



研究主题： 面向扩散模型的生成图水印注 入技术

本学期工作总结汇报

汇报人：陈建浩

01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划



01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划





一、研究背景与研究价值

研究背景

- **AIGC技术的兴起与发展**：AIGC技术在生成多模态内容方面展现出远超传统手段的高效率、低成本和高质量的优势，受到广泛关注。我国也积极对相关算法进行备案管理，推动其安全化和标准化发展。
- **AIGC技术带来的风险**：不加节制地使用生成内容可能会引发虚假信息传播、隐私侵犯、舆论误导、法律纠纷，甚至社会动荡等诸多问题，给社会带来潜在风险。
- **政策法规的引导与监管**：为应对AIGC技术带来的挑战，我国出台了相关指南《网络安全标准实践指南—生成式人工智能服务内容标识方法》，要求在AI生成内容中添加水印，以实现标准化的实践指引和监管。
- **生成图水印的现实需求**：在模型即服务的产业模式下，预训练模型作为重要资源，易遭受恶意攻击，导致知识产权被侵犯。同时，生成模型滥用引发的虚假信息传播等问题，也使得深度鉴伪和生成内容溯源变得迫切，而传统的被动取证方法已难以满足需求。



一、研究背景与研究价值

研究价值

- **版权保护**：通过在生成模型中嵌入水印，可实现对模型的版权保护，防止模型被恶意攻击和知识产权被侵犯，维护模型所有者的合法权益。
- **溯源与责任认定**：水印可用于追踪溯源生成内容，确定其来源，进而追究生成内容发布者和滥用者的法律责任，实现责任闭环，有效遏制生成模型的滥用行为。
- **深度鉴伪**：随着GANs和DMs技术的快速进步，生成的图像、文本和音视频内容越来越逼真和多样化，近年来致力于区分生成内容和自然内容差异的研究变得困难。因此，在模型生成内容的过程中主动嵌入水印，可从源头上实现生成内容与自然内容的鉴别，为深度鉴伪任务提供更主动、有效的手段，满足公众对鉴别生成内容真伪的需求，保障社会信息的真实性和安全性。



一、研究背景与研究价值

当前技术存在的不足之处

- **水印安全性不足：**（1）**深度学习水印算法的局限性：**大多数现有方法通过最小化像素值修改嵌入水印，存在安全风险，如恶意用户可能通过黑盒方式获取含水印图像并规避水印，或利用图像再生成方案逃逸水印嵌入，甚至结合生成式AI直接抹除或伪造水印。（2）**特定水印方法的缺陷：**树轮水印和MDM水印方法虽有创新，但通常只能由模型所有者验证，且可能无法检测水印是否被破坏。目前根据输出分布嵌入水印的研究有限，但该领域有巨大发展潜力，未来需更多探索和创新。
- **实验评估有限：**在生成图水印领域，水印的嵌入与模型生成能力的匹配程度需要进一步实验评估，反取证工作有限。已有工作对文本水印的可靠性进行评估，指出在自然假设下模型带有强水印方案的不可实现性，但图像水印缺乏相关理论分析及方案评估。例如，是否可以通过分析生成图的统计分布进行水印查询，是否能够通过联合文本反转操作去除触发词以删除图像水印。逆向攻击的探索对于深化对水印技术的理解及其有效性的提升具有重要意义，为未来研究开辟了新的路径。



一、研究背景与研究价值

当前技术存在的不足之处

- **缺少主动防御：**在生成图水印领域，除了常规的被动防御措施，模型设计者也可以考虑加入对抗性扰动形成一种主动防御策略。对抗样本已在艺术作品保护中得到应用，通过添加扰动防止模型识别或重用图像。同样的策略可用于生成图水印，以提高恶意编辑的难度，有效保护版权。主动防御不仅增强了版权保护，也为防止未授权使用提供了新的解决方案。
- **生成速度较慢：**由于扩散模型固有特性，其推理速度通常较慢，限制了其大规模应用。在生成图水印实现过程中，加速扩散过程成为关键考虑因素，同时需保证模型适应多种生成任务且水印对生成内容有效嵌入。
- **平衡多方面因素困难：**在生成图水印算法的设计中，平衡水印容量、不可感知性、鲁棒性和生成质量是一个核心挑战。算法需使水印嵌入符合生成数据分布，同时确保在低误报率下实现高精度水印提取。生成质量角度：嵌入水印会降低模型生成质量，需将水印嵌入转化为符合扩散模型学习过程的方式，以最小化影响。鲁棒性角度：可结合文本水印和模型水印领域的经验评估生成图水印算法鲁棒性，尤其是结合文本攻击方法增强其鲁棒性。容量和不可感知角度：相比文本，图像有更高冗余度，为嵌入多比特水印提供可能，通过结合隐写技术可实现更符合数据分布的水印编码，如结合生成式隐写和可证安全领域的研究有效隐藏水印信息，为未来研究提供新探索方向。

01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划





二、国内外研究成果调研

论文调研情况

- FENG W T, ZHOU W B, HE J Y, et al. AquaLoRA: toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: ACM, 2025: 13423 - 13444
- XIONG C, QIN C, FENG G, et al. Flexible and Secure Watermarking for Latent Diffusion Model[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 1668-1676.
- YANG Z J, ZENG K, CHEN K J, et al. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 12162-12171.
- REZAEI A, AKBARI M, ALVAR S R, et al. Lawa: Using Latent Space for In-generation Image Watermarking[C]// Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2025: 118-136.
- LIU G H, CHEN T R, THEODOROU E A, et al. Mirror Diffusion Models for Constrained and Watermarked Generation[C]// Proceedings of the 37th International Conference on Neural Information Processing System, New Orleans: MIT Press, 2024: 42898 - 42917.
- BUI T, AGARWAL S, YU N, et al. Rosteals: Robust Steganography using Autoencoder Latent Space[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver : IEEE, 2023: 933-942.



二、国内外研究成果调研

论文调研情况

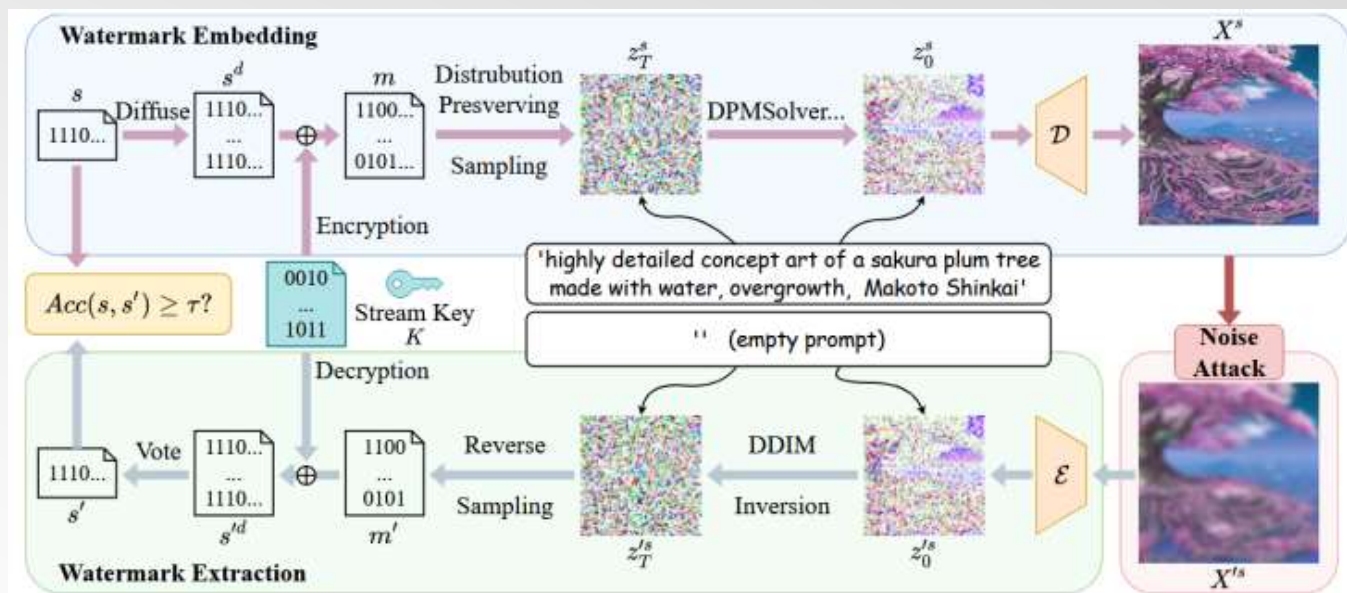
- FERNANDEZ P, COUAIRON G, JEGOU H, et al. The Stable Signature: Rooting Watermarks in Latent Diffusion Models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22466-22477.
- WEN Y X, KIRCHENBAUER J, GEIPING J, et al. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images[C]// Proceedings of the 37th International Conference on Neural Information Processing System, New Orleans: MIT Press, 2023: 58047 - 58063.
- CI H, YANG P, SONG Y R, et al. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-key Identification[C]// Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2025: 338-354.
- MIN R, LI S, CHEN H Y, et al. A Watermark-Conditioned Diffusion Model for Ip Protection[C]// Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2025: 104-120.
- ZHANG L J, LIU X, Martin A V, et al. Attack-Resilient Image Watermarking Using Stable Diffusion[C] //The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2024.
- Cui Y, Ren J, Xu H, et al. Diffusionshield: A watermark for copyright protection against generative diffusion models[J]. arXiv preprint arXiv:2306.04642, 2023.



二、国内外研究成果调研

Gaussian Shading

- YANG Z J, ZENG K, CHEN K J, et al. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 12162-12171.
- 该篇文章旨在解决扩散模型在版权保护和防止滥用方面的伦理问题。现有的水印技术要么降低模型性能，要么需要额外的训练，这对模型的操作者和用户都不理想。文章提出了一种名为Gaussian Shading的水印技术，该技术通过将水印映射到遵循标准高斯分布的潜在表示上，实现了无损性能的水印嵌入。具体方法包括三个主要步骤：水印扩散、随机化和保持分布的采样。
- 水印扩散将水印信息传播到整个潜在表示中，随机化确保水印的分布与原始潜在表示一致，而保持分布的采样则确保水印嵌入后图像的分布与正常生成的图像一致。此外，文章还提供了理论证明，证明了该方法的性能无损特性。

