中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721315

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	高保真的生成式内容水印技术研究

作　　者　杨之光
学科专业　信号与信息处理
导　　师　张新鹏教授
完成日期　2025 年 5 月

姓　　名：杨之光　　学号：22721315

论文题目：高保真的生成式内容水印技术研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名:

主　　席:

委　　员:

导　　师:

答辩日期:　　年　　月　　日

姓　　名：杨之光　　学号：22721315

论文题目：高保真的生成式内容水印技术研究

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下, 独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名:

日期:　　年　　月　　日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即: 学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名:　　导师签名:

日期:　　年　　月　　日　　日期:　　年　　月　　日

上海大学工学硕士学位论文

高保真的生成式内容水印技术研究

作　　者: 杨之光

学科专业: 信号与信息处理

导　　师: 张新鹏教授

通信与信息工程学院

上海大学

2025 年 5 月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Research on High-Fidelity Generative Content Watermarking Technology

Candidate: Zhiguang Yang
Major: Signal and Information Processing
Supervisor: Prof. Xinpeng Zhang

School of Communication & Information Engineering
Shanghai University
May, 2025

摘　要

人工智能驱动的内容生成技术为用户创作数字内容提供了巨大便利，但也引发了生成内容的版权鉴定、内容溯源和伪造检测等问题。为了应对这些安全挑战，研究人员运用数字水印技术监管生成式人工智能，通过在生成内容中嵌入水印，实现人工智能生成内容的主动标识、验证与追溯。现有工作主要利用人工智能模型本身或生成内容的冗余承载水印，能够有效地实现认证，但容易引入明显的失真，损害了水印的隐蔽性和安全性，并间接降低了水印的鲁棒性。在此背景下，本文针对图像和文本两种典型生成任务，设计了相适应的高保真水印技术，具体内容如下：

1）针对图像生成任务，本文提出一种抑制高频伪影的高保真鲁棒图像水印算法。该方法基于编码器-解码器结构，在生成图像阶段之后引入水印嵌入过程，采用后处理策略将水印信息直接嵌入生成图像，无需改动原始生成模型结构，从而避免了对结构复杂且多样化的生成模型进行干预，具备一定的通用性与兼容性。针对生成图像中高频伪影易暴露水印信息的问题，本文系统分析了伪影的产生机制，并设计了相应的扰动抑制策略，在嵌入过程中引导水印信息避开图像频域中的敏感区域，从而有效降低视觉伪影的产生。实验结果表明，该方法在面对压缩、噪声等常见攻击时仍具有良好的鲁棒性，同时提升了图像的感知质量与视觉自然度。

2）与图像具有较大的冗余空间、可嵌入难以察觉的噪声不同，文本数据冗余空间有限且高度结构化，稍有改动便可能破坏语义连贯性或被用户察觉。为解决文本水印嵌入在隐蔽性与保真性上的挑战，本文提出一种基于大语言模型参数空间的水印方法。该方法在无需修改模型结构的前提下，通过在部分关键权重中注入幅度可控、稀疏分布的微小扰动，实现水印信息的有效嵌入。其中，扰动根据预设的编码规则进行注入，不仅保证了水印的可验证性，也降低了对模型性能的影响。不同于通过修改输出内容实现水印嵌入的传统策略，本文方法在模型参数空间中引入可控扰动，以实现对文本生成行为的隐性控制。实验结果表明，该方法在较小影响模型生成性能的前提下，能够保持输出文本在可读性、连贯性与语义一致性方面与原模型一致，有效实现了水印的可靠嵌入与提取。

关键词：数字水印，图像水印，文本水印，人工智能生成内容

ABSTRACT

AI-driven content generation technologies offer significant convenience for users creating digital content, but they also raise challenges in copyright attribution, content traceability, and forgery detection. To address these security concerns, researchers are using digital watermarking to regulate generative AI by embedding watermarks into generated content. This enables proactive identification, verification, and traceability of AI-generated materials. Existing methods typically embed watermarks either within the AI model itself or in redundant components of the generated content. While effective for authentication, these approaches often introduce visible distortions, compromising the watermark's stealth and security and indirectly weakening its robustness. In this context, this dissertation designs high-fidelity watermarking techniques tailored to two typical generative AI tasks: image generation and text generation. The main contributions are as follows:

1. For image generation, we propose a high-fidelity and robust watermarking algorithm that suppresses high-frequency artifacts. Based on an encoder-decoder structure, the watermark embedding process is introduced after image generation, employing a post-processing strategy that embeds watermark information directly into the generated image without altering the original generative model. This avoids interfering with complex and diverse model architectures, ensuring broad generality and compatibility. To address the problem of watermark exposure due to high-frequency artifacts, we analyze the artifact formation mechanism and design a perturbation suppression strategy. This guides the watermark to avoid sensitive areas in the image frequency domain, thereby reducing visible artifacts. Experimental results indicate that the method maintains strong robustness against common attacks such as compression and noise, while significantly improving the perceptual quality and visual naturalness of the images.

2. Unlike images, which have large redundant space and can carry imperceptible noise, text data has limited redundancy and a highly structured form, where slight changes can disrupt semantic coherence or be easily noticed. To address the challenges of invisibility and

fidelity in text watermarking, we propose a watermarking method based on the parameter space of large language models. Without modifying the model architecture, small, sparsely distributed perturbations with controllable magnitudes are injected into key weights to embed watermark information. These perturbations follow predefined encoding rules, ensuring verifiability while minimizing performance impact. Unlike traditional approaches that modify output content, this method introduces controllable perturbations in the model parameter space to exert implicit control over text generation behavior. Extensive experimental results show that the method preserves readability, coherence, and semantic consistency of generated text while enabling reliable watermark embedding and extraction with minimal degradation of model performance.

Keywords: Digital Watermarking; Image Watermarking; Text Watermarking; Artificial Intelligence Generated Content

目　录

第一章　绪论

1.1　研究背景与意义

近年来，随着深度学习（Deep Learning）[1]和生成式模型（Generative Models）的快速发展，人工智能生成内容（Artificial Intelligence Generated Content, AIGC）技术在图像与自然语言处理等领域取得了突破性进展。从早期的门控循环单元（Gated Recurrent Units, GRUs）[2]与长短期记忆网络（Long Short-Term Memory, LSTM）[3]，到 2014 年生成对抗网络（Generative Adversarial Networks, GANs）[4]在图像生成中的广泛应用，人工智能模型的内容生成能力不断增强。2017 年，Transformer[5]为现代语言模型的发展奠定了基础，OpenAI 于 2018 年推出的 GPT（Generative Pre-trained Transformer）[6]系列的大语言模型（Large Language Models, LLMs）推动了自然语言生成在质量与通用性方面的飞跃发展。近年来兴起的扩散模型（Diffusion Models），如去噪扩散概率模型（Denoising Diffusion Probabilistic Models, DDPM）[7]、Stable Diffusion[8] 和 Imagen[9]，在图像生成的清晰度、细节保真度与可控性方面进一步超越了之前的模型，成为视觉生成的主流方法。人工智能内容生成技术正从“理解数据”迈向“创造内容”，不断拓展人工智能在创作领域的边界。

随着人工智能生成技术的不断进步，AIGC 技术正在逐步应用至教育、医疗、传媒、艺术和科研等多个领域，重塑内容生产与人机交互的方式。在教育领域，AIGC 被广泛用于生成个性化学习内容与测试题，例如美国医学考试协会曾借助 GPT-2 生成医学考试素材，以提升题目多样性与考试效率[10]。在艺术创作方面，图像生成工具如 Stable Diffusion[8] 和 Imagen[9]等，在视觉设计、广告创意与游戏开发中被广泛采用，降低了创作门槛，提升了创意效率。在编程领域，GitHub Copilot 和 Cursor 等辅助开发工具也提高了程序员的工作效率。与此同时，开源社区在 AIGC 技术发展中发挥了关键作用，通过共享模型权重、算法代码与训练框架，降低了技术获取门槛。近年来，多个具有代表性的模型陆续开源，例如 Meta 发布的 LLaMA[11-13]、Stability AI 推出的 Stable Diffusion[8]，以及国内团队 DeepSeek 开源的 DeepSeek R1[14] 等，不仅加快了相关技术的传播与落地，也促进了研究社区与产业界的协同发展。

尽管 AIGC 技术在多个领域展现出广泛应用潜力，但其快速发展也引发了版权

归属、内容可信性、伦理合规与技术可控性等多重挑战[15]。首先，人工智能生成内容的著作权界定尚不明确，尤其在缺乏人类直接创作参与的背景下，传统的“独创性”标准难以适用，导致其法律地位存在争议。与此同时，AIGC 技术被滥用于制造虚假新闻、高仿真度的深度伪造图像与视频，甚至用于诈骗、操纵舆论等非法活动，对公共安全构成威胁。在科研场景中，AIGC 技术也可能被用于自动生成虚假文献、伪造实验数据等，影响学术诚信。在技术层面，生成式深度神经网络模型普遍缺乏足够的可解释性、偏见控制机制与行为约束手段，使其在医疗、司法、金融等关键领域的可信部署面临挑战。

在 AIGC 技术迅速演进的背景下，内容的真实性、可识别性与可追溯性等问题日益凸显，全球范围内已在技术路径与政策框架层面达成强化生成内容标识与溯源机制的初步共识。国际上，以美国和欧盟为代表的主要国家和地区正积极推动相关规范的建立与完善。2023 年，美国白宫提出对 AI 生成内容进行明确标注的政策倡议，强调保障信息来源的透明性；欧盟则在其《人工智能法案》中，针对高风险 AI 系统设定了严格监管要求，其中包括对生成内容进行全生命周期可追溯的制度安排。在此基础上，Adobe、Microsoft、NVIDIA 等企业联合发起“内容真实性倡议”（The Coalition for Content Provenance and Authenticity, C2PA），探索通过嵌入元数据、数字水印等方式提升 AI 内容的可验证性与可控性。与此同时，中国也在加快国内相应制度建设，2025 年出台的《人工智能生成合成内容标识办法》首次以法规形式明确要求对人工智能生成的文本、图像、音频、视频等内容进行显性或隐性标识，并鼓励采用数字水印等技术作为标识的技术实现路径。这一举措不仅填补了相关领域的监管空白，也为可信 AIGC 生态构建提供了坚实的制度基础。全球主要经济体正通过政策引导与技术协同，推动形成一个“可识别、可追溯、负责任”的人工智能生成内容治理体系。

数字水印技术作为一种可嵌入、可验证且难以被感知与篡改的信息标识方式，在提升生成内容可追溯性与可信性方面展现明显优势。一方面，在应对深度伪造（Deepfake）等伪造内容检测方面，水印可作为隐性标识嵌入图像或视频中，支持生成内容真伪的后期验证与追踪；另一方面，在虚假信息传播检测中，水印可为人工智能生成的文本、图像等添加不可见的“数字签名”，辅助平台与监管机构实现对内容进行识别与溯源。此外，在版权保护领域，水印技术不仅可用于标识创作者身份和内容归属，支持生成内容的确权管理与侵权追溯，还可用于保护模型，通过对模型实例嵌入唯一性标识，实现对模型来源进行追踪，从而有效防范模型盗用或未经授权的分

发，提升开源环境下的内容安全管控能力。

尽管水印技术在内容标识与溯源方面展现出广阔前景，其在 AIGC 场景中的实际应用仍面临诸多技术挑战。尤其在图像和文本生成中，质量与用户感知体验的要求较高，水印一旦嵌入不当，容易产生视觉伪影或语义偏差，削弱生成内容的可用性。因此，如何在不损害内容质量的前提下，实现鲁棒、隐蔽且可验证的水印嵌入，成为当前 AIGC 水印研究的核心问题。本质上，这是对高保真水印技术的迫切需求。在图像与文本这两类典型生成任务中构建高保真的水印嵌入方法，不仅是推动合规标识落地的关键路径，也具有重要的学术价值与实际应用意义。

1.2 国内外研究现状

依据嵌入路径及干预层级的差异，现有 AIGC 水印技术通常可划分为外生水印与内生水印两类[16]，如图 1.1 所示。外生水印是在生成模型输出后，通过后处理方式将水印信息嵌入图像、文本或音频等载体中。该类方法将水印嵌入过程与内容生成解耦，对模型结构无依赖，无需针对具体架构进行定制，因而具备良好的通用性和适配性。相比之下，内生水印将水印嵌入过程与模型结构紧密耦合，使生成内容在生成阶段即携带特定标识。此类方法通常依赖对模型的深入操作，如参数微调或采样策略调整等，基于模型内部机制实现嵌入，能够在一定程度上缓解后处理带来的语义干扰，提升水印的隐蔽性与无害性。当前，国内外在 AIGC 水印技术的研究中均围绕外生水印与内生水印两类路径展开，形成了各具特点的发展趋势与代表方法。

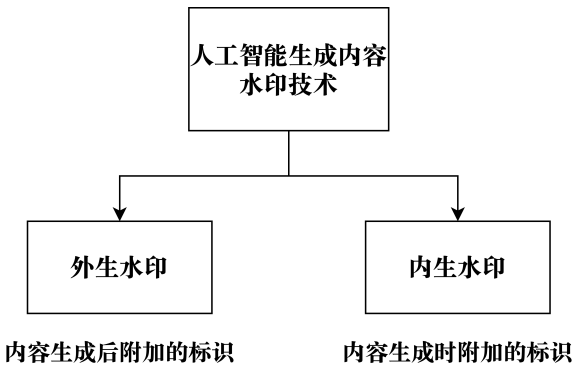

图 1.1 AIGC 水印方法分类

1.2.1 外生水印技术

外生水印技术作为数字水印研究中发展较早且相对成熟的，指在生成内容完成后，通过特定算法将水印信息嵌入图像或文本中。该技术最初源于数字媒体的信息隐藏需求，广泛应用于版权保护与内容认证等场景。早期的水印方法多基于空间域操作，如最低有效位（Least Significant Bit, LSB）替换法[17]，利用像素或字节中感知不敏感的比特位嵌入水印，虽然嵌入容量大、实现简单，但鲁棒性较差，易受压缩、滤波等基本图像处理攻击影响。为提升鲁棒性，研究逐渐转向频域方法，如离散余弦变换（Discrete Cosine Transform, DCT）、离散小波变换（Discrete Wavelet Transform, DWT）、傅里叶变换（Fourier Transform, FT）等，将水印嵌入频率系数中[18-20]，以增强水印在压缩、裁剪、缩放等操作下的稳定性，被广泛用于图像认证与归属验证任务，能够在保持图像质量的同时，实现较高的鲁棒性与隐蔽性。

频域水印因其较强的鲁棒性在传统图像水印技术中占据重要地位，而近年来基于深度学习的空间域水印方法亦展现出较大潜力。通过引入感知损失、对抗训练等机制，该类方法在不可感知性与鲁棒性之间实现了更优的权衡。Baluja 首次提出利用深度神经网络实现图像水印嵌入的思路[21]，为该方向奠定了基础；随后，Zhu 等人[22]提出 HiDDeN 框架，采用编码器-解码器结构，支持端到端的水印嵌入与提取训练，并结合感知损失与嵌入扰动模拟模块，有效提升了水印的不可见性与鲁棒性。在此基础上，Benz 等人[23]提出的 UDH 框架进一步增强了嵌入策略的稳定性与抗对抗攻击能力。此外，在特定应用场景中，如 Deepfake 检测，也发展出如 FaceSigns[24] 等半脆弱水印方法，通过构建对常规操作鲁棒、但对恶意篡改敏感的嵌入机制，有效支持对合成内容与伪造内容的识别与区分。

尽管已有方法在图像水印方向取得诸多进展，人工智能生成图像中常见的高频伪影，给水印设计带来了新的挑战。现有方法，尤其是在处理图像平滑区域或边缘细节时，易引入肉眼可见的伪影，进而损害图像的感知质量与保真性。因此，当前研究亟需探索更具细粒度控制能力、感知质量友好的水印嵌入机制，以在保障图像质量的同时提升水印的稳定性与不可见性。

在图像领域取得初步进展的同时，外生水印在文本中的应用亦受到广泛关注。现有研究主要围绕字符扰动、格式调整与语义保持改写等策略展开。常见的策略包括插入零宽字符、利用 Unicode 编码差异、控制标点符号模式或句子长度等方式进行信

息嵌入[25-26]。此类方法依赖对文本表层形式的控制，具备一定的隐蔽性，但在鲁棒性与可读性方面仍存在明显局限。为保障文本的可读性，部分研究采用同义词替换等语义保持策略进行水印嵌入[27]。近年来，随着预训练语言模型的兴起，部分研究探索通过深度学习手段进行同义词替换或语句改写，从而提升水印嵌入的上下文一致性与文本自然度[28-29]。

受限于文本数据的离散性与高度结构化特征，微小扰动往往容易破坏语义连贯性，或被用户察觉。现有文本外生水印方法普遍存在鲁棒性不足与语义一致性易受干扰的问题，易受到重写、翻译等常见自然语言处理操作的影响。尤其在保障语义一致性与文本流畅性的前提下，如何在嵌入容量与检测准确性之间实现有效平衡，仍构成该方向的重要研究难点。

1.2.2 内生水印技术

在外生水印方法不断发展的同时，另一种与生成过程深度融合的方案——内生水印，因其独特优势与研究潜力，正日益成为关注焦点。与外生水印不同，内生水印并非在生成完成后嵌入信息，而是在生成过程中将水印直接融入到待生成内容中。此类方法通常依赖于对生成模型内部结构或参数的深入操作，常通过模型微调或对生成策略施加约束实现水印嵌入。其优势在于嵌入过程与生成机制高度耦合，使水印具有较强的隐蔽性。

在图像生成模型中，内生水印方法已被广泛探索并应用。例如，在生成对抗网络中，可通过微调生成器参数或引入特定约束的方法，使其在生成图像的同时嵌入水印信息[30]。在扩散模型中，水印可直接嵌入噪声空间，使得去噪过程生成的图像自然携带可验证的水印标识[31]。近年来，基于潜在扩散模型（Latent Diffusion Models, LDMs）的生成方法受到广泛关注，LDMs 在预训练自编码器的潜在空间中进行扩散与去噪，生成高质量图像[8]。研究表明，在潜在空间中嵌入水印，可在较小影响图像质量的前提下，实现有效的版权保护与来源追踪。例如，Zhang 等人提出在扩散模型的潜在空间中嵌入不可见水印，以提升对抗攻击下的稳健性[32]。

与图像生成领域类似，内生水印在文本生成模型中的研究也逐渐展开，尤其在大语言模型上引发广泛关注。其中，直接微调模型参数是一种较为直观的实现方式，即在微调过程中引入水印相关的生成偏好，使模型在生成文本时自然携带可识别水印[33-34]。然而，该方法计算开销较高，实际部署受到一定限制。为降低水印嵌入的计

算成本，近年来也有研究提出更加轻量化的策略。例如，在每一步生成前预设一个由若干“绿色词汇”组成的候选集合，并在采样过程中引导模型优先选择这些词汇。通过对最终文本中绿色词汇的出现频率进行显著性统计检测，可有效识别生成来源[35]。该方法无需修改模型结构或参数，仅对采样过程进行控制，具备良好的可扩展性，并在文本质量与可检测性方面表现出一定潜力。

内生水印的核心优势在于其嵌入过程与内容生成过程高度耦合，能够在较小影响生成质量的前提下，实现较强的隐蔽性与鲁棒性。由于水印作为生成过程的一部分，模型在生成高质量内容的同时，以难以察觉的方式嵌入水印。然而，内生水印也面临多方面挑战。首先，水印策略需尽量避免影响模型生成的流畅性、连贯性与准确性，模型结构或参数的修改可能引发性能退化，因此需在嵌入效果与生成质量之间取得平衡。其次，大多数方法高度依赖特定模型架构，缺乏跨模型的通用性，限制了其推广能力。此外，部署成本也是一大障碍，许多方法需对预训练模型的微调，而这一过程代价过于高昂。如何在保持嵌入效果的同时降低训练复杂度，是当前研究的重要方向。最后，水印的鲁棒性仍面临挑战，攻击者可能通过模型蒸馏、剪枝等其他手段破坏水印。

1.2.3 面临的主要挑战

尽管外生水印与内生水印在技术路径上各具特点，它们在 AIGC 的标识与溯源任务中仍面临一系列共性挑战，尤其在保障高保真度这一核心需求下，问题尤为突出。当前研究的核心在于如何在不明显影响生成内容质量的前提下，实现水印的有效嵌入与稳健提取。以下几个方面构成了水印技术发展的关键难题：

- 保真度：水印嵌入需在尽可能保持生成内容自然性与完整性的前提下进行。在图像生成任务中，这意味着嵌入后的图像应避免出现伪影或失真，保持与原始图像的视觉一致性；在文本生成中，则需确保语义通顺、逻辑连贯，不影响可读性。高保真度不仅关系到用户体验，也直接影响水印方案的应用可行性。因此，如何在增强隐蔽性的同时最小化对生成质量的干扰，是水印设计中需重点权衡的问题。
- 鲁棒性：鲁棒性指水印在内容经历各种攻击后仍可被正确识别的能力。图像可能面临压缩、裁剪、旋转、加噪等变换；文本则可能经历翻译、重写或同义词替换等自然语言处理操作。若水印在这些扰动下易被破坏，其溯源与防伪能力

将受到削弱。因此，提升水印在多种干扰下的鲁棒性，已成为当前研究的关键方向之一。

- 容量与准确率：嵌入容量决定了水印可携带的信息量，而提取准确率反映了提取结果的精度与可靠性。在容量、不可感知性与鲁棒性之间存在固有的权衡关系：嵌入容量过高可能导致显著的质量退化或易被检测，而容量过低则难以满足实际应用中的信息嵌入需求。
- 效率与通用性：在实际部署场景中，水印嵌入与检测的效率决定其可扩展性，在大规模的生成任务中尤为关键。外生水印因脱离生成模型、嵌入灵活，便于快速部署，但在系统集成与运行效率上存在一定开销；内生水印嵌入过程效率较高，却往往依赖对特定模型的训练或推理优化。因此，提升水印算法在不同生成模型与任务之间的适配能力，也是推动其实际应用的核心方向之一。

高保真作为 AIGC 水印技术的核心诉求，对图像外生水印和文本内生水印都提出了严苛要求。尽管两类方法各有优势，外生水印在灵活性和部署便捷性方面更为突出，内生水印则在鲁棒性与安全性方面表现更强，但如何在保持生成质量的同时，实现隐蔽、稳健、可扩展的水印嵌入，仍是当前技术发展的关键瓶颈。未来研究亟需围绕高保真水印的系统设计展开深入探索。

1.3 本文的主要研究内容

本文围绕人工智能生成内容中的水印设计问题，面向图像与文本两类核心生成任务，开展了两项具有代表性的水印算法研究。鉴于图像与文本在数据表示形式、语义结构和对扰动容忍度方面存在差异，本文统一以“高保真度”为设计目标，分别采用外生水印与内生水印的技术路线，构建了两种具有适应性的嵌入方案，具体研究内容如下：

针对当前生成图像中普遍存在的高频伪影及其对水印检测造成干扰的问题，本文提出一种基于编码器-解码器结构的高保真鲁棒图像水印算法。该方法采用后处理策略，将水印嵌入过程与图像生成阶段解耦，无需修改原始生成模型结构，具备良好的通用性与部署灵活性。本文系统分析了高频伪影的成因及其在频域中的分布特征，设计了一种高频伪影抑制机制，引导水印避开高频敏感区域，从而降低伪影暴露风险并减轻对图像结构的干扰。实验结果显示，该方法在保证图像感知质量与视

觉自然度的同时，能有效抵御压缩、加噪等常见攻击，表现出良好的鲁棒性与隐蔽性，适用于生成图像的标识与溯源等实际应用场景。

考虑到文本生成对微小改动异常敏感，传统基于输出干预的水印方法往往难以兼顾隐蔽性与文本质量。为此，本文提出一种在模型参数空间嵌入水印的策略。该方法无需修改模型结构，通过在部分关键权重中注入稀疏且幅度可控的扰动，实现对生成行为的隐性引导。扰动按照设定编码规则生成，可结合密钥控制位置，提升水印的可验证性与安全性。不同于修改输出内容的方式，该方法直接作用于生成机制本身，避免对语义和语用造成干扰。实验表明，该方法对生成性能影响较少，嵌入水印后的文本在语义一致性、可读性和连贯性上与原模型保持一致，同时具备良好的识别准确性与应用稳定性。

1.4 本文的结构安排

本文共分为五章，各章内容安排如下：

第一章介绍了 AIGC 技术的发展背景，分析了其在版权标识、内容溯源等方面所带来的新挑战，指出在日益复杂的生成模型环境中，传统溯源手段难以满足高质量识别与防篡改的需求。在此基础上，进一步明确了数字水印作为实现人工智能生成内容可信性管理的重要技术路径的研究价值，并从高保真角度切入，界定了本文的研究目标、核心问题及主要工作。最后，对全文的结构安排进行了概述。

