中国科学 : 信息科学 2026 年 第 56 卷 第 11 期: 2935–2954
SCIENTIA SINICA Informationis
Agentic AI 赋能的 6G 融合通信专题 • 论文

Agentic AI 驱动的语义通信编解码模型水印动态嵌入与恢复技术研究

韩博宇[1], 李高磊[1,2*], 杨潇[1], 李建华[1,2], 张昊[3]

1. 上海交通大学计算机学院 (网络空间安全学院、密码学院), 上海 200240
2. 上海市信息安全综合管理技术研究重点实验室, 上海 200240
3. 中国信息安全测评中心, 北京 100085

* 通信作者. E-mail: gaolei_li@sjtu.edu.cn

收稿日期: 2026–03–26; 修回日期: 2026–10–26; 接受日期: 2026–11–12; 网络出版日期: 2026–11–18

国家自然科学基金 (批准号: 62572314, 62471301) 资助项目

摘要 面向 6G 语义通信中内容版权保护与按需授权的需求, 本文提出融合双层水印与上下文分级授权的语义通信框架. 在端到端语义通信链路中嵌入图像水印与 “用户名 + 时间戳” 文本水印, 用于版权标识与会话级时间标记. 上层采用 Agentic AI 启发的流程组织方式, 综合角色与认证结果、环境风险、请求意图、水印一致性和历史会话状态, 通过确定性约束选择重建图像、图像水印验证或双层水印解码模式. 实验结果表明, 在 30 dB AWGN 信道、压缩率 0.8 条件下, 语义通信重建达到 48.36 dB PSNR 和 99.86% SSIM, 图像水印恢复达到 38.48 dB PSNR 和 98.94% SSIM, 文本水印在 800 个测试图像与文本配对样本上实现 100% 精确恢复. 基于时间戳的历史会话重复利用防御实验中, 合法请求通过率为 99.88%, 重复利用检测率为 100%; 在仅监督公开语义输出、查询预算为 $N = 1000$ 的蒸馏设置下, 学生模型的图像水印恢复为 8.14 dB PSNR 和 22.91% SSIM, 文本水印恢复正确率为 0.00%. 所测 10 个受控场景中, 上下文约束策略与等价规则的模式符合率均为 100%, 结果支持上下文约束的授权作用, 不表明智能体相对等价规则具有额外性能优势.

关键词 6G, 语义通信, Agentic AI, 水印技术与版权保护, 智能调度

1 引言

面向 2030 年的第六代移动通信系统 (6G) 正由万物互联走向万物智联, 更加强调通信、计算与感知的深度融合以及网络智能的内生化 [1~3]. 不同于传统系统主要关注速率、时延和可靠性, 6G 更重视业务语义与任务目标 [4,5]. 在此背景下, 以比特级无差错传输为核心的传统通信范式已难以满足复杂场景下的多样化业务需求 [5,6]. 语义通信 (semantic communication, SemCom) 通过提取和传输与任务相关的关键信息, 将通信系统的优化目标由比特级准确传输拓展为语义保真与任务效用最大化, 因而被视为 6G 网络智能化的重要方向 [7~10].

随着语义通信的发展, 其安全性与可信性问题日益突出 [11]. 深度学习驱动的语义通信模型蕴含丰富的任务知识与数据信息, 易遭受模型窃取和知识蒸馏等攻击 [12]. 同时, 在开放环境下的内容生成与分发过程中, 如何在保证语义质量与传输效率的同时实现内容来源和模型归属的可验证与可追溯, 已成为亟待解决的问题 [13].

引用格式: 韩博宇, 李高磊, 杨潇, 等. Agentic AI 驱动的语义通信编解码模型水印动态嵌入与恢复技术研究. 中国科学: 信息科学, 2026, 56: 2935–2954, doi: 10.1360/SSI-2026-0102
Han B Y, Li G L, Yang X, et al. Research on dynamic watermark embedding and recovery for Agentic AI-driven semantic communication encoder–decoder models. Sci Sin Inform, 2026, 56: 2935–2954, doi: 10.1360/SSI-2026-0102

© 2026《中国科学》杂志社

数字水印技术为语义通信中的版权保护与安全审计提供了可行路径. 然而, 在端到端语义通信场景中, 水印设计不仅需抵抗信道噪声和压缩比波动等链路扰动, 还与语义重建质量, 任务性能及模型鲁棒性紧密耦合 [14~17]. 因此, 语义通信中的水印机制应同时服务于版权声明与安全审计, 即通过鲁棒内容水印支持来源验证与所有权声明 [18,19], 并结合用户标识和时间信息的文本水印实现会话标记、重复利用检测与责任追踪 [20].

与此同时, 6G 网络的高动态性与智能化服务特征, 对语义通信系统的访问控制与安全调度提出了更高要求 [21,22]. 仅依据角色分配固定权限的策略难以同时处理请求意图、环境风险与历史会话状态 [23]. 本文借鉴 Agentic AI 中的上下文组织与任务流程协调思路 [24~27], 将认证、信任评分、意图映射、会话校验和输出控制组织为可审计的分级授权流程. 本文采用确定性策略验证该流程的访问约束, 其作用是连接水印恢复结果与单次请求的权限需求, 而非证明智能体推理优于使用相同信息的规则系统.

基于此, 本文构建了一种面向 6G 语义通信的双层水印安全传输与上下文分级授权一体化框架. 在底层, 系统嵌入图像版权水印与包含用户名和时间戳的文本水印, 并通过语义编码器完成特征提取与信道传输; 在上层, 通过确定性权限约束融合角色、认证结果、环境风险、请求意图与会话状态, 选择当前可开放的恢复结果. 实验分别评价底层重建与水印恢复性能、会话重复利用检测以及授权策略功能, 避免将不同实验的指标共同解释为智能体独立性能优势.

1.1　6G 与语义通信

近年来, 6G 网络愿景与体系架构研究持续推进, 语义通信也由理论探索逐步走向系统实现. 在 6G 方面, 现有研究主要围绕网络智能管控、异构接入优化与频谱资源高效利用展开. 例如, Liao 等提出基于知识空间的 6G-ADM 管控体系, 通过知识驱动策略生成与异常检测实现闭环网络管理 [28]. Ma 等面向混合 LiFi/WiFi 场景设计无监督切换方案, 在降低复杂度的同时兼顾吞吐量、切换率和时延性能 [29]. Wang 等则综述了内生智能频谱管控在频谱感知、动态接入与安全保障中的关键技术 [30]. Han 等将隐写术与语义通信相结合, 提出可授权数据传输范式, 在载体中嵌入秘密数据并经语义通信链路传输, 由授权接收端恢复隐藏内容, 为语义通信中的隐蔽传输与差异化访问提供了研究思路 [31]. Yang 等提出凭证感知的语义通信知识保护框架 SemanAegis, 通过内置访问控制机制使授权输入获得正常响应, 并限制未授权输入对语义编解码知识的使用, 为语义通信中的知识泄露防护提供了研究思路 [32].

相较于上述网络架构, 语义通信更关注从编码范式层面重构传输目标. 现有研究主要聚焦于自适应编码、任务驱动传输、语义资源联合优化以及知识增强语义恢复. Zhang 等提出可变压缩率语义编码框架, 可根据重建质量约束实现样本级压缩率自适应调整 [33]. Zhang 等结合场景图建模与多智能体强化学习, 实现语义信息选择与资源分配的联合优化 [34]. Yi 等引入共享知识库增强语义编码与解码能力, 提升低信噪比条件下的恢复质量与传输效率 [35]. Peng 等通过多尺度语义提取与融合机制增强系统对语义扰动和复杂信道的鲁棒性 [36]. 总体来看, 现有语义通信研究主要面向传输效率与任务性能优化, 对版权保护及安全审计问题关注仍相对不足.

1.2　数字水印技术

数字水印技术通过在图像、模型参数或生成内容中嵌入隐蔽且可验证的标记, 实现版权确权、来源追溯与篡改检测, 已成为内容安全与模型安全的重要技术手段. 现有研究围绕鲁棒嵌入、不可感知性优化、大容量设计以及对抗压缩、屏摄和模型蒸馏等攻击展开. Tyagi 等提出融合黑盒触发验证与白盒特征约束的 DeepVerifier, 实现深度模型水印嵌入与版权认证 [37]. Guo 等提出 DoBMark, 通过双分支建模视觉保真性与屏摄鲁棒性, 显著增强对物理拍摄攻击的抵抗能力 [38]. Nie 等面向联邦学习提出 FedCRMW, 实现模型在压缩与分发过程中的可靠所有权验证 [39]. Xu 等提出 InvisMark, 可在高分辨率 AI 生成图像中实现高容量、不可感知且鲁棒的水印嵌入 [40]. Dathathri 等通过概率偏置设计可扩展的语言模型水印机制, 支持生成文本的统计检测与来源追溯 [18]. Liu 等提出基于卷积神经网络的零水印方法, 实现对多种图像处理与几何攻击的鲁棒版权验证 [41]. 尽管现有工作已在视觉内容、联邦模型和生成文本等场景取得进展, 但针对语义通信链路中跨信道传输与多层安全需求耦合的水印设计仍有待深入研究.

1.3　Agentic AI 与智能调度

随着 6G 网络向内生智能与服务化演进, 语义通信系统在复杂动态环境中的安全控制与资源调度问题日益突出, 传统基于静态规则的管理机制已难以满足多角色访问、多场景风险与多级服务需求. 因此, 引入具备感知、推理与决策能力的 Agentic AI, 为构建情境感知与自适应调度的语义通信架构提供了新的思路. Gu 等提出 Agentic 共识构建框架, 通过 Manager Agent 与多个 Subtask Agents 协同实现群体决策监测、任务分解与动态干预 [42]. Pandey 等从企业应用视角分析 Agentic AI 的体系结构与战略价值, 强调其在人机协同和多智能体协作中的作用 [43]. Gody 等提出 ConvoGen, 利用多智能体协作生成多角色合成对话数据, 为会话模型训练和数据增强提供支持 [44]. Duan 和 Lu 则围绕 Agent2Agent 协议分析智能体通信在边缘环境中的适应性与局限 [45]. 总体而言, 现有 Agentic AI 研究主要集中于协同决策、任务编排和智能交互, 但其与语义通信安全控制、水印开放策略及访问级别调度的深度结合仍较少, 这也为本文研究提供了切入点.

本文的主要贡献如下.

(1) 提出一种面向 6G 的语义通信与双层数字水印一体化安全传输框架, 在统一端到端网络中协同优化语义通信重建、整图版权水印与文本水印, 兼顾内容质量、嵌入隐蔽性与解码鲁棒性. 在 AWGN 与衰落信道下, 语义通信重建最高约为 48 dB PSNR, 第一层水印恢复约为 38.5 dB PSNR.

(2) 构建基于用户标识和时间戳的会话级文本水印机制, 并结合哈希承诺与时间窗口约束, 使水印由静态版权标识扩展为支持历史会话重复利用检测与责任追踪的安全元数据. 在 800 个测试图像与文本配对样本上, 第二层文本水印实现了 100% 的恢复率.

(3) 构建与底层水印网络解耦的上下文分级授权流程, 结合角色、认证、环境风险、请求意图与会话状态, 以显式权限约束组织多种水印恢复结果的开放. 在 10 个受控场景中, 确定性上下文策略的模式符合率为 100%, 静态角色策略为 50%; 使用相同信号与约束的等价规则同样为 100%. 该贡献在于双层水印与上下文授权的系统结合, 不主张智能体相对等价规则具有额外决策优势.