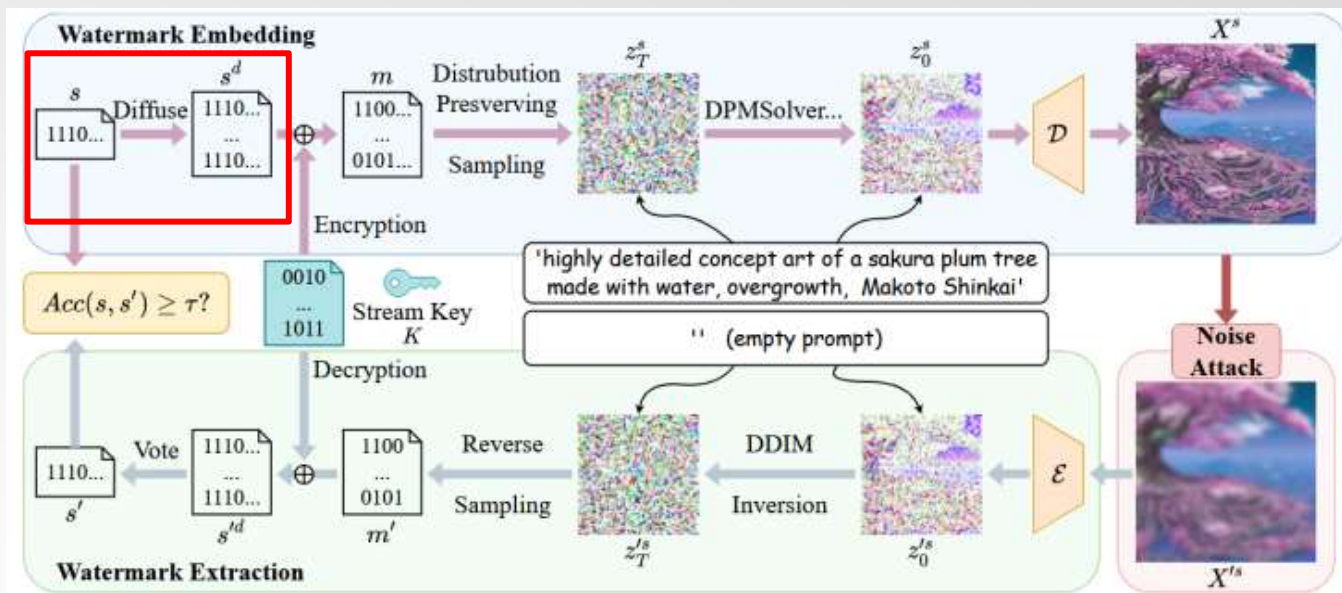
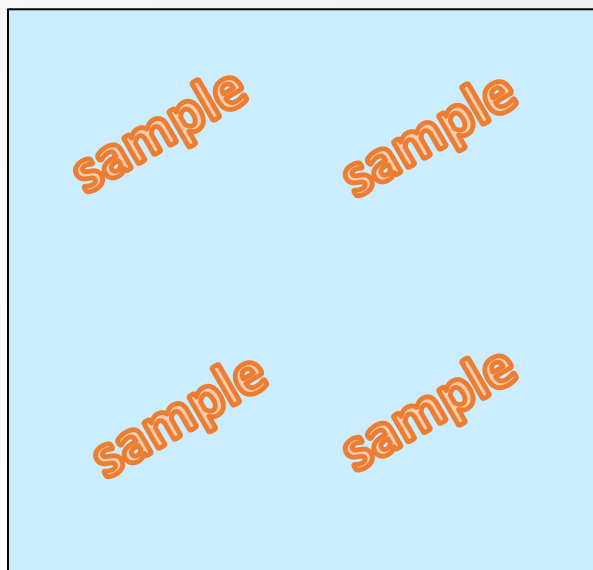




二、国内外研究成果调研

Gaussian Shading

- 水印扩散：**潜在表示的维度由 $c \times h \times w$ 给出，每个维度可以表示 l 位的水印。因此，水印容量变为 $l \times c \times h \times w$ 位。为了增强水印的鲁棒性，这里使用了一种**重复扩展小容量水印**的方式。具体来说使用了 $\frac{1}{f_{hw}}$ 作为高度和宽度的缩放因子，以及 $\frac{1}{f_c}$ 作为通道数的缩放因子，这样小水印 s 的容量就变为 $l \times \frac{c}{f_c} \times \frac{h}{f_{hw}} \times \frac{w}{f_{hw}}$ ，并重复扩展这种小容量水印 $f_c \cdot f_{hw}^2$ 次。因此，维度为 $l \times \frac{c}{f_c} \times \frac{h}{f_{hw}} \times \frac{w}{f_{hw}}$ 的水印 s 扩展为维度为 $l \times c \times h \times w$ 的扩散水印 s^d 。

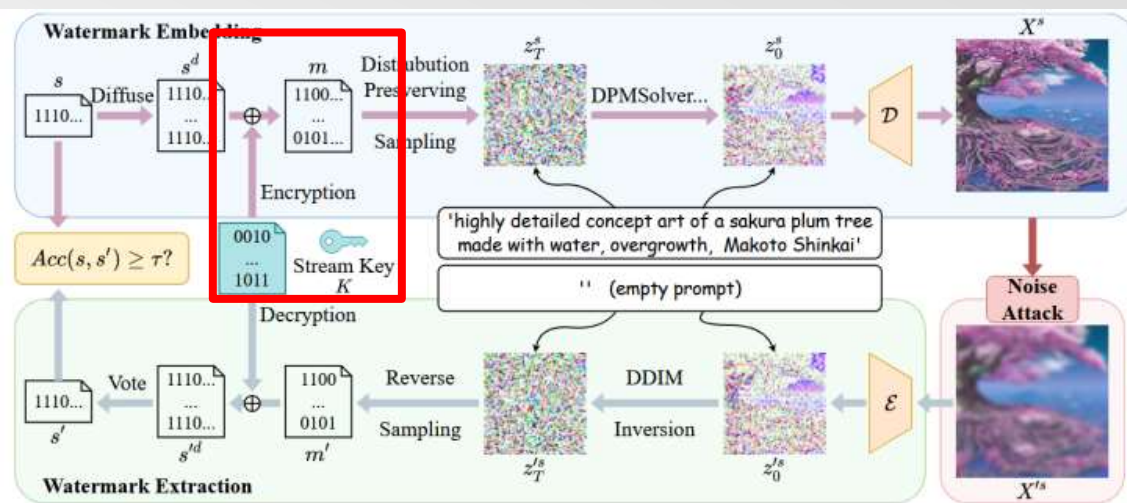




二、国内外研究成果调研

Gaussian Shading

- **水印随机化：**水印随机化：如果我们知道扩散水印 s^d 的分布，我们可以直接使用保持分布的采样来获得相应的潜在表示 z_T^s 。然而，在实际场景中，其分布总是未知的。因此，我们引入一个流密钥 K ，通过加密将 s^d 转换为分布已知的随机水印 m 。考虑使用计算安全的流密码，文中使用了ChaCha20加密算法。 m 遵循均匀分布，即 m 是一个随机的二进制比特流。
- 在Gaussian Shading方法中，水印随机化是一个关键步骤，目的是将水印信息转换成一种看似随机但实际上遵循特定分布的形式，以便它可以被嵌入到图像的潜在表示中，而不改变图像的整体统计特性。加密后的水印 m 现在遵循均匀分布，这意味着每个二进制位是0或1的概率是相等的，看起来完全随机。这样做主要可以保持图像质量（因为加密后的水印 m 看起来是随机的，它不会对图像的视觉质量产生明显的影响。）和提高安全性（由于水印是加密的，即使有人知道水印被嵌入了，他们也无法轻易地检测或移除它，因为不知道密钥就无法解密水印。）
- 例如假设有一个原始水印为01110111，它是一个特定的模式或信息，经过流密钥 K 加密后得到10110001，水印变成完全随机且0和1遵循均匀分布。





二、国内外研究成果调研

Gaussian Shading

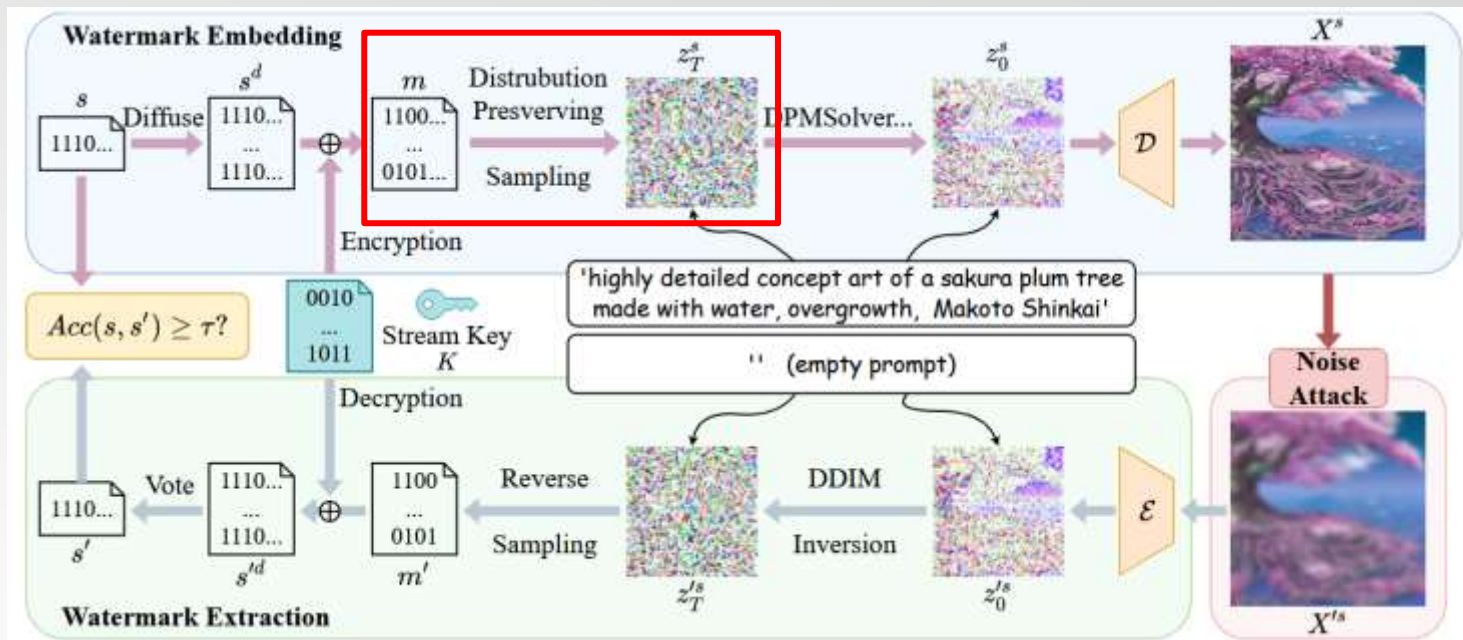
- **随机水印驱动分布保持采样：**分布保持采样的目的是将加密后的水印信息嵌入到图像的潜在表示中，同时确保嵌入水印后的潜在表示仍然遵循与原始潜在表示相同的分布。这样可以保证嵌入水印后的图像在视觉上与原始图像没有明显差异，从而不影响模型的性能。
- **整个过程可以大致分为五个步骤：**（1）随机水印（ m ）：通过水印随机化过程得到一个加密后的水印，它是一个二进制比特序列，长度是 $l \times c \times h \times w$ 。（2）将水印转换为整数，因为每一个水印位可以表示 l 位比特，因此它可以被视为一个整数，整数的取值范围是 $[0, 2^l - 1]$ 。例如 $l = 2$ ，则可表示的整数为 0、1、2、3。（3）高斯分布划分：将高斯分布划分为等概率区间，划分数位 2^l 。例如有四个区间，则每个区间的累计概率分布为 0-0.25、0.25-0.50、0.50-0.75、0.75-1.00。（4）采样过程：根据水印位对应的整数，从相应的高斯分布区间随机采样一个值，例如整数为 1，则从 0.25-0.50 的区间采样一个随机值。（5）生成水印的潜在表示：通过上述采样过程，生成嵌入水印的潜在表示 z_T^s 。这个 z_T^s 遵循与原始潜在表示相同的高斯分布。
- **采取保持分布采样的优势有：**（1）保持分布：通过从高斯分布的特定区间内采样，我们确保嵌入水印后的潜在表示 z_T^s 仍然遵循与原始潜在表示相同的高斯分布。这保证了嵌入水印后的图像在视觉上与原始图像没有明显差异。（2）鲁棒性：由于水印信息被扩散到整个潜在表示中，它对有损处理和擦除尝试具有很强的鲁棒性。（3）性能无损：因为嵌入水印后的潜在表示与原始潜在表示遵循相同的分布，所以模型的性能不会受到影响。



