华 中 科 技 大 学

网络空间安全学院

本科：多媒体数据安全课程报告

论文题目：*AUDIO WATERMARK: Dynamic and Harmless Watermark for Black-box Voice Dataset Copyright Protection*

姓　　名_____

班　　级_____

学　　号_____

联系方式_____

2025 年 5 月 20 日

课程报告要求

1. 报告不可以抄袭，发现雷同者记为 **0** 分。

2. 报告中不可以只粘贴大段代码或者大段文字，应是文字与图、表结合的，需要说明流程的时候，也应该用流程图或者伪代码来说明；如果发现有大段代码和文字粘贴者，报告打回重写。

3. 报告格式要求规范。

报告评分表

评分项目	分值	评分标准	得分
论文的研究背景与主要贡献	10	10-9：研究背景阐述全面，主要贡献提炼精准，逻辑严密，与背景形成呼应； 8-6：研究背景和贡献基本覆盖，无明显逻辑错误； 5-0：描述简单或逻辑混乱。	
系统架构	5	5：系统架构介绍清晰完整，解释到位。 4-3：系统架构介绍基本完整。 2-0：系统架构介绍错误或缺失。	
研究现状	10	10-9：文献引用充分，逻辑通顺，归纳合理，总结精炼； 8-6：文献引用基本覆盖，无逻辑错误； 5-0：描述简单或逻辑混乱。	
论文方案具体介绍	25	25-20：方案核心步骤与技术细节描述清晰； 19-12：方案概述完整但细节模糊； 11-0：描述简单或逻辑混乱。	
论文实验结果	10	10-8：实验环境、数据及对比结果描述完整，结论可靠； 7-5：实验结论正确，实验结果基本完整； 4-0：描述简单或逻辑混乱。	
论文阅读心得	30	30-25：思考深入，能结合自身知识提出批判性见解或扩展方向，体现独立思考； 24-18：心得较笼统，以总结为主； 17-0：心得过于简单或逻辑混乱。	
格式与表达规范	10	图、表的说明，行间距、缩进、目录等，一种不规范扣 1 分；要求达到 2500 字，每缺少一百字扣 2 分。	
总　分			
评分人：			

目 录

一. 研究背景与主要贡献

1.1. 研究背景

随着深度神经网络（DNNs）在语音识别、图像处理等领域的广泛应用，公开数据集（如 VoxCeleb、TIMIT 等）为研究人员提供了丰富的资源。然而，这些数据集通常限制仅用于学术或教育用途，禁止未经授权的商业使用。语音数据因其包含个人身份信息，具有较高的隐私风险，近年来数据所有者对数据滥用的担忧日益增加。传统的版权保护方法（如加密、成员推理和后门攻击）存在显著缺陷，例如限制数据访问、引入有害后门、降低音频质量或易被检测等。因此，开发一种动态、无害、适用于黑盒模型的语音数据集版权保护方法成为迫切需求。

论文 *AUDIO WATERMARK: Dynamic and Harmless Watermark for Black-box Voice Dataset Copyright Protection* 针对这一问题，提出了一种创新的语音数据集水印保护方案，旨在解决现有方法的局限性，确保数据所有者能够有效验证数据集使用情况，同时不损害模型性能或音频质量。

1.2. 主要贡献

论文提出了以下主要贡献：

1. 创新无害语音水印方案：这是首个针对语音数据集的动态、无害版权保护方法，通过使用域外（OOD）特征生成水印，确保水印不会改变原始标签，从而避免引入有害后门。

2. 动态水印设计：通过风格转换生成模型（Style Wave-U-Net）和随机参考风格，生成动态水印，增强水印的隐蔽性和抗攻击能力。

3. 黑盒设置下的高效性：采用双层对抗优化策略，训练广义代理模型以增强水印对多种目标模型的适应性，无需了解目标模型的具体架构或训练细节。

4. 多方面实验验证：在 2 个语音数据集（LibriSpeech 和 VoxCeleb）上测试了 10 个说话人识别模型，与 10 种现有保护方法进行比较，并针对 8 种攻击场景验证了水印的鲁棒性，取得了近 100%的良性准确率（BA）、95%的验证成功率（VSR）和对所有测试攻击的抵抗能力。

