$\psi_{u_1}\sim\psi_{u_2}$，便可以判断两个特征标是否属于同一共伴随轨道. 取$T\in V^t$，有

$$hTh^{-1}=\begin{pmatrix}0 & & \\ C+DB & 0 & \\ B & 0 & 0\end{pmatrix},$$

通过对V^t进行分类，G_E的特征标可分为以下三种情形：

1. 如果$B=0$. 取$u(C,0)$，它的稳定化子$H_{\psi_{u(C,0)}}$同构于H. 此情形对应q^6个 1维特征标，形如$\mathrm{Ind}_{H_{\psi_{u(C,0)}}\times A}^{G_E}\tau\otimes\psi_{u(C,0)}$.

2. 如果$r(B)=1$. 对于选定的B，任意C，通过选择恰当的D，可以使得

$$C+DB=\begin{pmatrix}* & * \\ 0 & *\end{pmatrix}=C'.$$

取$u(C',B)$，有$\psi_{u(C',B)}\sim\psi_{u(C,B)}$，它的稳定化子$H_{\psi_{u(C,B)}}$同构于$DB=0$的$H$. 此情形对应$q^4(q^2-1)(q+1)$个$q$维特征标，形如$\mathrm{Ind}_{H_{\psi_{u(C',B)}}\times A}^{G_E}\tau\otimes\psi_{u(C,B)}$.

3. 如果$r(B)=2$. 对于选定的B，通过选择恰当的D，可以使得

$$C+DB=\begin{pmatrix}* & * \\ 0 & 0\end{pmatrix}=C''.$$

取$u(C'',B)$，有$\psi_{u(C'',B)}\sim\psi_{u(C,B)}$，它的稳定化子$H_{\psi_{u(C,B)}}$为单位矩阵. 此情形对应$q^2(q^2-1)(q^2-q)$个$q^2$维特征标，形如$\mathrm{Ind}_{H_{\psi_{u(C'',B)}}\times A}^{G_E}\tau\otimes\psi_{u(C'',B)}$.

其它情形下，有着相同的结果.

为了进一步验证上述特征标构造与分类结果的正确性，我们采用有限群特征标维数平方和公式进行验证. 有限群特征标维数平方和等于群的阶数，这是有限群表示理论中的一个基本定理. 通过计算各类特征标的维数平方与个数的乘积之和，即

$$1\cdot q^6+q^2\cdot q^4(q^2-1)(q+1)+q^4q^2(q^2-1)(q^2-q)=q^{10}.$$

群G_E的阶数经过计算也为q^{10}.两者相等，这充分证明了我们所得到的特征标构造与分类结果的正确性.

4 结论

本文对GL_6对应划分为 (2,2,2) 的标准抛物子群的一类特殊幂幺根基G_E进行了探究，运用 Clifford 理论，利用共伴随轨道构造了G_E的所有的特征标. 在研究过程中，我们首先对群G_E的结构进行了剖析，刻画出矩阵形式以及各个分块矩阵需要满足的条件. 随后，提出了 pattern 群和共伴随轨道的概念并构造了研究其群表示的方法，利用 Clifford 定理并结合阿贝尔正规子群半直积的结构，得到了G_E所有的不可约特征标并对各维数的特征标个数给出了精确的结果.

这一完整的分类结果不仅揭示了该群表示的数量规律性，通过具体的计数公式清晰展现了不同维数特征标的数量与q的关系，更重要的是对 Higman、Lehrer 和 Isaacs 提出的著名猜想提供了新的验证实例. 通过特征标维数平方和公式的验证，我们进一步确认了所得结果的正确性。然而，本文也留有一些问题有待研究讨论. 例如，能否在其他类型的 pattern 群构造出类似的表示规律？不同结构的 pattern 群在特征标构造、维数分布以及与q的关系等方面可能存在差异，深入研究这些差异与共性，有助于我们更全面地理解 pattern 群的表示理论. 此外，本文所采用的方法是否可以推广到更广泛的代数结构中去？如更高阶的一般线性群、其他类型的有限群或代数结构等，将研究方法进行推广与拓展，能够进一步扩大研究的适用范围，为更多代数结构的表示理论研究提供新的思路与方法. 这些问题的研究将有助于我们更深入地理解有限群表示的内在规律.

