基于原理的校正以与采样器无关的方式引导生成轨迹满足水印约束。

3. 方法

3.1. ALIEN 框架

我们进一步实现了语义水印的适用性与理论可控性。如图2所示，我们的 ALIEN 将水印嵌入扩散过程的中间状态，通过精确校正概率流来操控生成轨迹。为生成难以察觉的水印残差，我们设计了不可感知潜在水印生成模块，在潜在空间中生成隐形且稳健的水印残差。为实现原理性嵌入，我们提出了解析SDE反向漂移校正方法，该方法针对方差保持随机微分方程（VP-SDE）（Song et al., 2020b）解析推导出扩散模型进行水印嵌入所需的概率流修改，从而为噪声预测提供明确的校正目标，并兼容随机和确定性采样过程。该推导所得目标通过噪声预测目标调制模块实现，该模块规定了模型噪声预测输出所需的调整，从而实现可控且难以察觉的语义水印。

3.2. 不可感知潜在水印生成

为了生成不可感知且鲁棒的水印残差，我们联合训练秘密编码器 E_s 和水印解码器 D_s。理想情况下，带水印的潜在表示z_w 应同时以输入潜变量 z_0 和消息 m 为条件，以增强不可感知性。然而，由于扩散模型的中间潜在状态无法轻易获取以进行确定性条件控制，因此采用依赖于 z_0 的水印并不切实际。因此，我们选择嵌入一种在潜在空间中显著的载体无关水印。具体而言，将秘密残差 $\delta_w = E(m)$ 加到输入潜变量 z_0上，形成带水印的潜变量$z_w = z_0 + \delta_w$。带水印图像 x_w 通过$x_w = \mathcal{D}(z_w)$生成，并将解码器 D_s 应用于 z_w来提取消息，得到 $m' = D_s(z_w)$。我们采用二元交叉熵损失，优化原始消息 m 与解码消息 m'之间的消息提取准确率。

为确保水印的视觉一致性，我们计算带水印图像 x_w 与重建图像 x_r之间的学习感知图像块相似度损失（Zhang etal., 2018年）和均方误差损失，而不是计算 x_w 与原始图像 x_o之间的损失。这样做很有必要，因为 VAE 压缩和重建已经在 x_o 与 x_r之间造成了可测量且不可恢复的质量差距。若针对 x_o 进行优化，就需要水印训练过程补偿 VAE 重建误差，这会增加复杂度并降低嵌入效果。我们的训练目标总结如下，其中 λ_1 和 λ_2 是系数：

$$\mathcal{L}_T = \mathcal{L}_{BCE}^{(m,m')} + \lambda_1 \mathcal{L}_{LPIPS}^{(x_r,x_w)} + \lambda_2 \mathcal{L}_{MSE}^{(x_r,x_w)}, \quad (1)$$

其中 $\mathcal{L}_{BCE}$是二元交叉熵损失，$\mathcal{L}_{LPIPS}$是学习感知图像块相似度损失（Zhang

算法 1 ALIEN 水印技术

第一阶段：嵌入

输入：预训练 U-Net θ、编码器 E_s、调度器 S，变分自编码器 $\mathcal{D}_{VAE}$、提示词 $\mathbf{c}$、秘密信息 $\mathbf{m}$、强度 λ、间隔 $[T_{start}, T_{end}]$

输出：带水印图像$\mathbf{x}_{wm}$

$\delta_{wm} \leftarrow E_s(\mathbf{m})$

$\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

对 $t = T$ **递减至**1**执行**

  $\epsilon_\theta^t \leftarrow$ 无分类器引导$(\theta, \mathbf{z}_t, S.\text{步}[t], \mathbf{c})$

  如果 $T_{start} \le t \le T_{end}$ **则**

    $\epsilon_\theta^t \leftarrow \epsilon_\theta^t - \lambda \cdot \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\right) \cdot \delta_{wm}$

  结束条件判断

  $\mathbf{z}_t \leftarrow S.\text{step}(\epsilon_\theta^t, t, \mathbf{z}_t)$.前一个样本

结束循环

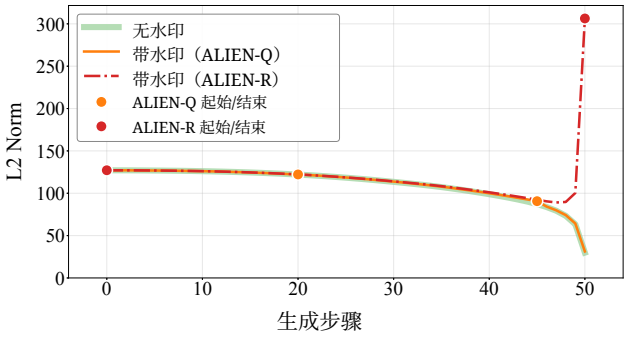
$\mathbf{x}_{wm} \leftarrow \mathcal{D}_{VAE}(\mathbf{z}_t)$

第二阶段：提取

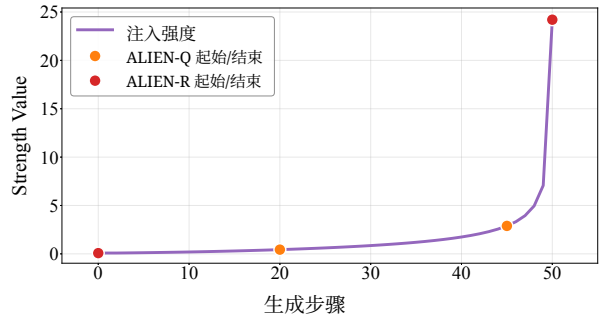
输入：图像 $\mathbf{x}_{wm}$、$\mathcal{E}_{VAE}$、解码器 D_s

输出：秘密信息$\mathbf{m}'$

返回 $D_s(\mathcal{E}_{VAE}(\mathbf{x}_{wm}))$



(a) 噪声预测范数差异



(b) 注入强度调度

图*3.* **ALIEN-R 与 ALIEN-Q 的比较**。(a) 扩散过程中噪声预测的 L2 范数。(b) 注入强度 $\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}$ 的演化。

等人，2018年），且 $\mathcal{L}_{MSE}$ 是均方误差损失。现有研究（Wang 等人，2025 年；Meng 等人，2024年）表明，在潜在空间中注入和检测水印能够天然抵御各种常见失真。我们遵循先前的做法，在训练时省略失真层（已在表 4中验证）。即使对手使用干净图像微调 U-Net θ 和 $\mathcal{D}$，再使用经过微调的潜空间解码器 $\mathcal{D}'$ 生成图像，水印的有效性仍不受影响（已在图 8中验证）。

3.3. 解析SDE反向漂移校正

为了在生成过程中实现精确的水印嵌入，我们利用 VP-SDE 解析推导出水印嵌入所需的精确概率流校正。

$\mathbf{z}_0$ 约束。 我们的目标是将预定的水印残差 δ_{wm} 嵌入最终去噪后的潜变量$\mathbf{z}_0$,，从而得到 $\mathbf{z}_0$ 空间约束 $\hat{\mathbf{z}}_0^{wm} = \hat{\mathbf{z}}_0^{orig} + \delta_{wm}$，其中 $\hat{\mathbf{z}}_0^{wm}$ 与 $\hat{\mathbf{z}}_0^{orig}$ 分别是 U-Net 针对当前潜在状态$\mathbf{z}_t$、在有水印和无水印情况下预测得到的无水印数据估计。

得分函数。 在 VP-SDE 下，扩散过程由正向随机微分方程定义，该方程将无水印数据逐步扰动为噪声：

$$\mathrm{d}\mathbf{z} = \mathbf{f}(\mathbf{z}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w},$$