二. 系统架构

论文提出的 AUDIO WATERMARK 系统架构包含三个主要组成部分，描述了水印生成、数据集构建和所有权验证的完整流程，解释了各部分的功能和相互关系，如下图所示：

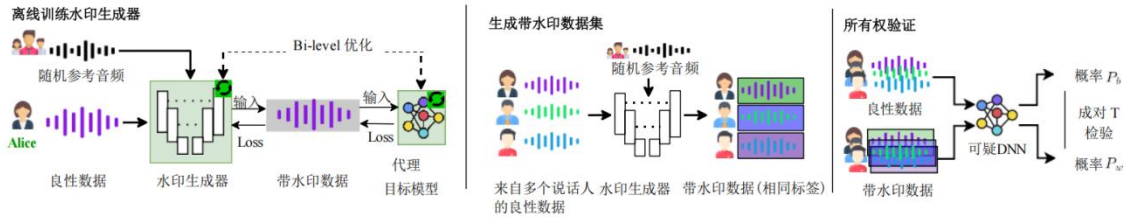

图 2-1 AUDIO WATERMARK 系统架构

1. 离线训练水印生成器：优化一个生成模型（Style Wave-U-Net）以产生动态水印，利用参考音频和原始音频生成具有风格转换的水印音频。

2. 生成水印数据集：使用训练好的水印生成器为部分音频样本添加水印，生成包含水印和干净样本的混合数据集，比例由中毒率（poison rate）控制。

3. 所有权验证：通过向可疑模型输入水印音频和干净音频，比较模型在两种输入上的预测概率（P_w和P_b），利用假设检验判断模型是否使用了水印数据集。

三. 研究现状

3.1. 现有研究综述

论文对语音数据集版权保护的相关工作进行了全面回顾，涵盖以下三类方法：

- **加密方法**：相关文献论述到加密方法通过加密整个数据集或敏感信息来保护数据，但需要解密密钥，限制了数据可访问性，影响开源数据集的可用性。
- **成员推理方法**：如 SIMASR 通过分析模型输出判断数据是否用于训练，但需要大量查询目标模型，且易受数据不平衡影响，导致较高的误报率。
- **后门水印方法**：包括脏标签后门（如 FreqTone、AdvBackdoo）和干净标签后门（如 Masterke），通过嵌入触发器验证数据集使用。然而，这些方法通常引入有害后门（如改变预测标签），降低音频质量，且固定触发器易被检测。

表 3-1 总结了这些方法的特性，指出它们在数据集可访问性、模型无关性、训练无关性、查询需求、无害性、适应性和攻击抵抗力等方面的局限性。相比之下，AUDIO WATERMARK 通过动态、无害的水印设计，克服了这些问题。

表 3-1 AUDIO WATERMARK 与其他方法的比较

分类	方法	数据集可访问性	模型独立性	训练独立性	最小查询	无害性	攻击适应性	语音鲁棒性	语音质量	验证准确性
加密	Encrypt Retrieval	✗	✓	✓	✓	✓	✗	✓	高	高
	Speech Encrypt	✗	✓	✓	✗	✓	✗	✗	高	高
成员推理	SLMIA-SR	✓	✓	✓	✗	✓	✗	✗	高	中
脏标签后门	FreqTon	✓	✓	✓	✓	✗	✗	✗	低	高
	UltraSoun	✓	✓	✓	✓	✗	✗	✗	高	高
	AdvBackdoor	✓	✗	✗	✗	✗	✓	✗	中	高
干净标签后门	Masterkey	✓	✓	✗	✓	✗	✓	✗	低	高
	AUDIO WATERMARK	✓	✓	✓	✓	✓	✓	✓	**中**	**高**

四. 论文方案具体介绍

4.1. 方案核心步骤

AUDIO WATERMARK 方案围绕解决三个主要挑战（C1: 无害水印、C2: 黑盒模型适应性、C3: 动态水印抗攻击性）展开，包含以下核心步骤：

1. 无害设计：

目标：确保水印不引入有害后门，即水印模型的预测$\hat{f}(x_i+\delta)=y_i$，与原始标签一致。

方法：利用域外（OOD）特征生成水印，使水印模型正确识别水印样本，而未训练过水印的良性模型会错误分类（$f(x_i+\delta)\neq y_i$）。

优化目标：最小化有害度H，通过下面的公式（1）优化水印δ，使水印模型准确预测，而良性模型误判。

$$\min_{\delta}\frac{1}{N}\left(\sum_{i=1}^{N}I\{\hat{f}(x_i+\delta)\neq y_i\}-\sum_{i=1}^{N}I\{f(x_i+\delta)=y_i\}\right) \tag{1}$$