第二章回顾了与本研究相关的技术基础。首先介绍了 AIGC 技术的典型生成机制及其主要模型架构；随后梳理了数字水印技术的发展历程与分类体系。本章特别分析了现有水印方法在隐蔽性、鲁棒性、容量与效率之间的权衡困境，指出当前技术在保持生成内容质量的同时实现可靠标识仍面临诸多挑战，为后续章节的研究方法奠定理论基础。

第三章提出了一种面向图像生成任务的高保真水印方法。该方法基于深度学习编码器-解码器架构，在生成图像之后通过后处理方式嵌入水印信息，避免直接干预生成模型结构，从而提升算法的通用性与部署灵活性。针对图像生成中常见的高频伪影问题，本文系统分析其成因并设计对应的抑制机制，引导水印避开敏感区域，降低视觉伪影。实验结果表明，该方法在 JPEG 压缩、加噪等常见攻击下保持良好的水印提取性能，同时提升了图像的感知质量。

第四章设计了一种面向文本生成的内生水印算法，通过调制大语言模型的参数空间实现水印嵌入。该方法无需修改模型结构或推理路径，仅在部分关键权重中注入稀疏、幅度可控的扰动，在较小影响生成质量的前提下，将水印信息隐式植入模型的生成过程之中。该方法从生成机制本身进行调控，具备一定的隐蔽性与稳定性。实验结果显示，该算法在保持语义一致性、可读性和连贯性的同时，实现了水印信息的有效嵌入与准确提取，具备较强的通用性与实用性。

第五章对全文研究工作进行了总结与展望。本文围绕图像与文本生成任务，设计并验证了两类兼具高保真性与鲁棒性的数字水印方法，系统探讨了外生与内生技术路径在不同生成任务下的适应性与应用优势。实验结果表明，所提出方法在保障生成内容质量的同时，能够实现可靠、隐蔽且高效的水印嵌入与检测。最后，针对当前存在的技术瓶颈，结合未来发展趋势，本文指出了在高鲁棒性算法设计、水印安全机制优化以及低干预提取策略等方面的后续研究方向。

1.5 本章小结

本章简要介绍了本研究的背景、目标、研究内容及论文结构安排。针对 AIGC 在版权标识与溯源中的技术需求，本文围绕图像生成与文本生成两个典型任务，设计了具有代表性的水印算法，探索了外生与内生两种实现路径。两类方法各具特点，面向不同生成场景，均以在保证内容质量的前提下提升水印的可验证性与鲁棒性为目标。后续章节将分别介绍所提方法的设计原理、关键机制与实验结果，并进一步分析当前技术面临的挑战与潜在改进方向。

第二章　人工智能生成内容与水印嵌入的相关技术基础

本章将系统介绍本文所依托的技术基础，包括主流生成模型的架构特点及发展脉络、数字水印的基本原理与分类，以及图像与文本水印中关键技术路径与典型挑战。通过对这些关键概念的梳理，为后续章节中具体水印算法的设计与实验提供理论支撑与技术背景。

2.1　生成式人工智能模型概述

生成式人工智能是一类以内容生成为核心目标的技术范式，其关键在于构建能够学习数据潜在结构与分布的模型，并据此生成具有新颖性与创造性，同时在统计特性上接近真实样本的内容。与主要完成分类、回归等任务的判别式模型不同，生成式模型强调对训练数据的分布建模，常通过采样或重构等方式生成新的数据样本[36]。为实现高质量生成，这些模型不仅需保证生成内容在语义上的逻辑一致性与上下文连贯性，还需在感知层面具备足够的逼真度，从而使生成结果在视觉或语言上与真实数据难以区分[1]。

近年来，生成式人工智能的发展得益于多个关键要素的协同推动：一方面，随着深度神经网络架构的不断演化，众多具有强表达能力的生成模型相继涌现，如生成对抗网络（Generative Adversarial Networks, GANs）[4]、变分自编码器（Variational Autoencoder, VAE）[37]、流模型（Flow）[38]、扩散模型（Diffusion Model）[7]和 Transformer[5]等，使模型在建模高维复杂数据分布方面的能力进一步增强；另一方面，开源社区和产业界推动了大规模、多模态、高质量数据集的建立，丰富了模型的训练语料基础[39]；同时，图形处理单元（Graphics Processing Unit, GPU）与分布式计算资源的发展，也为这些参数庞大、训练复杂的生成模型提供了必要的计算保障。

生成式模型在图像与文本两类模态中分别展现出不同的技术路径和代表性架构。为了系统梳理生成式模型在图像与文本两类模态中的技术演进与原理基础，接下来的内容将分别综述当前主流的图像生成模型与文本生成模型，为后续水印嵌入方法的设计提供理论支撑。

2.1.1 图像生成模型

图像生成模型旨在学习高维图像数据的潜在概率分布 $p(x)$，并基于该分布生成与真实图像在语义与视觉特性上高度一致的新样本。当前主流方法多基于隐变量建模范式，即引入隐变量 $z \sim p(z)$，通过生成器 $G(z;\theta)$ 实现从隐空间到图像空间的映射，从而得到条件分布 $p(x|z)$，实现生成图像的采样与重构。图像生成模型主要包括变分自编码器、生成对抗网络与扩散模型三大类，它们分别代表了生成式建模中不同的理论基础与技术路径。

VAE 是一种有向模型，它通过学习的近似推断来工作[37]，其目标在于最大化边际对数似然 $\log p(x)$。由于该项难以直接计算，VAE 引入近似后验分布 $q_\phi(z|x)$，并通过最大化证据下界（Evidence Lower Bound, ELBO）来进行变分推断：

$$\log p(x) \geq \mathbb{E}q_\phi(z|x)[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)\|p(z)) \tag{2.1}$$

其中，ϕ, θ 分别表示编码器与解码器的参数，$p(z)$ 为先验隐变量分布，通常设为标准正态分布，KL 散度项用于正则化隐变量空间，使学习到的后验分布不偏离先验。尽管 VAE 在训练稳定性和隐空间可解释性方面表现良好，但其生成图像往往存在模糊、细节缺失等问题，难以满足高质量图像合成的实际需求。

为弥补上述不足，Goodfellow 等人提出了生成对抗网络[4]。其核心思想是构建一个博弈框架，由生成器 $G(z)$ 与判别器 $D(x)$ 组成，优化目标为：

$$\min_G \max_D \ \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))] \tag{2.2}$$

生成器通过与判别器的对抗训练，不断提高生成图像的真实度。GAN 在图像细节还原方面表现突出，但训练过程易出现模式崩溃与梯度不稳定等问题。为缓解上述缺陷，后续研究提出了多种改进方法，并推动了生成对抗网络的快速发展。例如 WGAN[40] 采用 Wasserstein 距离缓解训练不稳定的问题，BigGAN[41]利用大规模参数与类别条件生成提升多样性，StyleGAN 系列则基于风格向量控制机制，实现了图像属性的可解释编辑[42-44]，增强了图像生成的可控性与细节表现力。

近年来，扩散模型在图像生成任务中表现出较强的性能。其核心思想是将一张真实图像逐步添加高斯噪声，形成一个前向的马尔可夫过程，并在训练阶段利用模型学习其反向过程，从噪声中逐步还原出原始的图像。典型代表是去噪扩散概率模型（Denoising Diffusion Probabilistic Models, DDPM）[7]。

在 DDPM 中，正向扩散过程被建模为如下形式的条件高斯分布：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2.3}$$

其中，x_t 表示在第 t 个时间步加入噪声后的图像、$\beta_t \in (0,1)$ 是在时间步 t 的噪声控制参数，控制每一步加入多少噪声；$\mathcal{N}(\mu, \Sigma)$ 表示均值为 μ、协方差为 Σ 的多维高斯分布、$\mathbf{I}$ 是单位矩阵，表示各维之间独立同分布的噪声。正向过程中，随着 t 的增加，原始图像逐步被高斯噪声所覆盖，最终趋近于各向同性高斯噪声。

在训练过程中，模型的目标是学习逆过程，即从噪声图像 x_t 恢复出原始图像 x_0。一种简化的做法是令神经网络 $\epsilon_\theta(x_t, t)$ 去预测在步骤 t 上添加的噪声 ϵ，从而推出原始的图像。对应的训练损失函数为：

$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{2.4}$$

其中，ϵ 服从标准高斯分布、$\epsilon_\theta(x_t, t)$ 是神经网络在输入为 x_t 和时间步 t 时输出的噪声估计、$\mathbb{E}_{x_0,t,\epsilon}$ 表示对原始图像、时间步和噪声的联合期望、$\|\cdot\|^2$ 是 L_2 范数的平方，用于评估预测误差。该损失函数旨在引导模型准确预测每一步中添加的噪声，从而实现逐步去噪，最终生成清晰的原始图像。

尽管 DDPM 在生成质量与训练稳定性方面相较 GAN 表现出明显优势，但其采样过程存在效率瓶颈。其主要原因在于，DDPM 的生成过程需依赖从高斯噪声开始、逐步执行数百至上千步的反向去噪迭代，以逐步还原出目标数据分布。每一步都需进行一次前向传播，导致总体生成时间较长，难以满足高效生成的实际应用需求。为此，后续研究提出了多种改进方案，以减少采样步骤、提升推理效率。

DDIM（Denoising Diffusion Implicit Models）作为一种重要改进方法被提出[45]，该方法将随机去噪过程转化为确定性过程，能够在不牺牲生成质量的前提下大幅减少采样步数，从而提升推理效率，使得扩散模型更适用于实际部署。得益于高效的推理效率，DDIM 被广泛应用于多种图像生成任务。此外，为解决扩散模型高计算复杂度的问题，潜在扩散模型（Latent Diffusion Models, LDMs）提出将扩散过程从像素空间迁移至编码器编码的潜在空间中，明显降低计算资源需求，同时保持图像质量[8]。以 Stable Diffusion 为代表的模型被广泛应用于图像生成等任务，在此基础上进一步发展出如 ControlNet 等分支模型[46]，提升了图像生成过程的可控性与灵活性。

2.1.2 文本生成模型

文本生成模型旨在建模自然语言中的语义结构与上下文依赖关系，以生成连贯、流畅且语义合理的文本序列。该任务通常被形式化为一个条件语言建模问题，其目标是在给定上下文的条件下，最大化目标词序列的概率分布，即：

$$P(x) = \prod_{t=1}^{T} P(x_t | x_{<t}; \theta) \tag{2.5}$$

其中 $x = (x_1, x_2, \ldots, x_T)$ 表示长度为 T 的词序列，θ 为模型参数。该目标的本质是学习拟合文本序列中的上下文依赖性和语义连贯性，确保生成文本不仅具备语法正确性，也具备内容一致性和逻辑合理性。

早期的文本生成模型主要依赖循环神经网络（Recurrent Neural Network, RNN）及其改进形式，如长短期记忆网络等。这类模型可以捕捉序列中的时间依赖性，但其在建模长距离依赖、训练并行化和梯度稳定性等方面存在明显的局限，难以支撑复杂的语言生成任务。

为突破传统序列建模在并行效率和长距离依赖建模方面的瓶颈，Vaswani 等人[5]于 2017 年提出了 Transformer 架构，彻底改变了自然语言处理的主流建模范式。Transformer 摒弃了传统的递归结构，转而采用多头自注意力机制（Multi-Head Self-Attention），能够直接建模输入序列中任意位置之间的依赖关系。此外，为缓解其不具备内在序列中感知相对位置的能力的问题，Transformer 引入位置编码，以显式编码序列的位置信息，从而弥补其对顺序结构建模能力的不足。该架构在增强模型并行计算能力与长程依赖建模能力的同时，为大语言模型的后续发展提供了重要支撑。

基于 Transformer 架构，近年来涌现出一系列预训练语言模型，推动了语言生成能力的提升。典型模型包括 BERT（Bidirectional Encoder Representations from Transformers），采用双向编码器结构，专注于语言理解任务[47]；GPT（Generative Pre-trained Transformer）系列基于单向解码器结构，采用自回归建模目标，是当前主流的文本生成模型[6]；T5 采用编码器-解码器等结构，将所有任务统一建模为文本到文本转换，提升了模型的适配性与泛化能力[48]。此后，LLaMA、GLM、DeepSeek、Gemma 等模型不断扩展参数规模、训练语料与结构设计，进一步增强了多语言、多任务场景下的生成能力[14,49-55]。

同时，为提升生成文本的可控性与表达质量，研究者还提出了一系列采样与控

制策略，包括 Top-k、Top-p、温度调节、长度惩罚等方法，使得生成内容在多样性与流畅性之间取得更优的平衡[56]。结合提示学习（Prompt Learning）和上下文注入机制等，文本生成模型已在问答系统、对话交互、智能写作、代码生成等场景中得到广泛应用。如图 2.1 所示，Chatbot Arena 基准测试的结果凸显了大语言模型对话能力的快速提升。该基准通过模拟真实世界中的交互场景评估模型的语言生成质量，结果表明模型在连贯性、上下文理解和语境响应方面均有显著的改进[57]。

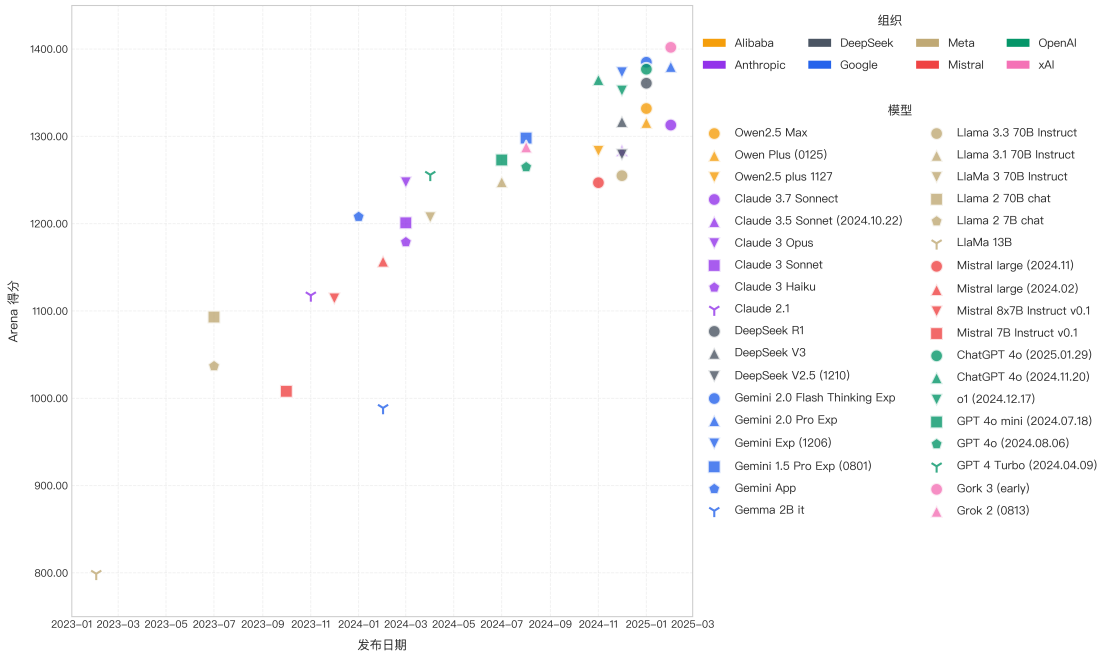

图 2.1 大语言模型在 Arena 上的排行榜得分

2.1.3 人工智能生成内容的安全风险与可信标识需求

生成式人工智能在图像、文本、音频、视频等多领域取得较大进展，提升了内容生产的效率与多样性。然而，其广泛应用也引发内容真实性、版权归属、恶意伪造等安全与伦理问题，威胁数字内容生态、社会信任与信息安全。

首先，人工智能生成内容具有高度的可伪造性与难辨识性。图像与视频生成模型能够合成逼真的伪造素材，常规检测手段难以识别，严重威胁新闻传播、法律证据、公共舆论等高可信度应用场景[58]。与此同时，大语言模型生成的文本通常语义连贯、逻辑严谨，使虚假新闻、伪造评论等问题更具隐蔽性和迷惑性[59]。随着技术门槛的持续降低，恶意内容的生成与传播变得更加高效和广泛[58]。其次，AIGC 在版权与隐私方面也面临挑战。训练数据来源与授权不透明，引发生成内容的归属权

与侵权争议[58]。同时，大语言模型也可能泄露个人身份信息等敏感数据[58-59]，加剧隐私泄露的风险。

当前，各国正探索监管路径，虽已有初步框架，如欧盟的人工智能法案、中国的标识办法等，但仍处于不断完善之中[60]。AIGC 的安全性与可控性，已成为其可持续发展所面临的关键约束。这一问题不仅具有技术复杂性，更对现有社会治理体系及国际合作机制提出了深层次挑战。

2.2 数字水印技术基础

数字水印作为保障数字内容安全与可追溯性的重要技术，致力于在内容中嵌入难以察觉且可验证的标识信息，广泛应用于版权保护、伪造检测与溯源等领域。随着生成式人工智能的快速发展，水印技术在标识生成内容来源、提升内容可信度方面展现出新的应用价值。然而，AIGC 水印技术对质量、自然性具有更高要求，水印不仅需要具备良好的鲁棒性，还需确保生成内容的高保真度，避免影响用户体验。本节将介绍数字水印技术的基本原理与评估指标，并重点阐述外生水印与内生水印两类主流技术路线，为后续研究提供理论基础。

2.2.1 数字水印技术

水印技术最早出现在 13 世纪的意大利，并随着造纸工艺的发展逐渐传播至欧洲各国[61]。早期的水印技术主要用于识别造纸商或纸模，具有一定的商标功能。到了 18 世纪，水印的用途变得更加广泛，不仅被用作纸张生产商的标识，还成为了防伪技术的重要组成部分，尤其是在货币和官方文件的印制中发挥了关键作用。随着印刷技术和造纸工艺的进步，现代水印技术不断演进，从传统的手工压制水印发展到数字水印技术。

数字水印技术的广泛关注，很可能源于人们对内容版权保护日益增长的重视。1993 年 11 月，Marc Andreessen 推出了 Mosaic 网页浏览器，使互联网变得更加用户友好[62]。随着这一变革，人们迅速意识到，用户不仅希望浏览网页，还渴望下载图片、音乐和视频。互联网作为数字媒体的分发渠道，凭借低成本和近乎即时的传输优势，迅速成为内容传播的理想平台。然而，这一便利性也让内容所有者意识到，互联网带来了严重的盗版风险。传统的加密技术虽能在传输过程中保护内容安全，却无法阻止合法用户在解密后对内容的非法传播。因此，数字水印技术应运而生，作为

加密技术的有力补充。它能够在数字内容中嵌入持久性的信息，即使经过解密、压缩或格式转换，依然能够保留，从而在版权保护、追踪溯源等领域发挥关键作用。

通用的水印系统通常由水印编码器和水印提取器两部分组成。如图 2.2所示，水印编码器的主要功能是将水印信息嵌入到数字内容载体中，它通常以原始载体和待嵌入的水印作为输入，生成带有水印的输出内容。在内容传输或存储过程中，水印可能会遭受各种有意无意的变换或攻击。水印提取器负责对接收内容进行分析，从中提取出嵌入的水印信息，实现验证或溯源。

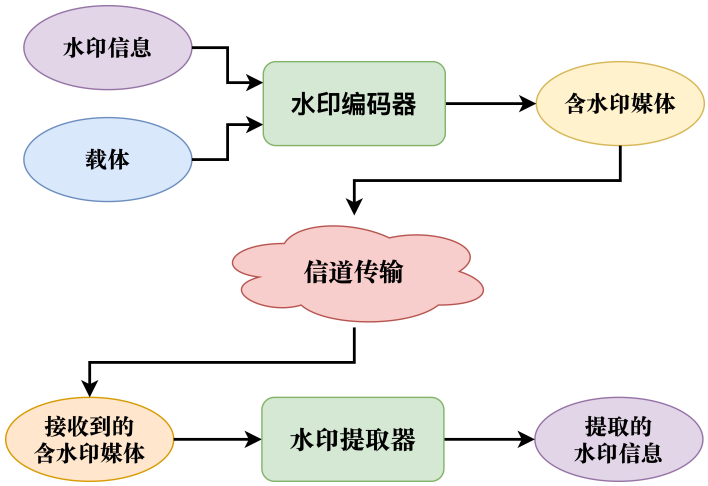

图 2.2　通用的水印系统框架

2.2.2　数字水印的评价指标

数字水印技术在 AIGC 场景中的设计与评估需兼顾生成内容质量、嵌入水印后的稳定性以及检测效果，通常从保真度、鲁棒性和准确性三个维度进行衡量。本节围绕这三项指标，详细阐述其评估方法及数学定义。

保真度衡量水印嵌入对生成内容质量的影响，是确保水印隐蔽性与生成内容自然性的核心指标。针对不同模态，保真度的评价指标有所不同。在图像领域，常用的指标包括峰值信噪比（Peak Signal-to-Noise Ratio, PSNR）和结构相似性指数（Structural Similarity Index, SSIM）。PSNR 主要用于衡量水印嵌入前后图像的变化程度，单位为分贝（dB），其定义如下：

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{\mathrm{MAX}^2}{\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - I_{\mathrm{w}}(i,j))^2} \right) \tag{2.6}$$

其中，MAX 为图像像素的最大值（通常为 255），$I(i,j)$ 和 $I_{\mathrm{w}}(i,j)$ 分别表示原始载

体图像与嵌入水印后的水印图像在位置 (i,j) 处的像素值，m 和 n 为图像的行列数。该指标用于衡量水印嵌入对图像造成的失真程度，值越大表示图像质量越接近原始图像，水印嵌入对视觉效果的影响越小。

SSIM 则从亮度、对比度和结构三个层面综合评估两张图像的相似性，定义为：

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.7}$$

其中，μ_x、μ_y 分别为图像 x、y 的均值，σ_x^2、σ_y^2 为方差，σ_{xy} 为协方差，C_1、C_2 为避免分母为零的常数。该指标用于衡量两幅图像在亮度、对比度与结构层面的相似性，其值越接近 1，表示水印图像与原始载体图像结构一致性越高，视觉质量越优。

在文本领域，保真度通常用困惑度（Perplexity, PPL）来衡量生成文本的自然性，定义为：

$$\text{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P(x_t|x_{<t})\right) \tag{2.8}$$

其中，T 为文本长度，x_t 为第 t 个生成的词，$P(x_t|x_{<t})$ 表示在上下文 $x_{<t}$ 条件下预测词 x_t 的概率，PPL 越低表示文本越自然、流畅。

准确率用于衡量提取出的水印信息与原始嵌入水印之间的一致程度，定义为：

$$\text{Accuracy} = \frac{N_{\text{correct bits}}}{N_{\text{total bits}}} \tag{2.9}$$

其中，$N_{\text{correct bits}}$ 表示成功提取并与原水印匹配的比特数，$N_{\text{total bits}}$ 为水印嵌入的总比特数。该指标用于衡量水印提取的准确性与完整性，准确率越高，说明水印在嵌入与提取过程中信息保留效果越好，系统具备更高的识别可靠性。

鲁棒性是指水印在经历各种失真或攻击操作后仍能被准确提取的能力，是衡量水印系统在实际应用中稳定性与可靠性的重要指标。在图像场景中，常见的攻击包括裁剪、压缩、旋转、缩放、加噪声与滤波等操作，鲁棒图像水印需在此类变换下保持嵌入水印信号的可识别性。在文本任务中，水印则需应对如同义改写、拼写修正、机器翻译、续写等语义层级的扰动，防止信息在语言变换中被抹除或削弱。鲁棒性通常通过在不同攻击条件下对水印提取结果进行评估，采用准确率、误检率或重构质量等指标加以量化，从而验证水印方案在复杂环境中的实用性。

2.2.3 外生水印与内生水印

针对 AIGC 水印嵌入的不同实现路径与干预层级，现有研究通常将水印技术划分为两大类：外生水印与内生水印。二者在嵌入时机、技术实现和应用特性上存在明显差异。

外生水印是指在生成模型输出内容之后，通过后处理方式在图像、文本、音频等生成数据中嵌入水印信息的方法。此类方法将水印嵌入过程与内容生成过程解耦，因而具备对模型结构的高度无关性，无需针对具体生成模型单独适配。外生水印通常通过在像素空间或频率空间对生成内容进行细粒度修改，如在图像中调整低感知敏感区域的像素值，或在文本中进行轻微的格式变化或同义词替换，以实现隐蔽信息的嵌入。外生方法在工程实现上具有较高的灵活性和部署便利性，适合批量化处理和现有系统的快速集成。然而，由于水印是生成后附加的，往往面临保真度与鲁棒性之间的权衡问题，且在遭遇诸如裁剪、压缩、重写、翻译等内容修改操作时，水印容易受到削弱甚至破坏。