其中，$\mathbf{f}(\mathbf{z}\ t)$：$\mathbb{R}^d \to \mathbb{R}^d$ 表示漂移系数，而$g(t) \in \mathbb{R}$ 表示标量扩散系数，其中 $\mathbf{w}(t)$ 表示标准维纳过程。扩散模型的反向生成过程由反向时间随机微分方程控制：

$$\mathrm{d}\mathbf{z} = \underbrace{[\mathbf{f}(\mathbf{z}, t) - g^2(t)\nabla_{\mathbf{z}} \log p_t(\mathbf{z})]}_{\text{Reverse Drift } \mathbf{F}_{rev}} \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \quad (2)$$

其中，$\bar{\mathbf{w}}$ 表示反向时间中的标准维纳过程。为实现水印 z_0 约束，我们精确控制概率流，这需要确定反向随机微分方程漂移项所需的校正量 $\Delta\mathbf{F}_{rev}$。由于 $\Delta\mathbf{F}_{rev}$ 依赖于分数函数$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)$，因此，解析量化分数函数的差异是推导 $\Delta\mathbf{F}_{rev}$的直接途径。分数函数与$\mathbf{z}_t$和其去噪估计$\hat{\mathbf{z}}_0$之间的差值具有精确的线性关系。该解析关系在数学上描述如下：

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t) = -\frac{1}{1-\bar{\alpha}_t}(\mathbf{z}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{z}}_0). \quad (3)$$

应用解析映射公式3，我们可以解析地确定由$\mathbf{z}_0$ 约束所导致的分数函数差异，从而将水印残差 δ_{wm} 转换为得分函数空间中所需的偏移。

概率流校正。 基于反向漂移项 $\mathbf{F}_{rev}$ 的 VP-SDE 解析结构，我们推导出

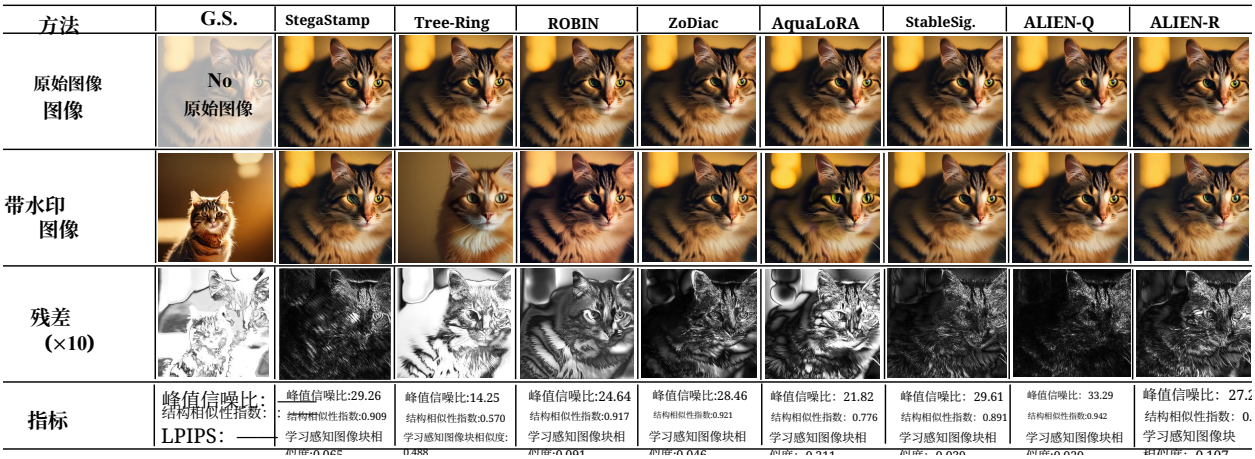


图*4*。 带水印样本与 10× 放大残差的定性比较。我们将 ALIEN 与涵盖后处理（StegaStamp）、潜变量修改（Tree-Ring）、优化（ROBIN、ZoDiac）和微调（AquaLoRA、StableSig。）的基线方法进行了比较。

表1。 不同水印方案的视觉质量与保真度定量比较。D.S.表示 DreamSim（Fu等人，2023年）指标。

方法	FID	CLIP	峰值信噪比	结构相似性指数	SIFID	D.S.
无水印	24.31	0.3368	—	—	—	—
StegaS.	24.56	0.3363	28.59	0.878	0.189	0.021
ZoDiac	—	—	28.01	0.922	0.121	0.022
AquaL.	24.79	0.3366	17.07	0.664	0.183	0.139
TreeR.	24.63	0.3370	12.76	0.429	0.741	0.291
G.S.	24.42	0.3378	—	—	—	—
ROBIN	24.61	0.3366	22.96	0.756	0.212	0.057
StableS.	24.56	0.3367	29.09	0.878	0.105	0.011
ALIEN-Q	24.29	0.3369	32.41	0.949	0.023	0.003
ALIEN-R	24.74	0.3366	20.42	0.745	0.227	0.061

用于满足水印约束的必要反向随机微分方程漂移校正 $\Delta\mathbf{F}_{rev}$ （详细信息见附录 A）。所得校正项 $\Delta\mathbf{F}_{rev}$ 由以下表达式解析定义：

$$\Delta\mathbf{F}_{rev}(\mathbf{z}_t,t) = -g^2(t)\frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\delta_{wm}. \tag{4}$$

这一解析推导式4 为 ALIEN 框架提供了理论基- 础。它精确量化了 $\Delta\mathbf{F}_{rev}$ 水印目标所需的校正 δ_{wm}，该校正位于随机微分方程概率流框架内。由于此校正针对潜变量分布的分数函数，因此证明了水印与采样器无关。

3.4. 噪声预测目标调制

为实现随机微分方程漂移校正 $\Delta\mathbf{F}_{rev}$的实际应用，我们在 VP-SDE 框架下，从解析上建立随机微分方程反向漂移校正与 U-Net 噪声预测偏移之间的联系（详见附录 A），推导出所需的补偿信号 $\Delta\epsilon_{wm}$ 以满足必要条件

表 2. 通过检测置信度衡量的稳定性评估。我们评估了不同采样器（DPM++ SDE、Euler a.、DPM2 a.）、推理步数（25,50）、引导尺度（10,20）和模型版本（v1.5、v2.1）下的鲁棒性。

方法	采样器			步数		尺度		模型	
	DPM-SDE	Euler a	DPM2 a	25	50	10	20	v1.5	v2.1
StegaS.	0.999	0.999	0.999						
StableS.	0.999	0.999	0.999	0.999					
TreeR.	0.000	0.000	0.000	0.963	0.963	0.963	0.963	0.958	0.963
G.S.	0.557	0.586	0.552						
AquaL.	0.952	0.953	0.939	0.945	0.955	0.940	0.939	0.954	—
ALIEN-Q	0.989	0.979	0.972	0.985	0.990	0.982	0.975	0.989	0.991
ALIEN-R	0.999	0.999	0.999						

以满足$\mathbf{z}_0$ 约束：

$$\Delta\epsilon_{wm} = -\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\delta_{wm}. \tag{5}$$

参见公式（5），为了在 δ_{wm}中实现固定的水印残差 $\mathbf{z}_0$ 空间，U-Net 将在其噪声预测中引入与水印残差成正比的时间相关补偿信号 $\Delta\epsilon_{wm}$。如图 3所示，注入系数在接近$\mathbf{z}_0$,时会激增，这意味着后期注入能够产生更强的约束，但也会损害视觉保真度。我们调整注入幅度和时间步范围，制定了两种配置：ALIEN-Q（质量导向）和 ALIEN-R（鲁棒性导向）。具体流程如算法1所示。