2. 动态水印生成：

Style Wave-U-Net：设计双通道生成模型，结合参考音频的风格嵌入和原始音频，通过下采样（DS）和上采样（US）网络生成动态水印，参考图 4-1。

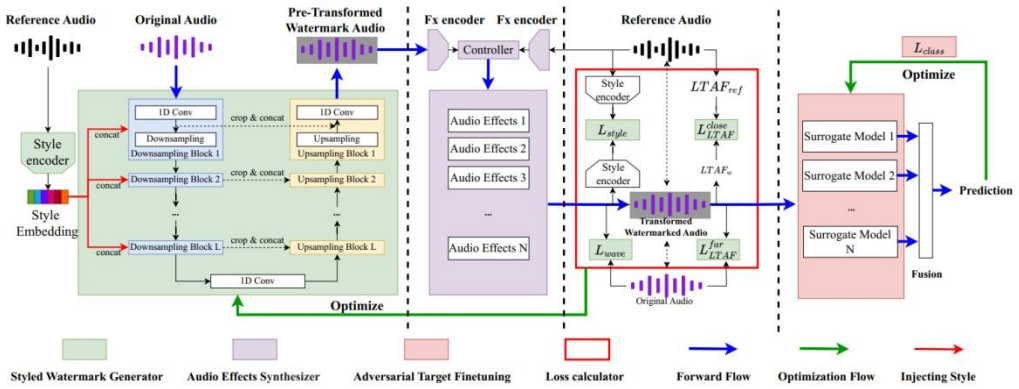

图 4-1 AUDIO WATERMARK 的设计组成

逆对比 LTAF 损失：优化水印生成器，使水印音频的 LTAF 特征与原始音频显著不同（同说话人时）或接近参考音频（不同说话人时），公式（2-4）。

$$L_{LTAF}^{far} = \left(\left|Norm\left(\mathcal{F}(\hat{x})\right) - Norm\left(\mathcal{F}(x)\right)\right| + \epsilon\right)^{-1}, \tag{2}$$

$$L_{LTAF}^{close} = \left|\left|Norm\left(\mathcal{F}(\hat{x})\right) - Norm\left(\mathcal{F}(r)\right)\right|\right|^{2}, \tag{3}$$

$$L_{LTAF} = \frac{1}{2N}\sum_{i=1}^{N}\left(d_i \cdot L_{LTAF}^{far} + (1-d_i)\cdot \max\left(0, L_{LTAF}^{close}\right)\right). \tag{4}$$

3. 音频质量保持：

频率均衡器：使用 DeepAFX-ST 调整水印音频的频率特性，确保与参考音频风格一致，公式（5）。

$$L_{style} = \left|\left|\mathcal{E}(\hat{x}) - \mathcal{E}(r)\right|\right|^{2}, \tag{5}$$

语义调控：通过最大均值差异（MMD）损失（公式 6）保持语义完整性。

$$L_{mmd} = \frac{1}{K}\sum_{k=1}^{K}\left(\left|\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\phi\left(x_i^k\right) - \frac{1}{n_k}\sum_{i=1}^{n_k}\phi\left(\hat{x}_i^k\right)\right|\right|^2\right), \tag{6}$$

波形调控：通过 MAE 损失（公式 7）限制波形变化，保持音频能量。

$$L_{wave} = \left|\,\hat{x} - x\,\right|, \tag{7}$$

综合损失：

$$L_{\text{total}} = L_{LTAF} + \alpha * L_{\text{style}} + \beta L_{\text{mmd}} + \gamma L_{\text{wave}} \tag{8}$$

4. 可迁移性设计：

采用双层对抗优化策略，训练水印生成器和多个代理模型（ResNet-18、VGG-M 等），通过L_{class}优化代理模型以提高水印对未知模型的适应性。

4.2. 技术细节

- **Style Wave-U-Net**：基于 Wave-U-Net 和 GST-Tacotron 的风格编码器，通过多层次风格嵌入生成动态水印，确保隐蔽性和多样性。
- **LTAF 特征**：利用长期平均场指纹（LTAF）表示说话人身份，通过逆对比损失优化水印的区分性。
- **双层优化**：水印生成器和代理模型通过对抗训练相互提升，水印生成器生成更隐蔽的水印，代理模型提高识别能力。