参考文献

[1] G. Higman, Enumerating p-groups. I: Inequalities[J]. Proceedings of the London Mathematica l Society, 1960, 3(1): 24-30.

[2] G. I. Lehrer, Discrete series and the unipotent subgroup[J]. Compositio Mathematica, 1974, 2 8(1): 9-19.

[3] I. M. Isaacs, Characters of groups associated with finite algebras[J]. Journal of Algebra, 1995, 177(3): 708-730.

[4] C. Nien, Characters of unipotent radicals of standard parabolic subgroups with 3 parts[J]. Jour nal of Algebraic Combinatorics, 2021: 1-9.

[5] S. M. Goodwin, T. Le , K. Magaard & et al, Constructing characters of Sylow p-subgroups of finite Chevalley groups[J]. Journal of Algebra, 2016, 468(1):395- 439.

[6] S. M. Goodwin, T. Le & K. Magaard, The generic character table of a Sylow p-subgroup of a finite Chevalley group of type D4[J]. Mathematics, 2017.

[7] S. M. Goodwin, Counting conjugacy classes in Sylow p-subgroups of Chevalley groups[J]. Jo urnal of pure and applied algebra, 2007, 210(1)

[8] T. Le, K. Magaard & A. Paolini, On the characters of Sylow p-subgroups of finite Chevalley groups G(pf)for arbitrary primes:, 10.1090/mcom/3488[P]. 2019.

[9] T. Le, K. Magaard & A. Paolini, The irreducible characters of the Sylow psubgroups of the Chevalley groups D6(pf)and E6(pf)[J]. Journal of Symbolic Computation, 2019, 95(NOV.-DE C.):68-99

[10] D. Ginzburg, On the representations of q unipotent groups over a finite field[J]. Israel Journ al of Mathematics, 2011, 181: 387-422.

[11] C. Nien, Characters of 2-layered Heisenberg groups[J]. Linear and Multilinear Algebra, 202 0: 1-10.

[12] C. Nien, Characters of unipotent radicals of standard parabolic subgroups with 3 parts[J].Jou rnal of Algebraic Combinatorics, 2021: 1-9.

[13] Y. Li & C. Nien, Characters arising from monomial coadjoint orbits for finite pattern groups [J]. Journal of Algebra, 2023, 619: 558-571.

[14] 李奕可, 若干有限典型群的特征标的研究[D].湖南师范大学, 2023.

[15] A. Kirillov, Merits and demerits of the orbit method[J]. Bulletin of the American Mathematic al Society, 1999, 36(4): 433-488.

[16] J. P. Serre, Linear representations of finite groups[M]. New York:springer, 1977.

 DOI: https://doi.org/10.12345/bdai.v6i6.34042

Research on generative AI security governance technology and platform construction

Shuang Han

Shanghai Digital Sec, Rity Technology Co., Ltd., Shanghai, 200040, China

Abstract

With the rapid advancement of generative AI technology, security risks such as deep fakes and automated misinformation generation have become increasingly prominent, posing serious threats to national security, social stability, and individual rights. This paper proposes a comprehensive “discovery-penetration-traceability-monitoring-assessment-hardening” full lifecycle governance solution to address these AI-related security challenges. By establishing three foundational databases, seven core technologies, and an integrated application platform, the solution achieves precise detection, reliable traceability, and effective governance of multimodal AI-generated content. Experimental results demonstrate that the platform achieves over 80% average detection accuracy for mainstream forgery techniques and can effectively identify 31 categories of security risks. This provides crucial technical support for public safety governance and the healthy development of the AI industry.