(4) 构建模块消融、不同查询预算的公开语义输出蒸馏与历史会话重复利用评价. 在限定的蒸馏监督下分别报告语义重建与水印输出, 并结合访问模式指标区分底层恢复质量、任务迁移和会话授权约束, 不将固定教师验证链路上的低恢复率等同于对自适应提取攻击的普遍保证.

本文后续组织结构如下. 第 2 节介绍 Agentic AI 驱动的双层水印语义通信系统模型与整体框架. 第 3 节给出实验设置与结果分析. 第 4 节对全文进行总结.

2　Agentic AI 驱动的双层水印语义通信系统模型

2.1　问题定义

本文研究面向不同用户需求与安全评分自适应调度的 6G 语义通信双层水印安全传输问题. 设待传输图像数据集为 $\mathcal{D}=\{\mathbf{x}_i\}_{i=1}^{N}$, 其中 $\mathbf{x}_i\in\mathbb{R}^{H\times W\times 3}$ 表示输入样本. 对于每个样本, 系统首先在图像域嵌入双层水印, 即将第一层图像水印 $\mathbf{w}_1$ 嵌入载体图像, 并将由用户标识与时间戳构成的第二层文本水印字符串 $s_i=u_i\mid t_i$ 编码为比特向量 $\mathbf{b}_i=\mathcal{B}(s_i)$, 进一步构造为局部水印表示后嵌入图像, 从而生成双层水印隐写图像 $\mathbf{x}_{2,i}$. 随后, 语义编码器 $\mathcal{E}_{\theta_E}(\cdot)$ 将其编码为语义表征 $\mathbf{z}_i=\mathcal{E}_{\theta_E}(\mathbf{x}_{2,i})$, 经信道模型 $\mathcal{C}(\cdot)$ 传输后得到接收端观测 $\mathbf{y}_i=\mathcal{C}(\mathbf{z}_i)$, 再由接收端恢复模块得到中间语义特征并进一步恢复复合图像、第一层图像水印和第二层文本水印.

在此基础上, 引入 Agentic AI 决策函数 $\mathcal{A}(\cdot)$, 对访问请求的认证结果、角色基线信任、终端与网络环境风险、自然语言请求意图以及文本水印校验结果进行联合判定, 并自适应选择访问模式 $m_j\in\{0,1,2\}$. 其中, $m_j=0$ 表示仅进行语义通信重建, $m_j=1$ 表示执行语义通信重建与第一层水印恢复, $m_j=2$ 表示执行语义通信重建与双层水印恢复. 本文目标是在不同访问等级约束下, 同时保证语义通信重建质量、水印恢复准确性与文本一致性, 并提升系统在模型提取攻击与历史会话重复利用场景下的整体鲁棒性.

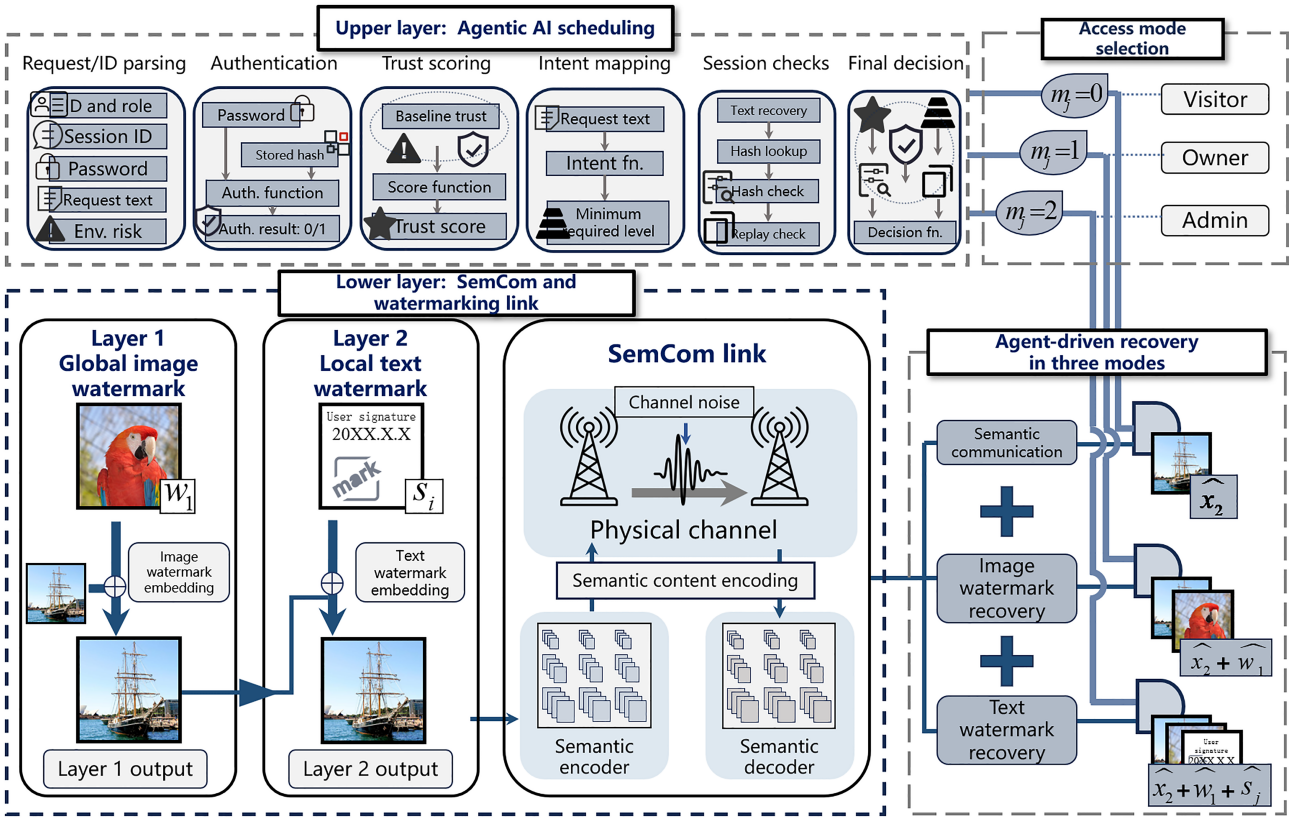

图 1　(网络版彩图) 双层水印语义通信与上下文分级授权的整体框架.

Figure 1　(Color online) Overall framework of dual-layer watermarking semantic communication and context-constrained hierarchical authorization.

2.2　框架实现流程

如图 1 所示, 本文框架由底层双层水印语义通信链路与上层 Agent 智能调度层组成. 发送端首先在载体图像中嵌入图像水印与文本水印, 再将复合图像编码为语义表征并通过无线信道传输. 接收端根据接收表征完成复合图像重建、第一层版权水印恢复和第二层文本水印解码. 在此基础上, Agent 结合身份认证结果、信任评分、访问意图、一致性校验结果以及历史会话重复利用检测结果, 自适应决定当前请求的访问等级与输出内容. 图中不同角色与访问模式的连接表示典型请求示例, 最终权限由认证、信任评分、访问意图与会话校验共同确定. 右侧的加号表示多项恢复结果同时开放, 对应后文的输出集合 $\mathcal{O}_j$, 不表示将水印重新叠加到已恢复的复合图像上.

2.2.1　语义编码器与双层水印嵌入的设计

双层水印在图像域完成嵌入后, 随复合图像进行语义编码与含噪信道传输. 接收端基于恢复的语义特征 $\hat{\mathbf{z}}_i$, 分别解码图像版权水印与会话级文本水印. 设输入载体图像为 $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$, 第一层整图图像水印为 $\mathbf{w}_1 \in \mathbb{R}^{3\times H\times W}$, 第二层文本水印字符串为 s_i. 发送端首先完成双层水印嵌入, 随后将嵌入结果送入语义编码器进行特征映射.

第一层全局图像水印嵌入. 其通过隐藏网络 $\mathcal{H}_A(\cdot)$ 在整幅图像范围内进行隐式融合, 得到中间隐写图像:

$$\mathbf{x}_{1,i} = \mathcal{H}_A(\mathbf{w}_1, \mathbf{x}_i). \tag{1}$$

该映射可理解为在全局空间域上对载体特征与水印特征进行非线性耦合. 若记 $\mathcal{H}_A$ 由参数 θ_A 表示, 则有

$$\mathbf{x}_{1,i} = \mathbf{x}_i + \mathcal{F}_A(\mathbf{x}_i, \mathbf{w}_1; \theta_A), \tag{2}$$

其中, $\mathcal{F}_A(\cdot)$ 表示由卷积网络学习得到的残差扰动项. 由于该扰动在整个空间域 Ω 上定义, 因此第一层水印属于全局嵌入形式, 不依赖任何局部掩膜约束.

载体图像与水印图像经共享的特征提取模块编码. 该模块由 3×3 卷积 $3\to64$、ReLU 及步长为 2 的 4×4 卷积 $64\to64$ 构成. 两路特征沿通道维拼接后, 依次经过 1×1 卷积 $128\to64$ 和 3×3 卷积 $64\to64\to32$, 再通过步长为 2 的 4×4 转置卷积 $32\to32$ 与 3×3 卷积 $32\to16\to3$ 恢复至输入分辨率. 普通 3×3 卷积采用步长 1 和填充 1, 下采样卷积与转置卷积的填充均为 1. 特征融合层采用批归一化与 ReLU, 输出层采用 Tanh. 两层隐藏网络结构相同, 参数独立.

第二层文本水印编码. 首先将字符串 s_i 编码为 UTF-8 字节序列, 再转换为二进制比特向量:

$$\mathbf{b}_i=(b_{i,1},b_{i,2},\ldots,b_{i,L_i}),\quad b_{i,k}\in\{0,1\}. \tag{3}$$

各字节按最高有效位优先顺序展开, 并依据嵌入区域容量进行补零或截断.

设角落嵌入区域尺寸为 $S\times S$, 容量为 S^2. 定义容量对齐后的比特向量:

$$\tilde{\mathbf{b}}_i=\begin{cases}(b_{i,1},\ldots,b_{i,L_i},0,\ldots,0), & L_i<S^2,\\ (b_{i,1},\ldots,b_{i,S^2}), & L_i\geqslant S^2,\end{cases} \tag{4}$$

并重排为二维比特特征图:

$$\mathbf{B}_i\in\{0,1\}^{1\times S\times S}. \tag{5}$$

将其嵌入右下角区域 $\Omega_2\subset\Omega$, 构造全尺寸文本水印图:

$$\mathbf{w}_{2,i}(p,q)=\begin{cases}\mathbf{B}_i(p',q'), & (p,q)\in\Omega_2,\\ 0, & \text{otherwise},\end{cases} \tag{6}$$

同时生成对应掩膜矩阵:

$$\mathbf{M}_2(p,q)=\begin{cases}1, & (p,q)\in\Omega_2,\\ 0, & \text{otherwise}.\end{cases} \tag{7}$$

第二层局部残差式嵌入. 为减小对整体视觉质量的影响, 第二层采用局部残差混合形式. 设第二层隐藏网络为 $\mathcal{H}_B(\cdot)$, 嵌入强度系数为 $\alpha\in(0,1)$, 则最终双层隐写图像为

$$\mathbf{x}_{2,i}=\mathbf{x}_{1,i}+\alpha\,\mathbf{M}_2\odot\left(\mathcal{H}_B(\mathbf{w}_{2,i},\mathbf{x}_{1,i})-\mathbf{x}_{1,i}\right). \tag{8}$$

进一步展开可得

$$\mathbf{x}_{2,i}=(1-\alpha\mathbf{M}_2)\odot\mathbf{x}_{1,i}+\alpha\mathbf{M}_2\odot\mathcal{H}_B(\mathbf{w}_{2,i},\mathbf{x}_{1,i}), \tag{9}$$

可见第二层扰动仅在掩膜区域 Ω_2 内生效, 在非嵌入区域保持 $\mathbf{x}_{2,i}=\mathbf{x}_{1,i}$, 从而实现局部可控嵌入.