4. 实验

4.1. 实验设置

水印基线方法。 我们将 ALIEN 与主流的潜空间语义水印方案进行比较，涵盖三大主要类别：初始潜变量修改

表 3. **不同再生成攻击下的鲁棒性评估**。每个单元格报告三种攻击下的性能：**再生成** → **Rinse-2X** → **Rinse-4X**。数值表示**检测置信度/ TPR@1% FPR**。

方法	调度器（单元格格式：再生成数据 |Rinse-2X 数据 |Rinse-4X 数据）			
	DDIM	DPM++ 2M 随机微分方程	Euler a	DPM2 a
StegaS.	0.692/0.957 | 0.578/0.347 | 0.531/0.102	0.636/0.575 | 0.562/0.182 | 0.502/0.000	0.624/0.631 | 0.542/0.180 | 0.513/0.000	0.615/0.591 | 0.556/0.120 | 0.495/0.000
ZoDiac	0.938/1.000 | 0.743/0.104 | 0.527/0.032	0.693/0.585 | 0.670/0.431 | 0.562/0.125	0.643/0.315 | 0.643/0.250 | 0.628/0.210	0.833/0.650 | 0.572/0.281 | 0.520/0.152
TreeR.	0.868/1.000 | 0.724/0.861 | 0.408/0.489	0.855/0.918 | 0.824/0.905 | 0.753/0.851	0.821/0.941 | 0.449/0.512 | 0.277/0.306	0.647/0.755 | 0.531/0.656 | 0.326/0.427
G.S.	0.978/1.000 | 0.898/1.000 | 0.845/1.000	0.943/1.000 | 0.953/1.000 | 0.926/1.000	0.946/1.000 | 0.847/1.000 | 0.695/0.905	0.923/1.000 | 0.858/1.000 | 0.711/0.915
AquaL.	0.757/0.742 | 0.627/0.145 | 0.573/0.000	0.879/0.942 | 0.833/0.858 | 0.696/0.486	0.663/0.225 | 0.618/0.078 | 0.551/0.000	0.679/0.371 | 0.604/0.086 | 0.536/0.029
ROBIN	— /0.638 | — /0.242 | — /0.153	— /1.000 | — /0.728 | — /0.275	— /0.980 | — /0.705 | — /0.486	— /0.075 | — /0.105 | — /0.100
StableS.	0.496/0.000 | 0.477/0.000 | 0.521/0.000	0.512/0.000 | 0.512/0.000 | 0.510/0.000	0.498/0.000 | 0.520/0.000 | 0.513/0.000	0.506/0.000 | 0.485/0.000 | 0.502/0.000
ALIEN-Q	0.848/1.000 | 0.752/0.865 | 0.618/0.342	0.935/1.000 | 0.898/1.000 | 0.842/0.942	0.866/1.000 | 0.661/0.581 | 0.563/0.163	0.857/1.000 | 0.673/0.636 | 0.581/0.124
ALIEN-R	0.999/1.000 | 0.908/1.000 | 0.829/1.000	0.999/1.000 | 0.989/1.000 | 0.967/1.000	0.988/1.000 | 0.908/1.000 | 0.731/0.727	0.989/1.000 | 0.878/1.000 | 0.742/0.875



图 5. 不同生成调度器（DDIM、DPM2 a、Euler a 和 DPM++ 2M 随机微分方程）下的视觉对比。

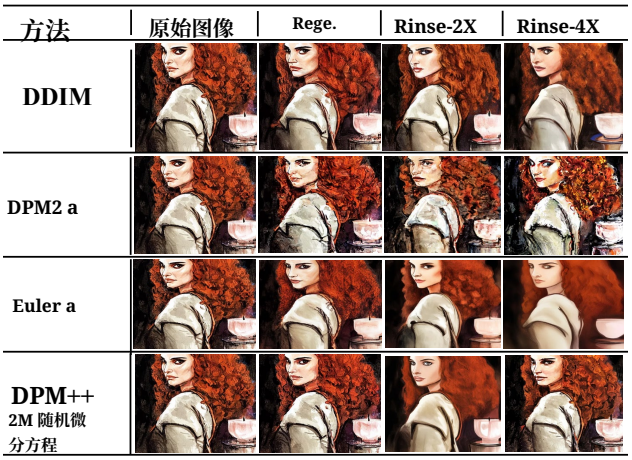


图 6. 再生成攻击下的视觉比较（DDIM、DPM2 a、Euler a 和 DPM++ 2M 随机微分方程）。

方法 Tree-Ring（Wen et al.，2024年）；受约束的采样方法 Gaussian Shading（Yang 等人，2024c）；以及迭代优化方法 ROBIN（Huang etal.，2024年）和 Zodiac（Zhangetal.，2024年）。此外，我们还选择了基于微调的方法 Stable Signature（Fernandezet al.，2023年）和 Aqualora（Feng et al.，2024年），以及后处理图像水印技术 StegaStamp

（Tancik 等人，2020）。实现细节详见附录B.2。

模型与任务。 遵循现有研究（Yang 等人，2024c），我们在 Stable Diffusion v1.5 和 Stable Diffusion v2.1 上实现 ALIEN 框架。我们针对两项任务评估所提出的方法：**(1) 检测**。 所有对比方法均配置为采用统一图案的单比特水印。我们针对每种方法分别设置检测阈值，使假阳性率（FPR）达到约 1%。**(2) 可追溯性**。 使用比特准确率评估多比特水印方法。由于 Tree-Ring、ZoDiac 和 ROBIN 等单比特方法不具备水印容量，因此不纳入评估。我们使用以下提示词

，这些提示词来自 StableDiffusion-Prompts（SDP）（Gustavosta，2023年），将引导尺度设置为 7.5，采样步数设为 50。TPR@1%FPR 和比特准确率均基于 350 张带水印图像计算。

4.2. 与基线方法比较

水印保真度。 我们采用 PSNR（Hore&Ziou，2010年）和 SSIM（Wang 等人，2004年）评估像素保真度，采用 LPIPS（Zhangetal.，2018年）和 DreamSim（Fu等人，2023年）评估感知相似度，采用 FID（Heusel etal.，2017）和 SIFID（Yang 等人，2024a）评估真实感，并采用 CLIP 分数（Radfordet al.，2021）评估语义对齐。如表1所示，ALIEN-Q 实现了最先进的不可感知性，显著优于所有基线方法。它以 32.41 dB 的 PSNR 实现了最高的像素保真度，并以 0.003 的 DreamSim 达到最佳感知质量，同时保持与无水印图像相当的 FID 和 CLIP 分数。

水印稳定性。 我们从三个维度验证水印稳定性：采样器无关性 涵盖

表4. 水印鲁棒性综合比较。性能以十进制刻度（0-1）报告。数值越高越好（↑）。**平均** 列表示全部11种攻击（不含无攻击）的均值性能。上半部分：TPR@1%FPR；下半部分：比特准确率。