内生水印将水印嵌入过程与内容生成过程紧密耦合，使模型在生成内容的过程中，天然携带可追溯的水印标识。内生水印通常通过调整模型参数、引入采样控制策略或在特征空间中施加隐式约束等方式来实现。与外生水印相比，内生水印方法在隐蔽性和鲁棒性上具有天然优势，因为水印嵌入与生成过程紧密耦合，即使内容在后续经历一定程度的编辑或攻击，水印信息也能部分保留。例如，在图像生成领域，可以通过微调扩散模型的潜在空间特性，使得输出图像隐含特定模式；在大语言模型生成文本时，可以通过控制采样策略，令生成文本在词频分布上携带隐式特征。尽管内生水印具备更高的安全性和稳定性，但其设计与实现往往依赖于对生成模型的深入干预，如参数修改或微调，因而在实际应用中成本较高，且可能对原任务产生影响。

总体而言，外生水印与内生水印在应用特性上各具侧重：前者强调通用性与部署灵活性，适用于对现有生成系统的非侵入式标记；后者则在隐蔽性与安全性方面具有潜在优势。在当前对生成内容高保真度要求不断提升的背景下，如何在两种嵌入路径之间权衡选择，设计兼具适应性与最小干扰的水印策略，正逐渐成为 AIGC 水印研究的重要议题之一。

2.3 图像水印的相关技术与基础

随着生成式模型在图像领域的广泛应用，人工智能生成图像在内容创作、媒体传播和数字版权等场景中发挥重要作用。但这类图像因高度仿真，易被滥用，带来伪造、侵权和不可控传播等风险。图像水印因此成为提升内容可追溯性与真实性的关键手段。尤其在应对多样攻击和保持图像质量的双重挑战下，实现鲁棒、隐蔽、可验证的水印嵌入，成为当前研究的核心问题。本节将围绕图像水印的核心实现机制与面临挑战展开讨论。首先回顾传统图像水印方法的发展脉络，为后续方法提供技术基础；随后系统梳理近年来深度学习驱动的图像水印技术，重点分析外生与内生策略的特点与适用场景；进一步，结合常见攻击手段与高频伪影的生成机制，探讨其对水印鲁棒性与高保真嵌入的影响，为后文工作的提出奠定理论基础。

2.3.1 传统的图像水印实现方法

图像数字水印技术最初主要聚焦于如何在不影响图像视觉质量的前提下，实现信息的隐蔽嵌入与稳定提取。这一阶段的研究主要依赖对图像基础表示的直接处理，形成了一系列经典的传统方法。根据水印嵌入所处的域不同，传统图像水印技术通常分为空域方法与变换域（频域）方法两大类。此类方法为后续基于深度学习的水印策略奠定了基础，至今仍在部分低成本或结构简单的应用场景中发挥作用。

空域方法通过直接修改图像像素值来嵌入水印信息，代表性技术为最不重要位（Least Significant Bit, LSB）替换法[17]。该方法将水印比特直接嵌入图像像素的最低有效位，由于其操作简便、信息容量高，在早期得到广泛应用。后续研究提出了多种变体，如自适应 LSB 嵌入、边缘区域优先嵌入等策略，以提升隐蔽性与抗攻击能力[63-65]。尽管如此，空域水印对压缩、裁剪、旋转、噪声等常见图像处理操作仍较为敏感，鲁棒性相对较弱。

为增强水印的稳定性与不可见性，研究逐步转向频域水印方法。此类方法首先对图像进行某种频域变换，如离散余弦变换（Discrete Cosine Transform, DCT）、离散小波变换（Discrete Wavelet Transform, DWT）或离散傅里叶变换（Discrete Fourier Transform, DFT），随后在变换系数中嵌入水印信息，再通过逆变换还原图像。其中，DCT 方法常选取中频系数作为嵌入位置，以兼顾人眼感知与抗压缩能力；DWT 则利用图像的多尺度分解特性，将水印分层嵌入到子带系数中，增强对多种攻击的鲁棒性；DFT 方法因其对旋转与缩放等几何变换的稳定性，常用于抗几何攻击的场

景[18-20]。此外，为进一步提高水印的稳定性，一些研究引入奇异值分解（Singular Value Decomposition, SVD）方法，将水印嵌入到奇异值矩阵中，从而实现较强的内容保真性与鲁棒性。

综上，传统图像水印方法在空域与频域之间已形成较为成熟的技术体系，为后续研究奠定了重要基础。然而，此类方法普遍依赖人工设计的嵌入规则与固定嵌入区域，缺乏对图像内容的自适应建模能力，难以在高保真度与高鲁棒性之间实现兼顾。随着深度学习技术的发展，研究者开始探索利用神经网络自动学习嵌入机制，在提升信息容量与安全性的同时，增强对抗攻击的鲁棒性与图像质量的保真性。

2.3.2 基于深度学习的图像水印技术

随着深度神经网络在图像建模与表示学习中的广泛应用，研究者开始将其引入水印系统，以突破传统方法在自适应性、容量与鲁棒性方面的限制。近年来，深度学习驱动的图像水印技术取得了可观的进展，尤其在外生水印场景中展现出较强的可扩展性与部署灵活性。外生水印方法指的是在图像内容生成完成之后，采用深度神经网络对图像进行后处理式的水印嵌入，不依赖于图像生成模型结构，因此在多种生成任务与下游应用中均具备良好的通用性与适配性。

早期研究中，Baluja 首次探索了使用深度神经网络实现图像级联嵌入的方法[21]。该方法采用端到端的卷积网络结构，直接在一幅图像中嵌入另一幅同尺寸彩色图像，通过反向传播优化嵌入图像的保真性与水印图像的可恢复性，在保持视觉质量的前提下实现了高容量的信息隐藏。

Zhu 等人提出的 HiDDeN 框架[22] 是深度学习水印技术的关键进展之一。该方法基于编码器-解码器架构构建嵌入与提取网络，在训练阶段引入的模拟攻击模块作为噪声层，对图像施加 JPEG 压缩、高斯模糊、裁剪、丢失像素等失真变换，以提升水印的鲁棒性。如图 2.3所示，外生水印方法通常由编码器、噪声层和解码器构成，通过端到端训练，实现对人工智能生成图像的水印信息进行稳健的嵌入与提取。在模型训练中，整体损失函数通常包括三部分：

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{rec}} + \lambda_2 \cdot \mathcal{L}_{\text{wm}} + \lambda_3 \cdot \mathcal{L}_{\text{percep}} \tag{2.10}$$

其中，$\mathcal{L}_{\text{rec}}$ 表示嵌入图像与原图之间的重构误差（如 L_2 损失），$\mathcal{L}_{\text{wm}}$ 为水印比特的提取损失，$\mathcal{L}_{\text{percep}}$ 为基于预训练的神经网络特征的感知损失；λ_1、λ_2、λ_3 为损失项权重

超参数，用于调节视觉质量与水印鲁棒性的权衡。噪声层用于模拟图像在实际传输过程中可能出现的失真。

在该工作基础上，Benz 等人进一步提出通用的水印架构[23]，揭示了频域结构在水印图像构建中的关键作用。其研究表明，频域分布的选择影响图像与水印之间的可分性，从而能够提升水印的鲁棒性与隐藏容量。

为进一步扩展水印容量与控制嵌入位置，Lu 等人设计了一种基于可逆神经网络的图像隐写方法[66]，实现了对大容量图像的可恢复嵌入。该方法在图像嵌入过程中构建了双向映射结构，不仅支持原始图像的无损重建，还在提取路径上增强了水印的解码能力。Guan 等人提出的 DeepMIH 框架[67] 在此基础上进一步实现了多图像级联嵌入，即通过级联多个水印子网络，在同一幅图像中嵌入多个水印实例，提高了系统的多样性与实用性。

此外，近期研究也关注水印在频域层面的调控。例如，一些方法在嵌入阶段显式引入傅里叶变换或小波变换模块，对不同频段信息施加权重，从而避开人眼敏感区域或压缩算法频率剪裁带，提高水印的不可见性与抗压缩能力。此类方法能够在保持高 PSNR 与 SSIM 的同时，实现对高斯噪声、JPEG、裁剪等攻击的鲁棒恢复。

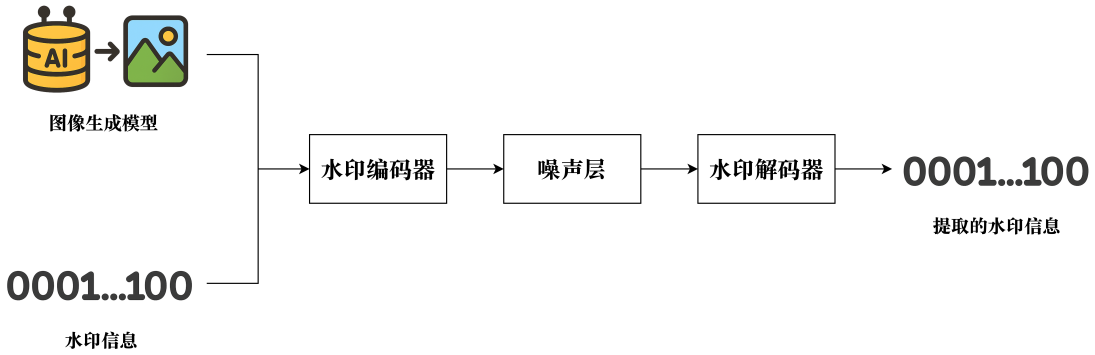

图 2.3　基于深度神经网络的外生图像水印系统结构示意图

相比于基于后处理的外生水印方法，内生水印通过在模型训练或推理过程中主动嵌入水印信息，使得生成内容在生产环节即自然携带可验证标识，具备更高的隐蔽性、溯源能力和安全性。当前主流的内生水印方法大致基于场景可划分为三种技术范式：白盒水印、黑盒水印和无盒水印，分别对应对模型访问权限的不同假设，具有各自的设计策略与应用场景[68]。

白盒水印方法假设水印提取者可以直接访问模型的完整参数信息，因此水印被

显式编码在模型权重中。Uchida 等人最早提出将水印嵌入权重向量的思想，在训练神经网络的同时引入水印嵌入约束[69]。其目标函数通常为：

$$\mathcal{L} = \mathcal{L}_{\text{task}}(f_\theta) + \lambda \cdot \|\mathcal{E}(\theta) - \mathbf{w}\|^2 \tag{2.11}$$

其中，$\mathcal{L}_{\text{task}}$ 表示原始任务的损失函数，$\mathcal{E}(\theta)$ 为从模型参数 θ 中提取水印信息的编码器，$\mathbf{w}$ 为目标水印向量，λ 为调节水印嵌入强度与任务性能之间权衡的超参数。该设计旨在在不影响模型原始任务表现的前提下，实现对水印信息的有效嵌入与提取。后续研究在不同层级（如特定卷积核、激活通道）中加入更灵活的嵌入机制，提升了水印容量与鲁棒性[70-72]。白盒方法的优势在于提取精度高、安全性强，但对部署环境要求较高，现实应用中往往不易满足对模型内部访问的假设。

为适应部署环境中无法访问模型参数的实际需求，研究者提出了黑盒水印方法。该类方法通过设计特定的“触发输入—触发输出”对，在模型训练阶段注入“后门”行为，从而在不破坏主任务性能的同时实现水印标识检测[73-75]。具体而言，在训练集中加入水印样本 (x_t, y_t)，使得训练后的模型满足：

$$f_\theta(x_t) \approx y_t \tag{2.12}$$

在测试阶段，通过输入触发样本 x_t 并检测其输出 y_t 是否与预设值一致，可用于验证水印的存在。该方法具有一定的隐蔽性与灵活性，适用于不公开模型参数的商业应用场景。然而，由于触发样本数量有限，且存在被绕过或篡改的风险，其鲁棒性与安全性仍是当前研究的重点方向。

随着生成模型结构日益复杂，输出内容越来越多样，无盒水印方法应运而生。该类方法不依赖于模型结构访问（白盒）或额外触发样本设计（黑盒），而是在生成过程中引导模型输出内容自然携带可识别水印信息，从而实现完全隐式的溯源控制。Wu 等人[30]提出无盒水印的框架，通过对生成模型进行微调，使模型在不知情状态下学习在输出图像中嵌入水印特征；Zhang 等人采用对抗训练机制提升了水印对模型压缩与迁移的鲁棒性[76-77]。此外，一些研究尝试通过微调模型或构造特定的采样策略，在保持生成性能不变的前提下实现水印嵌入。例如，Pivotal Tuning 方法[78] 通过对模型参数的局部子集进行微调，引导生成内容携带特定标识。进一步，Pierre 等人[79]提出在扩散模型的噪声输入空间中嵌入水印：通过对初始噪声 z_0 引入结构化扰动，使得扩散过程生成的最终图像 x_0 在保持感知质量的同时，隐式携带可识别的水

印特征。该方法不修改模型权重或采样算法，而是通过微调噪声生成器或设计噪声扰动分布，实现水印在图像生成路径中的自然注入。这种策略兼顾了图像质量与水印隐蔽性，适用于预训练扩散模型的水印增强任务，并在图像经过压缩、缩放等典型失真操作后仍表现出良好的鲁棒性。这类方法的核心在于将水印信号嵌入模型生成分布中，确保其在语义与视觉上不可感知、在统计意义上可检测。

综合来看，基于深度学习的图像水印技术，尤其是外生与内生策略的不断演进，地拓展了水印系统的应用边界。外生水印以其模型无关性和部署灵活性在多种图像生成任务中具有良好的通用性，而内生水印则借助模型生成机制本身，在提升隐蔽性与鲁棒性方面展现出独特优势。两类方法在目标定位、嵌入路径及设计约束上各具特点，但都面临如高保真嵌入、跨模型适配和抗破坏性增强等关键挑战。深入理解这两类方法的基本原理与技术路线，将为后续高保真、高鲁棒水印方法的设计提供坚实的理论基础与实践参考。

2.3.3 常见的攻击手段

在图像水印系统中，鲁棒性是衡量其实际可用性的重要指标。现实场景中，生成图像往往会在压缩、传输、编辑乃至恶意攻击中遭遇多种干扰操作，这些操作可能严重削弱水印的可提取性与可信度。综合现有研究，可将常见攻击手段大致分为两类：其一为像素级别攻击，主要包括缩放、裁剪、压缩、加噪与滤波等基础图像处理操作，这类攻击常见于内容传播链路中的无意干扰；其二为图像级别攻击，包括几何畸变与对抗攻击等更具结构性破坏性的扰动方式，通常用于蓄意规避水印检测或篡改认证机制。这两类攻击在形式与破坏机制上存在较大差异，给水印方法的稳定性提出了不同维度的挑战。

像素级别攻击主要通过对图像中局部像素值的直接干扰来削弱水印系统的鲁棒性。这类攻击在不改变图像整体语义或结构的前提下，往往会造成统计分布、频率特征或空间一致性的微妙扰动，从而影响水印的提取与验证。这些操作广泛存在于图像采集、压缩、编辑与传输等实际应用中。常见的像素级攻击包括但不限于图像平移、裁剪、压缩、加性噪声扰动与滤波操作。这些扰动虽然不直接更改图像语义内容，却往往破坏了水印嵌入依赖的局部结构或频率特性，导致水印信息在提取阶段无法正确恢复。

缩放操作通过改变图像的尺寸，导致嵌入水印的像素被压缩或插值重构，破坏

原有嵌入位点的结构特征；裁剪操作则直接移除图像的边界区域，一旦水印嵌入集中在被裁剪部分，将造成信息严重缺失，影响水印的完整性。

裁剪操作会截取原图的某个区域，若水印嵌入部分被裁剪掉，可能导致关键信息丢失，从而影响水印的完整性和提取准确率。

图像压缩（如 JPEG）则通过分块离散余弦变换并对高频系数进行量化以降低存储开销。其二维 DCT 可定义为：

$$C_{u,v} = \frac{1}{4} \sum_{x=0}^{7} \sum_{y=0}^{7} f(x,y) \cos\left[\frac{(2x+1)u\pi}{16}\right] \cos\left[\frac{(2y+1)v\pi}{16}\right], \tag{2.13}$$

其中 $f(x,y)$ 表示图像中像素值，$C_{u,v}$ 为变换后的频域系数。压缩过程中对高频部分的剪除会直接削弱频域水印的信号强度。

此外，加性高斯噪声会破坏嵌入的微弱扰动信号，其扰动形式为：

$$x' = x + n, \quad n \sim \mathcal{N}(0, \sigma^2), \tag{2.14}$$

其中，x 表示原始图像，x' 为加噪后的图像，n 是服从均值为 0、方差为 σ^2 的独立同分布高斯噪声。参数 σ 控制噪声的强度，即噪声的扰动幅度。由于水印通常以微弱扰动的形式嵌入图像中，因此较大的噪声可能掩盖或破坏这些关键信号，导致水印信息提取失败。

滤波操作（如均值滤波、高斯模糊）通过平滑局部像素变化来降低图像纹理细节，同样会削弱嵌入水印的信号特征，进一步降低其可提取性与稳定性。

图像级别攻击主要通过改变图像的几何结构或利用生成模型的脆弱性，破坏水印信息在更大尺度上的一致性与可辨识性。这类攻击往往不局限于局部像素修改，而是涉及整体图像的形态变换、语义重构或模型行为干扰，因而对水印的稳定性与可提取性构成更严峻的挑战。

常见的图像级攻击包括几何畸变与对抗攻击等。几何畸变通过旋转、非线性扭曲等方式干扰图像的结构一致性。以旋转操作为例，其可建模为二维坐标变换：

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{2.15}$$

其中 θ 表示旋转角度。此类变换会扰乱空间对齐关系，对依赖图像局部空间位置编码的水印系统具有较强破坏性。此外，仿射变换是图像级别攻击中另一种常见方式，

它通过线性变换与平移联合作用，实现图像的旋转、缩放、剪切与位移等操作。仿射变换在二维空间中可表示为：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{2.16}$$

其中 a_{ij} 表示线性变换矩阵系数，(t_x, t_y) 为平移向量。仿射变换保持图像的共线性与比例关系，但会破坏像素的局部排列与边界结构，进而影响水印系统依赖的空间一致性或频域特征稳定性，尤其对位置敏感的嵌入策略而言，其破坏效果尤为明显。

对抗攻击则利用深度神经网络的判别边界脆弱性，通过梯度计算，在输入图像中添加结构性扰动 δ，使其在视觉上无差异的情况下误导模型输出。典型的对抗扰动生成可通过最小化以下目标实现：

$$\delta^* = \arg\min_{\delta} \mathcal{L}(f(x+\delta), y) + \lambda \|\delta\|_p, \tag{2.17}$$

其中 $f(\cdot)$ 为目标模型，$\mathcal{L}$ 为损失函数，y 为原始标签，λ 控制扰动强度，$\|\delta\|_p$ 为 L_p 范数。当水印算法对输入图像较为敏感时，此类扰动易造成水印信号失效或误识别。

与像素级攻击相比，图像级攻击往往能在保持图像整体视觉可接受性的同时，实现对水印嵌入策略的精准破坏。针对该类攻击，提高水印的全局一致性与跨尺度鲁棒性，成为鲁棒水印设计的重要方向之一。

2.3.4 高频伪影与保真度挑战

在人工智能生成图像任务中，高频伪影是影响生成图像视觉质量与真实感的关键因素之一。该类伪影通常表现为边缘不自然、纹理重复以及细节区域中出现异常噪声等现象，在严重情况下甚至会降低图像的可用性并削弱用户的信任感。此类伪影的产生多源于生成模型内部机制的局限性或训练目标设定的不充分，尤其在以卷积神经网络为基础的生成模型中更为常见。

高频伪影的产生可主要归因于以下几个方面：首先，生成模型的频谱分布存在不均衡现象，模型在重建低频结构（如轮廓、背景）方面通常表现良好，但在高频细节（如纹理、边缘）处理上易出现噪声放大或模式拟合异常，从而在频域中引入非自然高频分量，造成视觉伪影。其次，优化目标设计不合理也是重要诱因。传统的损失

函数（如 L_1 或 L_2 范数）在度量图像质量时侧重像素级差异，而难以捕捉高层次的语义或感知一致性。这类损失函数往往忽略图像高频区域的感知特性，导致生成图像在细节区域呈现结构性噪声或断裂纹理。此外，神经网络中的上采样模块（如转置卷积、双线性插值）在重建图像分辨率的同时，也可能引入周期性重复模式，形成典型的棋盘状伪影。这类伪影在视觉上较为明显，不仅影响生成图像的自然性，也对后续的图像处理和感知任务造成干扰。

在此背景下，水印嵌入任务面临额外的挑战。许多水印系统，尤其是空域或频域嵌入方法，倾向于在高频区域进行信息植入，以避免干扰图像的结构性主干。然而，生成图像本身在高频部分的不稳定性，使得水印的嵌入更加复杂。一方面，图像中的伪影区域结构性扰动可能掩盖水印信号，导致水印提取精度下降；另一方面，若水印嵌入未能充分考虑原始图像的高频分布，可能进一步强化伪影，导致图像的整体保真度下降。

此外，图像的高频信息往往是非常易受干扰的。由于图像的高频区域包含了大量细节和纹理信息，这些区域通常是噪声最为集中的地方。因此，压缩、滤波或图像变换等操作都可能对高频信息产生影响，进而影响水印的稳定性和可提取性。由于高频信息在视觉感知中占据了重要地位，其一旦被破坏，水印信号便可能失效，尤其是在面对图像编辑、压缩或其他扰动时，水印的鲁棒性变得较差。这使得设计一种既能保证水印隐蔽性，又能在遭受攻击时仍能可靠提取的高保真水印方法，成为当前图像水印领域的重要挑战。

因此，为确保水印技术在人工智能生成图像中的有效性和适用性，研究者需在设计嵌入策略时充分考虑图像的频谱特性与感知质量，避免与模型固有的伪影特征发生冲突，特别是要提高对高频不稳定区域的适应性与鲁棒性。这不仅是提升图像视觉保真度的关键，也是实现高可靠水印嵌入的前提。

2.4 大语言模型水印相关技术与基础

随着大语言模型在自然语言处理领域的广泛应用，如何有效保护其生成内容的版权和溯源性成为一个亟待解决的问题。大语言模型水印技术作为一种重要的解决方案，通过在生成内容中嵌入可追溯的标识信息，以确保生成内容的所有权和来源可验证。本节将系统梳理与探讨大语言模型水印的核心技术与理论基础。首先，将

回顾大语言模型的基本架构及关键技术要素，包括 Transformer 结构、注意力机制以及大规模参数训练策略。随后，介绍当前主流的水印方法，并将其划分为两大类：无需训练的水印技术与基于微调的水印技术。最后，讨论当前水印技术所面临的主要挑战，特别是在确保生成质量、提升鲁棒性和降低计算资源开销等方面的技术瓶颈。

2.4.1 大语言模型技术

大语言模型在自然语言处理领域表现出卓越的性能，已被广泛认为是当前最具代表性和最强大的语言建模技术之一。其基础架构主要基于 Transformer 模型[5]，自 2017 年提出以来，迅速成为了各类 NLP 任务的核心架构。Transformer 的最大创新之一是引入了注意力机制，该机制通过并行处理输入数据而非传统的 RNN 中逐步处理序列的方式，提升了计算效率并且避免了长程依赖问题。

Transformer 模型的核心是自注意力机制，它使得每个输入词的表示都能够与其他所有词的表示直接关联，从而捕捉词语之间的长期依赖关系。具体而言，给定输入词的嵌入表示 $\mathbf{x} = [x_1, x_2, \ldots, x_n]$，Transformer 通过计算输入序列中各个词的注意力权重来决定它们之间的关联性。在注意力机制中，输入特征 $\mathbf{X}$ 会分别通过三组权重矩阵 $\mathbf{W}_{\mathrm{Q}}$、$\mathbf{W}_{\mathrm{K}}$ 和 $\mathbf{W}_{\mathrm{V}}$ 线性变换，生成查询（Query）、键（Key）和值（Value）矩阵：

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_{\mathrm{Q}}, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_{\mathrm{K}}, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_{\mathrm{V}} \tag{2.18}$$

其中，$\mathbf{X} \in \mathbb{R}^{n \times d}$ 表示输入的特征序列，n 是序列长度，d 是每个位置的特征维度。接着，模型通过计算查询 $\mathbf{Q}$ 和键 $\mathbf{K}$ 之间的点积相似度，经过 softmax 归一化生成注意力权重，最后用这些权重对值矩阵 $\mathbf{V}$ 进行加权求和，形成最终的注意力输出：

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.19}$$

其中，d_k 是键向量的维度，用于缩放点积的结果，以避免数值过大导致的梯度消失问题。注意力机制允许模型在处理每个词时，动态地关注输入序列中其他词的信息，从而捕捉到更丰富的上下文信息。通过堆叠多个这样的自注意力层，Transformer 架构能够有效捕捉序列中复杂的依赖关系，并且处理较长的序列，避免了传统 RNN 在长序列处理时遇到的梯度消失和爆炸问题。Transformer 中还引入了位置编码来弥补模型对顺序信息的缺失，确保输入序列中的词语位置信息能够被保留。如图 2.4 所示，Transformer 模型的架构包括了编码器和解码器两个部分，编码器负责处理输入