单通道文本水印图经通道复制后, 作为三通道输入送入第二层隐藏网络. 具体嵌入参数与预处理配置见第 3.1 小节.

语义编码映射. 完成双层水印嵌入后, 将隐写图像 $\mathbf{x}_{2,i}$ 输入语义编码器 $\mathcal{E}_{\theta_E}(\cdot)$, 得到语义特征表示:

$$\mathbf{z}_i=\mathcal{E}_{\theta_E}(\mathbf{x}_{2,i}),\quad \mathbf{z}_i\in\mathbb{R}^{C\times H'\times W'}. \tag{10}$$

2.2.2　信道传输的模拟

完成双层水印嵌入后, 复合图像 $\mathbf{x}_2\in\mathbb{R}^{3\times H\times W}$ 输入语义编码器 $\mathcal{E}(\cdot)$, 映射为待传输的语义表征:

$$\mathbf{z}=\mathcal{E}(\mathbf{x}_2),\quad \mathbf{z}\in\mathbb{R}^{C\times H'\times W'}. \tag{11}$$

为刻画无线环境中的动态波动, 本文在训练阶段对信噪比与压缩率引入样本级随机扰动. 对于第 i 个样本, 信噪比设为

$$\mathrm{SNR}_i = \mathrm{SNR}_0 + \Delta_i, \qquad \Delta_i \sim \mathcal{U}(-\Delta_{\mathrm{snr}}, \Delta_{\mathrm{snr}}), \tag{12}$$

压缩率按区间随机采样:

$$\mathrm{CR}_i \sim \mathcal{U}(\mathrm{CR}_0 - \Delta_{\mathrm{cr}}, \mathrm{CR}_0 + \Delta_{\mathrm{cr}}), \tag{13}$$

并通过

$$\mathrm{CR}_i \leftarrow \mathrm{clip}(\mathrm{CR}_i, \mathrm{CR}_{\min}, \mathrm{CR}_{\max}), \tag{14}$$

约束在合法范围内. 该机制提升了模型对不同信道质量与带宽约束的适应能力.

设 $\mathcal{C}(\cdot)$ 表示信道映射, 则接收端观测信号可写为

$$\mathbf{y} = \mathcal{C}(\mathbf{z};\, \mathrm{SNR}_i, \mathrm{CR}_i)\,. \tag{15}$$

在 AWGN 信道下,

$$\mathbf{y} = \mathbf{z} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}\big(\mathbf{0}, \sigma^2\mathbf{I}\big)\,, \tag{16}$$

其中, σ^2 由 SNR_i 决定. 在衰落信道下,

$$\mathbf{y} = \mathbf{h} \odot \mathbf{z} + \mathbf{n}, \tag{17}$$

其中, $\mathbf{h}$ 为随机衰落向量, $\odot$ 表示逐元素乘法.

2.2.3　语义解码与水印恢复的实现

接收端获得信道输出 $\mathbf{y}_i$ 后, 首先通过接收端特征恢复映射 $\mathcal{R}(\cdot)$ 得到中间语义特征:

$$\hat{\mathbf{z}}_i = \mathcal{R}(\mathbf{y}_i), \tag{18}$$

其中, $\hat{\mathbf{z}}_i \in \mathbb{R}^{C\times H'\times W'}$. 随后, 语义解码器 $\mathcal{D}_{\theta_D}(\cdot)$ 根据 $\hat{\mathbf{z}}_i$ 恢复复合图像估计:

$$\hat{\mathbf{x}}_{2,i} = \mathcal{D}_{\theta_D}(\hat{\mathbf{z}}_i), \qquad \hat{\mathbf{x}}_{2,i} \in \mathbb{R}^{3\times H\times W}. \tag{19}$$

第一层图像水印恢复. 第一层图像水印通过解码网络 $\mathcal{D}_A(\cdot)$ 基于接收端中间语义特征恢复, 即

$$\hat{\mathbf{w}}_1 = \mathcal{D}_A(\hat{\mathbf{z}}_i), \qquad \hat{\mathbf{w}}_1 \in \mathbb{R}^{3\times H\times W}. \tag{20}$$

其恢复质量通过与真实水印 $\mathbf{w}_1$ 计算 PSNR 和 SSIM 进行评估:

$$\mathrm{PSNR}_A = 10\log_{10}\left(\frac{1}{\mathrm{MSE}(\hat{\mathbf{w}}_1, \mathbf{w}_1)}\right), \tag{21}$$

$$\mathrm{SSIM}_A = \mathrm{SSIM}(\hat{\mathbf{w}}_1, \mathbf{w}_1). \tag{22}$$

第二层文本水印恢复. 第二层文本水印首先通过解码网络 $\mathcal{D}_B(\cdot)$ 恢复为特征图:

$$\hat{\mathbf{w}}_{2,i} = \mathcal{D}_B(\hat{\mathbf{z}}_i), \qquad \hat{\mathbf{w}}_{2,i} \in \mathbb{R}^{1\times H\times W}. \tag{23}$$

随后, 将其下采样至 64×64 得到

$$\tilde{\mathbf{w}}_{2,i} = \mathrm{Down}(\hat{\mathbf{w}}_{2,i}), \qquad \tilde{\mathbf{w}}_{2,i} \in \mathbb{R}^{1\times 64\times 64}, \tag{24}$$

并从对应嵌入区域 $\tilde{\Omega}_2$ 中提取比特向量:

$$\hat{\mathbf{b}}_i = \mathrm{vec}\left(\tilde{\mathbf{w}}_{2,i}|_{\tilde{\Omega}_2}\right) \in \mathbb{R}^{S^2}. \tag{25}$$

为实现文本级恢复, 引入字符预测头 $\mathcal{T}(\cdot)$, 将比特向量映射为字符概率分布:

$$\mathbf{P}_i = \mathcal{T}(\hat{\mathbf{b}}_i) \in \mathbb{R}^{L_{\max}\times|\mathcal{V}|}, \tag{26}$$

其中, $|\mathcal{V}|$ 表示词表大小, $L_{\max}$ 为最大文本长度. 通过逐位置取最大概率字符:

$$\hat{c}_{i,t} = \arg\max_k (\mathbf{P}_i)_{t,k}, \tag{27}$$

得到预测字符串 $\hat{s}_i$. 第二层文本恢复准确率定义为

$$\mathrm{Acc}_{\mathrm{text}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\hat{s}_i = s_i\right), \tag{28}$$

其中, $\mathbf{1}(\cdot)$ 为指示函数, N 为当前测试条件下的图像与文本配对样本数, 本实验取 $N = 800$; 该数量与模型提取实验的查询预算分别定义.

综上, 接收端通过语义解码与双支路水印恢复, 分别输出复合图像、第一层版权水印和第二层文本水印, 从而支持语义重建、版权验证与会话级追踪.

2.2.4　联合训练目标

为兼顾双层水印嵌入的视觉不可感知性、复合图像重建质量、第一层水印恢复精度以及第二层文本水印恢复可靠性, 本文先训练水印嵌入模块, 再固定嵌入网络, 对语义通信编解码模块及接收端水印恢复模块进行联合训练, 最后单独微调字符预测头. 对于训练样本 $(\mathbf{x}_i, \mathbf{w}_1, s_i)$, 经过双层水印嵌入、语义编码、信道传输与接收端恢复后, 分别得到双层隐写图像 $\mathbf{x}_{2,i}$、复合图像重建结果 $\hat{\mathbf{x}}_{2,i}$、第一层水印恢复结果 $\hat{\mathbf{w}}_1$ 以及第二层比特图恢复结果 $\hat{\mathbf{w}}_{2,i}$.

整体训练目标由以下 4 项组成.

(1) 隐蔽性约束项. 为保证双层水印嵌入后图像与原始载体图像在视觉上保持一致, 定义为

$$R_h(D_{\mathrm{train}}) = \lambda_h \cdot \mathbb{E}\left[\mathcal{L}_h(\mathbf{x}_{2,i}, \mathbf{x}_i)\right]. \tag{29}$$

隐蔽性约束用于嵌入网络预训练. 通信联合训练阶段固定嵌入网络参数, 仅优化语义通信与水印恢复网络, 因而设置 $\lambda_h = 0$.

(2) 语义通信重建项. 为提高复杂信道条件下的复合图像恢复质量, 定义为

$$R_r(D_{\mathrm{train}}) = \lambda_r \cdot \mathbb{E}\left[\mathcal{L}_r(\hat{\mathbf{x}}_{2,i}, \mathbf{x}_{2,i})\right]. \tag{30}$$

(3) 第一层整图水印恢复项. 为保证版权水印的稳定解码, 定义为

$$R_w(D_{\mathrm{train}}) = \lambda_w \cdot \mathbb{E}\left[\mathcal{L}_w(\hat{\mathbf{w}}_1, \mathbf{w}_1)\right]. \tag{31}$$

语义通信重建与第一层水印恢复均采用 MSE 和多尺度结构相似性 (MS-SSIM) 的加权损失, 分别定义为

$$\begin{aligned} \mathcal{L}_r(\mathbf{a}, \mathbf{b}) &= 0.5\,\mathrm{MSE}(\mathbf{a}, \mathbf{b}) + 0.5[1 - \text{MS-SSIM}(\mathbf{a}, \mathbf{b})], \\ \mathcal{L}_w(\mathbf{a}, \mathbf{b}) &= 0.7\,\mathrm{MSE}(\mathbf{a}, \mathbf{b}) + 0.3[1 - \text{MS-SSIM}(\mathbf{a}, \mathbf{b})]. \end{aligned} \tag{32}$$

(4) 第二层文本水印恢复项. 通信联合训练以第二层比特图为监督目标. 令 $\mathbf{G}_i^{(256)}$ 表示经最近邻插值获得的 256×256 单通道比特标签, $\mathbf{M}_2^{(256)}$ 表示对应掩膜, 则第二层恢复损失定义为有效嵌入区域内的归一化均方误差:

$$R_t(D_{\mathrm{train}}) = \lambda_t\, \mathbb{E}\left[\frac{\sum_{p,q} M_{2,pq}^{(256)} \left(\hat{w}_{2,i}(p,q) - G_i^{(256)}(p,q)\right)^2}{\sum_{p,q} M_{2,pq}^{(256)} + 10^{-8}}\right]. \tag{33}$$

综合上述各项约束, 整体联合训练目标为

$$R_{\text{total}}(D_{\text{train}}) = R_h(D_{\text{train}}) + R_r(D_{\text{train}}) + R_w(D_{\text{train}}) + R_t(D_{\text{train}}), \tag{34}$$

最终, 整体优化过程可表示为对联合损失函数 $R_{\text{total}}(D_{\text{train}})$ 的最小化.

设通信联合训练轮数为 $e \geqslant 1$, 线性预热系数为 $q(e) = \min\{1, e/100\}$. 各损失项权重设为 $\lambda_h = 0$、$\lambda_r = 1$、$\lambda_w(e) = 0.04q(e)$ 和 $\lambda_t(e) = 0.20q(e)$. 水印恢复损失权重在前 100 轮线性增加, 随后分别固定为 0.04 和 0.20.