五. 实验结果

5.1. 实验环境与数据集

实验在两个公开语音数据集上进行：

- **LibriSpeech**：包含 363.6 小时音频，921 个说话人。
- **VoxCeleb**：包含 100,000 条语音，1,251 个名人。 目标模型包括 10 种说话人识别模型（如 ResNet-18、ECAPA-TDNN 等），代理模型融合了 ResNet-18、VGG-M 等四种模型。

5.2. 评估指标

良性准确率（BA）：模型在干净测试集上的准确率。

验证成功率（VSR）：通过成对 T 检验比较水印和干净样本的预测概率（P_w和P_b），判断模型是否使用水印数据集。

有害度（H）：衡量水印引入的后门风险，公式（9）。

$$H = \frac{1}{N}\sum_{i=1}^{N} I\{\hat{f}(\hat{x_i}) \neq y_i\} \tag{9}$$

Mel 倒谱失真（MCD）：评估水印音频的失真程度。

5.3. 实验结果

1. 基准结果：

表 5-1 AUDIO WATERMARK 与现存的水印基准对比结果表

	保护方法	LibriSpeech				VoxCeleb			
		BA (%)	VSR (%)	有害度	MCD (dB)	BA (%)	VSR (%)	有害度	MCD (dB)
脏标签	FreqTone	98.5	100	0.98	8.1	92.5	100	0.99	7.5
	Backnoise	95.4	87.2	0.84	7.8	91.2	89.3	0.86	7.7
	AdvBackdoor	88.1	100	0.99	13.4	85.2	98.2	0.94	12.1
	BadNets	80.9	100	0.95	6.5	85.1	100	0.99	6.4
	Blended	90.4	61.8	0.74	7.2	91.2	64.2	0.75	6.8
	WaNet	91.6	19.4	0.24	7.5	90.8	24.3	0.29	6.6
	ReFool	80.6	73.5	0.79	7.9	82.1	74.6	0.82	8.2
干净标签	Label-Consistent	90.5	95.2	0.77	12.2	91.5	75.2	0.84	12.9
	Sleeper Agent	88.7	71.2	0.82	6.0	83.4	69.4	0.81	6.8
	Domain Watermark	12.6	78.4	0.05	15.9	15.5	85.1	0.04	14.5
	Audio Watermark	96.4	95.5	0.06	9.6	97.6	94.5	0.03	9.2

在 LibriSpeech 上，AUDIO WATERMARK 的 BA 为 98.7%，VSR 为 96.2%，H 为 0.01，MCD 为 9.3；在 VoxCeleb 上，BA 为 95.1%，VSR 为 98.5%，H 为 0.03，MCD 为 9.4。与其他方法（如 FreqTone、AdvBackdoor）相比，AUDIO WATERMARK 在 BA 和 VSR 上表现优异，有害度最低，MCD 适中。

2. 可迁移性：

可迁移性分为数据集可迁移性和模型可迁移性。数据集方面，水印生成器在 LibriSpeech 上训练后，成功应用于 VoxCeleb，VSR 达 98.5%。而模型方面，对 10 种模型均保持高 VSR（最低 75%），如图 5-1 所示。