二、国内外研究成果调研

Gaussian Shading

- 分布保持采样



- 水印提取:** 在水印提取之间通过图形变换模拟图片在现实传播中遇到的各种恶意或非恶意攻击。水印提取过程是水印嵌入过程的逆过程，通过反转扩散过程将图像还原成原始噪声，从噪声向量中采样 m' ，利用流密钥 K 解密得到扩散水印 s'^d ，根据缩放因子得到 s' ，将 s 与 s' 进行比特位比较，高于阈值 τ 则检测成功。



二、国内外研究成果调研

Gaussian Shading

- **创新性：**该篇文章的最大创新点在于提出了一种无损性能的水印嵌入技术Gaussian Shading。这是首个在扩散模型中实现性能无损水印嵌入的方法，并且提供了理论证明。Gaussian Shading通过将水印信息扩散到整个潜在表示中，实现了水印与图像语义的深度绑定，从而在不牺牲模型性能的情况下提供了强大的鲁棒性。此外，该方法无需对模型进行额外训练，可以即插即用，易于集成到现有的生成过程中。
- **不足：**尽管Gaussian Shading在性能无损水印嵌入方面取得了显著成果，但该篇文章也存在一些不足之处。首先，该方法依赖于DDIM反转，这限制了其在不使用基于ODE求解器的连续时间采样器的场景中的应用。其次，使用流密码需要在部署平台上进行适当的密钥使用和管理，这增加了实际应用中的复杂性。此外，文章假设模型不是公开可访问的，这在一定程度上限制了水印验证的灵活性。最后，Gaussian Shading容易受到伪造攻击，这需要运营商采取额外措施来保护模型参数。



二、国内外研究成果调研

Flexible and Secure Watermark

- XIONG C, QIN C, FENG G, et al. Flexible and Secure Watermarking for Latent Diffusion Model[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 1668-1676.
- **研究背景：**由于潜在扩散模型（LDMs）在图像生成领域的重大进展和开源支持，许多研究人员和企业开始微调预训练模型，以生成不同目标的专用图像。然而，犯罪分子可能会利用LDMs生成图像，然后进行非法活动。水印技术是解决这一问题的典型方案。但是事后水印容易被绕过以获得无水印图像，而现有的针对LDMs的水印微调方法只能嵌入固定的消息水印，即除非重训练，无法嵌入新的消息，这会在将模型分配给第三方使用者时需要再次训练或微调，消耗计算资源和时间。
- Xiong等人提出了一种用于LDMs的灵活且安全的水印方法。其目标包括：
 - (1) 实现无需重新训练或微调LDM即可灵活更改嵌入消息；
 - (2) 防止模型用户绕过消息嵌入，确保生成的图像中包含水印；
 - (3) 保持水印图像的高质量和鲁棒性，使其适用于图像认证和模型用户识别。

Flexible and Secure Watermark

- 文章提出了一种基于编码器-解码器和消息矩阵的端到端水印方法。具体方法包括：
 - 消息编码器 (Em) :将消息转化为消息矩阵，以便与LDM解码器的中间输出结果融合。
 - 消息解码器 (Dm) : 从水印图像中提取消息。
 - 消息嵌入：在LDM解码器中融合消息矩阵，生成水印图像。
 - 攻击层：模拟实际使用场景中的各种攻击，增强水印图像的鲁棒性。
 - 损失函数和安全机制：设计损失函数和训练策略，确保消息嵌入和提取的成功，并防止用户绕过消息矩阵的使用。

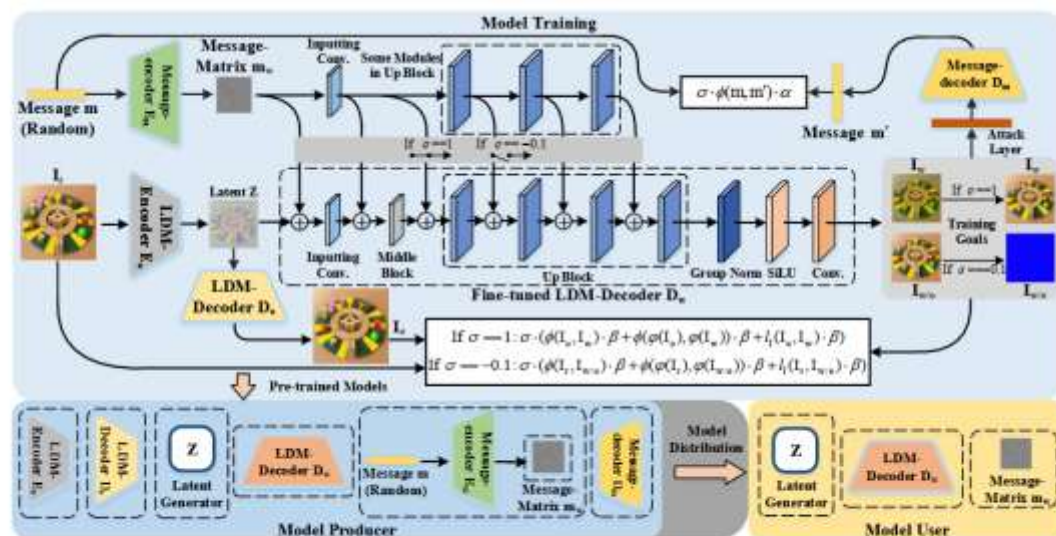


Figure 2: The training strategy of proposed watermarking method for Latent Diffusion Model. I_r is the real image, I_0 is the original generated image, I_w is the watermarked image and $I_{w/o}$ is the non-watermarked image.



二、国内外研究成果调研

Flexible and Secure Watermark

- 消息编码：**直接将二进制比特流消息 m 与 D_w 中的中间输出融合是困难的，因此设计了消息编码器 E_m 以将 m 转换为更适合融合的消息矩阵 m_w 。 E_m 的结构如图所示，主要由5个全连接（FC）层和6个卷积层组成。为了将向量 m 转换为矩阵 m_w ，设计的5个FC层首先用于预处理 m 以获得大小为 $1 \times (64 \times 64)$ 的向量 V 。然后，对 V 执行PyTorch的reshape和repeat操作以获得大小为 $1 \times 4 \times 64 \times 64$ 的矩阵 V' ，可以将其正常输入到其余的卷积层中。为了在 D_w 中更好地融合，设计的6个卷积层进一步处理 V' 以获得消息矩阵 m_w 。需要注意的是，在训练期间，应实时随机生成消息，以确保所有类型的 m_w 都可以在微调的LDM中使用。

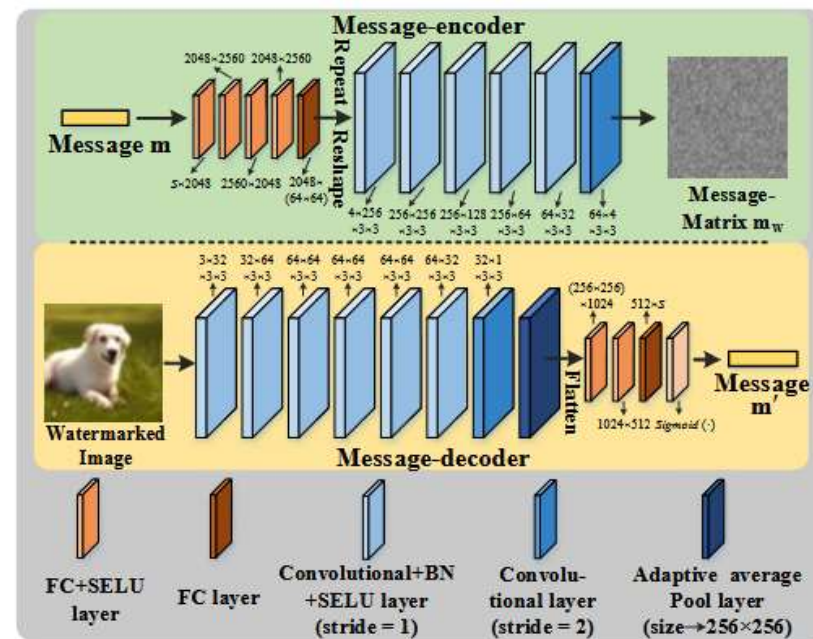


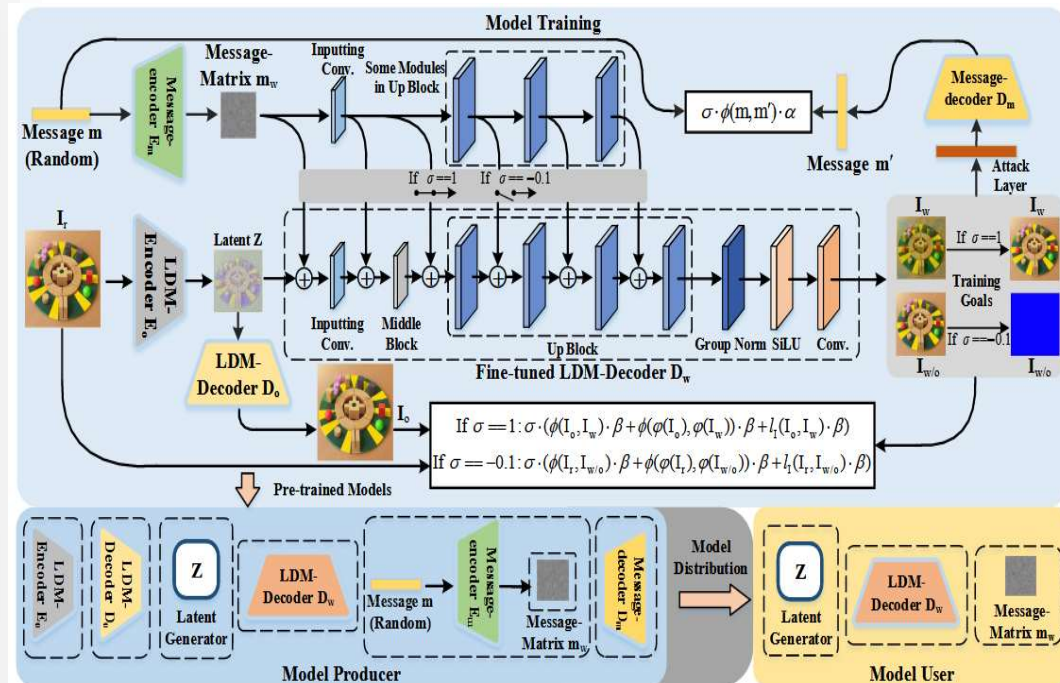
Figure 4: The structure of message-encoder E_m and message-decoder D_m .