字符预测头 $\mathcal{T}$ 在通信联合训练后, 使用训练载体对应的文本标签单独优化, 此阶段固定语义通信与水印恢复网络参数. 设有效字符标签序列为 $\{c_{i,t}\}_{t=1}^{\ell_i}$, 其中 ℓ_i 表示有效字符数. 排除 PAD 位置后, 字符预测损失定义为

$$\mathcal{L}_{\text{char}} = \frac{\sum_i \sum_{t=1}^{\ell_i} \mathcal{L}_{\text{CE}}((\mathbf{P}_i)_t, c_{i,t})}{\sum_i \ell_i}. \tag{35}$$

字符预测头采用 $256 \to 512 \to L_{\max}|\mathcal{V}|$ 的两层全连接网络, 隐藏层使用 ReLU. 该网络以 Adam 优化器微调 200 个 epoch, 学习率为 10^{-3}, 批大小为 4.

通过该联合训练策略, 系统能够在统一优化框架下兼顾水印隐蔽性、语义重建质量以及双层水印恢复可靠性, 从而为后续 Agent 智能调度提供稳定的底层支撑.

2.2.5　Agentic AI 驱动的访问模式调度

为实现双层水印系统的差异化授权与安全控制, 本文在语义通信与水印恢复模块之上构建 Agent 调度模块 $\mathcal{A}(\cdot)$. 该模块体现为上下文组织、校验调用、权限约束和结果路由的协调流程, 不参与底层网络训练. 本文受控实验采用确定性授权实现, 不将其表述为经独立验证的大模型自主推理机制. 对于每次请求, 流程依次完成身份认证与风险评分、意图到需求等级的映射、文本一致性与会话状态检查、最终权限计算及输出控制. 自然语言意图识别函数作为接口给出, 第 3.7 小节使用预先确定的意图等级, 不据此评价该接口的识别性能.

设第 j 次访问请求为 $\mathcal{R}_j$, 其包含用户身份 u_j、用户角色 r_j、输入密码 p_j、请求文本 q_j、待核验会话标识 κ_j 以及环境风险评分 $\rho_j \in [0,1]$. 系统为每个用户预存密码哈希值 h_{u_j}, 并维护历史合法会话记录. 认证结果定义为

$$\text{auth}_j = \mathbf{1}\left(\text{Hash}(p_j) = h_{u_j}\right), \tag{36}$$

目标会话参考承诺记为

$$\bar{h}_j = \mathcal{L}(\kappa_j), \tag{37}$$

其中, $\mathcal{L}(\cdot)$ 表示会话登记表查询映射.

综合用户角色基线信任值 τ_{r_j} 与环境风险, 定义请求信任评分:

$$T_j = \text{clip}\left(\tau_{r_j} + \eta_{\text{auth}}\,\text{auth}_j - \eta_{\text{risk}}\,\rho_j,\, 0,\, 1\right), \tag{38}$$

其中, $\eta_{\text{auth}}, \eta_{\text{risk}} > 0$ 分别为认证增益与风险惩罚权重, 与水印嵌入强度 α 区分. 为刻画访问意图, 引入意图识别函数:

$$I_j = \mathcal{F}_{\text{intent}}(q_j), \tag{39}$$

其输出 $I_j \in \{\text{view}, \text{ownership}, \text{forensic}\}$, 并映射为最低访问需求等级:

$$m_j^{\text{req}} \in \{0, 1, 2\}. \tag{40}$$

在访问控制阶段, 系统结合第二层文本水印的内部解码结果进行一致性校验与历史会话重复利用检测. 设恢复得到的第二层文本水印为 $\hat{s}_j = \hat{u}_j \mid \hat{t}_j$, 则一致性校验结果为

$$g_j = \mathbf{1}\left(\text{Hash}(\hat{s}_j) = \bar{h}_j\right). \tag{41}$$

设系统维护已使用会话集合 $\mathcal{S}_{\text{used}}$, 当前时间为 t_j^{now}, 容许时间偏差阈值为 Δ_T, 合法时间窗口为

$$\mathcal{T}_{\text{valid}}(t_j^{\text{now}}) = \left[t_j^{\text{now}} - \Delta_T,\ t_j^{\text{now}} + \Delta_T\right], \tag{42}$$

则历史会话重复利用标志定义为

$$\gamma_j = \mathbf{1}\left((g_j = 1)\ \wedge\ \left((\hat{t}_j \notin \mathcal{T}_{\text{valid}}(t_j^{\text{now}}))\ \vee\ (\bar{h}_j \in \mathcal{S}_{\text{used}})\right)\right). \tag{43}$$

最终访问等级由 Agent 决策函数确定:

$$m_j = \mathcal{A}\left(T_j,\ m_j^{\text{req}},\ \text{auth}_j,\ g_j,\ \gamma_j\right), \tag{44}$$

为使权限约束可直接复现, 令 $\theta_1 = 0.3$、$\theta_2 = 0.6$, 信任权限上限为

$$\pi(T_j) = \begin{cases} 0, & T_j < \theta_1, \\ 1, & \theta_1 \leqslant T_j < \theta_2, \\ 2, & T_j \geqslant \theta_2, \end{cases} \tag{45}$$

对应的确定性决策函数展开为

$$m_j = \begin{cases} 0, & \text{auth}_j = 0\ \vee\ g_j = 0\ \vee\ T_j < \theta_1, \\ \min\{m_j^{\text{req}}, \pi(T_j), 2 - \gamma_j\}, & \text{otherwise}, \end{cases} \tag{46}$$

其中, 请求需求等级由查看、版权验证和取证分别映射为 0、1 和 2; 认证失败或一致性失败优先触发模式 0, 重复利用命中仅将权限上限压至模式 1, 不会提升已有的较低权限. 该流程的决策完全由列出的信号与约束确定, 因而等价规则获得相同输出是预期性质. 其中, $m_j \in \{0, 1, 2\}$. 当认证失败, 信任评分过低或文本水印一致性校验失败时, 系统仅返回语义通信重建结果; 当请求满足版权验证需求且信任评分处于中等区间时, 允许访问第一层水印; 当信任评分较高, 请求满足安全取证条件, 且文本水印校验通过并未检测到历史会话重复利用时, 进一步开放第二层文本水印. 若检测到历史会话重复利用, 则将可开放等级上限限制为模式 1, 不改变认证失败等条件下的模式 0 输出. 为满足后文安全性分析的状态一致性前提, 对拟开放模式 2 的请求, 会话检查与已使用状态登记应作为原子操作在显式文本输出前完成; 仅在检查与登记成功后开放模式 2. 已使用会话集合的更新表示为

$$\mathcal{S}_{\text{used}} \leftarrow \mathcal{S}_{\text{used}} \cup \{\bar{h}_j\}. \tag{47}$$

因此, 接收端最终输出集合为

$$\mathcal{O}_j = \begin{cases} \{\hat{\mathbf{x}}_{2,j}\}, & m_j = 0, \\ \{\hat{\mathbf{x}}_{2,j}, \hat{\mathbf{w}}_1\}, & m_j = 1, \\ \{\hat{\mathbf{x}}_{2,j}, \hat{\mathbf{w}}_1, \hat{s}_j\}, & m_j = 2, \end{cases} \tag{48}$$

由此, 系统实现了语义通信层与访问控制层的结构解耦, 并通过身份认证、信任评分、意图识别、文本水印一致性校验与历史会话重复利用检测实现多级安全访问控制.

3　实验结果与分析

本节介绍实验设置, 并从重建质量、水印恢复、访问控制及攻击场景下的表现评估所提框架.

表 1　关键训练与水印参数.

Table 1　Key training and watermark parameters.

Watermark and channel		Training and optimization	
Parameter	Value	Parameter	Value
S (corner side)	16 (64×64 grid)	Joint training epochs/batch size	2500/4
α (embedding)	0.55	Joint training optimizer	Adam
$\lambda_h, \lambda_r, \lambda_w, \lambda_t$	0, 1, 0.04, 0.20	Initial learning rate	5×10^{-5}
$\mathrm{SNR}_0, \Delta_{\mathrm{snr}}$	30 dB, 3 dB	Minimum learning rate	5×10^{-6}
$\mathrm{CR}_0, \Delta_{\mathrm{cr}}$	0.8, 0.1	Watermark loss warm-up epochs	100
$\mathrm{CR}_{\min}, \mathrm{CR}_{\max}$	0.2, 0.95	Maximum gradient norm	1.0
$\lvert\mathcal{V}\rvert$	22 (including PAD)	Character-head epochs/batch size	200/4
$L_{\max}$	20 characters	Character-head optimizer/learning rate	Adam/10^{-3}

3.1　实验设置

为验证所提出框架的有效性, 本文在公开数据集 Tiny-ImageNet 上构建实验数据. 实验设置如下所示.

数据集与预处理. 实验使用 Tiny-ImageNet 数据集, 分别随机选取 800 张图像构成训练集和 800 张图像构成测试集, 共 1600 张原始载体图像. 两个集合按原始载体标识保持互斥, 同一载体及其水印派生样本仅归入其中一个集合. 图像与文本沿用同一划分: 训练载体及其对应文本用于训练, 测试载体及其对应文本用于评估, 不将测试图像或对应配对样本用于模型参数更新. 通信阶段图像分辨率统一为 256×256.

语义通信架构. 语义通信部分采用 JSCC-V 网络结构, 对完成双层水印嵌入的隐写图像进行语义编码与传输, 并通过两个解码分支分别恢复图像水印与文本水印.

训练策略. 固定嵌入网络后, 模型采用语义通信与水印恢复联合训练方式, 训练总轮数为 2500 个 epoch, 批大小为 4; 随后仅微调字符预测头 200 个 epoch. 嵌入网络训练、通信联合训练及字符预测头微调均沿用上述训练集, 测试阶段固定全部模型参数.

优化器与学习率. 使用 Adam 优化器进行参数更新, 初始学习率设为 5×10^{-5}, 并采用 CosineAnnealing 学习率调度策略.

信道扰动设置. 为增强模型对信道变化的适应能力, 训练过程中对信噪比与压缩率分别引入 ± 3 dB 与 ± 0.1 的随机扰动.

关键超参数见表 1. 表中 λ_w 和 λ_t 为线性预热结束后的权重, 字符预测头采用独立的优化配置.

水印嵌入与尺度配置. 嵌入阶段图像与掩膜尺寸均为 64×64, 角落区域边长为 $S = \max\{4, \lceil 64 \times 0.25 \rceil\} = 16$, 对应容量为 256 bit (32 byte). 嵌入区域位于图像右下角, 边距为 0, 零起始坐标范围为 $[48, 64) \times [48, 64)$, 嵌入强度设为 $\alpha = 0.55$. 隐写图像采用双三次插值调整至 256×256, 比特标签与掩膜采用最近邻插值. 尺度映射后, 16×16 比特区域对应右下角 64×64 像素区域, 有效载荷容量仍为 256 bit. 隐写图像保存前裁剪至 $[0, 1]$, 通信模型输入采用相同的数值范围.

训练阶段, 信噪比与压缩率分别在 $[27, 33]$ dB 和 $[0.7, 0.9]$ 内均匀采样, 压缩率裁剪区间设为 $[0.2, 0.95]$. 该裁剪区间仅约束训练采样; 消融评估的压缩率参数独立设置为 $[0.2, 1.0]$ 范围内的测试点. 文本载荷采用 "Boyu Han—MM.DD H:MM" 或 "Boyu Han—MM.DD HH:MM" 格式, 按 UTF-8 编码并对齐至 256 bit, 不足时补零, 超出时按字节截断. 字符预测头的最大输出长度为 20, 词表包含 21 个载荷字符及 1 个填充标记 (PAD). 字符索引映射由训练载荷确定, 并在测试阶段保持固定.