序列并生成其表示，而解码器则根据编码器输出的表示生成预测的目标序列。该结构支持并行计算，从而加速训练过程，并能够有效捕捉长期依赖。

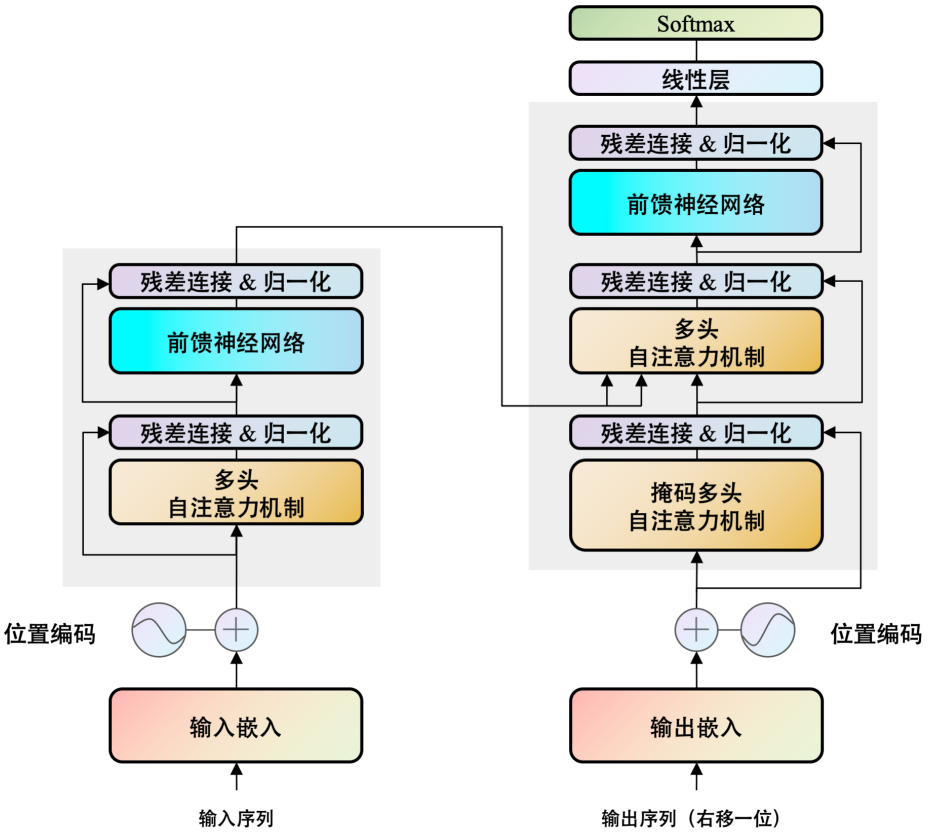

图 2.4　Transformer[5] 模型架构

在生成模型中，输出的每个词由一个词汇表中的所有词的概率分布决定。这一概率分布通过 softmax 函数计算得到，公式如下：

$$P(t_i|t_1,\ldots,t_{i-1}) = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^{|\mathcal{V}|}\exp(\mathbf{z}_j)} \tag{2.20}$$

其中，$P(t_i \mid t_1,\ldots,t_{i-1})$ 表示在已生成词 t_1 至 t_{i-1} 的条件下，生成下一个词 t_i 的概率；$\mathbf{z}_i$ 是词 t_i 在输出层对应的得分，$|\mathcal{V}|$ 表示词汇表的大小。Softmax 操作将所有词的得分转化为概率分布，并确保总和为 1。在实际生成过程中，模型依据 softmax 得到的概率分布选择下一个词。常见的采样策略包括贪心搜索、随机采样和温度采样等，这些方法对生成文本的多样性与流畅性具有较大影响，尤其在自然语言生成任务中尤为关键。

Transformer 架构的成功为大规模预训练模型的发展奠定了基础。为了提高语言

模型的能力，当前的做法是采用大规模的预训练，再通过微调等方法来适应特定的任务。预训练通常依赖海量的无标签数据，采用自回归或自编码方式，学习词汇、语法和语义等语言特征。例如，GPT 系列使用自回归方法进行预训练，训练目标是最大化给定上下文下预测下一个词的概率；而 BERT 采用双向编码结构，通过对输入中的部分词进行掩码，训练模型预测被遮挡的词汇。预训练完成后，模型可以针对特定任务（如文本分类、问答等）进行微调。在微调阶段，预训练模型的参数会根据任务的具体需求进行调整，通常采用有标签的数据对模型进行监督学习。预训练和微调的结合使得大语言模型能够高效地从大规模文本中学习一般语言特征，并在特定任务中获得优异的性能。

2.4.2 大语言模型水印方法

针对大语言模型生成内容的可追溯与可验证需求，研究者提出了多种文本水印策略以嵌入隐藏标识。由于文本生成的离散性与语义敏感性，直接迁移图像领域的嵌入方式往往效果有限，因而形成了适配语言模型自身特性的两类主要技术路径。一类方法在不修改模型结构的前提下，通过采样控制或输出重写策略，在生成文本中注入特定统计或语义特征，代表性方法包括语义编码等，本文称其为无需训练的水印方法。另一类方法则通过模型微调，使生成模型在输出内容的同时自然携带水印特征，具备更强的嵌入稳定性，本文称之为基于微调的水印方法。本节将按上述划分，分别探讨这两类方法的核心机制、代表技术与适用场景，并为后文提出的方法奠定技术基础。

无需训练的大语言模型水印技术是一类在推理阶段进行水印嵌入的方法，无需引入额外的训练过程，因而具有良好的实用性与工程可部署性。该类方法主要通过控制模型生成过程或对生成结果进行改写，以低成本、低侵入的方式完成水印信息的编码，适用于现有的黑盒模型场景。总体而言，无需训练的水印技术可大致分为两类：一类是基于已有文本的后处理嵌入策略，另一类则在文本生成阶段通过调整采样策略或引入辅助得分函数来实现隐式嵌入。两者在可实现性、保真度、嵌入容量与检测稳定性之间各具优势，构成当前无需训练水印研究的主要方向。

基于已有文本的水印嵌入技术通过对原始文本进行轻量级修改，以隐蔽方式植入可识别的水印信号，常见策略包括格式调整、词汇替换和语义改写。格式类方法通常在不影响可读性的前提下，对文本的排版、字体或间距等呈现形式进行微调，实

现对比特信息的编码[25-26]；词汇替换方法则选取语义接近的同义词或近义词替代原文中的特定词语，以在保持语义一致性的同时嵌入水印[28-29]；语义改写策略进一步扩展为对整句或段落进行等价表达的重构，使水印信息分布在更高层次的语言结构中，从而增强其不可察觉性和抗简单编辑能力[80]。这类方法无需访问模型内部结构，适用于静态文本内容的水印添加，具有实现简单、代价低廉的优势，但在鲁棒性和容量方面仍面临一定挑战。

除了对已有文本进行后处理外，另一类具有代表性的策略是在文本生成过程中直接干预模型的采样机制，以实现水印信息的隐蔽嵌入。该类方法同样不依赖于模型的微调，而是在推理阶段对输出概率分布或采样路径进行引导，从而嵌入可检测的标识。当前主要包括红绿列表采样方法与水印得分驱动方法两种实现形式[81]。

红绿列表采样方法通过对语言模型输出的 logits 分数进行调制[35]，使模型在自回归生成过程中倾向性地选择词汇表中的某一子集，从而在输出文本中嵌入水印信息。该方法将词汇表 $\mathcal{V}$ 按照哈希函数划分为绿色列表 $\mathcal{G}$ 与红色列表 $\mathcal{R}$，并向绿色词汇对应的 logits 添加一个偏置 δ，调整后的 logits 向量 $\mathbf{l}_{\mathrm{w}}$ 表达式如下：

$$\mathbf{l}_{\mathrm{w}} = \begin{cases} l_{\mathrm{o}} + \delta, & t_j \in \mathcal{G} \\ l_{\mathrm{o}}, & t_j \in \mathcal{R} \end{cases} \tag{2.21}$$

其中，$\mathbf{l}_{\mathrm{o}}$ 表示模型原始的 logits，t_j 为当前考虑的词。在生成阶段，模型因偏置引导倾向选择绿色词，从而实现水印的嵌入。水印的检测则可基于绿色词汇出现的频率进行统计检验，判断生成文本中是否存在水印。

水印得分方法则从另一个角度出发，通过引入外部水印得分函数对生成候选进行重排序，以在不明显改变 logits 分布的前提下实现水印嵌入[81]。该方法设计一个可控的水印打分函数 $\mathbf{s}$，与 logits 向量 $\mathbf{l}$ 联合构建输出概率分布 $\mathbf{p}$，其整体框架为：

$$\mathbf{p} = f(\mathbf{l}, \mathbf{s}) \tag{2.22}$$

其中，函数 $f(\cdot)$ 用于综合 logits 与水印得分信息，从而在保持语义自然性和语言流畅性的前提下完成水印的编码。OpenAI 研究员提出的伪随机函数方法即为此类代表，其利用伪随机函数 $f_s(\cdot)$ 以前 $n-1$ 个 token 组成的上下文序列 $(t_{i-n+1}, \ldots, t_{i-1})$ 作为

输入，生成伪随机得分 $\mathbf{r}_i$，并据此调整原始 softmax 概率 p_o^i，形成新的采样分布 p_w^i：

$$p_w^i \leftarrow r_i^{1/p_o^i}, \quad r_i = f_s\left(t_{i-n+1}, \ldots, t_{i-1}\right) \tag{2.23}$$

其中，p_o^i 表示模型原始输出的概率，p_w^i 为经过计算得到的含水印概率。该过程将模型的语言偏好 p_o^i 与水印信号 r_i 结合，从而在不显式更改语义的前提下嵌入水印信息。值得注意的是，只有当某一 token 对应的 p_o^i 和 r_i 分量均接近 1 时，其在 p_w^i 中的概率得到提升，因此更有可能在采样过程中被选为输出。该机制确保了水印嵌入的选择性与隐蔽性。

检测阶段可通过如下累加得分计算判断水印是否存在：

$$s_d = \sum_{i=1}^{T} \ln \frac{1}{1 - r_t'}, \quad \text{其中} \quad r_t' = f_s(t_{i-n+1}, \ldots, t_i) \tag{2.24}$$

在此基础上，一些研究进一步提出无失真水印策略[82]，保持嵌入后文本的生成分布与原始模型完全一致，通过对采样概率进行指数重参数化实现，如下所示：

$$\mathbf{p}_w^i \leftarrow \left(\xi^{(j)}\right)^{1/\mathbf{p}_o^i} \tag{2.25}$$

其中，$\xi^{(j)}$ 为随机采样的 logits 值，$\mathbf{p}_w^i$ 和 $\mathbf{p}_o^i$ 分别表示嵌入水印后和原始的概率分布。该方法通过对 logits 进行重参数化，使得嵌入后的文本生成分布与原始模型保持一致，从而实现无失真水印。

综上，这些无需训练的水印方法在不改变模型结构的前提下，有效实现了语言模型输出内容的标识与可追溯性，适用于大模型开放部署、黑盒验证等场景，为文本 AIGC 内容治理提供了高效可行的技术路径。

相较于推理阶段进行干预的无需训练方法，基于微调的大模型水印技术通过在训练或微调过程中引入水印信号，借助模型本身的学习能力实现更深层次的水印嵌入。这类方法虽在模型准备阶段具有更高的计算成本，但能够在不增加推理开销的前提下，在模型内部形成稳定而隐蔽的水印表示，具有更强的鲁棒性与适应性，适用于对安全性要求更高的场景。该类方法的核心策略主要包括两种路径：一是通过无监督微调等方式将水印直接融合进模型的参数中；二是通过引入外部水印解码器并配合优化机制，使得生成文本可由外部模块识别验证。

在无监督微调策略中，研究者通常设计端到端的水印嵌入—提取结构，通过在模型的训练目标中显式引入水印信息，从而在保持原有文本生成质量的同时实现可

识别的水印嵌入。Abdelnabi 等人提出的 Adversarial Watermarking Transformer（AWT）框架即采用 Transformer 编码器-解码器的结构，其中的隐藏网络负责在输入文本中编码二进制水印信息，而解码网络仅依据水印文本恢复原始消息[33]。为提高水印的不可见性与抵抗检测能力，该方法引入判别器并采用对抗训练策略，在最大化水印可恢复性、最小化文本扰动和欺骗检测器三者之间联合优化。Zhang 等人进一步优化了 AWT 框架中重参数化技术，提升了文本生成的连贯性和水印鲁棒性[83]。此外，Bertini 等人通过交叉注意力机制设计了嵌入式水印层，在不明显影响语言模型输出性能的基础上增强了水印一致性[84]。

另一种技术路线则引入外部解码模块，通过联合训练语言模型与水印识别器来实现水印的嵌入与验证。此类方法更具灵活性，适合于多任务模型或解耦部署场景。例如，Xu 等人提出利用强化学习中的近端策略优化（Proximal Policy Optimization, PPO）算法联合训练语言模型与水印解码器，使得模型生成的文本能够被外部模块可靠识别[34]。该策略可与模型对齐训练过程协同进行，降低额外开销。虽然其支持的水印信息容量有限，但在保持文本自然性的同时，具备较强的可验证性。为了提升水印容量，Xu 等人进一步提出通过构造重写模型分别嵌入 0 与 1 比特，并训练分类器进行识别[85]；同时，他们通过句子级分布式嵌入方式实现多比特扩展。

除显式设计嵌入策略外，部分工作借助大语言模型自身的生成能力进行水印策略学习。例如，研究者提出利用语言模型自动生成词汇替换规则、句式变换模板等嵌入手段，然后再借助外部模型进行检测[86-87]。这类“模型辅助生成—外部验证”的方法展现了大语言模型在水印设计过程中的潜力，尤其适用于水印对抗性提升与上下文敏感嵌入任务。

综上所述，基于微调的水印方法具有更强的集成性和水印稳定性，能够在不依赖推理阶段外部干预的情况下实现隐蔽、可验证的水印嵌入。但其训练开销与模型访问权限的要求相对较高，需在实际应用中结合模型规模、部署方式与安全需求进行综合考量。

2.4.3 大语言模型水印的难点与挑战

尽管近年来面向大语言模型的水印研究取得了初步进展，无论是外生水印还是内生水印方案，在实际应用中仍面临多方面挑战。总体而言，这些挑战集中体现在高保真度与生成质量、水印鲁棒性与安全性，以及模型微调过程中的资源消耗与性

能影响等方面。

首先，高保真度始终是文本水印技术面临的首要问题。不同于图像模态可以在感知域中接受一定程度的微扰，用户对文本的生成质量的敏感度较高。无论是基于词汇替换、格式嵌入的外生方法，还是通过采样引导或模型参数注入的内生方法，若嵌入策略影响语言流畅性、语法结构或语义一致性，都将严重损害用户体验，甚至暴露水印痕迹。因此，在保证语义完整、语言自然的前提下嵌入可识别信号，是构建实用文本水印系统的关键目标。尤其对于高质量文本生成任务，如教育、创意写作或自动新闻摘要等场景，水印方法更需兼顾输出内容的逻辑性与风格统一性。

其次，水印的鲁棒性与安全性也面临诸多挑战。攻击者可能通过重写、翻译、摘要等语义保持操作规避检测，或借助语言模型本身进行重写攻击来淡化水印特征。此外，若攻击者掌握水印嵌入策略，甚至可通过逆向分析或对抗优化构建“去水印”模型。对于外生方法，基于简单规则或嵌入词频偏移的机制常在轻微修改下失效；而内生方法如红绿列表采样与得分引导，在面对大模型的续写、再生成等操作时亦存在水印信息衰减的问题。

最后，从资源消耗的角度出发，内生水印方法尤其需要谨慎设计。多数基于微调的技术方案需要对原始模型进行微调，这在百亿级参数规模的大模型上代价昂贵。此外，若嵌入过程引入新结构或模块，也可能降低模型的推理速度，影响部署效率。在一些敏感应用中，如边缘部署或高频调用的 API 接口，复杂的水印机制甚至可能带来稳定性或延迟风险。因此，如何在模型无修改、推理无延迟的条件下实现可靠水印嵌入与检测，仍需进一步探索。

2.5 本章小结

本章首先从生成式人工智能的发展背景与安全挑战入手，介绍了图像生成与文本生成的主要技术框架，随后系统梳理了数字水印的概念、分类方法及常用评价指标，并结合 AIGC 场景下的特定挑战，阐述了外生水印与内生水印各自的原理与适用范围。在图像水印方面，本章列举了传统方法与基于深度学习的典型实现，并讨论了高频伪影在人工智能生成图像中的干扰影响。在文本水印方面，则概述了大语言模型的基本原理，介绍了主流的水印嵌入策略，如采样分布调整与伪随机函数引导的方法，并分析了当前方法在鲁棒性、不可感知性与检测效率之间的权衡问题。

第三章　抑制高频伪影的鲁棒图像生成水印技术

本章围绕图像水印的高保真嵌入与鲁棒性优化问题展开研究，关注于两个核心矛盾：一是如何在不破坏图像感知质量的前提下实现有效水印嵌入；二是如何提升水印在多种典型攻击下的稳定提取能力。针对当前人工智能生成图像中普遍存在的高频伪影现象，本章从频域特性出发，提出一种抑制高频伪影的鲁棒图像水印框架，以降低水印嵌入与图像高频结构之间的干扰，进而提升整体水印系统的性能。

3.1　引言

在人工智能生成图像广泛应用的背景下，实现对内容的追踪与溯源已成为亟需解决的问题。作为关键手段之一，图像水印技术需同时满足隐蔽性、鲁棒性与高保真度三方面要求。相比传统图像，人工智能生成图像具有更复杂的高频结构分布，且常伴随不可忽视的伪影噪声，这给水印的设计与评估带来了新的挑战。一方面，水印嵌入往往选择频域中不显著影响感知质量的区域，如高频系数。然而，人工智能生成图像的中高频区域的不稳定性可能使水印信号被误覆盖、干扰或放大，从而引发可感知伪影，影响水印的隐蔽性。另一方面，嵌入水印所依赖的频率信息在经过压缩、裁剪、滤波等攻击后易被破坏，严重影响水印的可提取性与鲁棒性。

本章构建了一种融合高频伪影抑制与对抗训练机制的鲁棒图像水印框架。通过对上采样模块的精心设计和端到端的训练策略，所提出的方法能够在有效抑制高频伪影的同时，提升了水印在多种失真场景下的鲁棒性，并保持了图像的视觉感知质量。本研究为实现具有高实用性和质量可控性的人工智能生成图像水印系统提供了一定的技术支撑。

3.2　总体框架

本章提出的抗高频伪影鲁棒图像水印模型总体架构如图 3.1 所示。该模型由四个可训练的模块，编码器、解码器、鉴别器、CNN-F 模块以及一个不可训练的噪声层组成，旨在通过端到端的学习框架，提高水印的隐蔽性、准确性以及对抗各种攻击的能力。模型的核心思路通过在载体图像和水印图像之间建立高效的映射关系，使

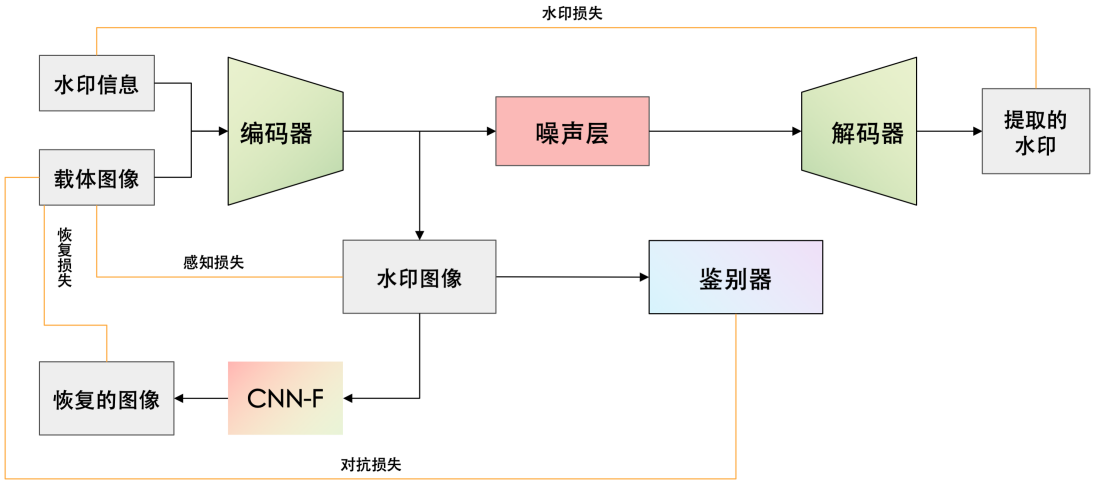

图 3.1 图像水印算法整体框架示意图

水印信息的嵌入对原始载体图像的视觉质量影响较小，同时确保水印可从受到干扰的图像中可靠提取。

模型的整体框架如下：首先，编码器采用 U-Net 结构[88]，接收载体图像和水印信息，并通过网络学习，在确保视觉质量和准确提取的前提下，将水印以一种不可见的形式嵌入载体图像，生成水印图像。随后，利用噪声层对该图像施加模拟攻击，以提升水印对现实场景中各种干扰（如压缩、噪声、滤波等）的抵抗能力。紧接着，解码器从可能已失真的水印图像中提取水印信息，确保水印在复杂环境下仍可被准确恢复。与此同时，鉴别器在多个频率域对含水印图像与无水印的载体图像进行判别，以增强水印的隐蔽性，使其难以被检测或移除。此外，引入 CNN-F 模块，以学习水印图像与原始载体图像之间的映射关系，从而进一步提升模型的性能。

不同于传统水印技术，本文在模型中引入了抗频谱混叠的上采样方法，以在增加特征尺寸时减少上采样过程中可能引入的高频伪影。该方法能够有效抑制水印嵌入过程中常见的伪影，使水印信息更为平滑地融合到载体图像之中，从而提升水印图像的不可感知性。此外，鉴别器还采用多频带分析机制，不仅在空间域对水印图像进行监督，同时在频域上的不同频段内施加约束，以确保水印在各个频率分量上均保持隐蔽性，从而增强水印的抗检测能力与鲁棒性。

为增强水印的抵抗真实环境的攻击，模型引入了一种基于对抗训练的优化策略，通过在训练过程中不断施加干扰并迭代优化水印信息，使模型在复杂现实场景下具

备更强的适应性和抗干扰能力。同时，在损失函数设计上，模型融合了感知损失与对抗损失，前者确保水印嵌入后载体图像的视觉质量不受明显影响，而后者则在鉴别器的监督下提高水印的不可感知性，使其更加隐蔽且难以检测。

整体而言，该框架结合深度神经网络的特性与频域特征，在一定程度上优化了传统水印技术在隐蔽性和鲁棒性方面的不足。后续章节将详细介绍各个模块的具体设计，包括编码器的抗频谱混叠上采样机制、多频带鉴别器的优化策略、CNN-F 模块的图像恢复，以及整体损失函数的设计，并通过实验验证其有效性。

3.3 水印嵌入

3.3.1 抗频谱混叠的编码器

在水印嵌入过程中，编码器承担着核心任务，即将水印信息无缝地融合到原始载体图像中，同时确保水印的不可感知性，使嵌入后的图像在视觉上与原始载体图像保持一致。此外，编码器还需保证水印的鲁棒性，使其在后续压缩、滤波、重采样等常见图像处理操作后依然能够被准确提取。为此，编码器要具备强大的图像重建能力，使其不仅能够嵌入水印，还能在高质量的图像合成过程中抑制伪影，优化水印的分布。

然而，传统基于转置卷积的生成器在执行上采样操作时，常常伴随频谱混叠效应的产生，从而在生成图像中引入周期性高频伪影。该现象主要源于转置卷积在低分辨率特征图上进行重构时无法有效抑制频率重叠，导致原本属于低频的信息在上采样后被错误映射至高频区域，形成结构性噪声与纹理异常。正如图 3.2 所示，在放大的细节区域，伪影尤为明显，尤其是在图像边缘与纹理分布密集的区域，高频干扰更加突出且不自然。这些高频伪影不仅严重降低图像的感知质量，还影响水印的隐蔽性和鲁棒性。一方面，伪影使得水印信息更容易暴露于图像表面，易被人眼察觉或通过图像处理手段检测；另一方面，由于伪影常与水印信号交叠，可能在经历压缩、滤波等操作后导致水印信息失真甚至丢失。

水印嵌入的核心在于确保水印信息能够隐蔽地融入图像，同时保持图像质量。然而，传统的转置卷积在上采样过程中容易引入混叠效应，导致高频伪影的产生。这种伪影在纹理复杂或边缘清晰的区域尤为明显，使得图像在视觉上呈现出不均匀的周期性噪声。混叠的主要原因在于，转置卷积在上采样时会将低频信号错误地复制