Keywords

Generative Artificial Intelligence; Security Governance; Deepfake Detection; Content Tracing; Risk Assessment

生成式人工智能安全治理技术研究与平台构建

韩霜

上海数字安全科技有限公司，中国·上海 200040

摘　要

随着生成式人工智能技术的快速发展，深度伪造、虚假信息生成等安全风险日益凸显，对国家安全、社会稳定和个人权益造成严重威胁。本文针对生成式人工智能带来的安全挑战，提出了一套完整的"发现-鉴伪-溯源-监测-测评-加固"全生命周期治理解决方案。通过构建三大基础数据库、七大核心技术和一体化应用平台，实现对多模态AI生成内容的精准检测、可靠溯源和有效治理。实验结果表明，该平台对主流伪造技术的平均检测准确率达到80%以上，能够有效识别31类安全风险，为公共安全治理和AI产业健康发展提供了重要技术支撑。

关键词

生成式人工智能；安全治理；深度伪造检测；内容溯源；风险测评

1 引言

近年来，以 ChatGPT、MidJourney、Stable Diffusion 为代表的生成式人工智能技术取得了突破性进展，在内容创作、教育培训、科学研究等领域展现出巨大的应用价值。这些技术能够自主生成文本、图像、音频、视频等多种形式的内容，为各行各业带来了革命性的变革。

然而，生成式人工智能技术的广泛应用也带来了前所未有的安全挑战。深度伪造（Deepfake）技术能够生成高度逼真的虚假音视频内容，被不法分子用于制造传播虚假新闻、实施网络诈骗、操纵舆论等违法犯罪活动。据统计，2023 年全球深度伪造相关案件数量同比增长超过 300%，造成的经济损失超过 50 亿美元。同时，AI 生成的虚假文本、图片等内容也在社交媒体平台上快速传播，严重扰乱了信息传播秩序，威胁社会信任体系。

面对生成式人工智能带来的安全风险，各国政府和国际组织纷纷出台相关政策法规。2024 年，中国发布《生成式人工智能服务安全基本要求》（TC260-003），明确了 31 类安全风险的治理要求；欧盟《人工智能法案》将生成式 AI 列为高风险应用，要求实施严格的安全评估；美国也制定了《深度伪造检测法案》，推动相关技术研发和应用。

当前，生成式人工智能安全治理面临着三大核心挑战：一是检测能力滞后，传统的内容安全检测方法对 AI 生成内容的识别准确率普遍偏低；二是溯源取证困难，缺乏有效的技术手段追踪伪造内容的生成工具和传播路径；三是治理体系不完善，缺乏覆盖全生命周期的一体化治理解决方案。

【作者简介】韩霜（1986–），男，中国江苏东海人，从事信息安全、网络安全、数据安全、AI算法安全研究。

本文针对上述挑战，提出了一套完整的生成式人工智能安全治理技术体系，主要贡献包括：

（1）构建了 "3+7+1+N" 的体系化治理架构，包括三大基础数据库、七大核心技术、一个一体化平台和 N 类应用场景；

（2）提出了基于多专家混合模型的多维度特征分析技术，通过空间理解、时序理解和多模态理解的有机结合，显著提升了 AI 生成内容的检测精度；

（3）研发了高泛化性能溯源技术，利用多尺度特征和频域分析，实现了对生成算法、工具软件和原始设备的精准追踪；

（4）建立了覆盖 31 类安全风险的测评体系，为 AI 模型安全评估提供了标准化方法。

2 相关技术背景

2.1 生成式人工智能技术发展现状

生成式人工智能技术基于深度学习模型，能够自主生成文本、图像、音频、视频等多种形式的内容。根据生成内容的模态不同，主要可分为文本生成、图像生成、音频生成和视频生成四大类。

在文本生成领域，以 GPT 系列、LLaMA、文心一言等为代表的大语言模型，通过大规模文本数据训练，具备了强大的自然语言理解和生成能力。这些模型参数规模从数十亿到数万亿不等，能够生成高质量的文章、代码、诗歌等内容。图像生成技术方面，基于扩散模型的 Stable Diffusion、MidJourney 等模型，能够根据文本描述生成高度逼真的图像。这些模型在艺术创作、设计、娱乐等领域得到广泛应用，但也被用于生成虚假图片和误导性内容。

音频生成和视频生成技术相对复杂，但近年来也取得了显著进展。基于 GANs、Diffusion 等技术的音频生成模型能够合成逼真的语音和音乐；视频生成模型则能够生成连贯的视频片段，甚至实现人物面部表情的实时驱动。