二、国内外研究成果调研

Flexible and Secure Watermark

- 消息嵌入**：通常，LDM的潜在生成器从噪声中采样潜在表示 Z ，原始LDM解码器 D_0 从 Z 中恢复图像。为了保持原始生成图像 I_0 和水印生成图像 I_w 之间的语义一致性，消息矩阵 m_w 在 D_w 的图像生成阶段融合，这是通过微调 D_0 获得的。融合的具体细节如图2所示，我们选择了一些层（即输入卷积层、中间块和前三个上采样模块）的中间输出来融合消息矩阵 m_w 。因此， m_w 逐渐在 D_w 中融合以生成水印图像 I_w 。需要注意的是，本工作中使用的LDM解码器是变分自编码器（VAE），它在稳定扩散模型中被广泛使用。因此，由于消息矩阵 m_w 在 D_w 中的使用方式，模型生产者可以根据需要灵活更改消息矩阵，无需再次训练或微调。

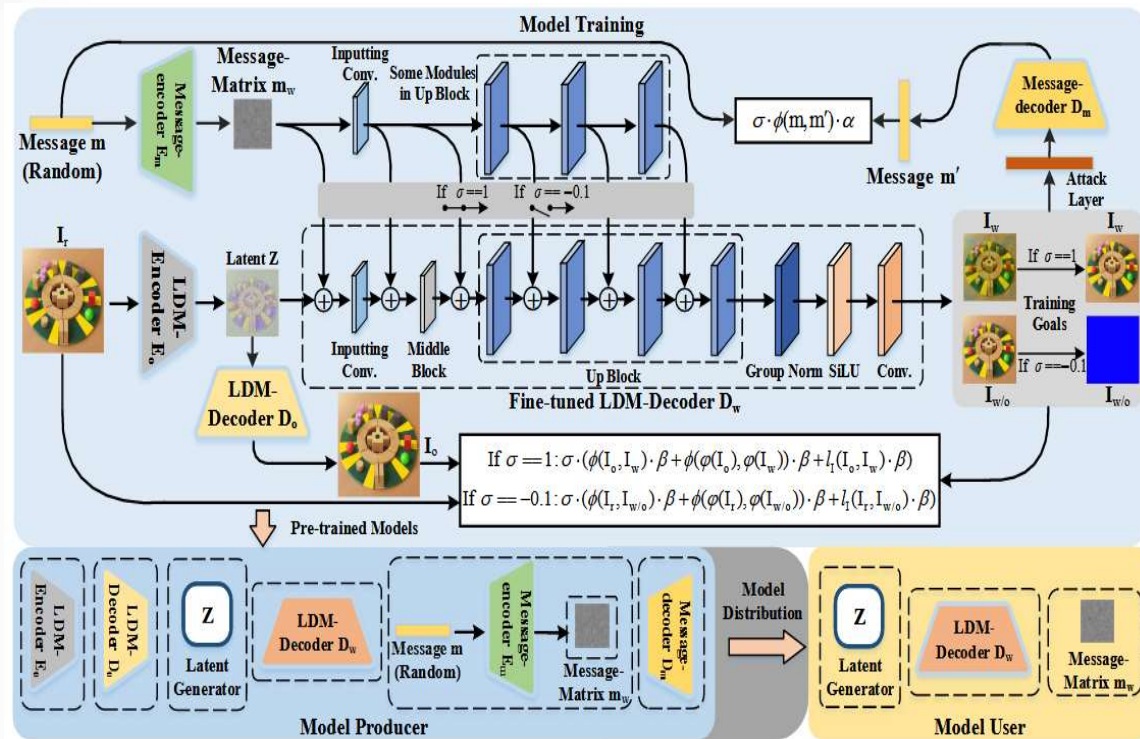




二、国内外研究成果调研

Flexible and Secure Watermark

- 消息解码：**使用消息解码器 D_m 从 I_w 中提取消息 m' 。 D_m 的结构如图4所示，主要由7个卷积层和3个FC层组成。由于 D_m 的输入是水印图像 I_w ，这些卷积层旨在从 I_w 中提取与消息相关的特征。然后，将这些特征展平并输入到最后3个FC层中以获得向量 m_t ，它与 m' 相似。为了增强消息解码器 D_m 的非线性拟合能力，并确保 m' 的每个元素接近0或1，我们使用 $Sigmoid(\cdot)$ 处理 m_t ，并获得消息 $m' = Sigmoid(m_t \times 10)$ 。因此，消息提取可以被视为消息嵌入的逆过程。

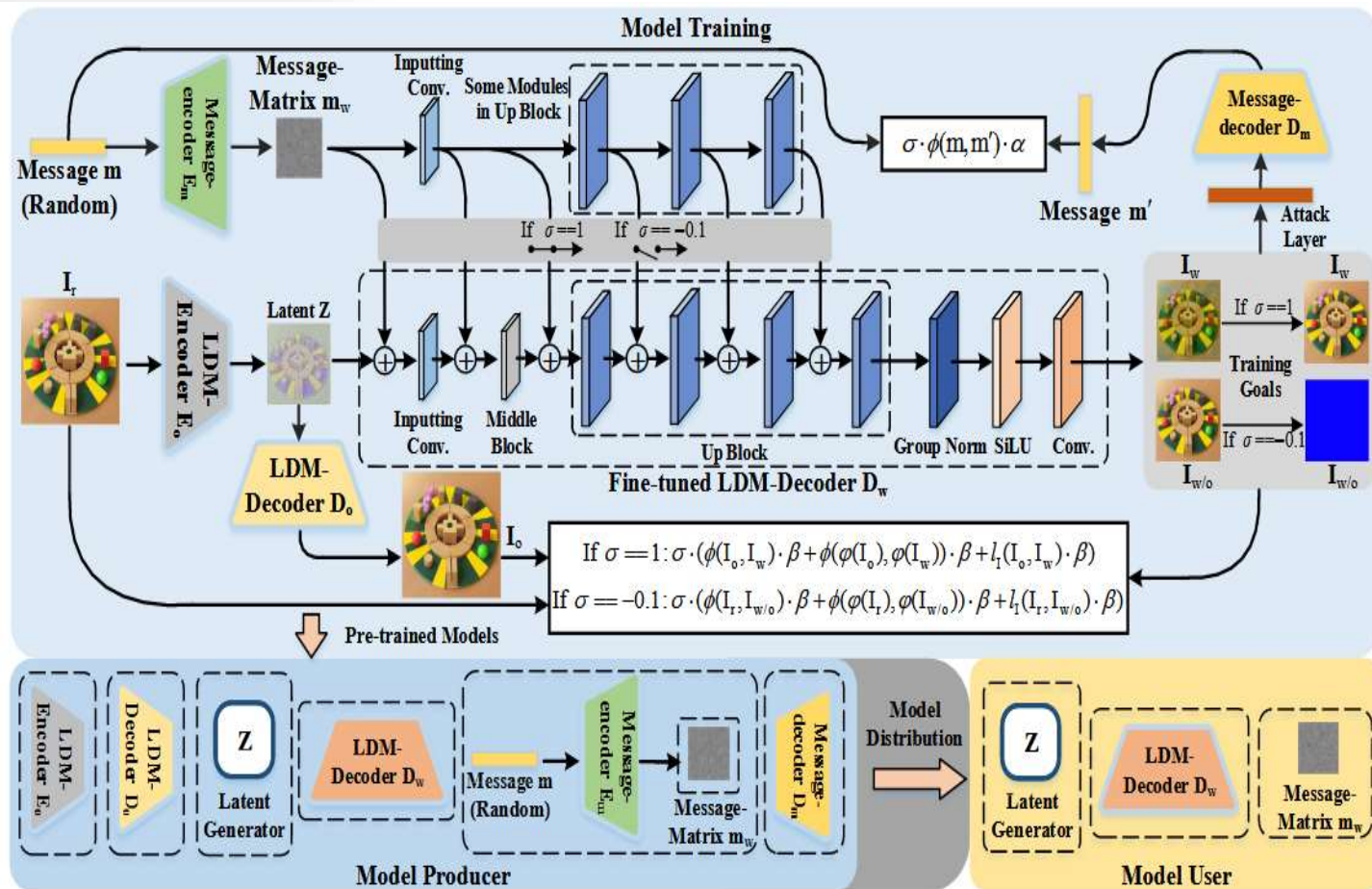




二、国内外研究成果调研

Flexible and Secure Watermark

- 攻击层：**由于在实际的图像使用场景中存在各种类型的攻击，我们使用攻击层在训练期间处理水印图像 I_w ，然后将其输入到消息解码器 D_m 中。这个攻击层涵盖了7种常见的攻击，即模糊、高斯噪声、亮度调整、对比度调整、饱和度调整、透视变形和JPEG。在训练过程中，攻击的强度逐渐增加到最大值。

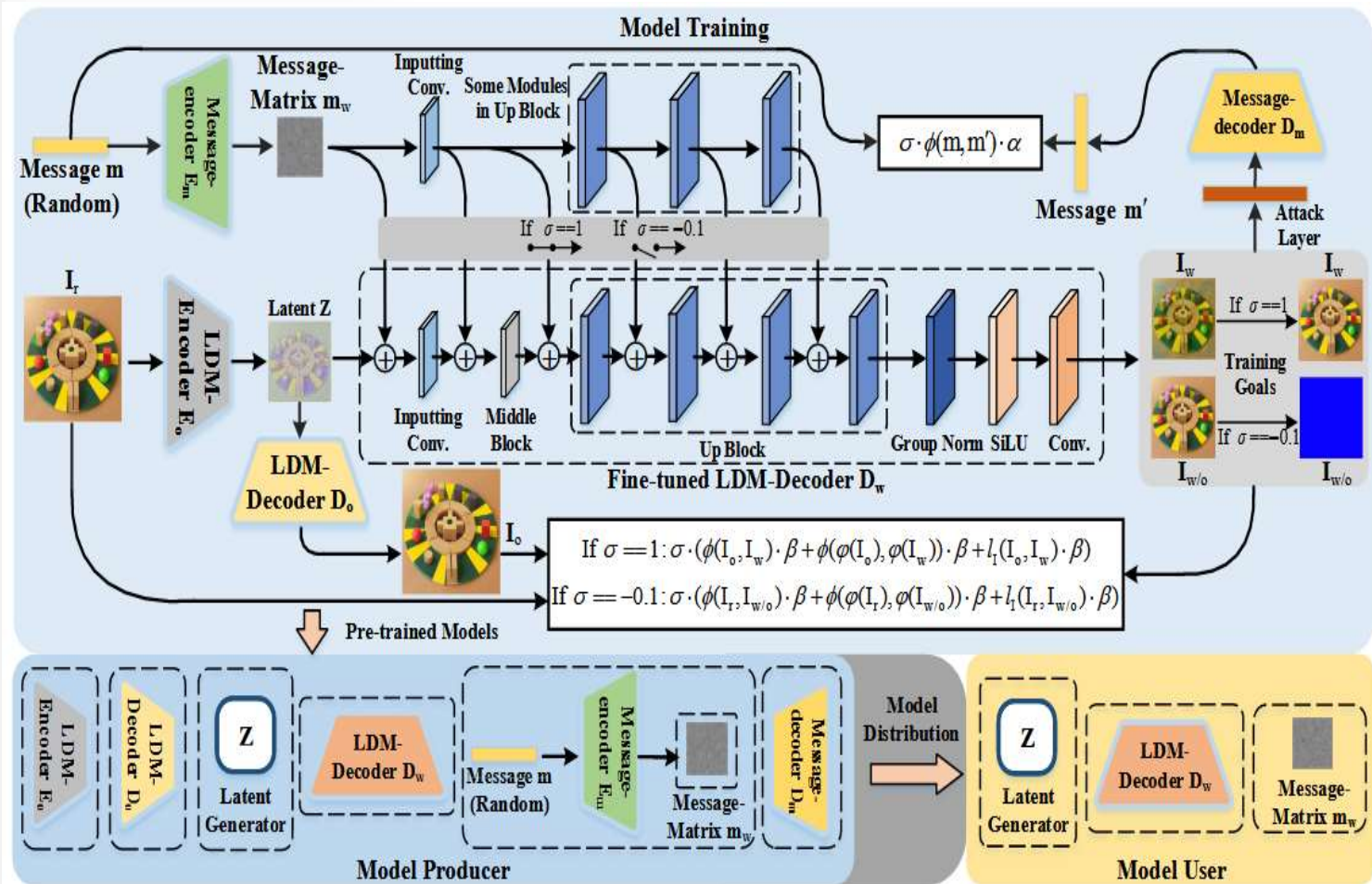




二、国内外研究成果调研

Flexible and Secure Watermark

- 损失函数和安全机制：
- 衡量预测值和真实值之间的差距：（1）消息嵌入的准确性：衡量输入消息和输出消息；（2）图像质量保持：
- 衡量原始图像和水印图像。调整模型训练方向：（1）安全机制的触发：当用户绕过水印时，触发安全机制，降低图像质量；（2）平衡不同目标：平衡损失函数中的不同部分，调整超参数。