指标	方法	No 攻击	光度		退化			几何			生成			**Avg.**
			亮度	对比度	JPEG	Blur	噪声	重新缩放	C.C.	R.C.	VAE-B	VAE-C	扩散	
TPR@ 1%假正例率阈值	StegaS.	1.000	0.964	1.000	1.000	1.000	1.000	1.000	0.964	0.975	1.000	1.000	1.000	0.991
	StableS.	1.000	0.958	1.000	0.781	0.973	0.042	0.935	1.000	1.000	0.000	0.079	0.000	0.611
	AquaL.	1.000	0.289	0.507	1.000	1.000	0.932	1.000	0.127	0.258	0.975	0.958	0.742	0.708
	TreeR.	1.000	0.775	0.909	1.000	1.000	0.954	0.968	0.013	0.021	0.954	1.000	1.000	0.778
	G.S.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.981	0.979	1.000	1.000	1.000	0.996
	ROBIN	1.000	0.913	0.684	0.921		0.013	0.935	1.000		1.000	1.000	0.638	0.826
	ZoDiac	1.000	0.535	0.375	0.959	0.979	0.357	0.965	0.017	0.021	0.660	0.685	1.000	0.592
	ALIEN-Q	1.000	0.783	1.000	0.954	1.000	0.645	1.000	0.153	0.311	0.965	0.970	0.945	0.793
	ALIEN-R	1.000	1.000	1.000	1.000	1.000	0.979	1.000	0.989	0.988	1.000	1.000	1.000	
比特准确率	StegaS.	0.999	0.811	0.850	0.999	0.999	0.821	0.999	0.667	0.681	0.837	0.816	0.692	0.834
	StableS.	0.998	0.892	0.851	0.712	0.845	0.535	0.878	0.991	0.993	0.506	0.541	0.489	0.748
	AquaL.	0.954	0.583	0.672	0.935	0.954	0.810	0.925	0.573	0.648	0.845	0.856	0.757	0.778
	G.S.	0.999	0.939	0.964	0.995	0.999	0.903	0.996	0.657	0.655	0.996	0.997	0.978	0.916
	ALIEN-Q	0.989	0.753	0.803	0.889	0.936	0.689	0.903	0.581	0.601	0.834	0.853	0.848	0.790
	ALIEN-R	0.999	0.949	0.992	0.999		0.855	0.999	0.826	0.835	0.979	0.983	0.999	

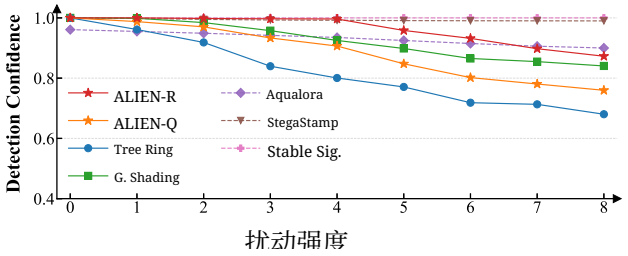


图7。 对 VAE 嵌入攻击的鲁棒性。结果取 KLVAE8、KLVAE16 和 SDXL 各变体的平均值。

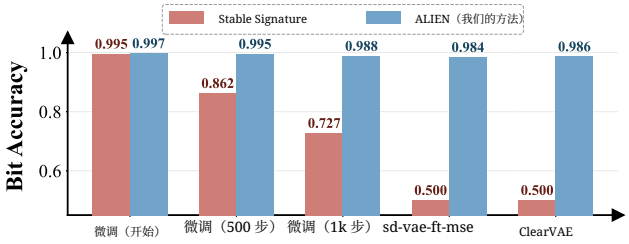


图 8. 在变分自编码器微调和替换情况下，与 Stable Signature 的比特准确率对比。

确定性（DDIM）和不可逆采样器（Euler Ancestral（Karras 等人，2022）、DPM++ 2M Ancestral（Karras 等人，2022）、DPM++SDE（Lu 等人，2025 年））；组件适应性 包括替换不同版本的变分自编码器和 U-Net；而超参数稳定性 则考察对引导尺度和推理步数的敏感性。如表2所示，现有免训练方法容易受到随机采样器的影响。具体而言，Tree-Ring 和 Gaussian Shading 在 DPM++ SDE、Euler a. 和 DPM2 a. 下完全失效，准确率徘徊在随机猜测基线附近。ALIEN 实现了最先进的采样器无关性，在所有测试的调度器下，比特准确率均保持在 0.97 以上。我们的方法在不同推理步数、引导尺度和模型版本下均表现出稳定性。

水印效率。 我们从两个主要维度评估计算开销：准备成本，即一次性的离线成本；以及运行时成本，即每张图像嵌入和提取的在线延迟。比较结果详见表7。ALIEN 的嵌入延迟为 0.079 秒，提取延迟为 0.023 秒，其速度与后处理基线方法相当，同时比基于优化的方法 ZoDiac 快几个数量级；后者每张图像需要超过 122 秒。

水印鲁棒性。我们考察水印对三类典型对抗攻击的鲁棒性，包括用于变换生成图像的图像处理攻击（Hore & Ziou，2010）、用于干扰水印验证的对抗攻击（An 等人，2024年），以及用于重新生成无水印图像的重构攻击（Zhao 等人，2023年）。详细参数设置请参阅补充材料。如表 4所示，ALIEN-R 在应对标准失真时表现出最先进的稳定性，真阳性率达到 0.996。在图7所示的对抗嵌入场景中，ALIEN-R 展现出卓越的抵抗能力，而潜在语义水印基线方法则迅速退化。表3显示，在面对重构攻击时，ALIEN-R 即使经过四轮重新生成，仍能保持高于 0.875 的置信度；相比之下，StegaStamp 和 Stable Signature 等方法的置信度则降至接近零。

表 5. 伪造攻击下的定量比较。数值越低表示抵抗能力越强（↓）。

配置	G.S.	Tree-Ring	ALIEN（我们的方法）
	准确率 / 峰值信噪比	准确率 / 峰值信噪比	准确率 / 峰值信噪比
场景 A：平均伪造			
10 张图像	0.952 / 12.23	0.882 / 18.86	0.539 / 20.06
50 张图像	0.969 / 12.79	0.908 / 20.34	0.605 / 25.29
100 张图像	0.969 / 12.65	0.910 / 20.95	0.708 / 26.48
场景 B：重提示伪造			
SD-V2.1	1.000 / —-	0.931 / —-	0.533 / —-

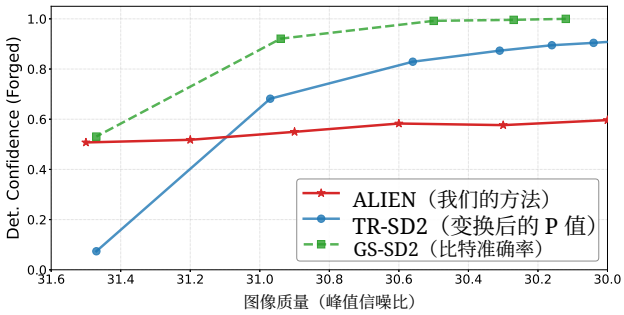


图 9. 针对印记伪造攻击的鲁棒性。

表 6. 不同水印强度下时间步范围的消融研究。我们比较了两种强度设置下的图像质量（峰值信噪比）和检测性能（检测/准确率）。

范围 $(t_{start} \to t_{end})$	强度 A($\lambda = 0.5$) 峰值信噪比	强度 A($\lambda = 0.5$) 检测 / 准确率	强度 B($\lambda = 1.0$) PSNR	强度 B($\lambda = 1.0$) 检测 / 准确率
$1 \to 50$	24.55	1.000 / 1.000	20.42	1.000 / 1.000
$25 \to 50$	30.53	1.000 / 1.000	26.27	1.000 / 1.000
$25 \to 40$	40.94	0.718 / 0.747	23.10	0.852 / 1.000
$10 \to 25$	31.54	0.614 / 0.421	26.83	0.681 / 0.721
$20 \to 45$	33.76	0.882 / 1.000	31.10	0.989 / 1.000