2.2 深度伪造检测技术研究进展

深度伪造检测技术主要分为基于视觉特征的检测、基于音频特征的检测和基于多模态融合的检测三大类。

基于视觉特征的检测方法主要分析图像或视频中的视觉伪影，如面部特征异常、光照不一致、边缘伪影等。早期的方法主要基于手工设计的特征提取器，如 LBPs、HOGs 等，结合传统机器学习算法进行分类。随着深度学习技术的发展，基于 CNN、Transformer 等深度学习模型的检测方法成为主流。这些方法能够自动学习图像中的深层特征，具有更强的表征能力和泛化性能。

基于音频特征的检测方法主要分析音频信号中的异常模式，如频谱特征、韵律特征、声纹特征等。这些方法在检测 AI 合成语音方面具有一定优势，但对复杂的音频伪造攻击抵抗力有限。基于多模态融合的检测方法结合视觉和音频信息，能够更全面地分析伪造内容的特征，提高检测的准确性和鲁棒性。

2.3 内容溯源技术研究现状

内容溯源技术旨在追踪数字内容的来源和传播路径，为内容真实性验证和责任追究提供技术支撑。目前主要的溯源技术包括数字水印、哈希校验、区块链存证等。

数字水印技术通过在内容中嵌入不可见的标识信息，实现内容的来源追踪。根据水印的可见性，可分为可见水印和不可见水印；根据水印的鲁棒性，可分为鲁棒水印和脆弱水印。哈希校验技术通过计算内容的哈希值，实现内容完整性验证。任何对内容的修改都会导致哈希值的变化，从而能够检测内容是否被篡改。

区块链存证技术利用区块链的去中心化、不可篡改特性，为内容的创作时间、作者信息等提供可信存证。这种技术在版权保护和内容溯源方面具有重要应用价值。

3 系统总体架构

本文提出的生成式人工智能安全治理平台采用 "3+7+1+N" 的体系化架构设计，如图 1 所示。该架构包括三大基础数据库、七大核心技术、一个一体化平台和 N 类应用场景，形成了覆盖 " 发现 - 鉴伪 - 溯源 - 监测 - 测评 - 加固 " 全生命周期的治理能力。

表 1 生成式人工智能安全治理平台架构组成

组成部分	主要内容	核心功能
三大基础数据库	伪造音视频图片样本库、开源算法工具库、大模型测评库	为算法训练、测试和评估提供数据支撑
七大核心技术	多维度特征分析、深度伪造检测、音频鉴伪、高迁移性识别、高泛化性能溯源、安全风险分类测评、安全加固	实现对 AI 生成内容的检测、溯源和治理
一体化平台	识别检测、溯源、安全测评、风险识别、监测、加固六大模块	提供统一的技术服务和管理界面
N 类应用场景	舆情监测、线索推送、大模型评测等	支撑不同领域的安全治理需求

3.1 三大基础数据库

构建三大关键库作为技术支撑。首先是伪造音视频图片样本库，依据多样性等四项原则，收集超 10 万条多模态高质量风险样本，为算法训练与测试提供数据支撑；其次是开源算法工具库，整合多类生成模型、伪造工具等资源，建立完整的算法指纹特征库，助力实现伪造内容溯源；最后是大模型测评库，内置合规标准与多维度题库，依据相关要求覆盖 31 类安全风险，保障大模型的合规性运行。

3.2 七大核心技术

七大核心技术是平台的技术支撑，包括基于多专家混

合模型的多维度特征分析技术、AI 生成文本的深度伪造检测技术、基于大模型的音频鉴伪检测技术、高迁移性识别技术、高泛化性能溯源技术、生成式人工智能安全风险分类及测评技术、生成式人工智能安全加固技术。这些技术相互协作，形成了完整的技术链条。

3.3 一体化平台架构

一体化平台采用微服务架构设计，包含六大核心模块：生成式人工智能识别检测模块、生成式人工智能溯源模块、生成式人工智能安全测评模块、生成式人工智能风险识别模块、生成式人工智能监测模块、生成式人工智能加固模块。各模块以标准化 API 服务对外提供跨模态、多层次的 AI 治理与安全分析能力。

4 核心技术详解

4.1 基于多专家混合模型的多维度特征分析技术

为解决传统方法对 AIGC 生成内容检测准确率低的问题，现提出基于多专家混合模型的多维度特征分析技术，以提升模型在空间、时序和多模态理解上的能力。空间理解技术借 Transformer 架构做像素级语义分析，识别异常区域；时序理解技术融合时空特征，弥补传统模型忽略时序逻辑的不足，综合判断视频真实性；多模态理解技术联合音画双模态，通过模态内、间特征差异，检测单模态真实性与音画一致性。