评估标准. 对于语义通信重建和第一层图像水印恢复, 采用峰值信噪比 (PSNR) 与结构相似性 (SSIM) 作为指标. 对于第二层文本水印, 采用恢复文本与原始的匹配正确率评估其恢复可靠性. 此外, 模型提取攻击实验中查询预算设置为 $N \in \{250, 500, 1000, 2000, 5000\}$, 其中 $N = 1000$ 为参考设置, 历史会话重复利用检测实验中合法时间窗口阈值设为 $\Delta_T = 10$ min. 安全性实验中的 800 次合法请求与 800 次重复利用请求分别按请求计数,



图 2　(网络版彩图) 双层水印框架的可视化结果. (a) 原载体图像; (b) 嵌入水印图像; (c) 恢复的嵌入图像; (d) 第一层图像水印; (e) 恢复的图像水印; (f) 第二层文本水印; (g) 恢复的文本水印.

Figure 2　(Color online) Visualization of the dual-layer watermarking framework. (a) Original carrier; (b) watermarked image; (c) reconstructed watermarked image; (d) image watermark; (e) recovered image watermark; (f) text watermark; (g) recovered text watermark.

与图像恢复实验的测试样本数区分.

各测试条件下均使用同一组 800 个测试图像与文本配对样本, 训练样本不参与测试统计. 图像重建与第一层水印恢复分别以双层隐写图像和版权水印为参考, PSNR 与 SSIM 逐样本计算后取平均值; 文本正确率以恢复的字符串与目标字符串完全一致为标准. 上述结果反映了既定文本格式在未载体上的恢复性能, 不用于推断未知用户名或文本格式的泛化能力.

3.2　语义通信重建与水印恢复结果展示

为直观展示所提语义通信双层水印框架的恢复效果, 本文给出了典型样本的可视化结果, 如图 2 所示. 图中从左至右依次为原载体图像, 嵌入水印后的图像, 经语义通信信道传输后恢复的嵌入图像, 第一层水印图像以及接收端恢复的水印图像, 右侧同时展示第二层文本水印及其恢复结果. 可以看出, 在嵌入双层水印后, 图像视觉质量与原始载体图像基本一致, 且接收端能够较好地恢复嵌入图像内容并稳定解析两层水印信息.

3.3　主实验分析

表 2 给出了在 Tiny-ImageNet 数据集上不同信道条件与 SNR 设置下的系统性能结果, 本表采用的压缩率为 0.8. 实验同时评估了语义通信模块与图像水印恢复模块的重建性能, 并统计第二层文本水印的恢复正确率.

(1) AWGN 信道下的语义通信与水印恢复性能. 在 AWGN 信道条件下, 随着 SNR 的提升, 语义通信模块的重建质量持续提升. 当 SNR 从 10 dB 提升至 30 dB 时, 语义通信模块的平均 PSNR 由 39.48 dB 提升至 48.36 dB, SSIM 由 99.17% 提升至 99.86%, 表明系统在高信噪比条件下能够实现接近无损的语义通信重建效果. 与此同时, 第一层图像水印的恢复性能也保持稳定增长, PSNR 最高达到 38.48 dB, SSIM 达到 98.94%, 说明嵌入的版权水印在语义通信传输过程中能够被可靠恢复.

(2) Rayleigh 信道下的鲁棒传输性能. 在 Rayleigh 衰落信道条件下, 系统整体性能略低于 AWGN 信道的情形, 但仍保持较高的重建质量. 当 SNR 为 30 dB 时, 语义通信模块的 PSNR 达到 47.46 dB, SSIM 达到 99.84%. 相较于相同信噪比下的 AWGN 信道, PSNR 下降 0.90 dB, SSIM 下降 0.02 个百分点, 这说明在该测试条件下, 衰落扰动对复合图像重建质量的影响较为有限. 同时, 第一层图像水印的恢复性能与 AWGN 信道基本一致, 其

表 2　不同信道与信噪比 (SNR) 条件下的系统性能. 粗体标示两种信道在 30 dB 信噪比下的结果. 此时两种信道的语义重建性能均最优.

Table 2 System performance under different channels and signal-to-noise ratios. Bold marks the best semantic reconstruction for each channel (30 dB).

Channel	SNR (dB)	Semantic reconstruction		Image watermark		Text accuracy (%)
		PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	
AWGN	10	39.48	99.17	36.54	98.46	100.00
	12.5	41.48	99.47	37.25	98.65	100.00
	15	43.33	99.64	37.69	98.76	100.00
	17.5	44.96	99.74	37.97	98.82	100.00
	20	46.29	99.80	38.16	98.86	100.00
	22.5	47.27	99.83	38.29	98.89	100.00
	25	47.90	99.85	38.38	98.91	100.00
	27.5	48.24	99.85	38.44	98.93	100.00
	30	**48.36**	**99.86**	**38.48**	**98.94**	**100.00**
Rayleigh	10	36.18	98.20	36.51	98.47	100.00
	12.5	38.36	98.88	37.22	98.62	100.00
	15	40.48	99.29	37.68	98.79	100.00
	17.5	42.42	99.53	37.96	98.80	100.00
	20	44.07	99.67	38.14	98.81	100.00
	22.5	45.53	99.75	38.27	98.82	100.00
	25	46.56	99.80	38.39	98.93	100.00
	27.5	47.33	99.83	38.43	98.94	100.00
	30	**47.46**	**99.84**	**38.43**	**98.91**	**100.00**

PSNR 和 SSIM 分别达到 38.43 dB 和 98.91%, 表明版权标识信息在衰落信道中仍能得到较完整的保留与恢复. 上述结果表明, 所提出的框架在存在衰落扰动的复杂无线环境中具有良好的鲁棒性与稳定性.

(3) 第二层文本水印的恢复可靠性. 第二层文本水印在表 2 所列信道与信噪比设置下均实现 100% 的恢复正确率, 包括 10 dB 的较低信噪比条件, 表明其在所测信道扰动下具有稳定的恢复性能. 准确恢复的用户名与时间戳用于后续一致性校验、为时间新鲜度检查和历史会话匹配提供了数据基础.

3.4　对比实验分析

为进一步验证所提框架的有效性, 本文在 JSCC-V [33], JSCC [46], JSCC-F [47] 和 JSCC-Q [48] 四种语义通信架构下开展对比实验, 并与未嵌入水印的模型变体及原始语义通信基线进行比较, 结果如图 3 所示. 其中, No embedding 表示去除水印嵌入模块后的模型变体, Baseline 表示对应架构的原始语义通信模型.

图 3 展示了各模型配置下记录的重建指标. 以 JSCC-V [33] 为例, 在 AWGN 信道下, 本文方法的 PSNR 达到 48.36 dB, 对应的 Baseline 和 No embedding 结果分别为 42.13 和 31.94 dB. 在 Rayleigh 信道下, 本文方法仍达到 47.46 dB PSNR 和 99.84% SSIM, 表明其在复杂信道条件下具有良好鲁棒性. 这些结果反映所列模型配置的恢复表现. 本文方法以双层隐写图像为重建参考; 不同方法间的性能归因仍需以重建目标、训练配置与信道资源一致为前提, 因而上述数值不单独证明嵌入水印带来的重建增益.

3.5　压缩率敏感性与模块消融分析

本节先保持模块组成不变, 在 AWGN 与 Rayleigh 信道下改变压缩率, 评价语义通信重建与双层水印恢复对资源设置的敏感性, 如图 4 所示; 随后固定信道与压缩率, 通过模块移除实验分析各组成部分的作用. 两类实验分别评价参数敏感性与模块贡献.

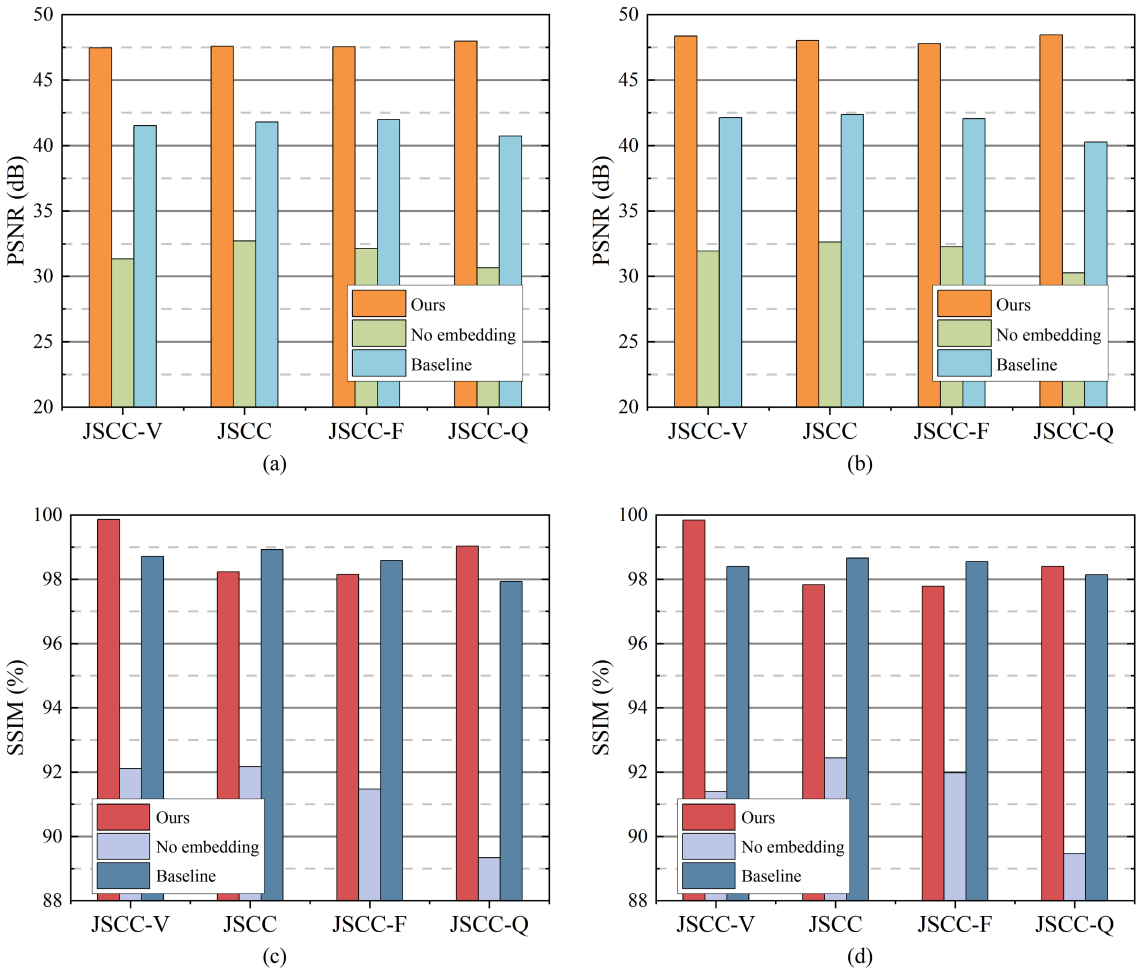

图 3　(网络版彩图) 不同语义通信架构下的重建性能对比. (a) AWGN 信道下的 PSNR; (b) Rayleigh 信道下的 PSNR; (c) AWGN 信道下的 SSIM; (d) Rayleigh 信道下的 SSIM.