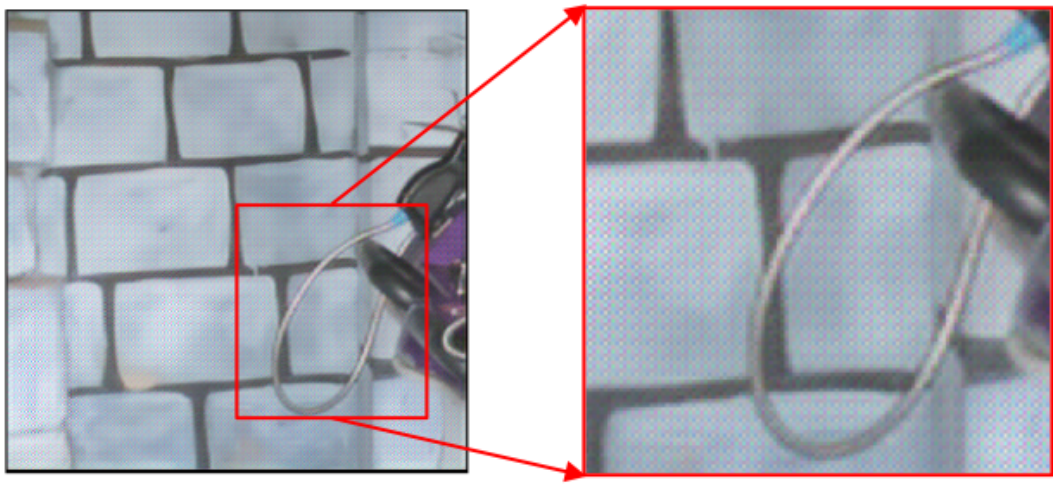

图 3.2 高频伪影示意图

到高频区域，从而导致图像细节失真，使水印嵌入后的图像质量下降，甚至暴露水印的嵌入痕迹，影响隐蔽性和鲁棒性。

为缓解转置卷积在上采样过程中所带来的混叠问题，本文设计了一种基于最近邻插值与标准卷积组合的上采样策略。该方法在结构上用最近邻插值替代了传统的反卷积操作，首先对低分辨率特征图进行尺寸扩展，再通过标准卷积提取局部空间结构与纹理特征。这种设计不仅规避了棋盘效应的产生，还能降低频谱混叠带来的高频伪影风险。具体而言，最近邻插值作为一种非学习型的上采样方式，具有较强的稳定性，能够在保留结构信息的同时，避免由权重学习引发的周期性伪影。而后续的标准卷积层则对插值结果进行进一步的细节建模与优化，使水印能够在图像中更自然地分布，减少对敏感高频区域的扰动，从而提升嵌入后的隐蔽性与图像整体的感知质量。图 3.3 展示了传统转置卷积的上采样策略与本文提出的抗频谱混叠上采样策略的结构对比。

从频域角度来看，转置卷积在上采样过程中会导致频谱混叠，将原有低频信息扩展至高频区域，形成额外的噪声分量，使得水印嵌入后的图像在频域上表现出异常峰值。这不仅降低了视觉质量，也可能成为水印检测和攻击的突破口。相比之下，最近邻插值结合标准卷积的方法则能有效减少这种影响，使水印嵌入后的图像在频域上更加平滑，降低了高频干扰，提高了嵌入水印的稳健性。

综上所述，所提出的抗频谱混叠上采样策略在提升水印隐蔽性的同时，有助于减少嵌入过程中的伪影，为后续的提取与恢复过程创造了更有利的条件。

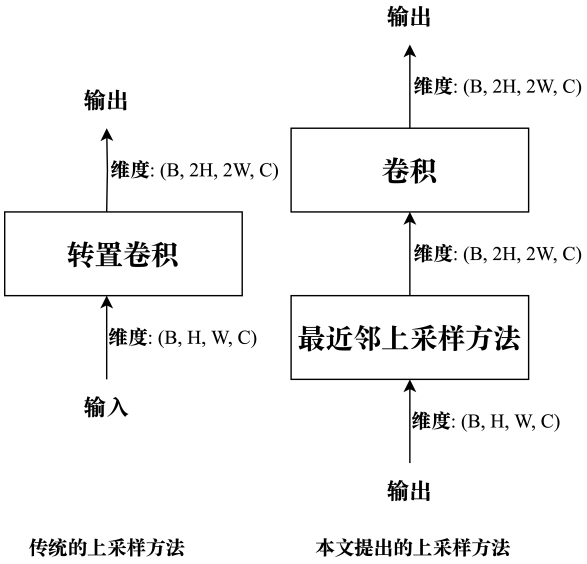

图 3.3　抗频谱混叠的上采样模块与转置卷积对比图

3.3.2　多频带鉴别器

生成对抗网络在水印嵌入任务中提供了强大的能力，通过引入鉴别器与生成器交替优化，使模型不断提升水印嵌入的质量。在传统生成对抗网络结构中，生成器通常以随机噪声作为输入，而在水印嵌入场景中，模型的输入则是原始载体图像。因此，生成器的任务不仅是生成带水印的图像，还需要确保水印的隐蔽性，使其在视觉上难以察觉。然而，由于水印嵌入过程对图像局部特征的影响，带水印图像与原始图像的分布之间依然存在一定的偏差，使得鉴别器能够识别出水印的存在。

早期的水印嵌入模型通常采用较为简单的鉴别器，仅基于图像空间域特征进行判断，而忽略了图像的频率特性。尽管这些方法在视觉质量上取得了一定的提升，但仍然难以避免在某些频率范围内出现失真，尤其是在高频区域。例如，高频伪影可能会导致边缘过度锐化或局部纹理异常，使水印嵌入的痕迹在某些场景下变得明显。由于人眼对不同频率信息的敏感程度不同，单一空间域的鉴别器无法全面衡量水印嵌入后的视觉一致性，导致水印隐蔽性仍存在一定提升空间。

为了解决这一问题，并缩小带水印图像与原始载体图像在不同频率范围内的分布差异，本文设计了一种多频带鉴别器，使鉴别器能够在不同频域层面上优化水印嵌入效果。该方法通过在鉴别器中引入不同频段的约束，使模型能够调整水印的嵌入方式，减少不同频率范围内的伪影，提高水印的隐蔽性和视觉一致性。

具体而言，本文所提出的方法在鉴别器中加入了低通滤波器和高通滤波器，分别

对带水印图像进行滤波，并将得到的低频和高频分量分别输入鉴别器进行判别。低通滤波器用于保留图像的整体结构信息，确保水印嵌入不会在低频部分造成明显的失真；高通滤波器则用于捕捉图像中的细节与边缘信息，避免水印嵌入过程中高频伪影的产生。通过在低频与高频域分别施加约束，鉴别器能够更精确地评估水印嵌入对图像不同频率成分的影响，进而引导生成器在各频域层面进行有针对性的优化，最终在空间域与频域实现水印图像与原始载体图像的良好一致性。

这种多频带鉴别器的设计，使得模型能够在一定程度上抑制水印嵌入过程中产生的高频伪影，同时确保水印的隐蔽性和不可感知性。相比传统的单一空间域鉴别器，该方法能够更全面地优化水印嵌入效果，提升水印在不同频率范围内的均衡分布，使得带水印图像在视觉上更加自然，难以察觉嵌入痕迹。

3.3.3 噪声层

在水印嵌入模型中，图像可能在实际应用场景中经历各种噪声干扰和攻击，例如压缩、模糊、色彩变化，甚至是因拍摄或截屏带来的几何变换。为了提升水印的鲁棒性，使其在真实世界的传播环境下依然能够被正确提取，本文在编码器和解码器之间引入了噪声层，以模拟真实世界可能遇到的各种干扰，并增强模型对这些干扰的适应能力。

这一设计灵感来源于 StegaStamp[89] 的研究，该研究通过在训练过程中引入一系列基本的图像扰动，提高模型在实际应用中的稳健性。受此启发，本文在水印嵌入模型的编码-解码流程中添加噪声层，使其在特征映射阶段即承受各种噪声扰动，从而使模型具备更强的抗干扰能力。在这一过程中，噪声层会对编码器输出的特征进行扰动，使得解码器在接收到受干扰的特征后，仍然能够准确恢复水印信息，从而提升整体模型对真实世界噪声的适应性。

在图像处理中，攻击大致可以分为像素级攻击和图像级攻击两类。噪声层通过模拟这两类攻击，在训练过程中进行对抗性优化，使水印具备更强的抗干扰能力。

像素级攻击主要指在不改变图像整体结构与位置的前提下，对像素值进行调整，常见于数字图像处理过程中。此类攻击包括多种形式：模糊（如高斯模糊和运动模糊）用于模拟图像在存储或传输过程中可能出现的模糊现象，考察水印在此类场景下的稳定性；亮度、对比度变化用于模拟不同显示设备所引起的色彩偏差，以增强水印在多样色彩环境下的鲁棒性；噪声添加（如高斯噪声、泊松噪声及椒盐噪声）用

于模拟传输干扰或感光器件噪声，验证水印在噪声干扰下的可提取性；JPEG 压缩则用于模拟图像在社交媒体平台或存储过程中的有损压缩，评估水印在压缩退化场景下的鲁棒性。

为了增强模型对这些像素级攻击的适应能力，本文在训练过程中将这些攻击操作施加到带水印的图像上，并将其作为解码器的输入。这种方式迫使生成器学习更加鲁棒的特征，以确保即便水印图像遭受这些像素级变换，解码器仍然能够成功提取水印信息。

相比像素级攻击，图像级攻击主要针对图像的空间变换，可能由于拍摄、截屏等因素导致水印图像发生位置偏移或几何变换。现实世界中，水印图像的传播往往不仅限于数字处理，还可能经历物理世界的干扰，如用户通过拍摄或屏幕截图保存带水印图像，从而产生以下问题：截屏误差，屏幕截图过程中可能会产生像素级的错位，导致水印信息轻微变形；拍摄引起的几何变换，当用户使用相机拍摄带水印图像时，视角不同可能导致水印图像发生透视变形，影响水印的正确提取。

为了应对这类图像级攻击，本文采用单应性变换进行对抗训练。单应性变换可以近似模拟图像在屏幕上的投影关系，如视角变化、透视变换等。具体而言，假设水印图像的四个角点坐标发生随机扰动，则可以计算出一个的单应性变换矩阵 $\mathbf{H}$，其形式如下：

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{3.1}$$

该矩阵用于将原始图像映射到透视变换后的图像，使模型学习在真实场景中适应不同视角下的水印提取。在训练过程中，通过随机扰动图像的四个角点，计算单应性变换矩阵 $\mathbf{H}$ 并应用于图像，从而得到不同角度和透视变化下的带水印图像。解码器需要在这些变化后的失真图像中依然正确提取水印信息，这使得模型在实际应用中对拍摄、截屏等攻击时更加鲁棒。

综上，噪声层通过像素级攻击（如模糊、噪声、压缩等）和图像级攻击（如截屏误差、透视变换等）的模拟，使模型在训练过程中不断增强对不同扰动的适应性。通过引入这些噪声扰动，模型能够学习到更加稳定的水印嵌入方式，使水印信息在复杂的实际环境下仍能保持可提取性。最终，这一机制在一定程度上提高了水印的鲁

棒性，使其在不同存储、传输和传播过程中都能够有效地抵抗各类噪声干扰，确保水印信息的完整性和可提取性。

3.4 水印提取

3.4.1 解码器

在以往的研究中，更多的关注点集中在编码器的优化上[22,89-90]，而对解码器的设计往往较为简单，通常仅由若干卷积层堆叠而成。现有的水印嵌入网络中，U-Net结构因其能够较好地保留网络前端的信息特性，并在保持较高图像质量的同时进行水印嵌入，被广泛应用于水印生成任务。然而，相较于编码器的优化，解码器的网络结构往往未经过精细设计，因此，水印提取的效果存在进一步提升的空间。

受到光学字符识别（Optical Character Recognition, OCR）技术的启发，本文可以将水印提取任务类比于 OCR 任务。OCR 旨在从图像中提取文本信息，而水印提取的目标则是从带水印的图像中恢复出嵌入的信息。近年来，OCR 领域的发展趋势是采用更加复杂的网络结构，以增强对不同干扰条件下文本的识别能力，从而提升模型的鲁棒性。同样地，水印提取任务也面临着复杂的干扰环境，因此，本文在设计解码器时，借鉴了这些先进方法，并针对水印提取的特点，构建了一种类似编码器的网络结构，以提高水印信息的恢复能力。

具体而言，本文在解码器的设计中采用了 U-Net 结构，以增强解码器的水印提取能力，同时保持解码器与编码器的结构对齐。U-Net 结构中的跳跃连接能够使解码器在水印提取过程中更好地恢复原始水印信息，从而提升水印提取的精度和鲁棒性。经过训练后的解码器在水印提取任务上表现出更强的抗干扰能力，能够在复杂条件下依然保持较高的提取准确率，为提升水印信息的准确恢复提供了技术保障。

3.4.2 CNN-F 模块

水印提取阶段的关键在于如何从带水印图像中准确提取出水印信息。为此，本文引入了 CNN-F 模块，借鉴了 CycleGAN[91] 的双向映射思想，构建了一个能够在水印嵌入与提取阶段保持信息一致性的逆向映射系统。该设计不仅提升了水印嵌入的稳定性，也能够在一定程度上提升水印的隐蔽性。

在传统的 GAN 结构中，模型通常学习输入源域和目标域之间的映射关系，以完

成特定任务。在水印嵌入问题中，本文采用互相映射的策略，在载体图像和水印图像之间建立双向映射，使水印嵌入和恢复过程保持一致性。具体而言，生成器 **G** 负责从原始载体图像 I_{c} 生成水印图像 I_{w}，即：

$$I_{\mathrm{w}} = \mathbf{G}(I_{\mathrm{c}}) \tag{3.2}$$

这一过程完成了水印的嵌入，使得生成的 I_{w} 在视觉上接近 I_{c}，同时携带隐蔽的水印信息。这一过程确保水印能够隐蔽地嵌入到图像中，然而仅有的前向过程难以确保生成水印图像的质量。因此，为了进一步确保水印的隐蔽性，即能够尽可能多得保留原始载体图像的特征，在逆向恢复过程中，本文设计了 CNN-F 模块（记作 **R**），用于从 I_{w} 中恢复原始图像。其映射关系表示为：

$$\hat{I}_{\mathrm{c}} = \mathbf{R}(I_{\mathrm{w}}) \tag{3.3}$$

这一双向映射过程在带水印图像域与原始图像域之间建立了严格的对应关系，使水印嵌入与恢复过程保持一致。这样的机制不仅确保了模型能够生成高质量的带水印图像，还使得在恢复阶段能够有效重建原始载体图像的关键特征，从而进一步提升水印图像的视觉质量和水印的隐蔽性。此外，逆向恢复过程为模型提供了额外的监督信号，有助于优化训练目标，使模型在水印嵌入和提取任务上更加稳定，同时加速整体收敛，提高训练效率。

3.4.3 损失函数设计

为了最小化水印图像与原始载体图像之间的差距，同时降低解码出的水印信息与原始水印信息之间的比特误码率（Bit Error Rate, BER），本模型的损失函数由四个部分组成，分别为编码器部分、CNN-F 部分、鉴别器部分和解码器部分。各模块针对不同优化目标设计，以增强水印系统在隐蔽性、鲁棒性、可提取性与信息完整性等方面的综合表现。

首先，编码器部分的损失函数用于约束水印图像与原始载体图像之间的视觉差异，以提高嵌入水印的隐蔽性，使水印在视觉上难以察觉。为此，本文采用 $\mathcal{L}_1$ 损失函数以及感知损失（Learned Perceptual Image Patch Similarity, LPIPS）[92] 来计算原始载体图像 I_{c} 与水印图像 I_{w} 之间的差异。L_1 损失用于衡量像素级的误差，而 LPIPS 损失则基于深度神经网络对图像的感知特性进行衡量，使得水印图像在高层次的特

征空间内与原始载体图像的特征相似：

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_1(I_{\text{c}}, I_{\text{w}}) + \mathcal{L}_{\text{LPIPS}}(I_{\text{c}}, I_{\text{w}}) \tag{3.4}$$

其次，CNN-F 部分用于辅助恢复原始载体图像，以加速网络收敛并提高水印的鲁棒性。在训练过程中，本文期望从水印图像 I_{w} 中恢复出接近原始载体图像 I_{c} 的结果 $\hat{I}_{\text{c}}$，以确保水印的嵌入过程不会对图像造成过大的破坏。本文采用 $\mathcal{L}_1$ 损失函数对 $\hat{I}_{\text{c}}$ 和 I_{w} 之间的误差进行约束，从而优化网络的学习目标，使得恢复的原始载体图像更接近于原始载体图像：

$$\mathcal{L}_{\text{cnn-f}} = \mathcal{L}_1(\hat{I}_{\text{c}}, I_{\text{w}}) \tag{3.5}$$

鉴别器部分用于区分水印图像和原始载体图像的不同频带特征，从而提升水印的抗攻击能力。为此，本文采用二元交叉熵（Binary Cross-Entropy, BCE）损失函数来衡量鉴别器的分类能力。具体而言，鉴别器被训练为能够识别输入图像是否嵌入了水印，并通过损失约束使得水印图像在鉴别器看来与原始载体图像相似，从而提高水印的隐蔽性：

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{bce}}(I_{\text{c}}, I_{\text{w}}) \tag{3.6}$$

解码器部分的损失函数用于衡量解码出的水印信息 $\mathbf{M}_{\text{ext}}$ 与输入的水印信息 $\mathbf{M}_{\text{in}}$ 之间的误差，以确保水印信息的可提取性和完整性。由于水印信息的嵌入过程会受到噪声、图像压缩及其他失真因素的影响，因此解码器的目标是在这些干扰存在的情况下，仍然能够正确提取水印信息。本文采用均方误差（Mean Squared Error, MSE）损失函数来计算解码出的水印信息与原始水印信息之间的误差，使得解码出的水印尽可能准确：

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{MSE}}(\mathbf{M}_{\text{in}}, \mathbf{M}_{\text{ext}}) \tag{3.7}$$

最终，本文所提出的模型通过最小化整体损失函数 $\mathcal{L}_{\text{total}}$ 进行优化，整体损失函数由上述四部分损失函数的加权和构成。为了平衡不同损失项在训练过程中的影响，本文为每个部分设置了对应的权重系数 $\lambda_1, \lambda_2, \lambda_3, \lambda_4$，以确保网络训练过程中各个目

标能够协同优化：

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{enc}} + \lambda_2 \mathcal{L}_{\text{cnn-f}} + \lambda_3 \mathcal{L}_{\text{dis}} + \lambda_4 \mathcal{L}_{\text{dec}} \tag{3.8}$$

上述损失函数的设计综合考虑了水印的隐蔽性、鲁棒性、可恢复性及信息完整性，为水印系统提供了稳定的优化目标。在训练过程中，可以通过调整各损失项的权重系数，使得模型在不同应用场景下表现出最佳性能。例如，在对抗篡改攻击的应用场景中，可以通过适当提高 λ_3 的权重，从而增强鉴别器对水印嵌入的约束能力；而在追求水印高还原度的应用中，可以增加 λ_4 的权重，以减少水印信息的误码率。这种灵活的优化策略使得本模型可以适应不同的实际需求，提升水印技术的应用价值。

3.5 实验结果与分析

本节旨在评估所提出的抗高频伪影图像水印方法在真实图像数据及多种干扰环境下的鲁棒性与隐蔽性表现。为此，本文在多个数据集上开展实验，系统分析模型的性能优势，并与现有主流水印方法进行了定量与定性对比。此外，还通过消融实验对各关键组件在整体性能中的作用进行验证。

3.5.1 实验相关设置

为保证评估的客观性与全面性，本文选取 ALASKA[93] 数据集中的 10,000 张图像作为训练集，另取 10,000 张作为验证集，并在 COCO[94] 与 ImageNet[95] 上各随机选择 1,000 张图像用于测试。嵌入的水印信息为长度为 $L = 64$ 的二进制序列 $\mathbf{M} \in \{0,1\}^L$，随机采样生成。

为综合评估水印系统的性能，本文采用以下指标：

水印提取准确率指标（Accuracy，Acc）：衡量在各种攻击扰动后能否准确提取水印。其定义如下：

$$\text{Acc} = \frac{1}{L} \sum_{k=1}^{L} \left| \mathbf{M}_{\text{ext}}^{(k)} - \mathbf{M}_{\text{in}}^{(k)} \right| \times 100\% \tag{3.9}$$

其中，$\mathbf{M}_{\text{ext}}$ 与 $\mathbf{M}_{\text{in}}$ 分别表示提取与原始水印。

不可见性指标（Peak Signal-to-Noise Ratio，PSNR）：用于衡量水印嵌入后图像质

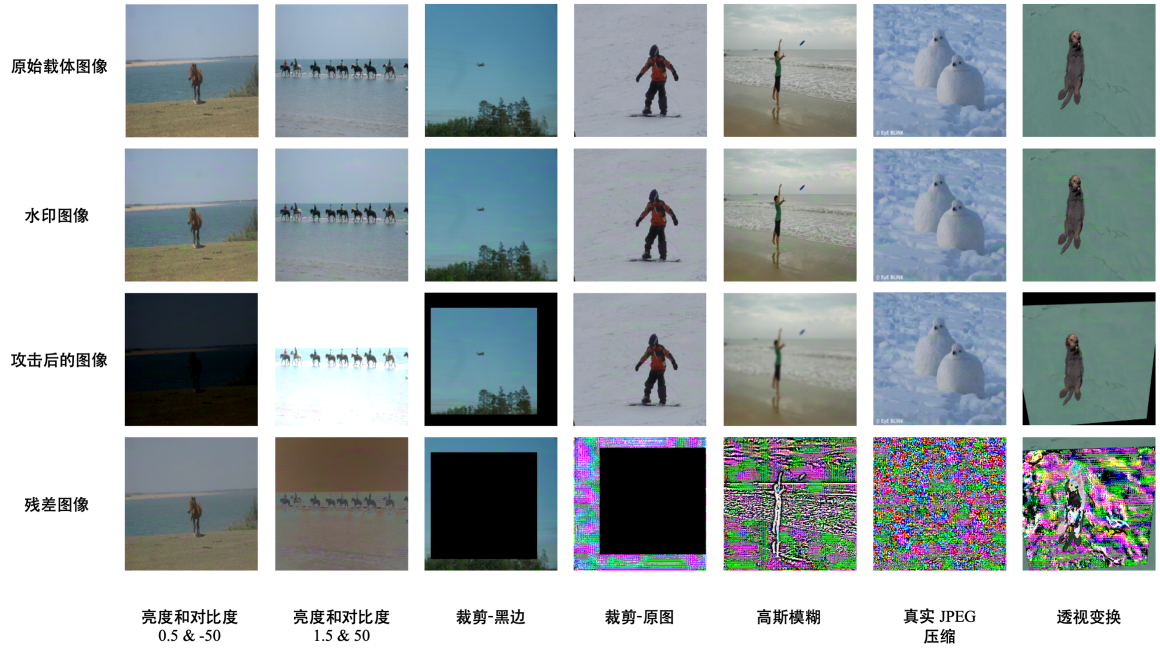

图 3.4 典型图像在不同扰动条件下的示意图

量的变化，单位为分贝（dB），定义如下：

$$\mathrm{PSNR}\left(I_{\mathrm{w}}, I_{\mathrm{c}}\right) = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}}{\sqrt{\mathrm{MSE}(I_{\mathrm{w}}, I_{\mathrm{c}})}}\right) \tag{3.10}$$

其中 I_{w} 与 I_{c} 分别为水印图像与原始图像，MAX 表示像素最大值。

3.5.2 实验结果与分析

为系统验证所提出水印方法在现实应用场景中的鲁棒性，本文在 COCO 与 ImageNet 两个主流图像数据集上设计并实施了一系列失真模拟实验，覆盖多种常见干扰类型，包括亮度与对比度调整、随机裁剪丢弃、高斯模糊、JPEG 压缩及透视变换等。这些扰动模拟了图像在编辑、传输和压缩等过程中的典型失真。

图 3.4 展示了典型图像在不同扰动条件下的视觉效果。第一行为原始图像，第二行为嵌入水印后的图像，第三行为遭受攻击后的图像，第四行为原图与嵌入图的残差图（$|I_{\mathrm{w}} - I_{\mathrm{c}}|$），直观反映了水印嵌入的隐蔽性及各类攻击对图像的影响。所采用的攻击方法包括对比度与亮度调整、裁剪、模糊、JPEG 压缩以及透视变换。

进一步地，表 3.1 总结了各类失真操作下的水印提取准确率（Acc）结果。从实验结果可见，所提出的方法在大多数干扰情境下均可实现稳定提取，尤其在高斯模糊与 JPEG 压缩等频域干扰中。同时，在面向裁剪、几何变换等空间结构扰动较强的场景中，所提出的方法仍保持 80% 以上的提取能力。上述结果验证了该方法良好的