二、国内外研究成果调研

Flexible and Secure Watermark

- **创新性：**灵活的消息嵌入（通过消息编码器生成消息矩阵，无需重新训练或微调LDM即可更改嵌入的消息）、安全机制（设计了一种安全机制，防止模型用户绕过消息矩阵的使用，确保生成的图像中包含水印）、高质量和鲁棒性（通过实验验证，所提出的水印方法生成的图像具有高质量和鲁棒性，适用于实际应用）
- **不足：**
 - **实验数据有限：**虽然实验结果表明了方法的有效性，但使用的数据集和实验场景可能有限，需要在更广泛的数据集和实际应用中进一步验证。
 - **计算资源消耗：**尽管方法无需重新训练或微调LDM，但初始的训练和微调过程仍然需要较大的计算资源。
 - **攻击类型的多样性：**虽然考虑了多种攻击类型，但在实际应用中可能还会遇到其他类型的攻击，需要进一步增强水印的鲁棒性。



二、国内外研究成果调研

Tree Ring

- WEN Y X, KIRCHENBAUER J, GEIPING J, et al. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images[C]// Proceedings of the 37th International Conference on Neural Information Processing System, New Orleans: MIT Press, 2023: 58047 - 58063.
- 研究目的:
- 提出一种新的水印技术，用于在扩散模型生成的图像中添加不可见且稳健的指纹，以追踪版权和防止 AI 生成内容的潜在危害。
- 解决现有水印方法在图像生成后进行事后修改导致的可见性问题，实现真正不可见的水印。
- 提高水印对常见图像变换（如裁剪、颜色抖动、扩张、翻转、旋转或噪声）的鲁棒性，以确保在日常使用和处理生成图像时水印的可靠性。
- 研究方法:
- Tree-Ring Watermarking 技术：在扩散模型的初始噪声向量的傅里叶空间中嵌入精心构建的模式（密钥），通过微妙地影响整个采样过程，生成对人类不可见的模型指纹。
- 傅里叶空间模式设计：利用傅里叶变换对周期信号的不变性属性，设计了 Tree-RingZeros、Tree-RingRand 和 Tree-RingRings 三种类型的密钥模式，以实现针对不同图像操作的鲁棒性。
- 水印检测方法：通过反转扩散过程来检索生成图像的初始噪声向量，然后在傅里叶空间中检查嵌入的密钥信号，计算 L1 距离或使用统计测试来判断水印的存在。



二、国内外研究成果调研

Tree Ring

- 文中设计了三种不同的水印模式，分别是Tree-RingZeros、Tree-RingRand、Tree-RingRings,它们各有各的好处，接下来将分别详细介绍它们。
- Tree-Ring_{Zeros}：选择掩码为圆形区域，以保持对图像空间中旋转的不变性。密钥被选择为零数组，这创造了对平移、裁剪和扩张的不变性。这个密钥对操作是不变的，但以严重偏离高斯分布为代价。它还阻止使用多个密钥来区分模型。
- Tree-Ring_{Rand}：从高斯分布中抽取一个固定密钥 k^* 。密钥具有与噪声数组的原始傅里叶模式相同的 iid 高斯性质，因此我们预计这种策略对生成质量的影响最小。这种方法还为模型所有者提供了拥有多个密钥的灵活性。然而，它对图像操作不是不变的。
- Tree-Ring_{Rings}：引入了一个由多个环组成的模式，并且每个环上的值是常数。这使得水印对旋转不变。我们从高斯分布中选择恒定的环值。这为多种类型的图像变换提供了一些不变性，同时确保整体分布仅从各向同性高斯中最小地偏移。



二、国内外研究成果调研

Tree Ring

- 模型框架

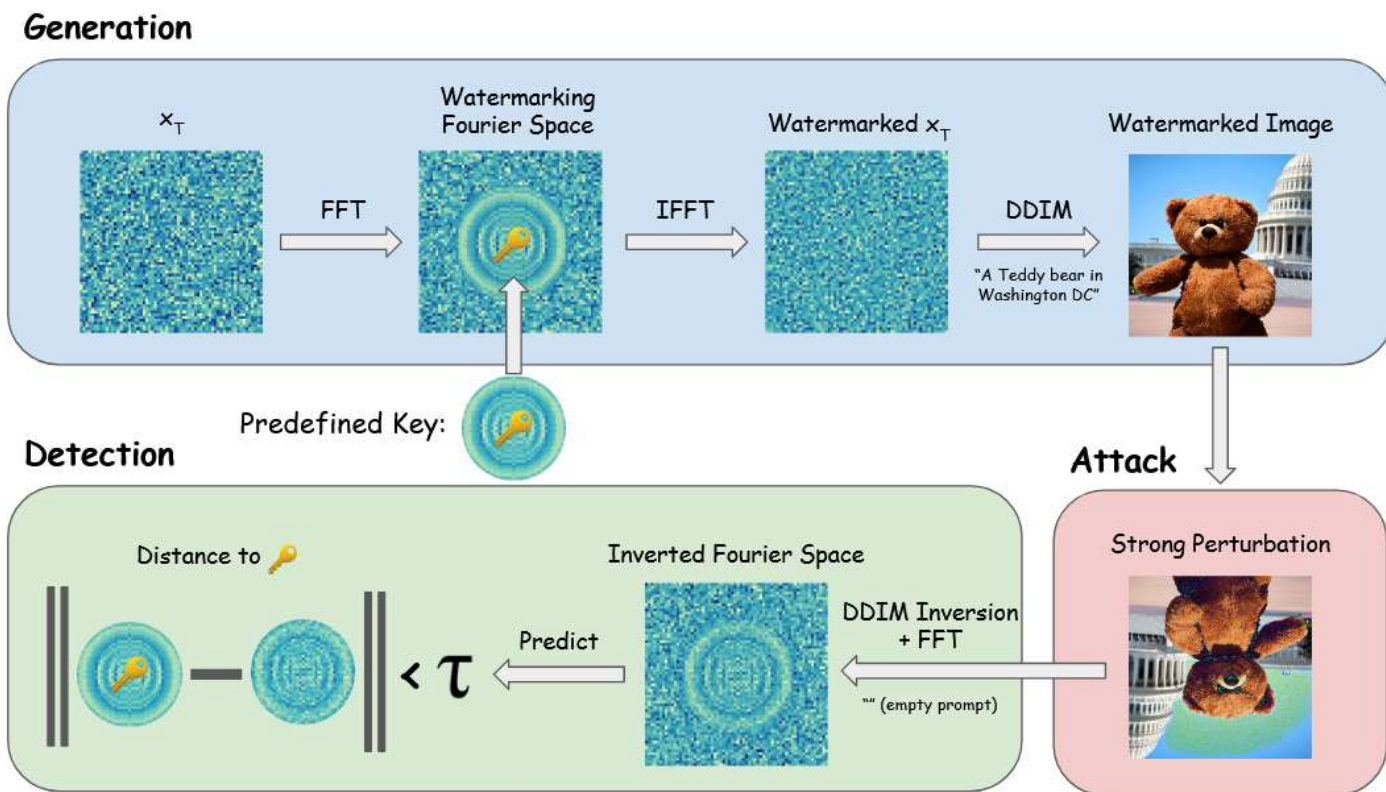


Figure 1: Pipeline for *Tree-Ring Watermarking*. A diffusion model generation is watermarked and later detected through ring-patterns in the Fourier space of the initial noise vector.



二、国内外研究成果调研

Tree Ring

- **创新性:**

- 不可见水印技术: 首次提出了一种真正不可见的水印技术, 通过在扩散模型的采样过程中嵌入水印, 而不是在生成图像后进行事后修改, 实现了对人类视觉的不可见性。
- 傅里叶空间密钥模式: 创新性地在傅里叶空间中设计了多种密钥模式, 利用傅里叶变换的不变性属性, 使水印对常见的图像操作具有鲁棒性, 提高了水印的可靠性和安全性。
- 无需额外训练或微调: Tree-Ring Watermarking 技术的实现不需要对扩散模型进行额外的训练或微调, 可以直接应用于现有的扩散模型 API, 降低了水印技术的应用门槛和成本。

- **不足:**

- Tree-Ring水印要求在推理时使用DDIM模型, 如果是其他模型, 需要调整水印以适应其他采样模式。
- 需要模型所有者来验证水印, 因为需要模型参数来执行反转过程, 这限制了第三方在不利于API的情况下检测水印。
- 水印的容量不足, 这是一种零比特水印, 只能用作模型版权校验, 无法为每一个用户分配唯一ID进行追踪溯源。



二、国内外研究成果调研

Ring-ID

- CI H, YANG P, SONG Y R, et al. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-key Identification[C]// Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2025: 338-354.
- 研究目的：
 - 重新审视树环水印：文章旨在重新审视树环水印方法，揭示了其在水印过程无意引入的分布偏移，并分析了其固有的设计缺陷，特别是在识别多个不同密钥的能力上。
 - 增强多密钥识别能力：提出RingID方法，以增强扩散模型水印在多密钥识别任务中的鲁棒性和准确性。
- 研究方法：
 - 深入分析：对Tree-Ring水印的鲁棒性进行深入分析，揭示其在验证和识别任务中的表现差异。
 - 数学建模与实验验证：通过数学推导和实验验证，分析分布偏移对水印鲁棒性的影响。
 - 提出RingID：设计RingID方法，采用多通道异构水印框架，结合离散化和无损印入等技术，提升水印的区分能力和鲁棒性。