Figure 3　(Color online) Reconstruction performance under different semantic communication architectures. (a) PSNR under AWGN; (b) PSNR under Rayleigh; (c) SSIM under AWGN; (d) SSIM under Rayleigh.

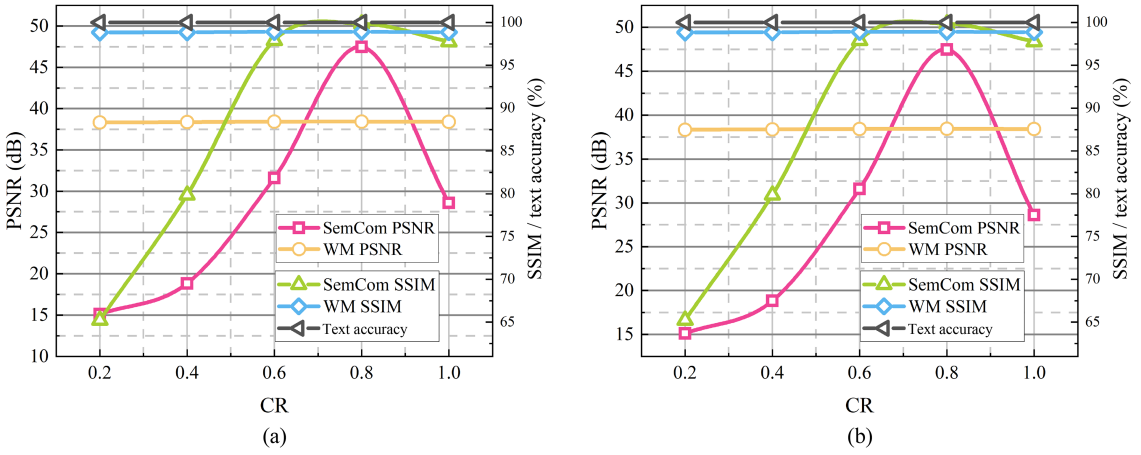

图 4　(网络版彩图) 不同压缩率条件下的敏感性分析. (a) AWGN 信道; (b) Rayleigh 信道.

Figure 4　(Color online) Sensitivity to the compression rate. (a) AWGN channel; (b) Rayleigh channel.

表 3　关键模块消融实验结果. 短横线表示该任务已移除, 相应指标不适用.

Table 3　Ablation results of key modules. A dash indicates that the corresponding task is removed and its metric is not applicable.

Model	Semantic PSNR (dB)	Semantic SSIM (%)	Watermark PSNR (dB)	Watermark SSIM (%)	Text accuracy (%)
Full model	48.36	99.86	38.48	98.94	100.00
Without image watermark	48.50	99.87	–	–	100.00
Without text watermark	48.55	99.87	38.60	98.96	–
Without character head	48.36	99.86	38.48	98.94	92.50
Without Agent scheduling module	48.36	99.86	38.48	98.94	100.00

(1) 压缩率对语义通信重建性能的影响. 实验结果表明, 压缩率对语义通信模块影响最为显著. 在 AWGN 信道下, 当压缩率由 0.2 提升至 0.8 时, 语义通信模块的平均 PSNR 由 15.12 dB 提升至 48.36 dB, 平均 SSIM 由 65.26% 提升至 99.86%. 在 Rayleigh 信道下也呈现相似趋势, 平均 PSNR 由 15.11 dB 提升至 47.46 dB, 平均 SSIM 由 65.23% 提升至 99.84%. 说明随着压缩率增大, 语义通信重建质量明显提升.

(2) 压缩率对双层水印恢复性能的影响. 相比之下, 第一层水印恢复模块对压缩率变化不敏感. 在两类信道条件下, 水印恢复 PSNR 始终保持在约 38.3 ∼ 38.5 dB, SSIM 基本维持在 98.8% ∼ 98.9% 之间, 说明第一层版权水印具有较强鲁棒性. 此外, 第二层文本水印在所有压缩率设置下均实现了 100% 的恢复正确率, 表明其具有稳定的结构恢复能力, 可持续支持会话级身份标识与时间戳追踪.

(3) 最优工作点分析. 需要指出的是, 当压缩率进一步增大至 1.0 时, 语义通信性能较 0.8 反而略有下降, 表明更高压缩率并不一定带来更优传输效果. 综合语义重建质量、水印恢复精度与传输稳定性, 压缩率 0.8 为当前系统的较优工作点, 因此后续实验均采用该设置.

(4) 关键模块消融. 在 Tiny-ImageNet 测试集上固定 AWGN 信道、30 dB 信噪比和 0.8 压缩率, 比较完整模型与去除第一层图像水印、第二层文本水印、字符预测头和 Agent 调度模块的配置, 结果见表 3. 去除任一水印层时, 同时移除该层嵌入与对应恢复任务, 并沿用第 3.1 小节的训练配置分别重新训练, 其余有效任务采用相同训练与测试划分. 去除字符预测头时, 固定底层网络, 对恢复比特作硬判决并按 UTF-8 解码; 去除 Agent 调度模块时同样保留底层权重. 删除任务的指标记为不适用, 不按零值统计. 图像重建指标以各配置对应的通信输入为参考, 因而跨配置的微小差异不单独用于估计相对原载体的嵌入失真.

完整模型的语义重建达到 48.36 dB PSNR 和 99.86% SSIM, 第一层水印恢复达到 38.48 dB PSNR 和 98.94% SSIM, 文本正确率为 100%. 去除第一层或第二层后, 对应通信输入的重建 PSNR 分别为 48.50 和 48.55 dB, 相较完整模型仅有小幅变化, 其余保留任务仍维持较高恢复水平. 去除字符预测头后, 图像与第一层水印指标保持不变, 文本正确率降至 92.50%, 比完整模型低 7.50 个百分点, 说明字符映射在所测恢复比特上改善了完整字符串恢复. 去除 Agent 调度模块不改变底层恢复指标, 符合该层仅控制结果开放、不参与网络恢复的结构设计; 其授权作用需通过第 3.7 小节的模式指标单独评价.

3.6　分级授权流程示例

为说明分级授权流程的输出行为, 本文列出 5 组给定用户角色、认证状态、访问意图及会话状态的请求示例, 如表 4 所示. 这些案例用于展示权限约束如何映射为结果开放模式, 不作为自然语言意图识别准确率或智能体推理能力的实验依据.

(1) 不同角色与访问意图下的分级授权. 表 4 表明, 所提出的 Agent 调度机制能够根据用户角色、认证状态、访问意图及安全检测结果实现差异化访问控制. 其中, 访问者仅获得 $m_j = 0$, 即只返回语义通信恢复结果; 普通所有者在认证通过且请求版权验证时获得 $m_j = 1$, 可访问第一层图像水印; 管理员在认证成功且请求安全取证时可获得最高访问等级 $m_j = 2$, 同时访问语义通信结果、第一层图像水印和第二层文本水印.

(2) 异常场景下的安全约束能力. 在异常场景下, 系统结合认证结果与历史会话重复利用检测实现访问约

表 4　典型访问场景下的分级授权结果.

Table 4　Hierarchical authorization under typical access scenarios.

Role	Auth.	Trust	Intent	Requested mode	Consistency	Replay detected	Output mode
Visitor	0	0.00	View	0	True	False	0
Owner	1	0.82	Copyright	1	True	False	1
Administrator	1	0.99	Forensics	2	True	False	2
Administrator	1	0.99	Forensics	2	True	True	1
Owner	0	0.30	Forensics	2	True	True	0

表 5　上下文授权策略及安全信号消融结果.

Table 5　Contextual authorization policies and security-signal ablations.

Policy	Mode conformity (%)	Excess access (%)	Insufficient access (%)
Full Agent scheduling module	100.00 (10/10)	0.00	0.00
Static role policy	50.00 (5/10)	40.00 (4/10)	10.00 (1/10)
Equivalent contextual rules	100.00 (10/10)	0.00	0.00
Without environmental risk	90.00 (9/10)	10.00 (1/10)	0.00
Without text consistency	90.00 (9/10)	10.00 (1/10)	0.00
Without session replay checks	90.00 (9/10)	10.00 (1/10)	0.00

束. 当管理员重复请求与历史合法会话对应的第二层文本水印时, 系统将访问模式由 $m_j = 2$ 降级为 $m_j = 1$; 当普通所有者密码认证失败且请求时间戳水印时, 系统则将其限制为 $m_j = 0$. 结果表明, 所提出的 Agent 调度机制能够实现分级授权控制, 并对水印信息释放进行有效约束.

3.7　上下文约束授权与静态角色策略的比较

按第 2.2.5 小节的评分与访问约束构建确定性策略实现. 设置 10 个受控场景, 依次为访客查看、所有者版权验证、管理员取证、管理员重复取证、所有者认证失败、管理员仅查看、管理员版权验证、文本一致性失败以及所有者低风险与高风险取证, 测试前指定的期望模式依次为 $(0, 1, 2, 1, 0, 0, 1, 0, 2, 1)$. 访客、所有者与管理员的角色基线信任值分别为 0.2、0.6 和 0.8, 认证增益 $\eta_{\text{auth}} = 0.25$, 风险惩罚权重 $\eta_{\text{risk}} = 0.3$, 信任分级阈值为 0.3 和 0.6; 各场景风险值依次为 $(1, 0.1, 0.2, 0.2, 1, 0.2, 0.2, 0.2, 0.1, 1)$. 除访客查看与所有者认证失败场景外, 其余场景认证通过; 除文本一致性失败场景外, 其余场景一致性校验通过. 认证失败、信任值低于 0.3 或一致性校验失败时输出模式 0; 否则, 信任值在 $[0.3, 0.6)$ 时权限上限为模式 1, 达到 0.6 时上限为模式 2, 最终取请求等级与权限上限的较小值. 重复利用命中时权限上限进一步限制为模式 1. 时间戳新鲜度阈值为 10 min, 各场景独立初始化状态; 管理员重复取证场景预置 5 min 前已使用的会话承诺. 静态角色策略在认证成功后按角色分别输出模式 0、1、2, 认证失败输出模式 0; 等价规则使用与本文上下文策略相同的安全信号和权限约束.

以实际输出模式等于、高于或低于期望模式, 分别定义模式符合率、过量开放率与不足开放率, 结果如表 5 所示. 静态角色策略仅依据认证结果和用户角色分配固定权限, 模式符合率为 50%, 其中 4 个场景存在过量开放, 1 个场景存在不足开放. 上下文约束策略进一步结合请求意图、环境风险、水印一致性与历史会话状态, 在所测场景中均输出符合预期的访问模式, 模式符合率提高至 100%, 过量开放率和不足开放率均降至 0%. 这表明, 相较于仅依据角色分配权限的静态策略, 引入上下文约束能够细化单次请求的授权范围, 减少不必要的水印信息开放及合法取证需求受限的情况. 采用相同安全信号与权限约束的等价规则获得相同结果, 因此上述改善体现了上下文约束的作用, 尚不能据此认定智能体相对等价规则具有额外性能优势. 本实验属于给定安全信号下的策略功能验证, 不涉及自然语言意图识别准确率或端到端授权性能的评价.

进一步在相同的 10 个场景中分别去除环境风险评估、文本一致性校验或历史会话复用检测, 其余策略条件保持不变. 文本一致性失败场景设为管理员取证请求. 三项变体的模式符合率均为 90%, 过量开放率均为 10%,

表 6　不同查询预算的模型提取结果与会话重复利用防御结果. 学生模型指标为 2026、2027 和 2028 三个随机种子下的均值 ± 样本标准差. 教师模型的指标为参考结果. SSIM 和文本正确率的标准差以百分点计. 会话重复利用防御按请求次数单独评估.