表7。 计算效率性能比较。 ‡ 表示根据原始论文估算的数据。

方法	准备时间（小时） 预训练	微调	优化时间（秒）	嵌入时间（秒）	提取时间（秒）
T.R.	—	—	—	0.011	3.575
G.S.	—	—	—	0.102	4.212
ROBIN	—	—	0.4	0.125	1.248
ZoDiac	—	—	—	122.4	3.601
StegaS.	30	—	—	0.108	0.028
S.S.	48‡	0.1	—	—	0.014
AquaL.	40‡	15‡	—	—	0.026
ALIEN	22	—	—	0.079	0.023

伪造抵抗能力。遵循现有研究（M¨uller 等人，2025 年；Yang 等人，2024b），我们评估了本方法抵御三种代表性伪造攻击的安全性：印记伪造、重提示伪造和平均伪造。印记伪造和重提示伪造依赖通过扩散反演进行潜变量优化。与基于初始潜变量的水印相比，ALIEN 对这两种攻击具有更强的抵抗能力。Tree-Ring 和 Gaussian Shading 等基线方法仅凭单张图像就可能被成功伪造，而 ALIEN 不受影响。平均伪造通过收集带水印图像与无水印图像的图像对来估计水印模式。Tree-Ring 等基线方法容易受到攻击，其置信度超过 0.9；而 ALIEN 将置信度抑制在 0.711，有效防止水印模式被快速提取。在现实世界的 API 场景中，获取大规模特定用户数据面临实际困难，这显著提高了安全门槛。

时间步范围的影响。 如表所示。6，我们对注入区间进行了消融研究，以探究不可感知性与鲁棒性之间的权衡。在从第 50 步到第 1 步的整个反向过程中注入水印，可确保完美检测，准确率达到 1.000，但会引入显著的感知质量下降，使得强度 B 下的峰值信噪比降至 20.42 分贝。将窗口缩小至较早的扩散阶段（例如第 45 步至第 20 步）可显著恢复图像质量，此时峰值信噪比在保持较高检测率的同时，强度 A 下提升至 33.76 分贝。

水印强度的影响。如表6所示，注入系数 λ 直接调节概率流校正的幅度。将强度从$\lambda = 0.5$ 强度 A 中的 $\lambda = 1.0$提高到强度 B 中的值，可在所有范围内持续提升检测性能。我们选择将范围设为 45 至 20，并采用适中的强度，作为 ALIEN-Q 的最佳配置，以在确保可靠检测的同时最大限度地提升质量。

5. 结论

不同于依赖启发式潜变量修改、采样约束或计算密集型优化的既有方法，我们提出了 ALIEN，并首次解析推导出时变调制系数。这种有理论依据的方法通过概率流调制，精确引导水印残差的扩散，实现与采样器无关的嵌入。ALIEN 克服了扩散反演和不可逆采样器带来的安全漏洞，同时确保较高的保真度和较低的推理成本。实验结果表明，ALIEN 在质量和鲁棒性方面均优于现有方法。未来研究将聚焦于开发内容感知潜在水印，以进一步增强抵御伪造攻击的安全性。

6. 影响声明

确保生成模型的可追责性和知识产权保护已成为当务之急。本文介绍了 ALIEN，这是一种有理论依据的水印框架，可为潜在扩散模型实现鲁棒且与采样器无关的来源追踪。通过克服现有方法易受不可逆采样器及多种攻击影响的脆弱性，我们认为，该方法能够有效降低版权侵权和恶意滥用等风险，从而助力开发更安全、更可信的生成式人工智能。

参考文献

An, B.、Ding, M.、Rabbani, T.、Agrawal, A.、Xu, Y.、Deng, C.、Zhu, S.、Mohamed, A.、Wen, Y.、Goldstein, T. 等。Waves：图像水印鲁棒性基准测试。*arXiv* 预印本 *arXiv:2401.08573*，2024年。Ball´e, J.、Minnen, D.、Singh, S.、Hwang, S. J. 和 Johnston, N. 基于尺度超先验的变分图像压缩。*arXiv* 预印本 *arXiv:1802.01436*，2018年。

Barrett, C. 识别并缓解生成式人工智能的安全风险。*arXiv* 预印本 *arXiv:2308.14840*，2023年。

拜登。关于安全、可靠和值得信赖地开发和使用人工智能的行政命令，2023年。网址 https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/07/fact-sheet-executive-order-on-artificial-intelligence/。

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. 使用离散化高斯混合似然和注意力模块的学习型图像压缩。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7939–7948 页，2020 年。

Ci, H., Song, Y., Yang, P., Xie, J., and Shou, M. Z. Wmadapter：为潜在扩散模型添加水印控制。*arXiv* 预印本 *arXiv:2406.08337*，2024年。

Ci, H., Yang, P., Song, Y., and Shou, M. Z. Ringid：重新思考树环水印，以增强多密钥识别能力。载于欧洲计算机视觉会议，第 338–354 页。Springer，2025 年。

Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. 数字水印与隐写术。Morgan kaufmann，2007 年。

Dhariwal, P. 和 Nichol, A. 扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34:8780–8794，2021年。

欧盟。人工智能法案：欧洲议会和理事会于2024年6月颁布的欧盟条例 2024/1689。网址：https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689，访问日期：2024-09-24。Feng, W.、Zhou, W.、He, J.、Zhang, J.、Wei, T.、Li, G.、Zhang, T.、Zhang, W. 和 Yu, N. Aqualora：通过水印 LoRA 为定制稳定扩散模型提供白盒保护。*arXiv* 预印本 *arXiv:2405.11135*，2024年。Fernandez, P.、Couairon, G.、J´egou, H.、Douze, M. 和 Furon, T. Stable Signature：将水印植入潜在扩散模型。在*IEEE/CVF* 国际计算机视觉会议论文集，第22466–22477页，2023年。

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., and Isola, P. Dreamsim：使用合成数据学习人类视觉相似性的新维度。*arXiv* 预印本 *arXiv:2306.09344*，2023年。

Gowal, S. and Kohli, P. 使用 SynthID 识别人工智能生成图像。https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid，2023年。访问日期：2023年9月23日。

Gunn, S., Zhao, X., and Song, D. 一种适用于生成式图像模型的不可检测水印。*arXiv* 预印本 *arXiv:2410.07369*，2024年。

Gustavosta. 稳定扩散提示词数据集，2023年5月。网址：https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts。

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。神经信息处理系统进展，30，2017年。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33：6840–6851，2020年。

Hore, A. 和 Ziou, D. 图像质量指标：峰值信噪比与结构相似性指数。在*2010*年第*20*届国际模式识别会议，第2366–2369页。电气电子工程师学会，2010年。

Huang, H.、Wu, Y. 和 Wang, Q. ROBIN：采用对抗优化的扩散模型鲁棒且不可见的水印。神经信息处理系统进展，37：3937–3963，2024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，35: 26565–26577, 2022。

Lee, S. J. 和 Cho, N. I. 重新构想语义水印：借助傅里叶完整性提升鲁棒性和生成质量。载于*IEEE/CVF* 国际计算机视觉会议论文集，第 18759–18769 页，2025 年。Li, K., Huang, Z., Hou, X. 和 Hong, C. GaussMarker：面向扩散模型的鲁棒双域水印。*arXiv* 预印本 *arXiv:2506.11444*，2025 年。Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P. 和 Zitnick, C. L. 微软 COCO：情境中的常见物体。载于计算机视觉–*ECCV 2014*：第 *13* 届欧洲会议，苏黎世，瑞士，*2014* 年 *9* 月 *6–12* 日，论文集，第 *V* 卷，第 740–755 页。Springer，2014 年。