二、国内外研究成果调研

Ring-ID

- Ring-ID从四个大方向考虑增强Tree-Ring，分别是多通道异构水印、旋转不变性、离散化和容量。接下来分别介绍每种方法。

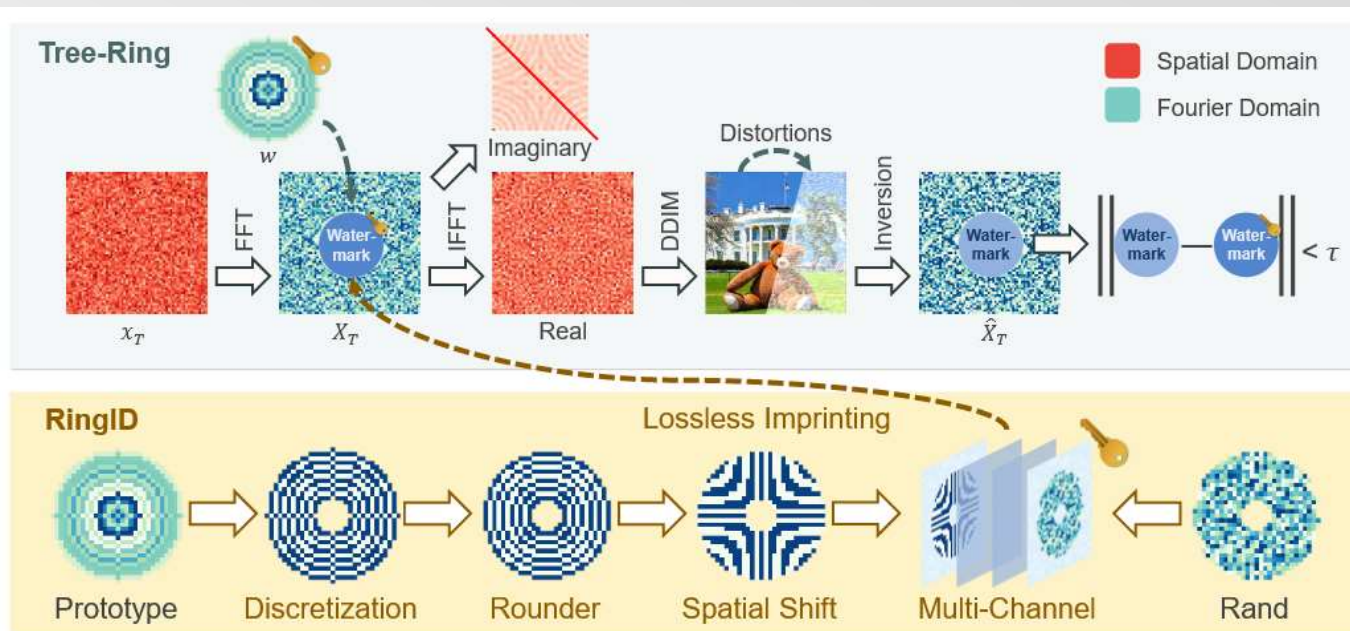


Fig. 2: Framework of the watermarking process. *RingID* introduces a series of approaches that can help to imprint a lossless and robust watermark. In contrast, *Tree-Ring* injects a lossy and less robust watermark.



二、国内外研究成果调研

Ring-ID

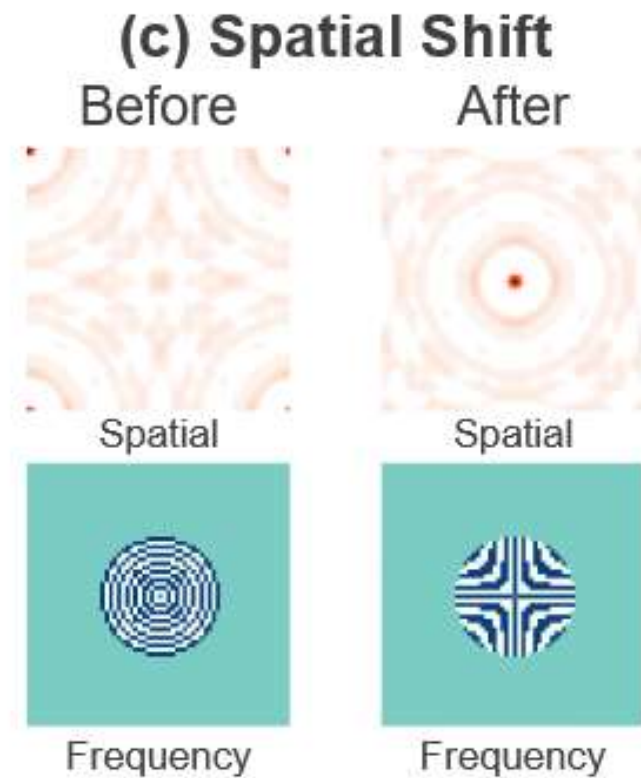
- 多通道异构水印 (MCH) 是基于一种向不同通道添加水印以融合多种水印各自优势的直觉, 设计而成的一种水印方法。在identification任务上遍历所有水印通道上的每一个可能水印, 将其与恢复得到的水印计算L1范数距离, 最小值对应的水印就是原始水印。这基于一种思想, 对某种攻击具有更强鲁棒性的水印会得到比较差鲁棒性水印更小的L1范数距离。(这增强了在不同攻击下选择最具鲁棒性水印的自适应性) MCH的总容量是每个水印的最小值, 因此建议在每一个水印通道上使用具有充足水印容量的水印模式。实验发现高斯噪声水印足有近乎无穷的容量并且和初始噪声具有相同的分布, 同时对非几何攻击有强鲁棒性, 和Tree-ring相辅相成, 是一个完美的选择。本文实验表明高斯噪声水印和Tree-ring的结合能完美融合两者的独特优势。



二、国内外研究成果调研

Ring-ID

- 旋转不变性又分为三个子模块，分别是角裁剪避免、无损嵌入和圆环增强。
- 空间移动以避免角裁剪：从图中可以看出，在频域中处于中心的Tree-ring水印，在空间域中分布到四个角上，造成了其在旋转过程中面对角裁剪的脆弱性。因此可以对空间域的图片在高和宽两个维度进行半长的循环移动，使得空间域中的水印分布到图像中心。这等于对频域水印直接乘以一个 $H[u,v]$ 给出的棋盘图案。但是这个操作也会导致在生成的图像中心产生一个圆形阴影。为了保证生成图像的质量，进一步对移动后的模式乘以一个影响因子，来降低中心峰值。实验发现，当影响因子在0.8~0.9之间时能在图像质量和水印鲁棒性之间取得一个较好的平衡。

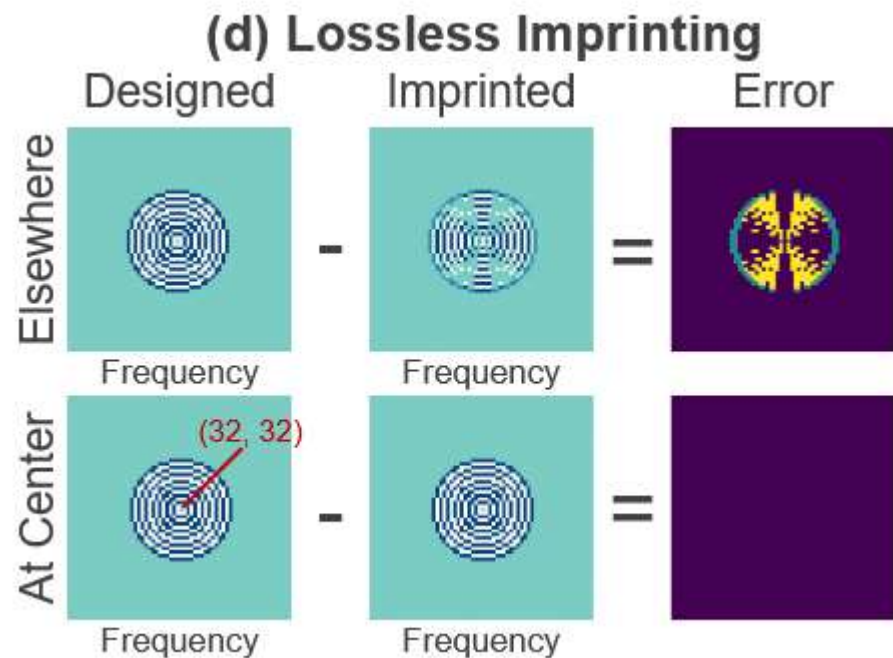




二、国内外研究成果调研

Ring-ID

- 无损嵌入：由图可知，在频域中嵌入水印后转换回空间域的过程中会丢失虚部内容，这不仅导致了L1-to-reference的分布改变，还破坏了水印模式，从而使水印模式失去旋转对称性（tree-ring在rotation上表现差）。水印噪声转回空间域的只有它的实部，参与图像生成的只有实部，这实际上和水印噪声的共轭是等效的。为满足环状模式的旋转对称性，需要消除变换过程中的旋转损失，方法是使水印噪声和其共轭噪声相等。文中指出实部需要关于傅里叶中心呈偶函数，虚部呈奇函数。为满足限制，将虚部内容变为空。加之，将tree-ring的中心从 (31, 32) 对齐到 (32, 32)。

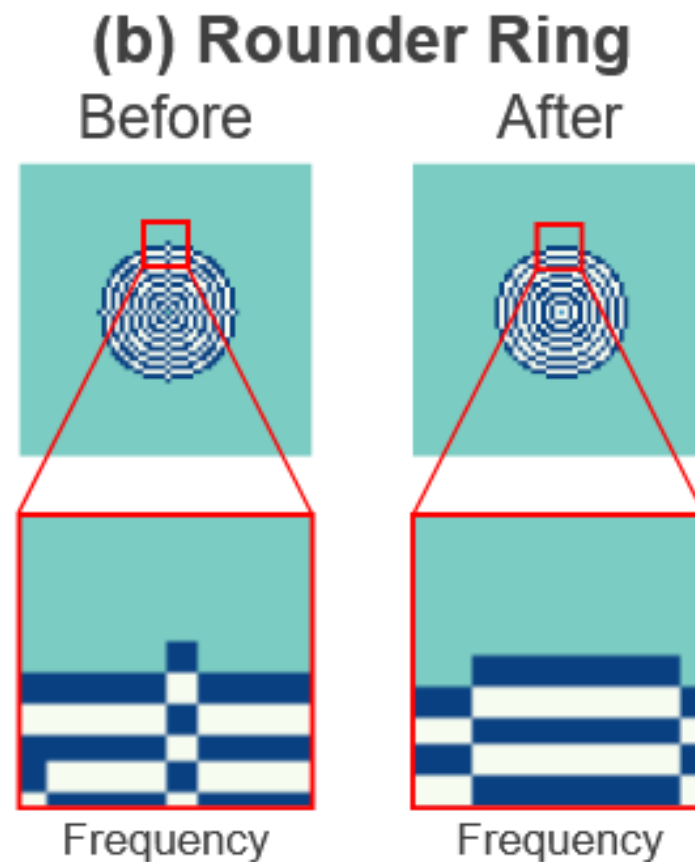




二、国内外研究成果调研

Ring-ID

- 设计一个更圆的环：在tree-ring中，环的设计是通过定位到给定圆心的等距像素来决定的。这是一种粗糙的方法，尤其在低分辨率的情况（例如 64×64 的image），生成的圆环不够圆润。解决方法是在距离旋转中心为 r 的地方，在黑色背景下放置一个白色像素，将低像素图片旋转360度并记录下白色像素的运动轨迹，获得一个半径为 r 的圆环。