4.2 AI 生成文本的深度伪造检测技术

AI 生成文本在词汇分布、句式结构和逻辑连贯性方面展现出可量化的非自然特征。基于自然语言处理技术提取文本的困惑度（Perplexity）和突发性（Burstiness）等特征，可以有效区分 AI 生成文本和人类撰写文本。

本技术采用基于 Transformer 的架构，通过文本预处理、特征提取、统计特征计算和分类判断四个步骤实现文本真实性检测。实验结果表明，该技术对 AI 生成文本的检测准确率达到 86.3%。

4.3 高迁移性识别技术

高迁移性识别技术是本项目的创新点之一，通过构建基于视觉 - 语言模型的初始检测器，引入低秩适配器进行高效微调，在保留预训练通用性的同时增强对伪造痕迹的适应性。该技术采用 K-FAC 方法减少知识遗忘，结合 LMC 技术提取不同生成模型的共性特征，通过线性插值实现跨任务迁移。

技术流程包括基础模型训练、适配器设计、知识蒸馏和持续学习四个阶段，使模型具备持续学习和快速适应新型伪造算法的能力。

4.4 高泛化性能溯源技术

高泛化性能溯源技术是本项目的另一重要创新点，利用 CLIP 和 Video MAE 等预训练模型提取通用特征，构建多尺度伪造痕迹识别基础。通过频域适配模块和傅里叶卷积捕捉频域异常纹理，结合 PRNU 特征实现设备级溯源。

溯源技术的核心在于利用多媒体内容的底层特征与特定的伪造方式紧密耦合性，研究具备跨伪造方式、高聚合度和高泛化能力的底层特征表达方法，提取高鉴别力特征。实验结果表明，该技术的平均溯源准确率达到 78.5%。

4.5 生成式人工智能安全风险分类及测评技术

紧密围绕《生成式人工智能服务安全基本要求》（TC260-003），针对生成模型的脆弱性，系统设计对抗性安全样本生成技术，通过提示注入、越狱攻击等方式构建覆盖 31 类风险场景的高质量、结构化测评题库。

测评技术采用数据增强与多模态语料合成技术，提升样本的覆盖面与真实性，构建面向中文大模型的多样化测试样本集。在模型推理阶段，通过模拟有害信息生成、隐私泄露、指令绕过等安全事件，量化评估模型在不同攻击下的安全表现与合规应对能力。

4.6 生成式人工智能安全加固技术

生成式人工智能安全加固技术构建 " 内生安全、外挂防护 " 的双层次加固体系。内生安全方面，通过指令微调和安全微调技术引入拒答或合规替代策略，训练基于人类与 AI 反馈的奖励模型以提升拒答稳定性。外挂式安全模块方面，构建多层次的实时检测引擎，结合规则过滤、语义分类和上下文感知评分器对输出内容进行快速而精准的风险识别。

5 实验验证与性能分析

5.1 实验环境与数据集

实验在配备 NVIDIA A100 GPU 的服务器上进行，使用 PyTorch 深度学习框架。实验数据集包括自建的多模态伪造样本库和公开的深度伪造检测数据集。

自建数据集包含 10 万 + 高质量风险样本，覆盖图像、视频、音频、文本等多种模态，涵盖主流的伪造技术和工具。公开数据集包括 FaceForensics++、DeepFakeDetection Challenge、Celeb-DF 等。

5.2 检测性能评估

本项目对提出的多维度特征分析技术进行了全面的性能评估。实验采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数作为评价指标。

实验结果表明，本项目提出的技术在不同模态和不同伪造技术上均取得了优异的检测性能：

表 2 不同方法检测性能对比

方法	准确率 (%)	精确率 (%)	召回率 (%)	F1 分数 (%)
传统 CNN 方法	72.3	71.8	72.1	71.9
单一 Transformer 方法	78.6	78.2	78.4	78.3
多模态融合方法	81.4	80.9	81.2	81.0
本文方法	85.2	84.7	85.0	84.8
性能提升	+12.9	+12.9	+12.9	+12.9