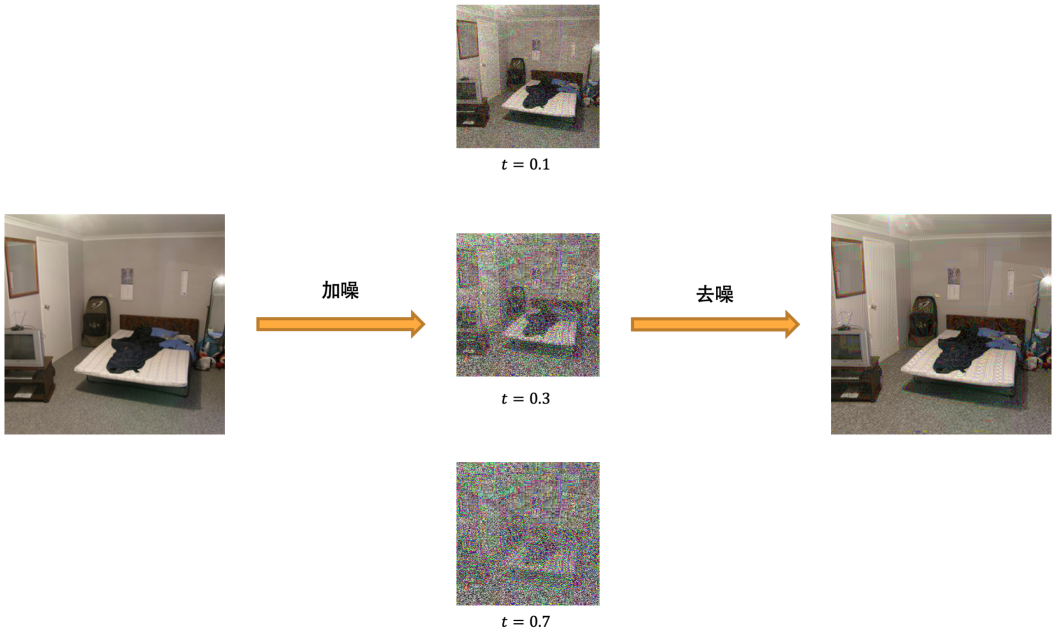

图 3.5　DiffPure 攻击流程图

抗干扰性与鲁棒性。

为进评估所提水印方法在极端对抗场景下的稳定性与安全性，本文还引入了两类具有高度破坏性的攻击策略：水印重写攻击与扩散去噪攻击（DiffPure）[96]。水印重写模拟真实应用中“二次嵌入”的情况，即在原始图像上使用另一种独立的传统水印算法（如基于 DWT-DCT 或 DWT-DCT-SVD）重新嵌入水印，从而扰乱原始水印的结构，意图削弱其可检测性与完整性。另一种攻击是基于扩散模型的图像重建方法 DiffPure[96]。其攻击流程如图 3.5 所示：攻击者首先在参数 t 控制下向图像注入高斯噪声，然后利用预训练的扩散去噪生成模型对其进行图像重建，试图在视觉恢复的同时清除潜在水印信号。这类方法通过图像的“再生成”能够擦除已有的水印信息，已被证明可有效破坏多种水印嵌入方式。尽管上述两类攻击具有较强的破坏性和现实威胁，实验结果表明，所提图像水印方法在该类场景下仍展现出良好的鲁棒性。这进一步验证了其在面对复杂数据重建攻击时，依然能够保持稳定且可靠的提取性能。

为全面评估所提出水印方法在不同失真条件下的综合性能，本文选取多个具有代表性的主流水印模型作为对比基线，包括 HiDDeN[22]、SSL[97]、HiNet[98]、StegaStamp[89] 以及 CIN[99]。在 COCO 和 ImageNet 两大公开图像数据集上构建统一测试环境，对比分析各方法在多种典型失真条件下的水印提取准确率与图像保真度表现。

表 3.2 汇总了各方法在 JPEG 压缩、高斯模糊、透视变换等常见攻击场景下的性能对比结果。可以观察到，本文提出的方法在大多数攻击类型下均表现出较高的水

表 3.1 所提出的方法在多种失真类型下的鲁棒性评估结果

攻击类型与参数		水印提取准确率	
攻击类型	参数说明	COCO	ImageNet
对比度与亮度	对比度因子 0.5，亮度降低 50	95.7%	95.3%
	对比度因子 0.5，亮度提升 50	99.9%	99.9%
	对比度因子 0.8，亮度降低 50	98.3%	98.0%
	对比度因子 0.8，亮度提升 50	99.9%	99.9%
	对比度因子 1.2，亮度降低 50	99.0%	98.5%
	对比度因子 1.2，亮度提升 50	96.0%	96.1%
	对比度因子 1.5，亮度降低 50	97.4%	97.0%
	对比度因子 1.5，亮度提升 50	91.7%	92.4%
裁剪（黑边）	保留 80%	90.3%	87.7%
	保留 90%	98.5%	98.4%
	保留 95%	99.7%	99.9%
裁剪（原图）	保留 80%	87.5%	90.6%
	保留 90%	97.1%	97.1%
	保留 95%	99.9%	99.9%
高斯模糊	核大小 5	99.9%	99.9%
	核大小 9	99.9%	99.9%
JPEG 压缩	质量因数 55	98.7%	98.5%
	质量因数 65	99.2%	99.0%
	质量因数 75	99.6%	99.4%
	质量因数 85	99.7%	99.7%
	质量因数 95	99.9%	99.9%
透视变换	变换因子 2	99.9%	99.9%
	变换因子 7	99.8%	99.8%
	变换因子 12	99.7%	99.6%
	变换因子 20	97.3%	97.2%
	变换因子 30	85.2%	84.8%
水印重写攻击	DWT-DCT 方法重写	99.9%	99.9%
	DWT-DCT-SVD 方法重写	99.9%	99.9%
DiffPure	时间步长 0.1	99.8%	99.8%
	时间步长 0.3	97.9%	98.2%
	时间步长 0.5	95.5%	96.5%

印提取准确率，特别是在透视变换等几何扰动下仍保持较高的水印提取准确率，优于其他对比方法。同时，在保真度方面，所提出方法的 PSNR 水平与现有主流模型相当，实现了不可见性与鲁棒性的有效平衡。值得一提的是，为验证方法在实际 AIGC 任务中的适应性，本文进一步在人工智能生成图像上（由 Prafulla 等人提出的高质量合成图像数据集[100]）进行了扩展实验。结果显示，该方法在生成图像上的嵌入与提取性能与自然图像保持一致，展现出良好的普适性与稳定性。这一发现进一步证实了所提水印框架在真实世界复杂应用场景中的可部署性与广泛适用性。

表 3.2 所提出的方法与现有方法在典型失真下的性能比较

数据集	方法	JPEG 质量因数 55	JPEG 质量因数 95	高斯模糊 核大小 5	高斯模糊 核大小 9	透视变换 变换因子 2	透视变换 变换因子 30	原图	PSNR (dB)
COCO	SSL	69.5%	95.3%	96.1%	94.3%	94.2%	81.0%	99.4%	41.9
	HiDDeN	56.5%	71.5%	81.2%	74.4%	87.1%	80.1%	89.0%	33.0
	StegaStamp	99.8%	99.9%	99.9%	99.8%	99.7%	76.7%	99.9%	28.4
	HiNet	56.1%	73.4%	53.2%	59.0%	46.8%	50.4%	**100%**	30.7
	CIN	**100%**	**100%**	99.9%	98.8%	70.6%	50.0%	100%	**42.0**
	Ours	98.7%	99.9%	**99.9%**	**99.9%**	**99.9%**	**85.2%**	99.9%	30.6
ImageNet	SSL	67.9%	94.0%	94.7%	92.5%	92.3%	70.3%	99.0%	**41.9**
	HiDDeN	57.6%	73.3%	81.0%	74.5%	86.7%	80.1%	88.0%	32.8
	StegaStamp	99.8%	99.8%	99.8%	99.7%	99.7%	76.7%	99.8%	28.2
	HiNet	55.7%	74.2%	53.0%	58.6%	48.1%	50.0%	**100%**	30.7
	CIN	**100%**	**100%**	99.9%	98.8%	70.5%	50.3%	100%	41.5
	Ours	98.5%	99.9%	**99.9%**	**99.8%**	**99.8%**	**84.8%**	99.9%	30.3
AIGC	Ours	98.6%	99.8%	99.9%	99.8%	99.8%	85.1%	99.9%	30.5

此外，为进一步验证本文方法在几何扰动条件下的稳定性与适应性，本文选取 CIN 模型[99] 作为代表方法进行对比实验。具体地，使用作者公开的代码在与本文相同的几何攻击参数下重新训练 CIN 模型，并对其生成的水印图像进行可视化分析。图 3.6 展示了该实验的结果，第一行为原始输入图像，第二行为重训练后嵌入水印的图像。可以明显观察到，CIN 模型在加入几何攻击后生成的图像在结构和纹理层面均出现较为严重的失真，存在明显的伪影问题，反映出其在鲁棒性增强下对图像质量与水印隐蔽性的降低。相比之下，本文提出的模型在相同攻击设置下仍能保持图像的整体结构清晰、纹理自然，且水印嵌入后对视觉质量影响较小。这表明所提出的方法保持了较好的视觉自然性和水印嵌入能力。

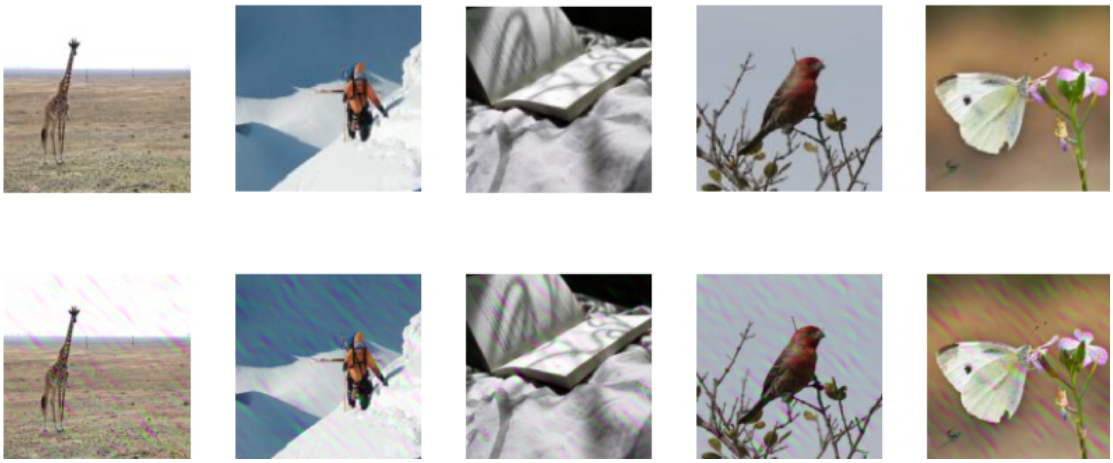

图 3.6　本文在加入几何变换扰动后重新训练 CIN 模型的可视化结果

3.5.3　消融实验

为进一步验证本文所提出各模块在提升水印系统性能方面的有效性，本节设计了两组消融实验，分别针对抗频谱混叠上采样层以及与编码器结构对齐的解码器进行定性分析。

首先通过实验对比了本文提出的抗频谱混叠上采样模块与传统转置卷积在图像质量上的表现差异。为消除对抗扰动因素对实验结果的干扰，本组实验在无攻击干预的设置下进行。图 3.7 所示为不同上采样策略在第 10 与第 30 个训练周期下生成的水印图像对比效果。其中，第一列为原始输入载体图像；第二与第三列分别为使用传统转置卷积在第 10 与第 30 个训练轮次下生成的结果；第四与第五列则为采用本文提出的抗频谱混叠上采样策略所得结果。

从图中可以明显观察到，由传统转置卷积作为上采样结构所生成的水印图像中存在明显的高频伪影，易对图像整体的自然感知造成干扰。相比之下，引入抗混叠上采样模块后，图像在结构清晰度与纹理自然度方面均得到改善，在一定程度上有效抑制了高频伪影的出现。该结果充分验证了所提方法在提升水印图像感知质量与嵌入隐蔽性方面的有效性，为后续模型鲁棒性与可用性奠定了基础。

为深入评估解码器结构对水印提取性能的影响，本文设计了对比实验，比较基于传统堆叠式卷积神经网络（Convolutional Neural Network，CNN）解码器与本文所提出的编码器对齐式解码器在不同扰动环境下的表现。实验过程中，逐步引入增强的对抗攻击以模拟真实场景中可能出现的噪声干扰，从而检验两种结构在水印提取性能方面的差异。

图 3.8 展示了不同解码器结构下的水印提取可视化结果。图中每列对应一种解

图 3.7　不同上采样方式下生成图像的视觉对比图

码器结构，左侧为传统 CNN 堆叠结构，右侧为编码器对齐的结构；第一行为在无噪声扰动条件下的提取结果，第二行为在加入对抗噪声后的提取效果。从图像可视化结果可见，传统 CNN 堆叠结构在面对对抗扰动时表现出明显的不稳定性，水印图案出现模糊、变形甚至信息丢失，难以实现可靠的提取。而编码器对齐式解码器在同等干扰条件下则能保持良好的提取性能，重建出的水印图案清晰完整，结构信息基本未受破坏。该结果充分表明，编码器结构对齐策略能够提升解码器对复杂扰动的适应能力，在保证提取准确性的同时增强整体系统的鲁棒性与实用性。

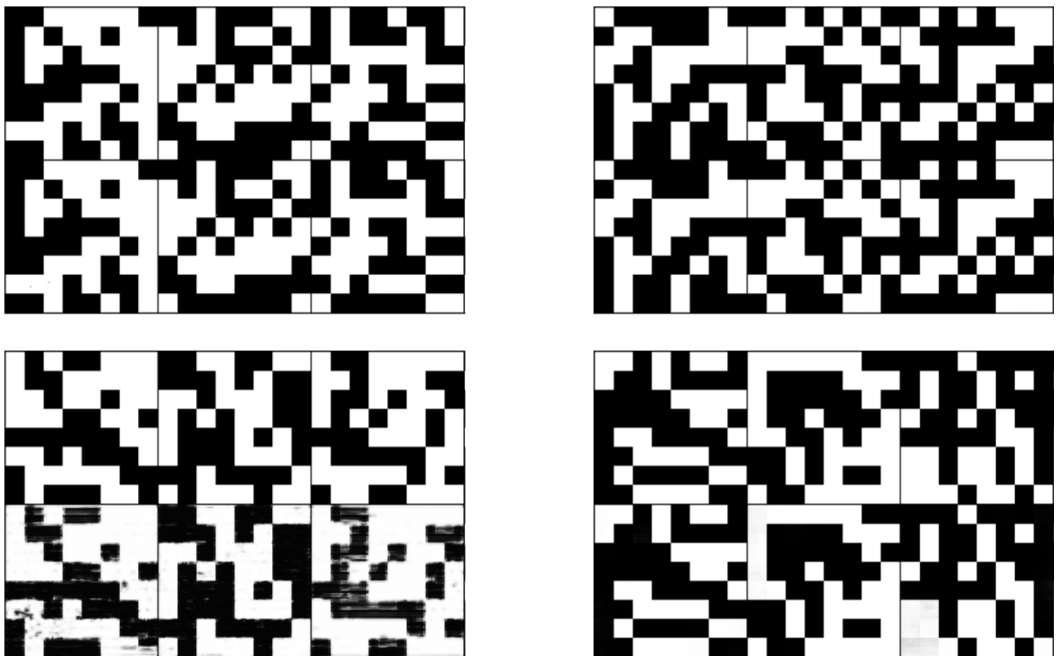

图 3.8　不同解码器结构下的水印提取效果对比图

综上所述，消融实验充分验证了本文在模型结构设计方面的改进对于提升图像质量、水印隐蔽性及提取鲁棒性所起到的关键作用。这些模块在构建端到端可训练的水印系统中发挥了互补协同的效果，是实现鲁棒图像水印的核心组成部分。

3.6 本章小结

本章围绕图像生成场景中的高保真鲁棒水印嵌入问题，提出了一种面向频域特性优化的外生水印方法。针对生成图像中普遍存在的高频伪影问题，本文设计了一种融合频谱感知机制的水印嵌入框架，通过抗频谱混叠上采样模块、多频段判别器和结构对齐解码器等核心组件，在提升水印鲁棒性的同时，在一定程度上缓解了高频伪影的产生，保障了图像的感知质量。

在实验部分，本文从多个维度系统评估了所提方法的有效性。在 COCO 和 ImageNet 两个通用图像数据集上进行的评估表明，本文所提出的水印方法具有良好的隐蔽性，同时在面对压缩、模糊、裁剪、几何变换等多种常见攻击下，水印提取准确率保持在较高水平，展现出较好的鲁棒性。与现有主流方法对比实验进一步验证了本方法在几何攻击下的优势。此外，通过一系列消融实验验证了各核心模块的有效性，表明所设计的结构在提升图像质量与增强水印稳定性方面具有积极作用。

第四章　面向大语言模型生成文本的水印技术

本章针对大语言模型（Large Language Models, LLMs）生成文本的可识别性与可追溯性问题，研究面向生成内容的水印技术。相较于传统外生水印方法存在的语义干扰与鲁棒性不足问题，本文提出一种基于参数扰动的内生水印技术框架，从模型层面实现对输出文本的隐式标记。该方法通过在模型参数中注入可控扰动，使生成文本在语义失真较小的同时携带特定水印信息，兼顾隐蔽性、鲁棒性与文本保真性。通过大量实验验证，该方法在面对删除、插入、修改等攻击时表现出较好的鲁棒性，适用于内容判别、模型版权保护等场景，为生成文本的可信管理提供技术支持。

4.1　引言

在大语言模型生成内容广泛应用的背景下，如何实现对生成文本的标识与溯源已成为亟待解决的关键问题。作为核心技术路径之一，文本水印需在保障语义连贯性与语言质量的同时，具备良好的隐蔽性与鲁棒性。相较于具备较高冗余度的图像数据，文本结构高度离散且语义敏感，微小改动即可能破坏语义逻辑或被用户察觉，从而使水印嵌入面临更严峻的挑战。

本章提出了一种基于大语言模型参数扰动的文本水印框架。该方法通过在模型的关键权重中注入结构化且稀疏分布的微小扰动，在对生成质量和语义一致性影响较小的前提下，实现了水印信息的隐蔽嵌入与稳定提取。该研究为构建性能可控、隐蔽性强的大语言模型生成文本水印系统奠定了技术基础。

4.2　相关技术基础

为了实现面向大语言模型的高保真水印嵌入与提取机制，有必要首先回顾大语言模型的基本结构。在本章所提出的方法中，水印信息以微弱扰动的方式注入模型参数，在尽可能对语言能力影响较小的前提下，实现对生成内容的稳定且隐蔽的标记。因此，深入理解语言模型的架构特性，是本章方法设计的基础。

大语言模型通常采用 Transformer 架构，已成为当前自然语言生成任务中的主流模型形式。其基本结构如图 4.1 所示，包括输入嵌入层、若干层堆叠的 Transformer 模

块以及最终的输出线性层。模型首先将离散的文本 token 映射为嵌入向量表示，经多层 Transformer 逐步提取上下文语义特征，最终通过输出层生成词表上每个 token 的概率分布，从而实现对下一词的预测与语言序列的生成。

图 4.1 大语言模型的基本结构图

在输入阶段，原始文本首先被映射为离散的 token 序列 $t_1, t_2, \ldots, t_n$，并转化为嵌入向量 $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$，即：

$$\mathbf{e}_i = \text{Embedding}(t_i), \quad i = 1, \ldots, n \tag{4.1}$$

随后，这些嵌入向量 $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ 经过若干层 Transformer 块进行后续的处理。具体来说，模型利用自注意力机制（Self Attention）与前馈网络（Feed-Forward Networks, FFN）捕捉上下文之间的相互依赖关系，生成一系列上下文感知的表示向量 $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n$。对于每一层 Transformer 模块，其结构可形式化描述为：

$$\mathbf{h}_i = \text{FFN}\left(\text{Self Attention}(\mathbf{h}_i)\right) \tag{4.2}$$

最终，在输出阶段，模型将最后一层的隐藏状态 $\mathbf{h}_n$ 通过线性变换映射为与词表大小一致的 logits 向量 $\mathbf{l}$：

$$\mathbf{l} = \mathbf{W} \cdot \mathbf{h}_n + \mathbf{b} \tag{4.3}$$

其中，$\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$ 为输出层的权重矩阵，$|\mathcal{V}|$ 表示词表大小，d 为隐藏状态维度，$\mathbf{b}$ 为偏置项。

随后，logits 向量 $\mathbf{l}$ 经由 softmax 归一化处理，得到每个候选 token 的生成概率分布 $\mathbf{p}$：

$$\mathbf{p} = \text{softmax}(\mathbf{l}) \tag{4.4}$$

在自回归生成框架中，模型按顺序生成每一个 token，当前时间步 i 的预测仅依赖于前 $i-1$ 个 token，通过最大化条件概率 $P(t_i | t_1, \ldots, t_{i-1})$ 实现序列的逐步生成。

上述权重矩阵 **W** 和偏置项 **b** 在每个时间步共享，用于生成下一个词的概率分布，是模型输出线性层的核心参数。

4.3 水印嵌入

为在不改变模型主体结构的前提下，在生成文本中嵌入可检测的水印信号，本文提出了一种基于输出线性层扰动的轻量级文本水印方法。该方法仅对语言模型输出层中的线性变换参数施加微小扰动，通过略微提升词表中部分 token 对应的投影向量权重，从而在采样过程中提高其被选中的概率，实现在对语义干扰较小的情况下嵌入水印信息。由于扰动仅作用于输出层，整个嵌入过程无需重新训练模型，对生成质量与语义一致性影响较小，具备良好的保真性、实用性和鲁棒性。基于上述设计理念，以下将系统介绍该方法的核心技术框架，包括水印嵌入的总体思路及具体实现机制。

4.3.1 水印嵌入方法概述

大语言模型的输出通常通过一层线性变换将隐藏状态映射到词表维度的 logits 分布，随后经 softmax 得到最终的生成概率。设隐藏状态为 $\mathbf{h} \in \mathbb{R}^d$，词表大小为 $|\mathcal{V}|$，输出层的权重矩阵为 $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$，偏置为 $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$，则输出 token 的 logits 表达式为：

$$\mathbf{l} = \mathbf{W} \cdot \mathbf{h} + \mathbf{b} \tag{4.5}$$

在以 Transformer 架构为基础的大语言模型中，Transformer 块通常由多层自注意力机制与前馈网络堆叠而成，核心功能在于捕捉输入序列中的上下文依赖关系与语义结构。经过多层变换后得到的隐藏状态 **h**，已高度融合了序列级的语言特征和潜在语义信息。相比之下，输出线性层仅在模型末端起到映射作用，将语义空间中的表示 **h** 投影至词表维度，生成对应的 logits 分布以用于词的采样。由于其仅承担语义向分布空间的转换职责，与主干 Transformer 模块在建模目标上相对独立，因此对该层参数进行轻微扰动，对模型整体的语言建模能力影响较小，为在保持语义流畅性的前提下嵌入外部信息提供了可行性基础。在此基础上，若将水印嵌入设计为直接修改 Transformer 主体结构，往往会破坏模型原有的语义构建机制，影响文本生成质量。相比之下，输出线性层因其参数规模有限、功能相对独立，成为更适合引入细粒度干预的部位。此外，从安全性角度看，该层的局部性特征也增强了水印的隐蔽性。

通过对投影矩阵中部分向量的有选择性扰动，可有效降低水印嵌入被察觉或逆向破解的风险，为实现轻量、鲁棒且不可感知的文本水印提供了结构层面的支撑。

基于上述分析，本文提出一种通过对输出层权重矩阵 $\mathbf{W}$ 中部分向量引入微小扰动以实现水印嵌入的方法。具体地，选取一组目标 token 集合 $\mathcal{T} \subset \mathcal{V}$，并对其对应的投影向量 $\mathbf{w}_t$ 施加缩放操作，即乘以一个略大于 1 的缩放因子 $\alpha > 1$。该操作在对模型生成质量和语义一致性影响较小的前提下进行，提升目标 token 在采样过程中的被选中概率，从而实现隐式的水印信息植入。具体的数学表达如下：

$$\mathbf{w}_t^{\text{wm}} = \alpha \cdot \mathbf{w}_t, \quad \forall t \in \mathcal{T} \tag{4.6}$$

由于这些 token 的 logits 得分在 softmax 后具有更高的采样概率，生成文本中出现的频率将相对提升，从而可作为水印信号。这一修改过程只需一次性调整输出层的部分参数，无需训练或反向传播，具备高度的嵌入效率与部署灵活性。同时，为防止攻击者直接推测扰动模式，token 子集 $\mathcal{T}$ 的选择采用伪随机函数生成，结合嵌入密钥进行控制，增强水印的安全性与抗移除能力。

在生成文本后，检测器通过统计 $\mathcal{T}$ 中 token 在文本中的出现频率，并使用 z-score 进行统计显著性分析进行检测，即可判断该段文本是否包含水印。该策略可应用于内容追踪等任务场景，具有较强的可扩展性。