二、国内外研究成果调研

Ring-ID

- 离散化以增强区分性：tree-ring在从高斯分布中为每个环采样数值，以保持其和初始噪声的分布一致性。然而这种随机采样策略极大的增加了不同key之间的区分难度（identification task）。文中的解决方法是规定每个环的数值不是 α 就是 $-\alpha$ ，使数值离散化（理论上会造成生成质量的降低）。这样做会减少水印容量的理论上限，但它显著增强了其有效容量，确保可用插槽的最佳利用。文中实验表明将 α 设置为初始噪声的标准差可以在确保不同key间的区分度的同时将对生成质量的影响降到一个很小的水平。
- 增加容量：文中考虑了两种增加容量的方法：调整单个通道中环的数量、在多个通道中嵌入环。总结：在单个通道中增加水印可以有效增加水印的容量，但过多的水印数量会明显降低鲁棒性和生成质量；在多个通道中扩展水印指数级的增加了容量，然而，它可能会在生成的图像中心引入环形伪影。最终，本文决定在单个通道内调整容量。



二、国内外研究成果调研

Ring-ID

- **创新性：**
 - 多通道异构水印框架：提出多通道异构水印框架，有效融合不同水印的优势，提升水印在多密钥识别中的鲁棒性和准确性。
 - 离散化和无损印入技术：引入离散化和无损印入技术，显著提高了水印的区分能力，减少了图像质量损失。
 - 系统性解决方案：提供了一个系统性的解决方案，从理论和实践两个层面解决了Tree-Ring在多密钥识别中的局限性。
- **不足：**
 - 对某些攻击的鲁棒性仍有提升空间：尽管RingID在多密钥识别中取得了显著进步，但在裁剪和缩放攻击下的鲁棒性仍有提升空间。
 - 复杂度和计算成本：多通道异构水印框架可能会增加水印的复杂度和计算成本，需要在实际应用中进行优化平衡。

01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划





三、代表性工作复现进展

工作复现1 “Tree-Ring”

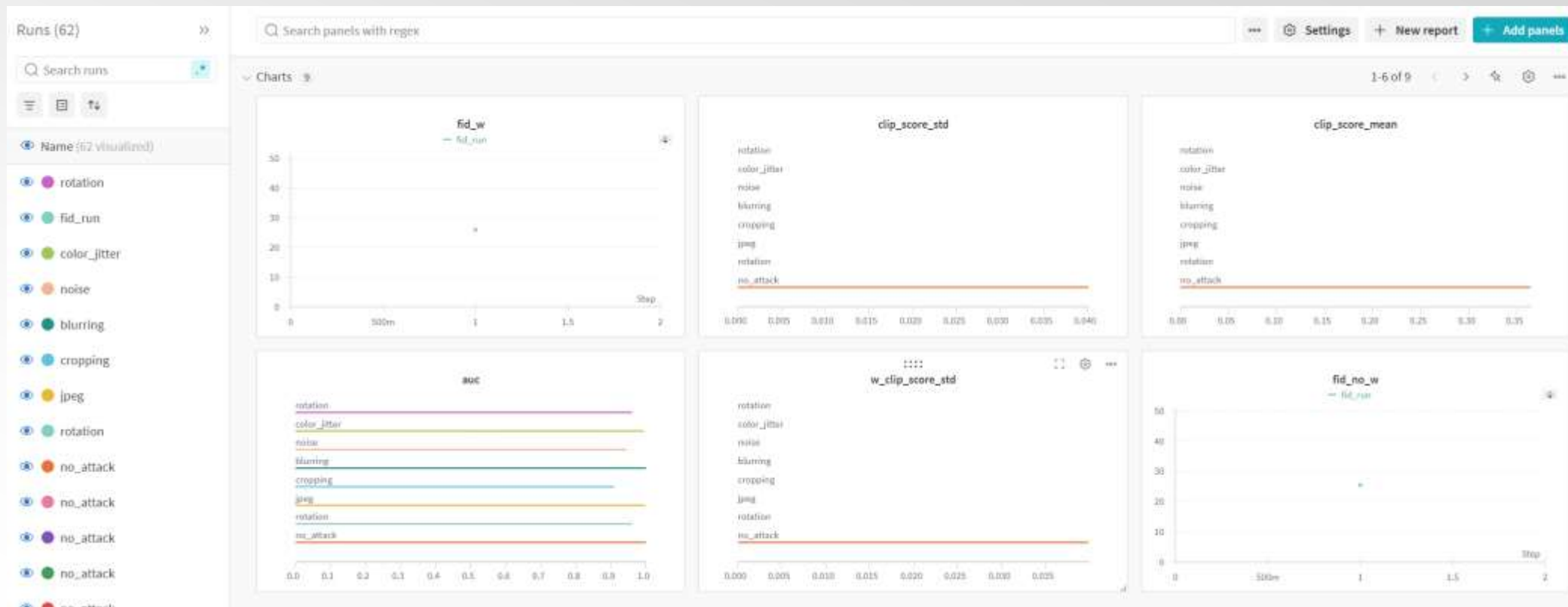
- 为了基准测试我们水印的鲁棒性，我们专注于记录其在 6 种流行的行为攻击使用的数据增强下的性能。这些包括 75° 旋转、25% JPEG 压缩、75% 随机裁剪和缩放、 8×8 滤波器大小的高斯模糊、 $\sigma = 0.1$ 的高斯噪声和亮度因子在 0 和 6 之间均匀采样的颜色抖动。此外，我们还进行了消融研究，以研究这些攻击的不同强度的影响。我们在平均情况下报告 AUC 和 TPR@1%FPR，其中我们在干净设置和所有攻击之间平均指标。在所有消融研究中，我们报告平均情况下的结果。

Table 4: AUC under each Attack for the ImageNet model, showing the effectiveness of *Tree-RingRings* over a number of augmentations. Cr. & Sc. refers to random cropping and rescaling.

Method	Clean	Rotation	JPEG	Cr. & Sc.	Blurring	Noise	Color Jitter	Avg
DwtDct	0.899	0.478	0.522	0.433	0.512	0.365	0.538	0.536
DwtDctSvd	1.000	0.669	0.568	0.614	0.947	0.656	0.535	0.713
RivaGan	1.000	0.321	0.978	0.999	0.988	0.962	0.924	0.882
<i>T-R</i> _{Zeros}	0.999	0.953	0.806	0.997	0.999	0.938	0.775	0.921
<i>T-R</i> _{Rand}	0.999	0.682	0.962	0.997	0.999	0.986	0.956	0.940
<i>T-R</i> _{Rings}	0.999	0.975	0.940	0.994	0.999	0.979	0.861	0.966



• 实验结果











三、代表性工作复现进展

工作复现1 “Tree-Ring”

- 实验效果图

runs.summary[*Table*]			
	gen_no_w	gen_w	prompt
23			a large campaign trailer parked in a parking lot.
22			A fish eye view of a bus rounding a curve on a city street.
25			A street with two busses and people walking.



三、代表性工作复现进展

工作复现2 “Ring-ID”

- Ring-ID的实验设置参照Tree-Ring,在六种攻击下做鲁棒性实验，同时增设了水印key鉴别实验

Table 1: Comparison with *Tree-Ring* in the verification task. The table shows the ROC-AUC values under various image distortions and CLIP Scores.

Methods	Ring Radius	Clean	Rotate	JPEG	C&S	Blur	Noise	Brightness	Avg	CLIP Score
<i>Tree-Ring</i>	0-10	1.000	0.935	0.999	0.961	0.999	0.944	0.983	0.975	0.364
<i>RingID</i>	0-10	1.000	1.000	1.000	0.979	0.994	0.969	0.991	0.990	0.359
	3-14	1.000	1.000	1.000	0.987	0.989	0.998	0.994	0.995	0.365

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      AUC      | Clean | Rot 75 | JPEG 25 | C&S 75 | Blur 8 | Noise 0.1 | Brightness [0, 6] | Avg |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| L1 |a-b|      | 1.0000 | 1.0000 | 1.0000 | 0.9654 | 0.9999 | 0.9966 | 0.9951 | 0.9939 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| TPR @ 1% FPR | Clean | Rot 75 | JPEG 25 | C&S 75 | Blur 8 | Noise 0.1 | Brightness [0, 6] | Avg |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| L1 |a-b|      | 1.0000 | 1.0000 | 1.0000 | 0.4700 | 0.9900 | 0.9200 | 0.9700 | 0.9071 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Average Quality Metrics
CLIP No Watermark: 0.3658
CLIP Fourier Watermark: 0.3662
```



三、代表性工作复现进展

工作复现2 “Ring-ID”

- 2048种key的鉴别

Table 2: Comparison with *Tree-Ring* in the identification task. We report the identification accuracy under various distortions for different numbers of keys.

Methods	#Keys	Clean	Rotate	JPEG	C&S	Blur	Noise	Brightness	Avg	Avg _{noC&S}
<i>Tree-Ring</i>	32	0.790	0.020	0.420	0.040	0.610	0.530	0.420	0.404	0.465
	128	0.450	0.010	0.120	0.020	0.280	0.230	0.170	0.183	0.210
	2048	0.200	0.000	0.040	0.000	0.090	0.070	0.060	0.066	0.077
<i>RingID</i>	32	1.000	1.000	1.000	0.530	0.990	1.000	0.960	0.926	0.992
	128	1.000	0.980	1.000	0.280	0.980	1.000	0.940	0.883	0.983
	2048	1.000	0.860	1.000	0.080	0.970	0.950	0.870	0.819	0.942

```
-----
Ring capacity = 2048
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|      Id Acc      | Clean | Rot 75 | JPEG 25 | C&S 75 | Blur 8 | Noise 0.1 | Brightness [0, 6] | Avg |
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| L1 |a-b|          | 1.000 | 0.890 | 0.990 | 0.080 | 0.970 | 0.980 | 0.920 | 0.833 |
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