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. 和 Zhu, J. Dpm-solver：一种用于扩散概率模型采样的快速常微分方程求解器，可在约 10 步内完成。神经信息处理系统进展，35:5775–5787，2022。

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. 和 Zhu, J. Dpm- solver++：用于扩散概率模型引导采样的快速求解器。机器智能研究，第 1–22 页，2025 年。

Meng, Z., Peng, B. 和 Dong, J. 潜在水印：在潜在扩散空间中注入和检测水印。*arXiv* 预印本 *arXiv:2404.00230*，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。载于计算机视觉与模式识别大会论文集，第 20937–20946 页，2025 年。

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. 从自然语言监督中学习可迁移视觉模型。发表于国际机器学习大会，2021年。

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., andChen, M. 使用对比语言-图像预训练潜变量进行分层文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*，1(2):3，2022。

Ren, K., Yang, Z., Lu, L., Liu, J., Li, Y., Wan, J., Zhao, X., Feng, X., and Shao, S. 知识系统化：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。

Rezaei, A., Akbari, M., Alvar, S. R., Fatemi, A. 和 Zhang, Y. LAWA：利用潜在空间进行生成过程中图像水印。载于欧洲计算机视觉会议，第 118–136 页。Springer，2024年。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. 和 Ommer, B. 利用潜变量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*），第 10684–10695 页，2022 年 6 月。Song, J., Meng, C. 和 Ermon, S. 去噪扩散隐式模型。*arXiv* 预印本 *arXiv:2010.02502*，2020a。Song, Y. 和 Ermon, S. 改进基于分数的生成模型训练技术。神经信息处理系统进展，33:12438–12448，2020。

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Er- mon, S., and Poole, B. 通过随机微分方程进行基于分数的生成建模。*arXiv* 预印本 *arXiv:2011.13456*，2020b。

Tancik, M., Mildenhall, B., and 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，第 2117–2126 页，2020。

Wang, Z., Bovik, A. C., Sheikh, H. R., and Si moncelli, E. P. 图像质量评价：从误差可见性到结构相似性。*IEEE*图像处理汇刊，13(4):600–612，2004。

Wang, Z., Guo, J., Zhu, J., Li, Y., Huang, H., Chen, M., and Tu, Z. Sleepermark：致力于抵御针对文生图扩散模型微调的稳健水印。载于计算机视觉与模式识别大会论文集，第 8213–8224 页，2025 年。

Wen, Y., Kirchenbauer, J., Geiping, J., 和 Goldstein, T. 树轮水印：扩散图像的不可见指纹。神经信息处理系统进展，36，2024年。

Yang, J., Liu, H., Guo, W., Rao, Z., Xu, Y., 和 Niu, D. Sifi d：使用 LLM 重新评估摘要事实不一致检测。*arXiv* 预印本 *arXiv:2403.07557*，2024a。

Yang, P., Ci, H., Song, Y., 和 Shou, M. Z. 简单平均能否击败现代水印？神经信息处理系统进展，37：56644–56673，2024b。

Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., 和 Yu, N. Gaussian Shading：面向扩散模型的可证明无性能损失图像水印。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12162–12171 页，2024c。

Zhang, L., Liu, X., Martin, A. V., Bearfield, C. X., Brun, Y., 和 Guan, H. 使用 Stable Diffusion 实现抗攻击的图像水印。神经信息处理系统进展，37:38480–38507，2024年。

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., 和 Wang, O. 深度特征作为感知度量的惊人有效性。载于 *IEEE*计算机视觉与模式识别会议论文集，第 586–595 页，2018年。

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生成式人工智能以可证明的方式移除。*arXiv* 预印本 *arXiv:2306.01953*，2023年。

Zhao, X.、Zhang, K.、Su, Z.、Vasan, S.、Grishchenko, I.、Kruegel, C.、Vigna, G.、Wang, Y.-X. 和 Li, L. 生成式 AI 可被证明能够移除不可见图像水印。神经信息处理系统进展，37:8643–8672，2024年。

补充材料

A ALIEN 的理论分析 13

A.1 水印约束与分数差异.13 A.2 反向 SDE 漂移项修正的证明.13 A.3 U-Net 目标噪声的推导.13

B 实现细节 14

B.1 算法伪代码.14 B.2 超参数与训练设置.14

C 实验设置与扩展实验 15

C.1 数据集设置.15 C.2 鲁棒性评估设置.16 C.3 扩展实验.16

D 扩展相关工作 17

D.1 水印方案分类.17

A. ALIEN 的理论分析

本节将严谨推导$\mathbf{z}_0$空间中的水印约束与概率流漂移项所需修正之间的解析联系。该推导证明，ALIEN 在理论上建立于 VP-SDE 框架之上，且本质上与采样器无关。

A.1. 水印约束与分数函数差异

1. 水印约束定义 我们的目标是将固定的水印残差 δ_{wm} 嵌入潜在表示中。我们将其定义为对估计的干净数据流形施加的几何约束。令 $\hat{\mathbf{z}}_0$ 表示由模型 θ得到的原始估计，令 $\hat{\mathbf{z}}_0^{wm}$表示目标带水印估计：

$$\hat{\mathbf{z}}_0^{wm} = \hat{\mathbf{z}}_0 + \delta_{wm}. \tag{6}$$

2. 分数函数差异的推导 Δ分数 在 VP-SDE 框架下，Tweedie 公式确立了分数函数 $\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)$ 与去噪估计 $\hat{\mathbf{z}}_0$ 之间的线性双射关系：

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t) = -\frac{\mathbf{z}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{z}}_0}{1-\bar{\alpha}_t}. \tag{7}$$

为满足水印约束（式 6），分数函数必须调整为 $\nabla_{\mathbf{z}_t} \log p_t^{wm}$。将约束条件$\hat{\mathbf{z}}_0 - \hat{\mathbf{z}}_0^{wm} = -\delta_{wm}$代入后，可推导出分数差异 Δ分数：

$$\begin{aligned} \Delta\text{Score} &= \nabla_{\mathbf{z}_t} \log p_t^{wm} - \nabla_{\mathbf{z}_t} \log p_t \\ &= -\frac{1}{1-\bar{\alpha}_t}\left(-\sqrt{\bar{\alpha}_t}(\hat{\mathbf{z}}_0^{wm} - \hat{\mathbf{z}}_0)\right) \\ &= \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\delta_{wm}. \end{aligned} \tag{8}$$

A.2. 反向 SDE 漂移项修正的证明 $\Delta\mathbf{F}_{rev}$

反向漂移的定义 $\mathbf{F}_{rev}$ 扩散模型中的生成轨迹由反向时间 SDE 控制。该过程的确定性部分称为漂移项 $\mathbf{F}_{rev}$，其定义如下：

$$\mathbf{F}_{rev}(\mathbf{z}_t, t) = \mathbf{f}(\mathbf{z}_t, t) - g^2(t)\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t). \tag{9}$$

其中，$\mathbf{f}(\mathbf{z}_t, t)$ 是前向漂移，$g(t)$ 是扩散系数。

漂移修正的推导 $\Delta\mathbf{F}_{rev}$ 我们计算为适应水印而需要对漂移项进行的修改。由于前向物理过程 $\mathbf{f}(\mathbf{z}_t, t)$ 保持不变，漂移修正仅取决于分数的偏移：

$$\Delta\mathbf{F}_{rev} = \mathbf{F}_{rev}^{wm} - \mathbf{F}_{rev}^{orig} = -g^2(t)\left(\Delta\text{Score}\right). \tag{10}$$

代入 方程（8）到上式中，可得水印漂移力的显式形式：

$$\Delta\mathbf{F}_{rev}(\mathbf{z}_t, t) = -g^2(t)\frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\delta_{wm}. \tag{11}$$