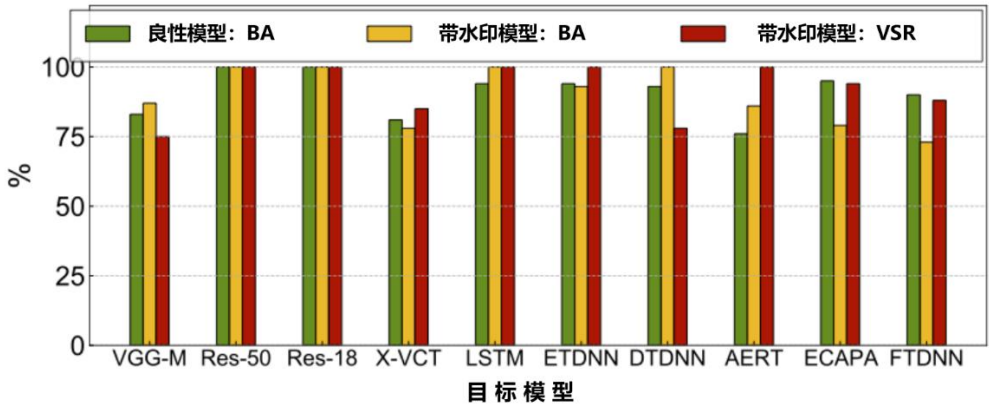

图 5-1 不同模型间的水印可迁移性

3. 消融研究：

中毒率(%)即水印样本占完整水印数据集的比例。在数据集保护流程中，防御者仅向水印数据集注入少量水印样本。通常而言，较低的中毒率会导致保护成功率下降。若某种保护方法在低中毒率下仍能生效，则表明该保护机制具有强效性。如图 5-2（a）所示，中毒率对 VSR 影响显著，10% 中毒率下所有模型 VSR 均较高。噪声水印实验表明，噪声水印在 10 种模型上仅对 3-4 种有效，且易泛化导致误报，如图 5-2（b）所示。

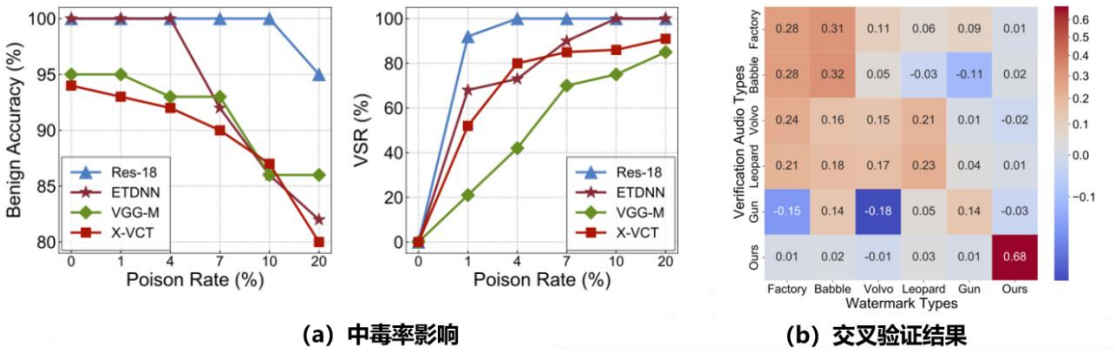

图 5-2 消融研究

4. 鲁棒性：

该论文验证了对下面四种场景下的鲁棒性。模型级攻击：对抗模型微调、剪枝和 Neural Cleanse，VSR 保持稳定，如图 5-3（a）所示；数据级攻击：对抗噪声去除、STRIP 等攻击，AUROC 约 0.5，表明攻击无效，如图 5-3（b）所示；任务迁移攻击：在说话人验证任务中，ResNet-18 的 VSR 仍达 80%，如图 5-3（c）所示；物理世界测试：通过扬声器和麦克风测试，VSR 达 90%。