4.3.2 输出线性层结构与扰动机制

在当前主流的大语言模型中，输出层通常由一组参数化的线性变换构成，其功能是将 Transformer 模块输出的隐藏状态 $\mathbf{h} \in \mathbb{R}^d$ 映射至词表维度的 logits 分数空间。具体而言，输出层包含权重矩阵 $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$ 与偏置向量 $\mathbf{b} \in \mathbb{R}^{|\mathcal{V}|}$，生成的 logits 表达式如公式（ 4.5）所示。在此基础上，通过 softmax 函数可获得 token 对应的概率分布：

$$p_i = \frac{e^{l_i}}{\sum_{j=1}^{|\mathcal{V}|} e^{l_j}}, \quad t_i \in \mathcal{V} \tag{4.7}$$

本文提出的水印嵌入方法正是基于该输出层结构进行设计：在冻结模型主体参数（包括 Transformer 块与嵌入层）的前提下，仅对输出层 $\mathbf{W}$ 中部分目标 token 对应的列向量进行微小扰动，以最小代价实现水印信号的注入。具体而言，设定一个由密钥 key 控制的伪随机函数 $f_k(\text{key})$，用于从词表中选取一组目标 token 构造 $\mathcal{G}$，其对应的向量乘以微扰因子 α。其中，$f_k(\text{key})$ 是由密钥 key 控制的伪随机函数，用于决

定哪些 token 作为水印的承载元素，确保水印嵌入的稀疏性与不可预测性。α 为扰动因子，确保在 softmax 的指数放大作用下，目标 token 的采样概率获得可控提升，但不会对语言模型整体输出质量造成较大影响。

通过这种扰动，水印 token 在语义合理的上下文中更可能被采样，生成文本在统计层面隐含偏向性。这种微弱偏移无法被肉眼识别，且不会影响生成文本的语义连贯性、可读性与上下文一致性。更重要的是，由于模型主体参数保持不变，原有任务性能（如问答、摘要等）影响较小，具备良好的保真性与实用性。

此外，该方法具有良好的可扩展性。在不同模型架构、不同大小的语言模型上均可采用统一策略，仅需对输出层进行定向操作，便可实现模型级别的水印标识，满足在大规模模型部署场景下对各模型副本进行归属识别与认证的需求。

4.3.3 水印嵌入的具体流程

为实现结构最小化修改和部署便捷性，本文所提出的水印嵌入方法仅作用于语言模型的输出层参数，即输出 logits 前的线性映射权重矩阵 $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$，而不修改模型主体结构。其基本思路为：在保持 $\mathbf{W}$ 主体不变的前提下，选取其中一部分 token 对应的列向量 $\mathbf{W}[j]$，施加一个微小的线性扰动因子 $\alpha > 1$，从而提升对应 token 在 softmax 输出中的采样概率，并实现水印信号的隐式注入。

为提升水印的不可预测性与安全性，本文引入密钥控制的伪随机函数 $f_k(\text{key})$，用于决定嵌入 token 的集合 $\mathcal{G}$，即：

$$\mathcal{G} = \text{TokenSelect}\{\mathcal{V}, f_k(\text{key}), \gamma\}, \quad \text{and} \quad |\mathcal{G}| = \gamma|\mathcal{V}| \tag{4.8}$$

其中 $\gamma \in (0,1)$ 控制嵌入的稀疏性。对 $\mathcal{G}$ 中的每一个 token，对应列向量 $\mathbf{W}[j]$ 乘以扰动系数 α，实现如下变换：

$$\mathbf{W}[j] \leftarrow \alpha \cdot \mathbf{W}[j], \quad \forall j \in \text{Index}(\mathcal{G}) \tag{4.9}$$

在 softmax 层的指数放大作用下，这些目标 token 在采样过程中被选中的概率将相应提升，从而在不破坏语义自然性与输出连贯性的前提下，在统计分布层面显性偏向某些水印 token。扰动强度 α 的设置需在隐蔽性与检测性之间权衡。

该嵌入过程无需额外训练或梯度更新，仅通过一次性地调整输出层权重矩阵中的部分列向量即可完成水印的注入，操作简便高效。该策略不依赖模型架构细节，适

用于多种规模和类型的大语言模型，具备良好的通用性与可移植性。此外，由于改动范围受控，对模型原有功能和性能影响较小，便于在实际部署中灵活集成。其整体实现流程如算法 4.1 所示，输入为输出层权重矩阵 $\mathbf{W}$、扰动因子 α、嵌入比例 γ 以及密钥 key，输出为嵌入水印后的权重矩阵 $\mathbf{W}_{\text{wm}}$。

算法 4.1: 输出线性层水印嵌入算法

输入: 输出层权重矩阵 $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times d}$;

扰动因子 $\alpha > 1$；嵌入比例 γ；密钥 key

输出: 嵌入水印后的权重矩阵 $\mathbf{W}_{\text{wm}}$

初始化伪随机函数 $f_k(\text{key})$ // 由密钥 key 控制

初始化绿色列表 $\mathcal{G} \leftarrow \text{TokenSelect}(\mathcal{V}, f_k(\text{key}), \gamma)$ // 构造绿色词表，且 $|\mathcal{G}| = \gamma|\mathcal{V}|$

for 每个 *token* 索引 $j \in \mathit{Index}(\mathcal{G})$ **do**

  $\mathbf{W}[j] \leftarrow \alpha \cdot \mathbf{W}[j]$

end

return $\mathbf{W}_{\text{wm}} \leftarrow \mathbf{W}$

4.4 水印检测

在完成语言模型水印嵌入后，如何从模型生成的文本中准确且稳定地检测出水印信号，是实现内容可追溯与归属验证的核心技术环节。该检测任务不仅需具备良好的鲁棒性与低误报率，还应适配在线部署与大规模自动处理场景。为此，本文采用基于统计显著性判别的检测框架，通过分析生成文本中绿色列表 token 的频率分布是否偏离无水印情况下的统计期望，从而判断文本是否来源于已被水印化的模型。

水印嵌入阶段通过对输出层中绿色列表 $\mathcal{G}$ 中 token 对应的权重施加微小放大因子，从而使其在 softmax 后具有略高的选择概率。这种扰动虽然在单个 token 级别不可感知，但会导致绿色 token 在整段文本中的出现频率升高，为后续检测提供可测量的信号。检测阶段则以此为依据，统计生成文本中绿色 token 的出现次数，并将其与在无水印假设下的理论分布进行比较，以判断偏差是否具有统计显著性。

设词表大小为 $|\mathcal{V}|$，水印密钥控制下的绿色列表比例为 γ，待检测文本长度为 T，其中绿色 token 的实际出现次数为 X。在原假设 H_0（即待检测文本不含水印）成立的情况下，X 服从参数为 (T, γ) 的二项分布。由于文本长度 T 通常较大，根据中心极限定理，X 可以近似建模为正态分布：

$$X \sim \mathcal{N}(T\gamma,\, T\gamma(1-\gamma)) \tag{4.10}$$

由此可得标准化的统计检测量（z-score）：

$$z = \frac{X - T\gamma}{\sqrt{T\gamma(1-\gamma)}} \tag{4.11}$$

该 z-score 反映了绿色 token 出现频次相对于无水印分布的偏离程度。由于水印的嵌入策略设计为仅提升绿色 token 的出现频率。若计算所得 z-score 超过某一显著性阈值 z_α，则拒绝原假设 H_0，认为该文本含有水印。该方法具有无需访问原始模型、无需比对未加水印样本的优点，适用于在线部署与大规模检测任务。

为清晰呈现该过程，本文形式化描述了整个检测流程，如算法 4.2 所示，输入待检测文本序列 s、水印密钥 key、绿色列表比例 γ 以及显著性阈值 z_α，输出判断是否检测到水印。该算法首先构造绿色列表 $\mathcal{G}$，然后统计文本中绿色 token 的出现次数 X，最后计算 z-score 并进行显著性检验。

4.5 实验结果与分析

为全面评估本文所提出的大语言模型内生水印方法在实际生成任务中的性能表现，本文从保真性、鲁棒性与检测准确性三个核心维度开展系统实验。首先，通过设置不同的扰动强度与扰动比例，对水印嵌入策略进行参数化测试，从而评估其在生成质量与水印可检测性方面的影响；同时选取不同主流语言模型版本进行横向对比，探讨模型结构差异对水印嵌入效果的适应性与稳定性。此外，为进一步验证方法的实用性与性能优势，实验引入多种具有代表性的现有文本水印方法，并在统一评估框架下，对各方法在不同模型与场景中的表现进行对照分析。实验结果将围绕语义保真、抗干扰能力及水印检测准确率等关键指标展开，为本文方法在实际应用中的可行性与综合优势提供实证支撑。

算法 4.2: 基于 z-score 的水印检测算法

输入: 待检测文本序列 $s = t_1, t_2, \ldots, t_T$；水印密钥 key；绿色列表比例 γ；显著性阈值 z_α

输出: 是否检测到水印（True/False）

初始化绿色列表 $\mathcal{G} \leftarrow \text{TokenSelect}(\mathcal{V}, \text{key}, \gamma)$ // 构造绿色词表

$X \leftarrow 0$ // 绿色 token 计数器

foreach $t_i \in s$ **do**

  if $t_i \in \mathcal{G}$ **then**

    $X \leftarrow X + 1$;

  end

end

计算 z-score 统计量：

$$z \leftarrow \frac{X - \gamma T}{\sqrt{T\gamma(1-\gamma)}}$$

return $z > z_\alpha$

4.5.1 实验相关设置

本实验选用 C4（Colossal Clean Crawled Corpus）英文语料库[101]作为数据基础，从中随机抽取 400 条长度适中的语句构建验证集，以兼顾语言多样性、自然分布和统计效率。模型方面，选用 Meta 发布的 LLaMA3 系列，包括基础版本 LLaMA3-8B 以及其指令微调版本 LLaMA3-8B-Instruct（以下简称 LLaMA3-8B-it），用于测试方法在不同模型配置下的通用性。

在评价指标方面，本文采用两项核心指标对水印嵌入效果进行定量分析。首先，困惑度（Perplexity, PPL）用于衡量语言模型生成文本的自然程度与语言质量。PPL 值越低，表示模型生成的文本更符合语言模型的预测分布，即语言流畅性与语义合理性越高。因此，PPL 是评估水印嵌入是否干扰语言生成性能的重要依据。较低的 PPL 值表明水印嵌入对模型输出的影响较小，体现出该水印方案在保持语言生成保真性方面的优势。

其次，本文引入 z-score 统计量用于衡量水印信号在生成文本中的显著性。该指

标反映绿色列表中 token 的实际出现频率相对于其理论期望的偏离程度。z-score 值越高，说明偏离越显著，即水印信号越强，从而提升检测的置信度与可靠性。作为一种标准化的显著性指标，z-score 为水印的可验证性提供了直观且量化的分析依据。

为便于水印检测过程中阈值的设定与检测性能的量化分析，表 4.1 给出了若干常用 z-score 值及其对应的 p-value 与置信度。其中，p-value 表示在原假设成立（即无水印存在）的前提下，观察到当前 z-score 或更极端统计量的概率，而置信度反映了当前检测结果为“有水印”的可信程度。该对照表可为实际应用中水印检测阈值的选择提供直观依据，支持在不同应用需求下灵活设定。

表 4.1 z-score 与 p-value 及置信度对照表

z-score	p-value	置信度（%）
1.64	0.050	95.0
2.33	0.010	99.0
2.58	0.005	99.5
3.09	0.001	99.9
3.72	0.0001	99.99
4.00	3×10^{-5}	99.997

4.5.2 实验结果与分析

本节从可验证性、鲁棒性与效率三方面全面评估所提出水印方法的性能表现。首先，为验证整体可检测能力，选取目前主流的四种典型文本水印算法作为对比基线，分别与本文方法在 LLaMA3-8B 与 LLaMA3-8B-it 两个不同模型上进行横向对比。所有实验均在统一硬件环境下完成，验证语料来自 C4 数据集中随机抽取的 400 条样本。评估指标包括水印检测显著性（以 z-score 衡量）与语言困惑度（PPL），以检验水印是否对生成质量构成干扰。

随后，为评估该方法在真实应用场景中的鲁棒性，进一步设计了删除、掩码和插入三类典型文本扰动操作，并在不同攻击强度下测试各算法的抗干扰能力。同时，为分析方法在实际部署中是否具备工程可用性，本文还测量了各方法在相同推理配置下的平均生成时延，评估其对服务性能的影响程度。

为验证所提出水印嵌入方法在实际生成任务中的可行性与有效性，表 4.2 展示了多个具有代表性的生成示例。每组样例均给出了相同输入 prompt 在无水印（NW）

与嵌入水印（W）条件下模型的输出结果，并同时报告了对应的水印检测 z-score 以及生成文本的困惑度（PPL）指标。

从表中结果可以看出，嵌入水印后的生成文本在语义合理性、语言自然性方面均未出现较大的退化，内容质量与无水印版本保持相对一致。与此同时，所有嵌入水印后文本的 z-score 均有提升，表明水印信号已成功注入且可被准确检测。在不同设置下，PPL 指标基本保持稳定，进一步表明本文方法在较小影响生成性能的前提下，能够有效保持输出文本的语言质量。综合而言，该方法具备良好的实用性、可控性与通用性，适用于现有主流语言模型的水印部署需求。

表 4.2 若干典型输入下的语言模型水印嵌入效果示例（NW 表示未嵌入水印，W 表示嵌入水印）

输入提示	无水印的续写（NW）	嵌入水印的续写（W）	z-score（NW/W）	PPL（NW/W）
... One of the key roles of the human resources department is to keep the workforce	well trained and motivated. Training is one of the main ways to keep employees motivated and productive ...	happy and engaged, which helps increase productivity and, ultimately, profits ...	-2.01/2.47	3.82/4.06
... the largest casino and hotel operator on the Las Vegas Strip, is looking for a rebound in the famed	U.S. gambling hub after the coronavirus pandemic forced it to temporarily close its resorts in March ...	gambling destination as it tries to recover from a weak second quarter, hurt by lower gambling revenue and a drop in room rates ...	-1.98/3.52	2.46/2.98

表 4.3 汇总了各水印方法在无攻击干扰条件下的语言模型困惑度（PPL）与水印检测显著性（z-score）表现，旨在全面评估各方法在保证生成质量前提下的可验证能力。对比方法涵盖四种具有代表性的文本水印策略，包括 KGW[35]、SWEET[102]、EWD[103] 以及 Google 提出的 SynthID[104]。为确保公平性与可比性，所有对比方法均在相同的语言模型架构，即 LLaMA3-8B 上进行部署与测试，并使用统一的评估语料与环境配置进行实验。

从实验数据中可以观察到，本文方法在 LLaMA3-8B 与 LLaMA3-8B-it 两个模型

实例上均表现出良好的水印可检测性，所对应的 z-score 值稳定且处于较高水平。其中，LLaMA3-8B 模型下的 z-score 达到 6.20，已超过常用显著性检测阈值，充分说明嵌入的水印信号在统计意义上具有可辨识性。虽然在部分设置下，本文方法的 z-score 略低于部分对比方法，但在语言模型的困惑度方面，其 PPL 优于其他对比方法，说明水印对模型生成质量的影响较小。

这一平衡特性充分体现了本文方法“高可验证性、低干扰性”的核心优势，在实现有效水印嵌入的同时，一定程度地保留了语言生成的自然性与连贯性，能够缓解现有方法中常见的语义偏移与内容失真等问题。因此，本文方法在保障检测性能的同时，具备较好的实用性，尤其适用于对语言质量要求较高的实际应用场景。

表 4.3 不同水印算法的 PPL 和 z-score 检测对比

方法	PPL ↓	z-score ↑
无水印	3.82	-1.42
KGW	5.08	7.34
SWEET	5.05	8.20
EWD	5.03	8.50
SynthID	5.07	2.55
Ours(LLaMA3-8B)	4.72	6.20
Ours(LLaMA3-8B-it)	5.23	5.4

表 4.4 展示了在 LLaMA3-8B 模型下，本文方法与典型水印方法在面对三类常见文本扰动——随机掩码（Masked, M）、随机删除（Deleted, D）与随机插入（Inserted, I）——时的性能表现，涵盖语言模型困惑度（PPL）与水印检测显著性（z-score）两项指标。其中，PPL 用于衡量扰动后文本的语言自然性与语义连贯性，数值越高表示文本在语言模型眼中的“困惑度”越强，即与正常语言分布偏离越远，反映出更高程度的语义失真；而 z-score 用于评估水印信号在扰动条件下的显著性，数值越高说明水印依然具有较强的可检测性。

观察表中数据可以发现，本文提出的方法在大多数扰动场景下依然维持较高的 z-score，表明即便生成文本经历一定程度的结构变动，水印信号仍具备较强的可检测性。这一结果充分体现出水印算法对常见文本编辑操作的良好适应能力，为水印技术在开放式应用环境中的稳定部署与可靠使用提供了技术支撑。

在实际应用场景中，语言模型的水印方案不仅应具备良好的可检测性与鲁棒性，

表 4.4 不同扰动场景下多种水印方法的鲁棒性对比结果，以 z-score 和 PPL 为评估指标

扰动类型	KGW		SWEET		EWD		SynthID		Ours	
	z-score	PPL	z-score	PPL	z-score	PPL	z-score	PPL	z-score	PPL
原始文本	8.01	5.49	8.50	5.52	8.63	5.57	2.56	5.87	6.52	5.27
M (10%)	3.88	7.87	6.00	7.89	6.20	7.78	2.53	8.14	6.50	5.82
D (10%)	6.57	11.33	6.33	12.03	6.77	11.45	2.54	11.37	6.11	8.63
I (10%)	6.63	6.61	6.17	6.62	6.39	6.69	2.53	6.78	6.76	4.90
M (30%)	-3.16	7.83	4.61	8.23	4.52	7.29	1.51	7.29	6.53	5.78
D (30%)	4.45	37.10	4.29	37.96	4.05	36.33	1.52	37.21	5.41	27.6
I (30%)	4.83	7.94	5.01	7.73	5.42	7.51	1.52	7.96	6.76	7.90
M (50%)	-9.43	7.49	5.32	8.44	5.64	8.65	0.50	8.28	6.66	5.90
D (50%)	3.27	110.03	2.61	113.65	2.90	118.48	0.50	107.85	5.76	88.90
I (50%)	3.13	8.02	5.09	8.83	5.05	8.85	0.51	8.08	4.76	8.90

还应避免对推理效率造成负面影响。特别是在大语言模型逐步部署于低延迟场景的背景下，推理延迟成为系统可用性的关键约束。因此，水印方案在保持有效性的前提下对原模型推理效率的影响，是衡量其实用性与可扩展性的关键因素之一。

为此，本文评估了所提出方法与多种代表性无训练水印策略在相同硬件环境与数据设置下的推理时延表现，结果如图 4.2 所示。图中所有方法的推理耗时均已归一化为相对值，其中未嵌入水印的原始模型设为基准值 1。可以观察到，本文方法在 LLaMA3-8B 与 LLaMA3-8B-it 两个模型版本上均未引入明显的额外计算成本，推理耗时与原始模型接近一致。

这一优势得益于本方法的设计策略：仅在输出层静态地扰动部分权重参数，无需引入新的模块、动态控制逻辑或判断分支，因此推理路径与原始模型差异较小，具备一定的效率友好性。相比之下，KGW、EWD 和 SWEET 等策略需在每一步生成中执行动态采样控制等操作，造成较大的推理开销。而 SynthID 则依赖复杂的嵌入与检测机制，其整体推理效率下降更为显著。

综上所述，本文方法在实现水印嵌入与有效检测的同时，对模型推理效率影响较小，较好地满足了大语言模型在实际部署中对延迟与计算开销的基本要求。得益于其轻量级的设计，本方法具有良好的工程适配性，适合于对推理性能较为敏感的在线生成任务与实际应用场景。

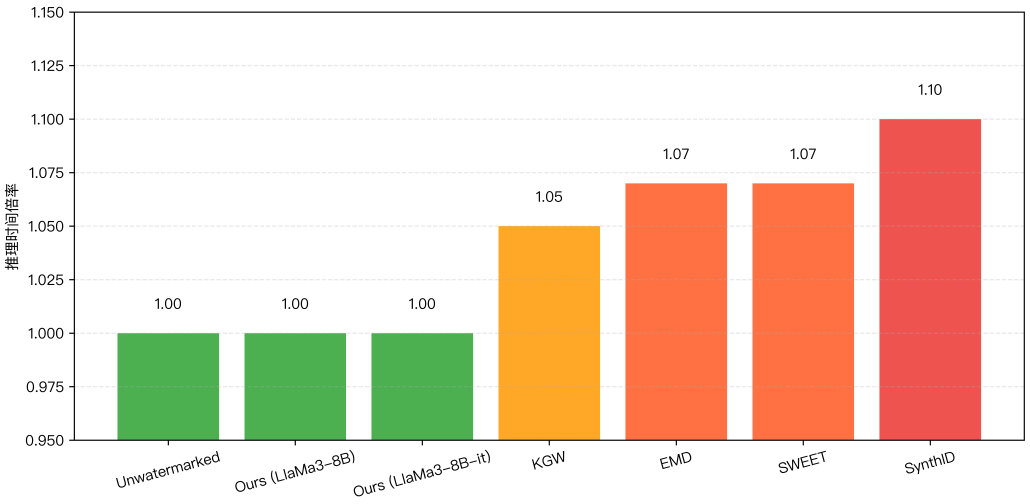

图 4.2 不同水印算法的推理相对时延

4.5.3 嵌入强度与比例参数对性能影响与分析

为了评估水印嵌入强度对模型性能与可检测性的影响，本文设计了两组超参数实验，分别考察扰动缩放因子 α 与嵌入比例 γ 对模型输出困惑度（PPL）和水印检测显著性（z-score）的影响。通过系统量化不同参数设置下的嵌入效果，旨在为实际应用场景中的水印部署提供可调节、可控的超参数选择依据。

在第一组实验中，本文考察了扰动缩放因子 α 对模型性能与水印可检测性的影响。α 控制了输出层中权重向量的缩放幅度，从而调节水印信号在生成文本中的嵌入强度。如图 4.3 所示，随着 α 从 1.05 增加至 1.5，z-score 呈上升趋势，表明水印信号的统计显著性增强，模型的可检测能力稳步提升。值得注意的是，当 α 接近 1.3 后，z-score 增幅趋于平缓，检测性能进入饱和阶段，进一步增大扰动所带来的收益逐渐递减，体现出典型的边际效益下降规律。

同时，图 4.4 显示，生成文本的困惑度（PPL）在 α 增大过程中整体呈现缓慢上升趋势，数值范围维持在 3.4 至 5.0 之间，说明语言模型的生成能力在此扰动范围内仍可基本保持稳定，水印嵌入对文本质量影响有限。

进一步地，本文分析了参与扰动的参数比例 γ 对模型性能与水印显著性的影响。如图 4.5 所示，随着扰动比例的增加，z-score 呈现先升后降的趋势，说明在扰动比例较低时，嵌入的水印信号尚不足以在统计层面形成显著偏移；而当比例过高时，扰动间的重叠可能导致信号干扰，反而削弱整体的可检测性，体现出水印设计中对扰动稀疏性与表达强度的平衡需求。

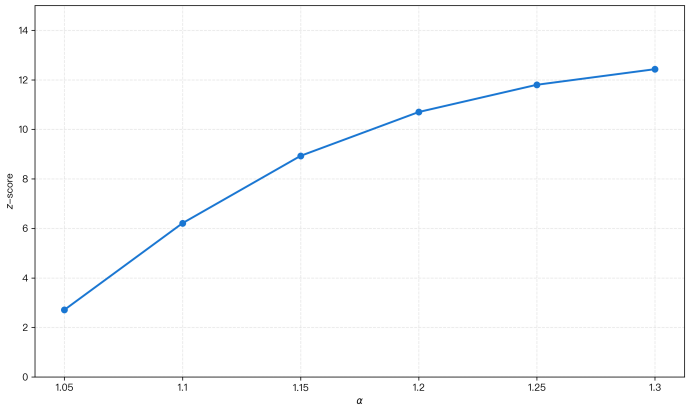

图 4.3 不同缩放因子对 z-score 的影响

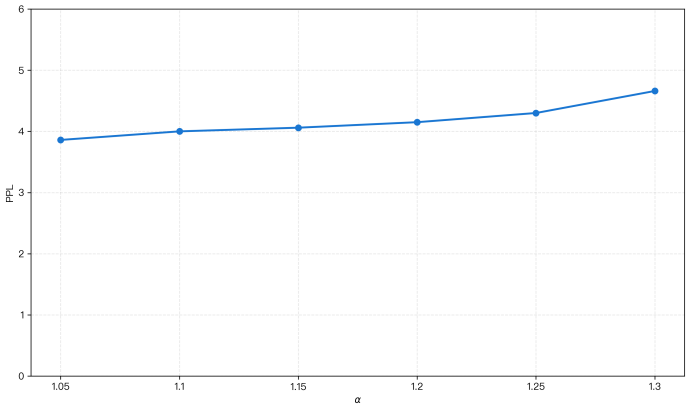

图 4.4 不同缩放因子对困惑度的影响

与此同时，从图 4.6 可见，困惑度（PPL）在整个扰动比例变化区间内波动幅度较小，整体维持在合理范围，说明生成文本的语言质量未受到较大影响。该结果表明，所提方法在一定扰动比例范围内具备较强的语义保真性与鲁棒性，能够适应不同水印嵌入密度的实际需求。