定理： 空间约束 δ_{wm} 会在概率流上施加一个恒定的确定性作用力 $\Delta\mathbf{F}_{rev}$。由于 $\mathbf{F}_{rev}$是随机逆向随机微分方程和确定性概率流常微分方程（PF-ODE）的共同驱动项，因此这一修正能够确保水印嵌入在不同采样器下均具有鲁棒性。

A.3. 噪声预测目标的推导 ϵ^{target}

为了在实践中实现理论漂移校正 $\Delta\mathbf{F}_{rev}$，我们对 U-Net ϑ的输出进行调制。

分数-噪声关系 神经网络 ϵ_ϑ 通过以下关系近似分数函数：

$$\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t) = -\frac{\epsilon_\vartheta(\mathbf{z}_t, t)}{\sqrt{1-\bar{\alpha}_t}}. \tag{12}$$

算法 2 ALIEN 水印技术

1：阶段 I：水印嵌入 2：输入：预训练 U-Net θ、编码器 E_s、调度器 S、VAE 解码器 $\mathcal{D}_{VAE}$、提示词 c、秘密信息 m、注入区间$[T_{start}, T_{end}]$、强度 λ3：输出：带水印图像 x_{wm}4：$\delta_{wm} \leftarrow E_s$(m) 5：$\mathrm{z}_t \sim \mathcal{N}$(0, I) 6：对于 $t = T$，从大到小遍历至 1 7：t $\leftarrow S$.timesteps$[t]$8：$\epsilon_\theta^t \leftarrow$ 无分类器引导(θ,z_t, t, c) 9：如果 $t \le T_{end}$ 且 $t \ge T_{start}$ 则 10：// 调节预测目标 11：$\epsilon_\theta^t \leftarrow \epsilon_\theta^t - \lambda \cdot \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\right) \cdot \delta_{wm}$12：结束 if 13：$\mathrm{z}_t \leftarrow S$.step($\epsilon_\theta^t$, t, z_t).prev sample 14：结束 for 15：$\mathrm{x}_{wm} \leftarrow \mathcal{D}_{VAE}(\mathrm{z}_t)$ 16：阶段 II：水印提取 17：输入：带水印图像 x_{wm}、VAE 编码器 $\mathcal{E}_{VAE}$、水印解码器 D_s18：输出：提取的秘密信息 m′19：// 将图像重新编码为带水印的潜变量 20：z $_0 \leftarrow \mathcal{E}_{VAE}(\mathrm{x}_{wm})$ 21：// 从潜变量中提取秘密消息 22：$\mathrm{m}' \leftarrow D_s(\mathrm{z}_0)$ 23：返回 m′

将映射漂移校正到 ϵ-空间我们将根据漂移要求得出的分数差与噪声预测差相等。使用 $\Delta\text{Score} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}(\epsilon^{target} - \epsilon_\vartheta) = -\frac{\Delta\epsilon}{\sqrt{1-\bar{\alpha}_t}}$，可得：

$$-\frac{\Delta\epsilon}{\sqrt{1-\bar{\alpha}_t}} = \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\delta_{wm}. \tag{13}$$

最终更新规则 求解校正项 $\Delta\epsilon$:

$$\Delta\epsilon = -\sqrt{1-\bar{\alpha}_t} \cdot \frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\delta_{wm} = -\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\delta_{wm}. \tag{14}$$

因此，强制施加水印所需的最终目标噪声 ϵ^{target} 为:

$$\epsilon^{target} = \epsilon_\vartheta + \Delta\epsilon = \epsilon_\vartheta - \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\right)\delta_{wm}. \tag{15}$$

B. 实现细节

B.1. 算法伪代码

ALIEN 水印框架的核心逻辑详见算法2。该算法概述了在反向扩散过程中进行水印嵌入，以及随后从潜在空间中提取水印的完整过程。

B.2. 超参数和训练设置

水印生成模块训练。 我们使用 AdamW 优化器训练不可感知潜在水印生成模块，学习率为 5×10^{-5}，权重衰减为 1×10^{-5}。我们构建了一个合成训练集，包含

表 *8*。 **详细的鲁棒性评估设置**。 我们将攻击分为四个不同的类别。请注意，为节省篇幅，图像处理攻击按类型（光度、几何、退化）分组。

类别	方法与参数
图像处理	**光度**：亮度（$\times 6.0$）、对比度（$\times 12.0$） **几何**：缩放（$s = 0.5$）、中心裁剪（$s = 0.4$）、随机裁剪（$s = 0.4$） **退化**：高斯噪声 ($\sigma = 0.25$)、模糊 ($radius = 1.5$)、JPEG ($Q = 50$)
重建类	**VAE 压缩**：VAE-B（Ball´e 等人，2018年）（$Q = 1$），VAE-C（Cheng 等人，2020）（$Q = 1$） **生成式攻击**：Regen-Diff（Zhao 等人，2024年）（SD v1.5，强度 $S = 0.2$），Rinsing（$N = 2,\ 4$）
对抗类	**Adv-Emb**：基于PGD的潜空间攻击（An 等人，2024年）
伪造	**印记**：通过反演进行潜空间优化（M¨uller 等人，2025 年） **重新提示**：将图像反演为初始潜变量并重新生成（M¨uller 等人，2025 年） **平均**：通过对残差取平均来估计水印模式（Yang 等人，2024b）



(a) 原始图像 *(b)* 亮度 *(c)* 对比度 *(d)* JPEG *(e)* 模糊

(f) 噪声 *(g)* 中心裁剪 *(h)* 随机裁剪 *(i)* VAE-B *(j)* VAE-C

图 *10.* **视觉鲁棒性示例**。 带水印图像在各种攻击下的对比：(a) 原始图像，(b) 亮度，(c) 对比度，(d) JPEG 压缩，(e) 高斯模糊，(f) 高斯噪声，(g) 中心裁剪（C.C.），(h) 随机裁剪（R.C.），(i) VAE 压缩（BMSHJ），以及 (j) VAE 压缩（Cheng）。

10,000 张基于 COCO2014（Lin 等人，2014）提示词生成的图像。该模型以 16 的批量大小训练 50,000 步。为确保训练稳定性与感知质量，我们将最大梯度范数设为 10^{-2}。总目标结合了秘密恢复损失（$\mathcal{L}_{sec}$）、逐像素重建损失（$\mathcal{L}_{mse}$）以及感知损失（$\mathcal{L}_{lpips}$，使用 AlexNet 骨干网络）。损失权重设为 $\lambda_{sec} = 1.0$、$\lambda_{mse} = 30.0$ 和 $\lambda_{lpips} = 0.3$。值得注意的是，λ_{mse} 和 λ_{lpips} 在前 5,000 步中从 0 线性上升至各自峰值，以促进早期训练阶段的稳定收敛。

注入设置。 在推理阶段，ALIEN-Q 会在第 20–45 步的采样区间内注入水印（注入强度为 $\lambda = 0.85$），而 ALIEN-R 则将注入窗口扩展至第 0–50 步（$\lambda = 1.0$），以增强鲁棒性。

C. 实验设置与扩展实验

C.1. 数据集设置

评估数据集。 我们采用两个不同的数据集，全面评估水印性能和图像生成质量：为衡量检测准确率（TPR@1%FPR）和载荷容量（比特准确率），我们

表*9*。 **性能提升明细**。 我们将 ALIEN 与各项指标/条件下表现最佳的基线方法（最佳当前最优方法）进行比较。**A面板**计算了 5 项质量指标的平均相对提升（33.1%）。**B面板** 展示了 15 种不同条件下的鲁棒性提升（12 种生成 + 3 稳定性）。最终的 14.0% 是这 15 种条件下的加权平均提升。