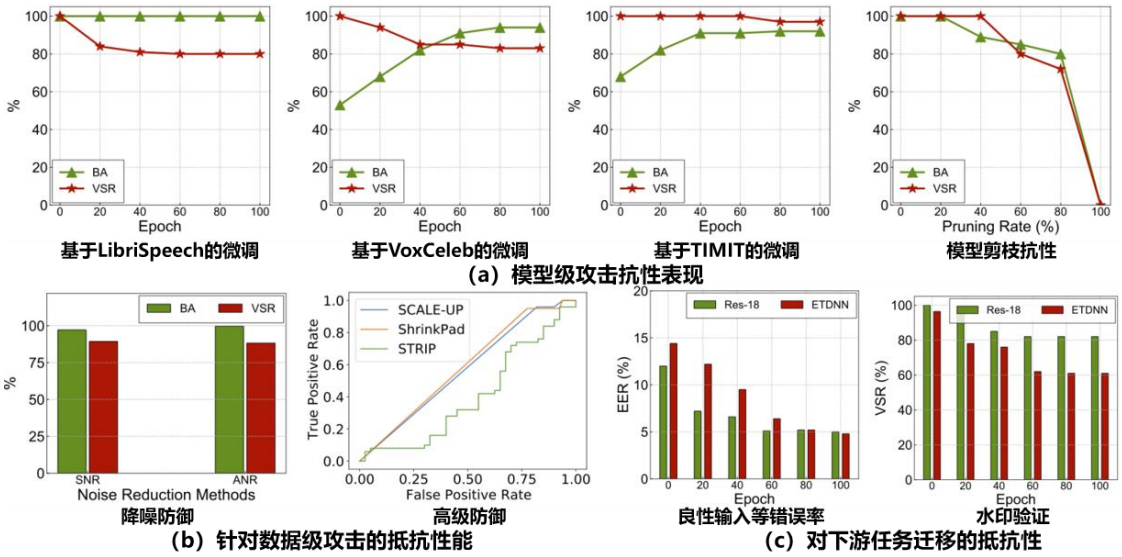

图 5-3 鲁棒性验证

六. 阅读心得

论文提出了一种创新的语音数据集版权保护方法，其核心创新在于通过动态、无害的水印设计，结合风格转换和逆对比 LTAF 损失，实现了对黑盒模型的高效验证，同时保持音频质量和模型性能。这种方法在技术上具有突破性，还在伦理上符合公平性和有益性原则，避

免了有害后门的引入。从技术角度看，Style Wave-U-Net 的设计巧妙地结合了语音风格转换和生成模型的优势，通过多层次风格嵌入生成动态水印，避免了固定触发器的易检测性。逆对比 LTAF 损失的引入是另一亮点，通过优化水印音频的说话人身份特征，确保其难以被其他数据集泛化，从而降低了误报率。双层对抗优化策略进一步增强了水印的适应性，使其在黑盒设置下表现出色。

尽管论文在技术上取得了显著成果，论文中也指出存在一些值得探讨的局限性，例如其计算复杂性较高，Style Wave-U-Net 和双层优化需要较高的计算资源，可能对资源有限的数据集所有者构成挑战，未来可探索轻量化模型以降低计算成本；其次实验表明，如上文中图 5-3（c）所展示的，水印在说话人验证任务中的 VSR 略有下降，说明跨任务迁移仍需改进，论文也提到多任务支持是未来研究方向，可考虑引入跨任务共享特征的生成模型；然后是对于物理世界鲁棒性，虽然验证了物理场景下的水印有效性，但仅测试了单一距离（30cm），实际环境中，噪声、距离和设备差异可能进一步影响水印性能，需更广泛的测试。

基于论文的启发，通过查阅资料和目前的研究现状我有一些浅薄的不成熟的看法，我认为以下方向值得进一步研究，首先在目前机器学习多模态的研究盛行下，音频水印的风格可以结合语音和图像数据集的水印技术，开发适用于多模态数据集的统一保护框架；其次，我们可以进行测试和验证的数据集总归是有限的，但是现实场景中的数据是千变万化的，为了更好的模型普遍适应性，我们可以引入在线学习机制，使水印生成器根据数据集特征动态调整，适应不同类型的数据分布；另外水印的隐私保护也是非常重要的，通过本课程《多媒体数据安全》所学的知识我们可以针对性地提出和制定一些隐私技术，进一步降低隐私泄露风险，这个方向也是将我们的专业理论灵活运用于实际的多媒体场景中，希望之后我们有机会可以更加深入地和细致地对相关领域进行研究和探索。

同时在精读学习的过程中，对于科研论文的行文结构、实验设计以及理论创新有了新的理解，首先对创新点如何进行论证，论证分为理论论证和实验论证，博古通今在前人的基础上提出自己更优的方案，而如何通过合理全面的实验对我们的理论效果进行可信的展示也是具有技巧的，比如这篇论文设计了大量的有效实验，我们可以通过广泛阅读来学习，同时我也尝试对它的开源代码进行了复现，考虑学习它的相关部分对自己的科研项目进行推动。

原创性声明

本人郑重声明本报告内容，是由作者本人独立完成的。有关观点、方法、数据和文献等的引用已在文中指出。除文中已注明引用的内容外，本报告不包含任何其他个人或集体已经公开发表的作品成果，不存在剽窃、抄袭行为。

已阅读并同意以下内容。

判定为不合格的一些情形：

（1） 请人代做或冒名顶替者；

（2） 替人做且不听劝告者；

（3） 课程报告内容抄袭或雷同者；

（4） 课程报告内容直接用翻译软件翻译得到者；

（5） 课程报告内容借用大模型生成者。

作者签名：