Table 6 Model extraction at different query budgets and session replay defense. Student values are the mean ± sample standard deviation over seeds 2026, 2027, and 2028. Teacher values are reference results. Standard deviations of SSIM and text accuracy are in percentage points. Session defense is evaluated separately by request counts.

Model	Semantic PSNR (dB)	Semantic SSIM (%)	Watermark PSNR (dB)	Watermark SSIM (%)	Text accuracy (%)
Teacher	47.46	99.84	38.43	98.91	100.00
Student ($N = 250$)	36.20 ± 0.13	96.90 ± 0.07	7.60 ± 0.08	19.80 ± 0.36	0.00 ± 0.00
Student ($N = 500$)	39.11 ± 0.12	98.08 ± 0.06	7.93 ± 0.09	21.48 ± 0.35	0.00 ± 0.00
Student ($N = 1000$)	41.44 ± 0.08	98.76 ± 0.02	8.14 ± 0.05	22.91 ± 0.18	0.00 ± 0.00
Student ($N = 2000$)	43.20 ± 0.13	99.14 ± 0.04	8.50 ± 0.08	24.60 ± 0.36	0.00 ± 0.00
Student ($N = 5000$)	44.66 ± 0.16	99.44 ± 0.04	9.12 ± 0.09	27.30 ± 0.46	0.00 ± 0.00
Legitimate request pass rate ($\Delta_T = 10\,\text{min}$)				99.88% (799/800)	
Session replay detection rate				100.00% (800/800)	

不足开放率均为 0%. 去除风险评估使高风险取证请求获得超过预定上限的输出; 去除一致性或会话复用检测分别使相应异常请求获得过量开放. 因此, 所列安全信号分别约束不同类型的请求, 但这一功能性消融仍不说明智能体相对等价上下文规则具有额外优势.

3.8　安全性指标与理论分析

在认证、会话登记表、时钟及输出接口可信, 且会话检查与状态更新具有原子性的前提下, 开放第二层文本水印须满足认证成功、信任评分达到取证阈值、一致性校验通过及未命中历史会话重复利用检测, 即

$$m_j = 2 \implies \text{auth}_j = 1 \ \wedge \ T_j \geqslant \theta_2 \ \wedge \ g_j = 1 \ \wedge \ \gamma_j = 0, \tag{49}$$

其中, θ_2 为取证权限阈值. 对于一致性校验通过但时间戳无效或已使用的会话, 有 $\gamma_j = 1$, 输出权限上限为模式 1. 该约束针对显式水印解码结果, 模式 0 返回的复合图像仍包含嵌入水印.

采用合法请求通过率 (LPR)、历史会话重复利用检测率 (RDR) 及重复利用请求误接受率 (FAR) 评价时间戳防御, 定义为

$$\begin{aligned} \text{LPR} &= \frac{N_{\text{legit,pass}}}{N_{\text{legit}}}, \\ \text{RDR} &= \frac{N_{\text{replay,detect}}}{N_{\text{replay}}}, \\ \text{FAR} &= \frac{N_{\text{replay,pass}}}{N_{\text{replay}}}, \end{aligned} \tag{50}$$

其中, N_{legit} 和 N_{replay} 分别为合法请求与重复利用请求总数, 三个分子依次为合法请求通过数、重复利用请求检出数及误接受数; 通过或误接受以是否开放模式 2 为判据.

表 6 显示, 合法请求通过率为 99.88%, 历史会话重复利用检测率为 100%, 对应所测重复利用请求的经验误接受率为 0%. 该结果限于测试请求集合, 未单独估计时间窗内重复利用的检测性能. 在 $N = 1000$ 的查询预算下, 学生模型的第一层水印恢复 PSNR 和 SSIM 相较于教师模型分别下降 30.29 dB 和 76.00 个百分点, 文本正确率由 100% 降至 0%, 上述差异对应仅监督语义输出的蒸馏配置, 水印指标来自验证者冻结的教师恢复链路, 其与学生表征可能不兼容, 因而该对比本身不构成对其他提取策略的安全保证. 所有权判定的误接受率仍需通过无水印及错误水印等负例另行评估.

3.9　鲁棒性实验分析

为验证所提双层水印语义通信框架在安全场景中的鲁棒性, 本文设计了模型提取攻击与基于时间戳的历史

表 7　模型提取攻击实验配置.

Table 7　Configuration of the model extraction experiments.

Parameter	Setting
Teacher	Full dual-layer watermarking JSCC-V
Student	Independently initialized JSCC-V; same layer hierarchy, kernel sizes, strides, and residual blocks as the teacher
Student width and output	50% of teacher hidden channels; output projection preserves the semantic feature dimensions
Query budgets	$N \in \{250, 500, 1000, 2000, 5000\}$ distinct samples
Query carriers	Tiny-ImageNet images disjoint from teacher training and test carriers
Watermark embedding	No pre-embedded watermark in query carriers; embedding occurs inside the teacher service
Attacker-visible output	Mode 0 reconstruction; no watermark labels or private teacher watermark decoder parameters
Distillation target	Public semantic reconstruction outputs of the teacher
Distillation loss	$0.8\,\mathrm{MSE} + 0.2(1 - \text{MS-SSIM})$
Optimizer/initial learning rate	Adam/10^{-4}
Learning rate schedule	CosineAnnealing
Epochs/batch size	300/8
Channel/compression rate	30 dB Rayleigh/CR = 0.8
Test set	The same 800 held-out image–text pairs for every run
Query counting and caching	One teacher call per distinct sample; cached responses reused for all 300 student epochs
Ownership verification	Frozen teacher watermark recovery chain held by the verifier; not available to the attacker or used in distillation
Repeated runs	Seeds 2026, 2027, and 2028; nested budgets within each run and disjoint query pools across runs
Aggregation	First average over 800 samples within each run, then report the three-run mean and sample standard deviation

会话重复利用防御两类实验, 结果如表 6 所示. 其中, 模型提取攻击实验在 30 dB Rayleigh 信道下进行.

模型提取实验配置汇总于表 7. 查询预算 N 定义为独立查询样本数量, 每个样本对应一次教师前向调用. 在每组重复实验内, 先确定 5000 个查询载体, 再构造大小为 250、500、1000、2000 和 5000 的嵌套查询集. 查询载体来自独立于教师训练集和测试集的 Tiny-ImageNet 图像, 本身不含预嵌入水印, 双层水印由教师服务内部加入. 攻击者只获得模式 0 的复合图像重建输出, 无法获得显式水印标签或教师内部水印解码器参数; 模式 0 不等同于图像内部没有水印. 教师对同组每个不同查询样本只执行一次推理并缓存输入及响应, 较小预算复用对应子集的缓存. 因此, 对给定预算 N, 学生可用教师响应数为 N, 后续 300 轮训练均复用该缓存, 不产生额外教师调用.

学生以 JSCC-V 为基础独立初始化, 保留教师的网络层级、卷积核尺寸、步长及残差块配置, 将内部隐藏层通道数缩减为对应教师层的 50%, 并通过输出投影保持与教师一致的语义特征维度. 蒸馏训练仅拟合教师公开的语义重建响应, 不使用图像或文本水印的显式恢复标签. 所有权验证由验证者在训练后执行: 验证者使用其保存的冻结教师水印恢复链路评价学生语义表征, 该链路不向攻击者开放, 也不参与学生蒸馏优化. 因而水印指标反映固定验证链路与学生表征的兼容性. 记第 i 个查询输入为 $\mathbf{x}_i^q$, 教师可见响应为 $\mathbf{y}_i^T$, 学生语义重建为 $\hat{\mathbf{y}}_i^S$, 蒸馏目标定义为

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{N}\sum_{i=1}^{N}\left[0.8\,\mathrm{MSE}(\hat{\mathbf{y}}_i^S, \mathbf{y}_i^T) + 0.2\left(1 - \text{MS-SSIM}(\hat{\mathbf{y}}_i^S, \mathbf{y}_i^T)\right)\right]. \tag{51}$$

学生以 Adam 优化 300 个 epoch, 批大小为 8, 初始学习率为 10^{-4}, 采用 CosineAnnealing 调整学习率. 使用随机种子 2026、2027 和 2028 进行三组独立重复实验; 每组独立选择查询载体并初始化学生, 三组最大查询集互不重叠, 每组内部各预算的查询集保持嵌套, 即组内嵌套、组间互斥. 所有学生均在同一组 800 个测试图文对上评估, 测试载体不参与任何查询集或模型训练. 每组先逐样本计算 PSNR、SSIM 和字符串完全匹配结果, 再在 800 个

样本上汇总. 表 6 中学生模型的各项指标均以三组独立重复实验结果的均值 ± 样本标准差表示. 计算样本标准差时, 分母取 $n-1$, 其中 $n=3$ 为独立重复实验次数. SSIM 和文本正确率的标准差以百分点计. 三组均使用同一测试集, 不将其合并计为 2400 个独立测试载体.

(1) 模型提取攻击下的鲁棒性分析. 如表 6 所示, 随着查询预算由 250 增至 5000, 学生语义重建 PSNR 由 36.20 dB 增至 44.66 dB, SSIM 由 96.90% 增至 99.44%, 表明增加可见语义样本有助于拟合公开重建任务. 当 $N=1000$ 时, 学生语义重建为 41.44 dB PSNR 和 98.76% SSIM, 第一层水印输出为 8.14 dB PSNR 和 22.91% SSIM, 文本正确率为 0.00%. 当 $N=5000$ 时, 图像水印指标为 9.12 dB 和 27.30%, 文本正确率仍为 0%. 这些结果描述了无显式水印监督条件下的任务迁移差异; 由于验证者冻结的教师恢复链路可能与学生表征不兼容, 低水印指标不能单独归因于所提水印的抗提取能力, 亦不能据此推断在获得水印监督、增加提取器训练或采用其他攻击策略时仍有相同表现.

(2) 历史会话重复利用防御分析. 在基于时间戳的历史会话重复利用防御实验中, 系统通过对第二层文本水印中的时间戳信息和会话记录进行一致性校验与复用检测, 实现对异常请求的识别与拦截. 结果表明, 在 800 次合法请求中, 系统正确通过 799 次, 合法请求通过率为 99.88%; 在 800 次模拟历史会话重复利用请求中, 系统检测并拦截全部异常请求, 检测率达到 100.00%. 该结果表明, 所设计的时间戳文本水印机制能够有效识别会话级敏感信息的重复利用行为.

综合来看, 所提框架在所测条件下兼顾复合图像重建与双层水印恢复, 上下文授权约束限制敏感恢复结果的开放. 提取实验说明了仅监督公开语义输出时的学生任务表现, 固定验证链路上的水印输出受蒸馏监督范围及表征兼容性共同影响, 不据此推断模型提取已被阻止.

4　结论

本文面向语义通信中的版权保护、会话级安全追踪与按需授权需求, 提出融合双层水印与上下文分级授权的一体化框架. 第一层水印用于版权标识, 第二层文本水印用于用户与会话时间标记. 通过固定嵌入网络后的语义通信与水印恢复联合训练以及字符预测头微调, 系统在所测信道条件下实现了较高的复合图像重建质量与稳定的水印恢复. 上层通过显式认证、信任与会话约束控制恢复结果的开放, 实验支持这些上下文约束在给定请求场景中的作用, 不支持智能体相对等价规则的额外性能优势. 模块消融表明, 在既定协议下去除字符预测头降低了文本恢复正确率, 而移除 Agent 调度模块不改变底层恢复指标, 其作用体现为授权模式的变化. 不同查询预算的蒸馏实验报告了公开语义任务拟合程度与固定教师恢复链路上的水印输出, 相关结论限于所述攻击设置. 后续仍需通过获得水印监督或针对恢复分支优化的攻击评价其安全边界.