Average Quality Metrics
CLIP No Watermark: 0.3658
CLIP Fourier Watermark: 0.3627
```



三、代表性工作复现进展

工作复现2 “Ring-ID”

- 效果图





三、代表性工作复现进展

工作复现3 “Gaussian Shading”

在主要实验中，Gaussian Shading的设置 $fc = 1$ 、 $fhw = 8$ 、 $l = 1$ ，实际容量为256位。我们选择了五种基线方法：三种由SD官方使用的，即DwtDct、DwtDctSvd和RivaGAN，一种多比特水印称为Stable Signature，以及一种无需训练的不可见水印称为Tree-Ring。鲁棒性评估为了评估鲁棒性，我们选择了图4中所示的九种具有代表性的噪声类型。我们按照图4中的噪声强度进行实验。评估指标。在检测场景中，我们计算与固定误报率（FPR）相对应的正确率（TPR）。在可追溯性场景中，我们计算比特准确率。为了衡量模型性能的偏差，我们计算了10批水印图像的FID和CLIP-Score，并对水印图像和无水印图像的平均FID和CLIP-Score进行了t检验。所有实验都使用PyTorch 1.13.0框架进行，运行在单个RTX 3090 GPU上。

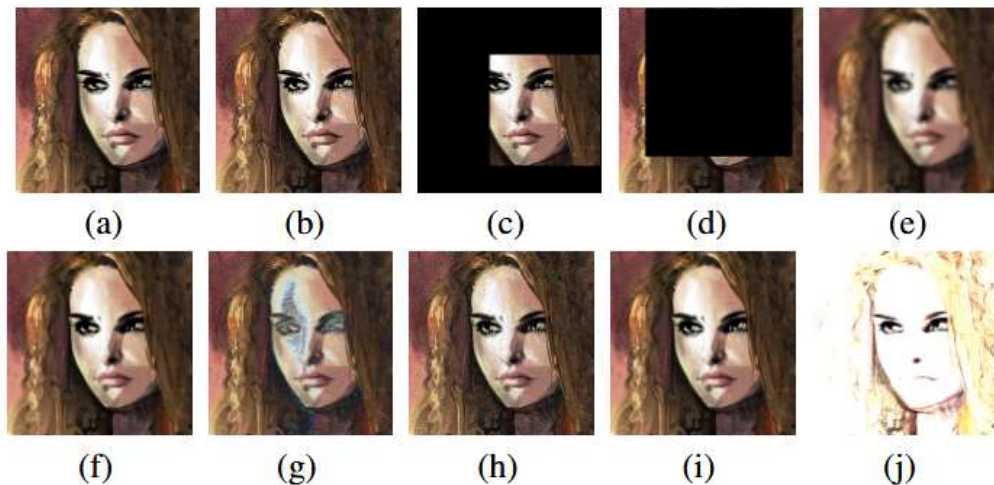


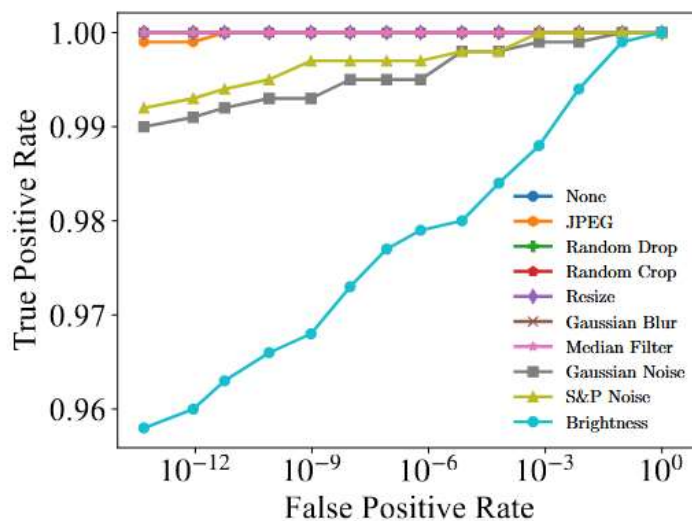
Figure 4. Watermarked image is attacked by different noise. (a) Watermarked image. (b) JPEG, $QF = 25$. (c) 60% area Random Crop (RandCr). (d) 80% area Random Drop (RandDr). (e) Gaussian Blur, $r = 4$ (GauBlur). (f) Median Filter, $k = 7$ (MedFilter). (g) Gaussian Noise, $\mu = 0$, $\sigma = 0.05$ (GauNoise). (h) Salt and Pepper Noise, $p = 0.05$ (S&PNoise). (i) 25% Resize and restore (Resize). (j) Brightness, $factor = 6$.



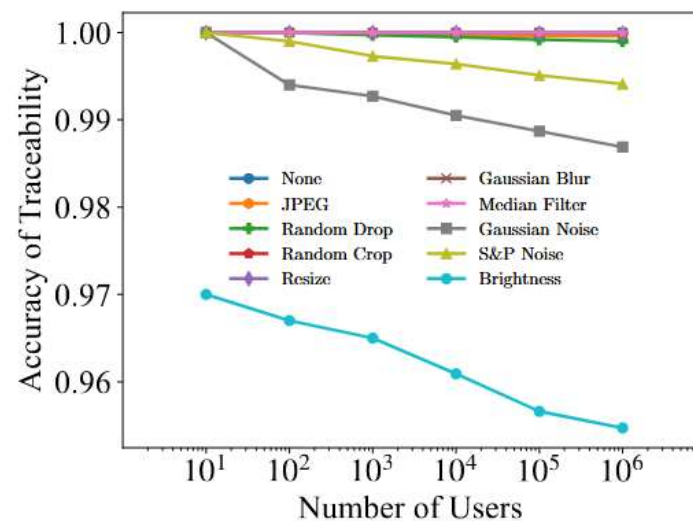
三、代表性工作复现进展

工作复现3 “Gaussian Shading”

- 检测和溯源结果



(a) Detection results.



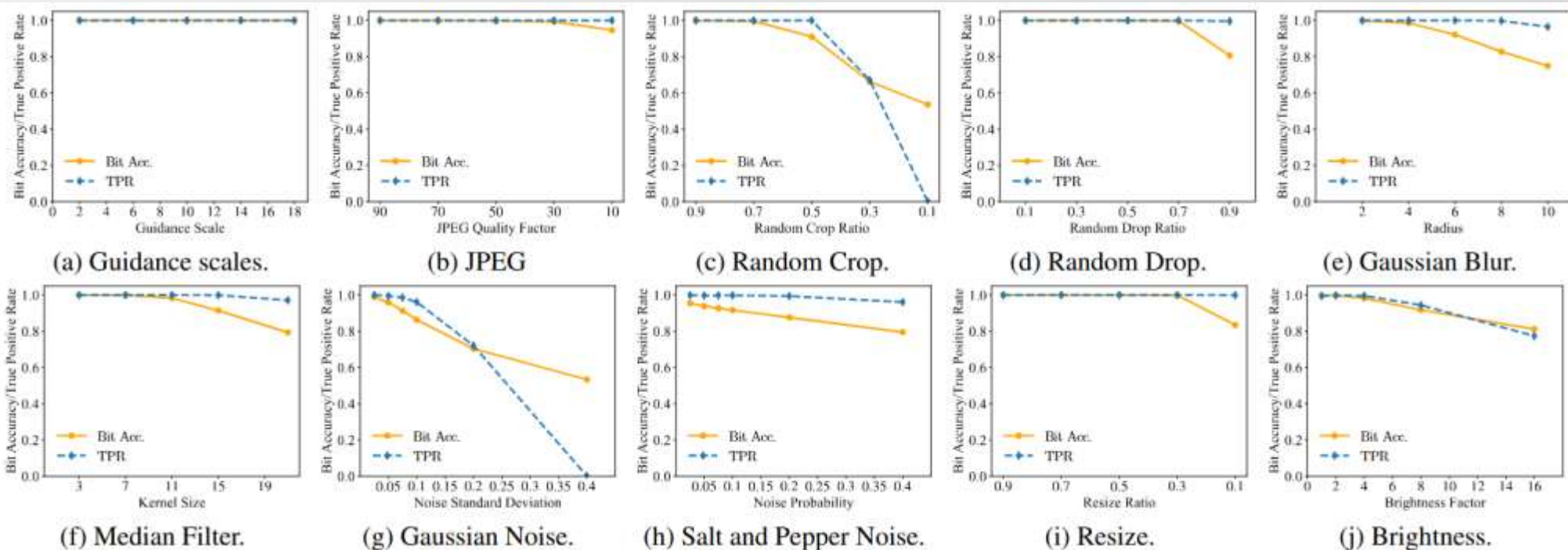
(b) Traceability results.



三、代表性工作复现进展

工作复现3 “Gaussian Shading”

• 消融实验



01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划





四、创新目标和初步思路

一些创新的想法

• 1. 多层水印嵌入

- **思路：**在不同的潜在表示层嵌入多个水印，每个水印可以有不同的强度和特征。这样可以增加水印的鲁棒性，使其在面对复杂的攻击时更难被移除。
- **实现：**可以在生成模型的不同阶段（例如，不同的迭代步骤或不同的特征层）嵌入多个水印。每个水印可以有不同的加密密钥和分布保持采样策略。

• 2. 自适应水印强度

- **思路：**根据图像内容的复杂度自适应地调整水印的强度。对于内容复杂的图像，可以嵌入更强的水印；对于内容简单的图像，可以嵌入较弱的水印。
- **实现：**可以使用图像复杂度度量（如熵）来评估图像的复杂度，然后根据复杂度调整水印的嵌入强度。例如，可以使用图像的梯度信息或纹理特征来决定水印的强度。

• 3. 动态水印更新

- **思路：**定期更新水印内容，使其更难被攻击者预测和移除。可以使用时间戳或用户特定的信息来生成动态水印。
- **实现：**在水印生成过程中引入时间戳或用户特定的动态信息，例如用户的会话ID或生成时间。这样，每次生成的水印都是唯一的，增加了水印的不可预测性。

01

研究背景与研究价值

02

国内外研究成果调研

03

代表性工作复现进展

04

创新目标和初步思路

05

下学期工作开展计划





五、下学期工作开展计划

进度安排

进 度	具 体 安 排
阶段1	<p>确定创新点（2月中旬-3月中旬）</p> <p>目标：全面了解扩散模型及生成图水印嵌入技术的研究现状、前沿动态及存在的问题，为创新点的挖掘提供依据，并提出可行创新点。</p> <p>任务：1、每天安排至少3小时查阅国内外相关学术期刊、会议论文、学位论文等文献资料，重点关注近3年的研究成果。2、按照研究主题对文献进行分类整理，建立文献数据库，记录文献的关键信息，如作者、发表年份、研究方法、主要结论、创新性、不足等。3、继续完成文献综述，总结现有研究的成果与不足，梳理研究脉络，明确研究空白点，重点关注扩散模型在生成图水印嵌入方面的应用现状和挑战。4、对提出的创新点进行初步筛选，结合研究领域的发展趋势、实际应用需求等因素，确定3-5个较为有价值的创新点作为备选。5、针对每个备选创新点，详细阐述其创新性、可行性、预期成果及可能面临的挑战，形成创新点分析报告。</p>
阶段2	<p>进行实验（3月中旬-5月底）</p> <p>目标：根据确定的创新点，制定科学、严谨、可操作的实验方案，确保实验能够有效验证创新点的可行性，规范开展实验操作。</p> <p>任务：1、依据创新点的具体内容，明确实验目的、实验对象、实验变量（自变量、因变量、控制变量）等关键要素。例如，实验目的可能是验证新水印模式的鲁棒性。2、选择合适的实验方法和技术手段，如实验设计（对照实验、随机区组实验等）、数据采集方式（生成图像的水印检测等）、数据分析方法（统计分析、机器学习算法等）。3、按照实验步骤，分阶段、分批次开展实验，严格控制实验条件，确保实验数据的准确性和可靠性。在实验过程中，实时监测实验进展，及时处理出现的问题或异常情况，并详细记录实验过程中的各种现象、数据及操作细节。</p>
阶段3	<p>总结实验成果并撰写论文（6月初-6月底）</p> <p>目标：依据实验成果，撰写高质量的科研论文，遵循学术规范，突出创新点和研究成果，经过多轮修改完善，确保论文质量达到发表标准。</p> <p>任务：1、根据学术论文的结构要求，撰写论文初稿，包括标题、摘要、关键词、引言、文献综述、实验方法、实验结果、讨论、结论与展望、参考文献等部分。在撰写过程中，注重语言的准确性、简洁性、逻辑性和学术性，突出研究的创新点和核心贡献。</p>

2025.1.16