A面板：质量提升（ALIEN-Q 与最佳当前最优方法对比）

指标	描述	最佳当前最优方法	方法	ALIEN-Q	改进
FID ↓	Fr´echet Inception 距离	24.42 (G.S.)	数值越低越好	24.29	+0.5%
峰值信噪比 ↑	峰值信噪比	29.09 (StableSig.)	数值越高越好	32.41	+11.4%
结构相似性指数 ↑	结构相似性	0.922（ZoDiac）	越高越好	0.949	+2.9%
SIFID ↓	单图像FID	0.105 (StableSig.)	越低越好	0.023	**+78.1%**
DreamSim ↓	感知相似度	0.011（StableSig.）	越低越好	0.003	**+72.7%**
			平均质量提升		**+33.1%**

B面板：鲁棒性提升（ALIEN-R vs. 免训练最先进水平）

条件类别	具体内容	最佳当前最优方法	指标	ALIEN-R	Gain
1. 生成式变体（12 个条件）	12 个条件下的平均准确率（4 个调度器 × Regen/Rinse）	∼0.84	Acc.	∼0.90	+6.5%
2. 采样器稳定性（3 个条件）	3 种随机采样器的平均准确率 (DPM++ 随机微分方程、Euler a、DPM2 a)	∼0.56	Acc.	∼1.00	**+44.0%**
		计算：$(12 \times 6.5\% + 3 \times 44.0\%)/15$	**平均鲁棒性提升**		**+14.0%**

表*10*。 不可感知性与鲁棒性比较。我们比较了48位和128位载荷。保真度 以均值 ± 标准差报告。鲁棒性 以检测准确率报告。缩写：无攻击 （无攻击）、压缩 （压缩）、组合 （组合攻击）。

载荷	不可感知性指标			鲁棒性（比特准确率）									
	峰值信噪比(↑)	结构相似性指数(↑)	学习感知图像块相似度(↓)	无攻击	模糊	噪声	JPEG	缩放	锐化	亮度	对比度	Sat.	组合
48 位	$31.17_{\pm 0.90}$	$0.62_{\pm 0.03}$	$0.097_{\pm 0.01}$	0.99	0.99	0.92	0.99	0.99	0.99	0.99	0.99	0.99	0.87
128 位	$30.79_{\pm 0.85}$	$0.59_{\pm 0.03}$	$0.151_{\pm 0.01}$	0.99	0.99	0.90	0.98	0.99	0.99	0.99	0.99	0.99	0.82

采用 Stable Diffusion-Prompts（SDP）数据集（Gustavosta，2023年）。我们从该数据集中随机选取 350 条提示词，生成图像对（带水印与无水印），用于分别评估。为评估生成保真度，具体为 Fr´echet Inception Distance（FID），我们使用 MS-COCO 数据集（Lin 等人，2014年）。我们从 MS-COCO 验证集中随机采样 5,000 条图像描述，生成 5,000 张带水印图像，并对照相应的真实参考图像计算 FID 分数，以确保标准化的质量比较。

生成配置。 我们在两种具有代表性的潜在扩散模型上进行实验：Stable Diffusion v1.5 和 Stable Diffusion v2.1。除非另有说明，我们使用 DDIM采样器，推理步数为 50。无分类器引导尺度设为 7.5。所有生成图像的分辨率均为 512×512。所有实验均在单块 NVIDIA RTX 3090 GPU 上完成。

C.2. 鲁棒性评估设置

C.3. 扩展实验

性能增益详细分析。 为证实摘要中的主张，我们详细分解了性能提升。表 9 展示了 **33.1%质量提升** 以及 **14.0%鲁棒性提升**相较于最先进（SOTA）方法的计算过程。

载荷大小的影响。我们进一步研究载荷大小对水印容量、不可感知性和鲁棒性之间权衡的影响。如表 10所示，将载荷容量从 48 位提高到 128 位，仅会导致保真度指标略微下降，例如峰值信噪比从 31.17 dB 小幅降至 30.79 dB。这一结果验证了 ALIEN 能够有效支持高容量嵌入，同时保持与低容量设置相当的视觉质量。

表 *11.* 固定阈值法对假阳性率的影响（ROBIN方案）。使用 ROBIN 水印方法时，对比干净图像的均值（μ_{clean}）与裁剪的最优阈值（τ_{crop}）。对于 SD v1.5 和 v2.1，模糊和 JPEG 等良性退化会使 μ_{clean} 低于 τ_{crop}（标记为 ×）。

攻击	Stable Diffusion v1.5（$\tau_{crop} = 39.45$）			Stable Diffusion v2.1（$\tau_{crop} = 55.80$）		
	无水印均值	最优值 τ	固定 τ 风险	无水印均值	最优值 τ	固定 τ 风险
无攻击	40.08	38.28	✓	55.90	52.74	✓
裁剪	40.25	39.45	-	55.86	55.80	-
模糊	32.38	30.31	×	52.65	50.24	×
JPEG	37.22	36.91	×	55.65	53.57	×
颜色抖动	36.91	36.85	×	54.04	51.39	×
噪声	41.63	40.02	✓	57.07	55.62	✓

基于距离的方法的阈值确定。我们评估了图像退化导致的分布偏移对基于距离的方法 ROBIN 的阈值确定所产生的影响。如表 11所示，在其他失真情况下使用根据具有挑战性的场景（裁剪，$\tau = 39.45$）确定的固定阈值存在风险。我们观察到，对于抑制高频信息的退化，指标空间中出现了明显的分布偏移。例如，模糊情况下未带水印图像的平均指标降至 32.38，低于裁剪场景下 39.45 的固定阈值。由于指标低于阈值时会触发检测，这种偏移会导致假正率（FPR）上升，使无害的低质量图像被误判。

D. 扩展相关工作

本节详细回顾并分类了现有生成模型水印方法。

D.1. 水印方案分类

我们根据嵌入阶段，将现有生成模型水印方法分为五大类：随机种子修改、U-Net 修改、VAE 修改、潜在空间修改和后处理。

随机种子修改（初始潜变量）。此类方法通过在初始潜变量的采样过程中进行修改来嵌入水印。例如，**Tree-Ring** (Wenetal.，2024年）在傅里叶域中修改初始噪声向量，以嵌入环形图案，该图案可通过扩散反演检测。**Gaussian Shading** (Yang 等人，2024c）采用约束采样策略，将特定图案应用于初始潜变量。

模型修改。这些方法通过微调生成模型的特定组件来嵌入水印，确保水印在生成过程中得以保留。**Stable Signature** (Fernandez et al.，2023年）微调变分自编码器解码器，在潜变量到图像的解码阶段将水印嵌入像素空间。**Aqualora**(Fenget al.，2024年）修改 U-Net，在迭代去噪过程中注入水印。

潜空间优化。 与静态编码方法不同，这些方法将水印嵌入表述为潜空间中的一个优化问题。 **ZoDiac** (Zhang et al.，2024年）通过迭代优化初始潜变量，确保生成输出具有较高的可检测性。 **ROBIN**(Huang et al.，2024年）专注于优化扩散过程中中间的潜在表示。它会引入可学习的提示词，使带水印的潜变量与文本条件保持一致。

后处理方法。 这些技术在图像完全生成后应用数字水印算法，独立于生成流程运行。 **StegaStamp** (Tanciket al.，2020年）是一种基于深度学习的编码器-解码器框架，可将不可见超链接或比特串嵌入最终图像输出中。