参考文献

1 Wang C X, You X, Gao X, et al. On the road to 6G: visions, requirements, key technologies, and testbeds. IEEE Commun Surv Tut, 2023, 25: 905–974

2 Chen W, Lin X, Lee J, et al. 5G-advanced toward 6G: past, present, and future. IEEE J Sel Area Comm, 2023, 41: 1592–1619

3 Zhang Y C, Zhang P, Wei J B, et al. Semantic communication for intelligent devices: architectures and a paradigm. Sci Sin Inform, 2022, 52: 907–921 [张亦弛, 张平, 魏急波, 等. 面向智能体的语义通信: 架构与范例. 中国科学: 信息科学, 2022, 52: 907–921]

4 Banafaa M, Shayea I, Din J, et al. 6G mobile communication technology: requirements, targets, applications, challenges, advantages, and opportunities. Alexandria Eng J, 2023, 64: 245–274

5 Cui Q M, You X H, Wei N, et al. Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. Sci China Inf Sci, 2025, 68: 171301

6 Xu X D, Han S J, Zhao Y, et al. 6G wireless coverage extension technologies. Sci Sin Inform, 2024, 54: 1144–1176 [许晓东, 韩书君, 赵芸, 等. 6G 无线覆盖扩展技术. 中国科学: 信息科学, 2024, 54: 1144–1176]

7 Yang W, Du H, Liew Z Q, et al. Semantic communications for future internet: fundamentals, applications, and challenges. IEEE Commun Surv Tut, 2022, 25: 213–250

8 Luo X, Chen H H, Guo Q. Semantic communications: overview, open issues, and future research directions. IEEE Wireless Commun, 2022, 29: 210–219
9 Chaccour C, Saad W, Debbah M, et al. Less data, more knowledge: building next generation semantic communication networks. IEEE Commun Surv Tutor, 2025, 27: 37–76
10 Sagduyu Y E, Erpek T, Yener A, et al. Will 6G be semantic communications? Opportunities and challenges from task oriented and secure communications to integrated sensing. IEEE Netw, 2024, 38: 72–80
11 Yang Z, Chen M, Li G, et al. Secure semantic communications: fundamentals and challenges. IEEE Netw, 2024, 38: 513–520
12 Yang X, Lai Y, Li G, et al. Toward polymorphic backdoor against semantic communication via intensity-based poisoning. IEEE Trans Inf Foren Secur, 2026, 21: 4505–4520
13 Saad W, Bennis M, Chen M. A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Netw, 2020, 34: 134–142
14 Hosny K M, Magdi A, ElKomy O, et al. Digital image watermarking using deep learning: a survey. Comput Sci Rev, 2024, 53: 100662
15 Awasthi D, Tiwari A, Khare P, et al. A comprehensive review on optimization-based image watermarking techniques for copyright protection. Expert Syst Appl, 2024, 242: 122830
16 Sharma S, Zou J J, Fang G, et al. A review of image watermarking for identity protection and verification. Multimed Tools Appl, 2024, 83: 31829–31891
17 Singh H K, Singh A K. Digital image watermarking using deep learning. Multimed Tools Appl, 2024, 83: 2979–2994
18 Dathathri S, See A, Ghaisas S, et al. Scalable watermarking for identifying large language model outputs. Nature, 2024, 634: 818–823
19 Yang X, He Y, Li G, et al. WatCom: unconscious watermarking for semantic communication intellectual property protection. In: Proceedings of IEEE International Conference on Communications (IEEE ICC), 2025. 1181–1187
20 Yu Y, Yang W, Ding W, et al. Reinforcement learning solution for cyber-physical systems security against replay attacks. IEEE Trans Inf Foren Secur, 2023, 18: 2583–2595
21 Lu K, Zhou Q, Li R, et al. Rethinking modern communication from semantic coding to semantic communication. IEEE Wireless Commun, 2022, 30: 158–164
22 Tang F, Chen X, Zhao M, et al. The roadmap of communication and networking in 6G for the metaverse. IEEE Wireless Commun, 2022, 30: 72–81
23 Du H, Wang J, Niyato D, et al. Rethinking wireless communication security in semantic Internet of Things. IEEE Wireless Commun, 2023, 30: 36–43
24 Sapkota R, Roumeliotis K I, Karkee M. AI agents vs. agentic AI: a conceptual taxonomy, applications and challenges. Inform Fusion, 2025, 126: 103599
25 Hosseini S, Seilani H. The role of agentic AI in shaping a smart future: a systematic review. Array, 2025, 26: 100399
26 Salameh A I, El Tarhuni M. From 5G to 6G—challenges, technologies, and applications. Future Internet, 2022, 14: 117
27 Chen B, Li G, Wu J, et al. AgentChain: blockchain-empowered multi-agent coordination for trustworthy LLM question-answering systems. IEEE Trans Depend Secure, 2026, 23: 8501–8518
28 Liao J X, Fu X Y, Qi Q, et al. 6G-ADM: knowledge based 6G network management and control architecture. J Commun, 2022, 43: 3–15 [廖建新, 付霄元, 戚琦, 等. 6G-ADM: 基于知识空间的 6G 网络管控体系. 通信学报, 2022, 43: 3–15.]
29 Ma G, Khalili M, Parthiban R, et al. A low-complexity handover scheme using unsupervised learning techniques for 6G multi-networking. In: Proceedings of International Conference on 6GNet (6GNet), 2023. 1–5
30 Wang X M, Ren Y Z, Jiang T Y, et al. Survey on key technologies for native-intelligence-driven spectrum management towards 6G. J Commun, 2026, 47: 1–19 [王先梅, 任语铮, 姜天宇, 等. 面向 6G 的内生智能频谱管控关键技术研究. 通信学报, 2026, 47: 1–19.]
31 Han B, Li G, Yang X, et al. When steganography meets semantic communication: an authorizable data transmission paradigm. IEEE Trans Mobile Comput, 2026, 25: 16633–16650
32 Yang X, Lai Y, Li G, et al. SemanAegis: toward credential-aware semantic communication against knowledge leakage threats. IEEE Trans Mobile Comput, 2026, 25: 5032–5049
33 Zhang W, Zhang H, Ma H, et al. Predictive and adaptive deep coding for wireless image transmission in semantic communication. IEEE Trans Wireless Commun, 2023, 22: 5486–5501
34 Zhang W, Wang Y, Chen M, et al. Optimization of image transmission in cooperative semantic communication networks. IEEE Trans Wireless Commun, 2023, 23: 861–873
35 Yi P, Cao Y, Kang X, et al. Deep learning-empowered semantic communication systems with a shared knowledge base. IEEE Trans Wireless Commun, 2023, 23: 6174–6187
36 Peng X, Qin Z, Tao X, et al. A robust image semantic communication system with multi-scale vision transformer. IEEE J Sel

Area Comm, 2025, 43: 1278–1291
37 Tyagi T, Singh K N, Singh A K, et al. DeepVerifier: robust watermarking of deep neural networks based on black-box and white-box reasoning. IEEE Trans Comput Social Syst, 2025. 1–10
38 Guo D, Zhu X, Li F, et al. DoBMark: a double-branch network for screen-shooting resilient image watermarking. Expert Syst Appl, 2024, 246: 123159
39 Nie H, Lu S. FedCRMW: federated model ownership verification with compression-resistant model watermarking. Expert Syst Appl, 2024, 249: 123776
40 Xu R, Hu M, Lei D, et al. InvisMark: invisible and robust watermarking for AI-generated image provenance. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025. 909–918
41 Liu G, Xiang R, Liu J, et al. An invisible and robust watermarking scheme using convolutional neural networks. Expert Syst Appl, 2022, 210: 118529
42 Gu W, Li Z, Hasegawa S, et al. An agentic framework to achieve systematic facilitation in group consensus building. In: Proceedings of IEEE International Conference on Agentic AI (ICA), 2025. 252–254
43 Pandey K, Narnag A. The agentic enterprise: a strategic analysis of advanced agentic workflows and collaborative AI. In: Proceedings of the International Conference on Foundation and Large Language Models (FLLM), 2025. 394–404
44 Gody R, Goudy M, Tawfik A Y. ConvoGen: enhancing conversational AI with synthetic data: a multi-agent approach. In: Proceedings of IEEE Conference on Artificial Intelligence (IEEE CAI), 2025. 252–257
45 Duan Q, Lu Z. Agent communications toward agentic AI at edge—a case study of the Agent2Agent protocol. In: Proceedings of IEEE International Conference on Edge Computing and Scalable Cloud (IEEE EdgeCom), 2025. 150–155
46 Bourtsoulatze E, Kurka D B, Gündüz D. Deep joint source-channel coding for wireless image transmission. IEEE T Cogn Commun, 2019, 5: 567–579
47 Kurka D B, Gündüz D. DeepJSCC-F: deep joint source-channel coding of images with feedback. IEEE J Sel Areas Inf Theor, 2020, 1: 178–193
48 Tung T Y, Kurka D B, Jankowski M, et al. DeepJSCC-Q: constellation constrained deep joint source-channel coding. IEEE J Sel Areas Inf Theor, 2022, 3: 720–731

Research on dynamic watermark embedding and recovery for agentic AI-driven semantic communication encoder–decoder models

Boyu HAN[1], Gaolei LI[1,2*], Xiao YANG[1], Jianhua LI[1,2] & Hao ZHANG[3]

1. *School of Computer Science (School of Cyber Science and Engineering, School of Cryptography), Shanghai Jiao Tong University, Shanghai 200240, China*
2. *Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China*
3. *China Information Technology Security Evaluation Center, Beijing 100085, China*
* Corresponding author. E-mail: gaolei_li@sjtu.edu.cn

Abstract To address copyright protection and on-demand authorization in 6G semantic communications, this paper proposes a semantic communication framework integrating dual-layer watermarking with context-constrained hierarchical authorization. An image watermark and a username + timestamp text watermark are embedded in the end-to-end communication link for copyright identification and session-level temporal tagging. The upper layer follows an Agentic AI-inspired workflow organization and applies deterministic constraints to user roles, authentication, environmental risk, request intent, watermark consistency, and session history to select image reconstruction, image watermark verification, or dual-layer watermark decoding. Under a 30 dB AWGN channel with a compression-rate parameter of 0.8, semantic reconstruction achieves 48.36 dB PSNR and 99.86% SSIM, while image watermark recovery achieves 38.48 dB PSNR and 98.94% SSIM. The text watermark is recovered exactly on all 800 test image–text pairs. The timestamp-based session replay experiment yields a legitimate request pass rate of 99.88% and a replay detection rate of 100%. Under public-output-only semantic distillation with a query budget of $N = 1000$, and with fixed owner-side watermark decoders applied to student representations, image watermark recovery reaches 8.14 dB PSNR and 22.91% SSIM, with text recovery accuracy of 0.00%. In 10 controlled authorization scenarios, both the context-constrained policy and equivalent rules achieve 100% mode conformity. These results support the role of context constraints in authorization, without establishing an additional advantage of an agent over equivalent rules.

Keywords 6G, semantic communication, Agentic AI, watermarking and copyright protection, intelligent scheduling