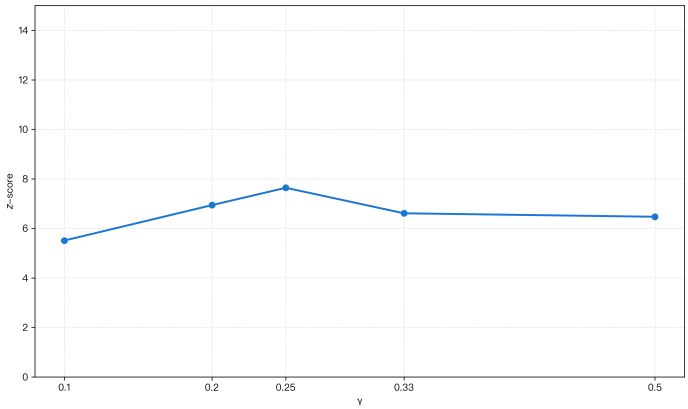

图 4.5 不同扰动比例对 z-score 的影响

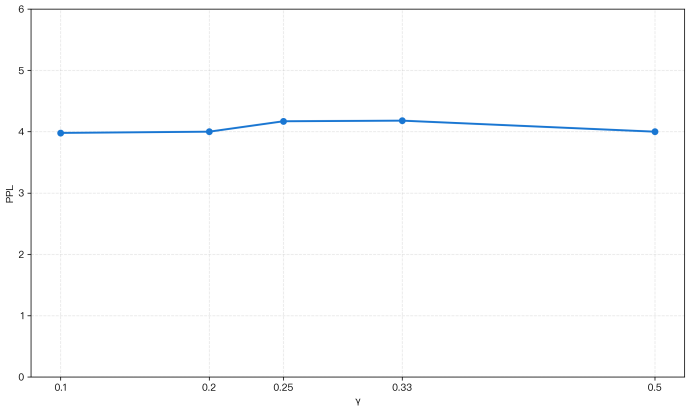

图 4.6　不同扰动比例对困惑度的影响

4.6　本章小结

本章围绕文本生成场景下的大语言模型内生水印技术展开研究，提出了一种基于参数级微扰的轻量级水印嵌入方案，在对模型原有性能影响较小的前提下，实现了对生成文本的有效标记。

首先，本文系统分析了大语言模型的结构特点，将水印嵌入位置聚焦于输出层的线性变换模块，并提出对输出权重中部分向量施加缩放扰动，从而在不破坏语义生成质量的情况下，有针对性地调整生成词汇的概率分布。随后，引入基于密钥控制的稀疏扰动机制，以增强嵌入过程的不可感知性与安全性。在检测阶段，设计了一种基于统计特征的检验机制，用于判断生成文本中是否包含嵌入信号，实现快速且可靠的水印识别。

在实验部分，本文基于 LLaMA3 系列模型进行了系统评估，结果表明所提方法在保真性、可验证性与鲁棒性方面均表现出较好的综合性能。对比分析进一步显示，该方法在对推理开销影响较小的前提下，能够实现稳定且有效的文本水印嵌入，并在面对插入、删除、掩码等典型扰动操作时仍保持可检测性，体现出较好的实用性与抗干扰特性。

第五章　总结与展望

5.1　总结

随着生成式人工智能技术的飞速发展，人工智能生成内容的可追溯性与版权保护问题日益凸显。水印技术，作为实现内容标识与溯源验证的重要手段，正在成为确保人工智能生成内容可信性与可监管性的关键研究方向。本文聚焦于当前主流的图像生成与文本生成两大任务场景，从保真度、鲁棒性与检测准确性等实际需求出发，分别设计并实现了两套具有代表性的水印嵌入方案。

在图像生成任务中，考虑到现有图像生成模型在视觉层面存在高频伪影的问题，影响水印隐蔽性与图像质量，本文提出了一种基于编码器-解码器的外生水印方法。该方法以后处理的方法完成水印的嵌入，通过特殊设计的抑制高频伪影的上采样机制，从源头抑制高频伪影。结合频域对抗训练机制与多频段鉴别器，提升了水印在复杂攻击（如压缩、模糊、裁剪）下的稳定性。实验结果表明，该方法在较小影响视觉自然度与主观质量的前提下，实现了水印的高保真嵌入与鲁棒提取。

在文本生成任务中，语言内容对微小扰动更为敏感，传统水印方法容易破坏语义流畅性和语言逻辑性。为此，本文设计了一种参数扰动的内生水印机制，仅对输出层部分关键的权重施加幅度微小、位置稀疏的扰动，从而在较小影响生成语义连贯性与可读性的前提下，改变采样分布以传递水印信息。该方法在不改变模型主体结构的前提下，于统计层面引入可测信号，同时减小对语义流畅性和生成质量的影响。实验结果显示，该方法在多种模型与任务设置下均表现出较好的水印嵌入效果和检测性能，且在语言连贯性、自然度及系统扩展性方面具有良好表现。

总体而言，本文提出的图像外生水印与文本内生水印方法，虽针对不同的生成任务，但均以高保真嵌入为核心目标，力求在对生成内容的语义完整性、视觉质量及任务性能影响较小的前提下，实现水印信息的有效嵌入。所提出的方法能够在保持内容自然性与功能一致性的同时，实现水印的高保真嵌入与准确提取。实验结果表明，所提出的方法在多种干扰场景下展现出良好的鲁棒性、可验证性与实际可用性，为生成内容的可信标识与溯源追踪提供了一种高保真、可靠的技术路径。

5.2 展望

本文所提出的图像与文本水印技术在高保真嵌入与有效检测方面已取得良好实验效果，但在生成式人工智能不断演进的背景下，水印系统仍面临诸多亟待突破的研究难题。未来的研究可围绕以下几个方面进一步深化与拓展。

首先，提升水印嵌入后的保真性依然是水印系统设计中的核心目标。尽管现有方法在整体感知质量上已较好地控制了扰动强度，但在更精细的表达层级上仍可能引入潜在偏差，这些偏差虽不易被察觉，却可能对用户体验或下游任务造成实质性影响。随着生成内容质量的持续提升，用户对语言流畅性与图像细节的感知阈值愈发敏感，对水印嵌入的隐蔽性也提出了更为严苛的要求。因此，未来的研究应聚焦于开发面向感知机制优化的嵌入策略，在对生成内容自然性影响较小的前提下，确保水印信号的有效嵌入。

其次，水印系统的鲁棒性对于其在实际环境中的可靠部署具有决定性作用。在真实应用场景中，生成内容往往会经历各种形式的后处理操作，包括编辑、压缩、改写甚至语义等价的结构变换，这些操作可能对嵌入的水印信息造成不可逆的破坏。特别是在自然语言文本中，诸如改写、重组或分句等语言层级的调整，易削弱水印信号的结构完整性与可检测性。因此，未来研究需进一步拓展鲁棒建模的深度与广度，不仅要覆盖传统的信号级扰动与噪声攻击，也应将语义一致性下的文本扰动纳入防御范畴。同时，还应探索更为高效与自适应的检测机制，以确保水印在部分受损或退化的情况下，依然具备较强的可验证性与信息恢复能力。

接着，水印嵌入的容量限制在自然语言生成任务中表现尤为突出。与图像相比，文本具有更高的信息密度和更强的语义与语法结构约束，使得在保持一定语义连贯性与语言自然性的前提下，水印信息的嵌入空间有限。因此，如何在控制语言质量损失的同时提升单位文本长度下的水印容量，是当前亟待解决的研究难题之一。

最后，水印系统的安全性是构建可信生成内容机制的根本保障。在面向开放环境的实际应用中，一旦水印机制被逆向解析或恶意伪造，不仅会导致模型归属难以判定，还可能引发严重的信任危机与法律风险。因此，未来的水印设计应将安全性作为系统性要素贯穿于整个嵌入与提取的流程之中。

参考文献

[1] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. The MIT Press, 2016.

[2] CHO K, VAN MERRIENBOER B, GÜLÇEHRE Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL, 2014: 1724-1734.

[3] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//GHAHRAMANI Z, WELLING M, CORTES C, et al. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. 2014: 2672-2680.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 5998-6008.

[6] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020: 1877-1901.

[7] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020: 6840-6851.

[8] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022: 10674-10685.

[9] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[C]//Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022. 2022: 36479-36494.

[10] 王蕾. 人工智能生成内容技术在教育考试中应用探析[J]. 中国考试, 2023, 08: 19-27.

[11] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[12] TOUVRON H, MARTIN L, STONE K R, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[13] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.

[14] DEEPSEEK-AI, GUO D, YANG D, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv:2501.12948, 2025.

[15] LOTH A, KAPPES M, PAHL M O. Blessing or curse? a survey on the impact of generative ai on fake news[J]. arXiv preprint arXiv:2404.03021, 2024.

[16] 郭钊均, 李美玲, 周杨铭等. 人工智能生成内容模型的数字水印技术研究进展[J]. 网络空间安全科学学报, 2024, 2(1): 13-39.

[17] VAN SCHYNDEL R, TIRKEL A, OSBORNE C. A digital watermark[C]// Proceedings of 1st International Conference on Image Processing: volume 2. Austin, TX, USA, 1994: 86-90.

[18] CHEN B, ZHOU C, JEON B, et al. Quaternion discrete fractional random transform for color image adaptive watermarking[J]. Multimedia Tools and Applications, 2018, 77(16): 20809–20837.

[19] WANG X, WANG C, YANG H, et al. A robust blind color image watermarking in quaternion fourier transform domain[J]. Journal of Systems and Software, 2013, 86 (2): 255–277.

[20] LI J, LIN Q, YU C, et al. A qdct- and svd-based color image watermarking scheme using an optimized encrypted binary computer-generated hologram[J]. Soft Computing, 2018, 22(1): 47–65.

[21] BALUJA S. Hiding images in plain sight: Deep steganography[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 2069-2079.

[22] ZHU J, KAPLAN R, JOHNSON J, et al. Hidden: Hiding data with deep networks[C]// Lecture Notes in Computer Science: volume 11219 Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV. Springer, 2018: 682-697.

[23] BENZ P, ZHANG C, KARJAUV A, et al. Robustness may be at odds with fairness: An empirical study on class-wise accuracy[J]. arXiv preprint arXiv:2010.13365, 2020.

[24] NEEKHARA P, HUSSAIN S, ZHANG X, et al. Facesigns: Semi-fragile watermarks for media authentication[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(11): 1-21.

[25] BRASSIL J, LOW S, MAXEMCHUK N, et al. Electronic marking and identification techniques to discourage document copying[J]. IEEE Journal on Selected Areas in Communications, 1995, 13(8): 1495-1504.

[26] RIZZO S G, BERTINI F, MONTESI D. Content-preserving text watermarking through unicode homoglyph substitution[C]//Proceedings of the 20th International Database Engineering & Applications Symposium, IDEAS 2016, Montreal, QC, Canada, July 11-13, 2016. ACM, 2016: 97-104.

[27] TOPKARA U, TOPKARA M, ATALLAH M J. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions[C]//Proceedings of the 8th workshop on Multimedia & Security, MM&Sec 2006, Geneva, Switzerland, September 26-27, 2006. ACM, 2006: 164-174.

[28] YANG X, ZHANG J, CHEN K, et al. Tracing text provenance via context-aware lexical substitution[C]//Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, 2022: 11613-11621.

[29] ZHOU W, GE T, XU K, et al. Bert-based lexical substitution[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics, 2019: 3368-3373.

[30] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(7): 2591-2601.

[31] WEN Y, KIRCHENBAUER J, GEIPING J, et al. Tree-rings watermarks: Invisible fingerprints for diffusion images[C]//Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023. 2023: 58047-58063.

[32] ZHANG L, LIU X, MARTIN A V, et al. Attack-resilient image watermarking using stable diffusion[C]//Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024. 2024.

[33] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding[C]//42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021. IEEE, 2021: 121-140.

[34] XU X, YAO Y, LIU Y. Learning to watermark llm-generated text via reinforcement learning[J]. arXiv preprint arXiv:2403.10553, 2024.

[35] KIRCHENBAUER J, GEIPING J, WEN Y, et al. A watermark for large language models[C]//Proceedings of Machine Learning Research: volume 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 17061-17084.

[36] BISHOP C M. Pattern recognition and machine learning (information science and statistics)[M]. Berlin, Heidelberg: Springer-Verlag, 2006.

[37] KINGMA D P, WELLING M. Auto-encoding variational bayes[C]//2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014.

[38] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[C]// JMLR Workshop and Conference Proceedings: volume 37 Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR.org, 2015: 1530-1538.

[39] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of Machine Learning Research: volume 139 Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. PMLR, 2021: 8748-8763.

[40] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of Machine Learning Research: volume 70 Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. PMLR, 2017: 214-223.

[41] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[C]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[42] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217-4228.

[43] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020: 8107-8116.

[44] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks [C]//Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. 2021: 852-863.

[45] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[C]//9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[46] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023: 3813-3824.

[47] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 4171-4186.

[48] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[49] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.

[50] TOUVRON H, MARTIN L, STONE K R, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[51] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.

[52] DU Z, QIAN Y, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 320-335.

[53] DEEPSEEK-AI, LIU A, FENG B, et al. Deepseek-v3 technical report[J]. arXiv preprint arXiv:2412.19437, 2024.

[54] MESNARD G T T, HARDIN C, DADASHI R, et al. Gemma: Open models based on gemini research and technology[J]. arXiv preprint arXiv:2403.08295, 2024.

[55] RIVIERE G T M, PATHAK S, SESSA P G, et al. Gemma 2: Improving open language models at a practical size[J]. arXiv preprint arXiv:2408.00118, 2024.

[56] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration [C]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[57] CHIANG W, ZHENG L, SHENG Y, et al. Chatbot arena: An open platform for evaluating llms by human preference[C]//Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. JMLR.org, 2024: 8359-8388.

[58] BENGIO Y, MINDERMANN S, PRIVITERA D, et al. International ai safety report [J]. arXiv preprint arXiv:2501.17805, 2025.

[59] MA X, GAO Y, WANG Y, et al. Safety at scale: A comprehensive survey of large model safety[J]. arXiv preprint arXiv:2502.05206, 2025.

[60] STIX C, PISTILLO M, SASTRY G, et al. Ai behind closed doors: a primer on the governance of internal deployment[J]. arXiv preprint arXiv:2504.12170, 2025.

[61] COX I J, MILLER M L, BLOOM J A, et al. Chapter 1 - introduction[M]//The Morgan Kaufmann Series in Multimedia Information and Systems: Digital Watermarking and Steganography (Second Edition). Second edition ed. Burlington: Morgan Kaufmann, 2008: 1-13.

[62] ABBATE J. The electrical century: inventing the web[J]. Proceedings of the IEEE, 1999, 87(11): 1999-2002.

[63] LEE G J, YOON E J, YOO K Y. A new lsb based digital watermarking scheme with random mapping function[C]//2008 International Symposium on Ubiquitous Multimedia Computing. 2008: 130-134.

[64] FAZLI S, KHODAVERDI G. Trade-off between imperceptibility and robustness of lsb watermarking using ssim quality metrics[C]//2009 Second International Conference on Machine Vision. 2009: 101-104.

[65] DEHKORDI A B, ESFAHANI S N, AVANAKI A N. Robust lsb watermarking optimized for local structural similarity[C]//2011 19th Iranian Conference on Electrical Engineering. 2011: 1-6.

[66] LU S, WANG R, ZHONG T, et al. Large-capacity image steganography based on invertible neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021: 10816-10825.

[67] GUAN Z, JING J, DENG X, et al. Deepmih: Deep invertible network for multiple image hiding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 372-390.

[68] 吴汉舟, 张杰, 李越等. 人工智能模型水印研究进展[J]. 中国图象图形学报, 2023, 28(6): 1792-1810.

[69] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017. ACM, 2017: 269-277.

[70] ROUHANI B D, CHEN H, KOUSHANFAR F. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2019, Providence, RI, USA, April 13-17, 2019. ACM, 2019: 485-497.

[71] WANG J, WU H, ZHANG X, et al. Watermarking in deep neural networks via error back-propagation[J]. Electronic Imaging, 2020, 32: 1-9.

[72] FERNANDEZ P, COUAIRON G, FURON T, et al. Functional invariants to watermark large transformers[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024. IEEE, 2024: 4815-4819.

[73] ADI Y, BAUM C, CISSÉ M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]//27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018. USENIX Association, 2018: 1615-1631.

[74] ZHAO X, WU H, ZHANG X. Watermarking graph neural networks by random graphs [C]//9th International Symposium on Digital Forensics and Security, ISDFS 2021, Elazig, Turkey, June 28-29, 2021. IEEE, 2021: 1-6.

[75] LIU Y, WU H, ZHANG X. Robust and imperceptible black-box dnn watermarking based on fourier perturbation analysis and frequency sensitivity clustering[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(6): 5766-5780.

[76] ZHANG J, CHEN D, LIAO J, et al. Model watermarking for image processing networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34 (07): 12805-12812.

[77] ZHANG J, CHEN D, LIAO J, et al. Deep model intellectual property protection via deep watermarking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4005-4020.

[78] LUKAS N, KERSCHBAUM F. PTW: pivotal tuning watermarking for pre-trained image generators[C]//32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023. USENIX Association, 2023: 2241-2258.

[79] FERNANDEZ P, COUAIRON G, JÉGOU H, et al. The stable signature: Rooting watermarks in latent diffusion models[C]//IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023: 22409-22420.

[80] ATALLAH M J, RASKIN V, CROGAN M, et al. Natural language watermarking: Design, analysis, and a proof-of-concept implementation[C]//Lecture Notes in Computer Science: volume 2137 Information Hiding, 4th International Workshop, IHW 2001, Pittsburgh, PA, USA, April 25-27, 2001, Proceedings. Springer, 2001: 185-199.

[81] YANG Z, ZHAO G, WU H. Watermarking for large language models: A survey[J]. Mathematics, 2025, 13(9).

[82] KUDITIPUDI R, THICKSTUN J, HASHIMOTO T, et al. Robust distortion-free watermarks for language models[J]. Transactions on Machine Learning Research, 2024.

[83] ZHANG R, HUSSAIN S S, NEEKHARA P, et al. REMARK-LLM: A robust and efficient watermarking framework for generative large language models[C]//33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024. USENIX Association, 2024: 1813-1830.

[84] BALDASSINI F B, NGUYEN H H, CHANG C, et al. Cross-attention watermarking of large language models[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024. IEEE, 2024: 4625-4629.

[85] XU X, JIA J, YAO Y, et al. Robust multi-bit text watermark with llm-based paraphrasers[J]. arXiv preprint arXiv:2412.03123, 2024.

[86] PANG K, QI T, WU C, et al. Modelshield: Adaptive and robust watermark against model extraction attack[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1767-1782.

[87] ZHONG X, DASGUPTA A, TANVIR A. Watermarking language models through language models[J]. arXiv preprint arXiv:2411.05091, 2024.

[88] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Lecture Notes in Computer Science: volume 9351 Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Springer, 2015: 234-241.

[89] TANCIK M, MILDENHALL B, NG R. Stegastamp: Invisible hyperlinks in physical photographs[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation / IEEE, 2020: 2114-2123.

[90] ZHANG C, BENZ P, KARJAUV A, et al. UDH: universal deep hiding for steganography, watermarking, and light field messaging[C]//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. 2020: 10223-10234.

[91] ZHU J, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 2017: 2242-2251.

[92] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society, 2018: 586-595.

[93] COGRANNE R, GIBOULOT Q, BAS P. Alaska#2: Challenging academic research on steganalysis with realistic images[C]//12th IEEE International Workshop on Information Forensics and Security, WIFS 2020, New York City, NY, USA, December 6-11, 2020. IEEE, 2020: 1-5.

[94] LIN T, MAIRE M, BELONGIE S J, et al. Microsoft COCO: common objects in context[C]//Lecture Notes in Computer Science: volume 8693 Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Springer, 2014: 740-755.

[95] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, 2009: 248-255.

[96] NIE W, GUO B, HUANG Y, et al. Diffusion models for adversarial purification [C]//Proceedings of Machine Learning Research: volume 162 International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. PMLR, 2022: 16805-16827.

[97] FERNANDEZ P, SABLAYROLLES A, FURON T, et al. Watermarking images in self-supervised latent spaces[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, 2022: 3054-3058.

[98] JING J, DENG X, XU M, et al. Hinet: Deep image hiding by invertible network[C]// 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE, 2021: 4713-4722.

[99] MA R, GUO M, HOU Y, et al. Towards blind watermarking: Combining invertible and non-invertible mechanisms[C]//MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022. ACM, 2022: 1532-1542.

[100] DHARIWAL P, NICHOL A Q. Diffusion models beat gans on image synthesis[C]// Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. 2021: 8780-8794.

[101] DODGE J, SAP M, MARASOVIC A, et al. Documenting large webtext corpora: A case study on the colossal clean crawled corpus[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, 2021: 1286-1305.

[102] LEE T, HONG S, AHN J, et al. Who wrote this code? watermarking for code generation[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. Association for Computational Linguistics, 2024: 4890-4911.

[103] LU Y, LIU A, YU D, et al. An entropy-based text watermarking detection method[J]. arXiv preprint arXiv:2403.13485, 2024.

[104] DATHATHRI S, SEE A, GHAISAS S, et al. Scalable watermarking for identifying large language model outputs[J]. Nature, 2024, 634: 818-823.

攻读硕士学位期间取得的研究成果

[1] Liu Q, **Yang Z**, WU H. JPEG steganalysis based on steganographic feature enhancement and graph attention learning[J]. Journal of Electronic Imaging, 2023, 32(3), 033032 (SCI)

[2] **Yang Z**, Zhang Y, Zhang X, WU H. Robust and high-fidelity image watermarking with mutual mapping and antispectral aliasing[J] Journal of Electronic Imaging, 2024, 33(2), 023006 (SCI)

[3] **Yang Z**, ZHAO G, WU H. Watermarking for large language models: A survey[J]. Mathematics, 2025, 13(9). (SCI, invited paper)

致　谢

在岁月悄然流转中，学业的篇章迎来了一个新的节点。回望来路，那些在求索途中给予的点滴帮助与真挚鼓励，早已化作我心中最为珍贵的风景。幸得良师益友一路帮助与支持，使我能够不断前行。怀揣感恩之心，愿将这份厚重的谢意，献给所有在我人生旅途中留下温暖印记的人们。

在此，首先要衷心感谢我导师张新鹏教授。张老师以严谨的治学态度和深邃的学术洞见，为我树立了榜样，也为我带来了许多新的思考视角，使我在专业素养和学术视野上都获益良多。同时，我也要感谢课题组的吴汉舟老师。吴老师在课题研究的各个环节都给予了我细致入微的指导，无论是学术上的疑惑还是实验中的难题，吴老师总能耐心解答、悉心点拨。吴老师的关怀与鼓励，让我不断成长与进步。

在学习和科研的过程中，师兄师姐们的耐心指导和无私分享，为我提供了宝贵的经验与支持；师弟师妹们的朝气与热情，也让我常常感受到集体的活力与温暖。与同门们交流讨论、携手前行，是我学业旅程中宝贵的收获。此外，还要感谢在开源社区和邮件交流中给予我帮助的朋友们，你们无私的分享精神与耐心解答，让我受益良多，也为我的成长提供了坚实的支持。

在我成长和求学的道路上，家人和朋友始终给予我最深的理解与支持，是我不断前行的坚强后盾。父母无条件的关爱和包容，以及家人的陪伴，让我拥有了前行的勇气与动力。无论顺境还是逆境，你们总是默默守护在身后，让我无后顾之忧。朋友们的陪伴与鼓励，也为我的学业和生活增添了许多温柔与力量。你们总是在平凡的日子里给予我真诚的关心与陪伴，无论是分享喜悦，还是倾听烦恼，你们的存在都让我的生活更加温暖和丰富。每一次相聚与交流，都是我人生中弥足珍贵的记忆。衷心感谢你们一路上的陪伴，让我始终怀揣信心，勇敢前行。

同时，诚挚感谢本论文的审稿人，百忙之中细致审阅论文并提出宝贵意见和建议，帮助我完善和提升了本研究的质量。

未来的道路依然充满未知与挑战，但我会始终怀揣感恩之心，不断学习和成长。希望自己能够将这一路上获得的支持与鼓励转化为前行的动力，继续努力追求学术与人生的进步。不负时光，不负所托，愿在新的征途上，勇敢、坚定地迈步前行。