文章类型：同行评审文章

本文已提交至 IEEE，拟供其考虑发表。版权可能在不另行通知的情况下转让，此后此版本可能不再提供访问。

不要猜测，升级处理：迈向可解释且经过不确定性校准的人工智能取证智能体

Giulia Boato，*Truebees*，意大利特伦托，邮编 *38122*；特伦托大学，意大利特伦托，邮编 *38123*Andrea Montibeller，*Truebees*，意大利特伦托，邮编 *38122*；特伦托大学，意大利特伦托，邮编 *38123*Edward Delp，美国西拉斐特，印第安纳州 *47907*，普渡大学Luisa Verdoliva，那不勒斯费德里科二世大学，意大利那不勒斯，邮编 *80125*Daniele Miorandi，*Truebees*，意大利特伦托，邮编 *38122*

arXiv:2512.16614v1 [cs.MA] 18 Dec

摘要——人工智能正在重塑多媒体取证领域。我们提出人工智能取证智能体：可靠的编排器，可选择并组合取证检测器，识别溯源和上下文，并提供不确定性感知评估。我们指出当前解决方案中的缺陷，并提出一个统一框架，以改进真实性验证流程。

关键词：多媒体取证、人工智能生成媒体、人工智能代理、媒体可信度。

Multimedia 取证旨在确定图像、视频和音频录音等数字内容的真实性和溯源。多年来，该领域一直依赖人工设计的分析线索——噪声模式、压缩伪影、传感器指纹和时间不一致性 [1]。当采集流程得到充分了解时，这些方法能提供坚实的基础。例如，如果所有图像都来自有限的一组相机型号，并经过可预测的压缩和分发链路，那么最终的数据分布就足够受限，可以对设备特定伪影进行建模并加以利用。同样，当威胁模型较为简单——例如剪切粘贴伪造、单帧插入和换脸——算法便能针对相应场景进行有效优化。在这些受控情境下，人工设计的线索在区分真实内容与经篡改或合成的内容方面仍具有很强的判别力。

随着生成式人工智能的兴起，这一领域迅速发生了变化。大型视觉、语言和音频模型如今能够大规模生成高度逼真的合成图像，视频和语音。它们生成的内容对长期以来关于哪些取证线索仍值得信赖的假设提出了挑战。此外，过去只有专家操作者才能掌握的能力，如今通过易于使用的编辑和生成工具，已广泛普及。因此，法证科学必须跟上对抗性环境的步伐，而这一环境正以迅速且难以预测的方式演变。

我们认为，多媒体取证领域现在必须从设计和实现孤立的单一用途检测器，转向整体性的人工智能取证智能体：能够自主编排多种工具，以检测、表征和解读人工智能生成内容的系统。这一需求主要由两个因素驱动。首先，如今的生态系统十分碎片化：大多数检测器都局限于特定模态或篡改类型，因而难以泛化到新兴威胁或跨模态威胁。其次，在缺乏编排的情况下，溯源链、元数据、多模态相关性和不确定性估计等关键信息仍未得到充分利用。现有的检测器组合尝试通常依赖多数投票、加权聚合或简单的分数融合规则。尽管此类集成方法可能优于单个模型，但其最终完整性分数往往不透明，几乎无法说明哪些取证线索影响了决策。

XXXX-XXX © 2021年 IEEE 数字对象标识符
10.1109/XXX.0000.0000000

本文立场文章阐述了可信人工智能取证智能体的愿景。我们指出了关键的科学与技术鸿沟，包括整合异构取证痕迹，以及设计可解释方法，使其既能报告完整性分数和经校准的置信度分数，又能在证据不足时abstain。我们设想，人工智能智能体编排器能够将专用检测器的低级取证输出转化为人类可理解的论据，以及围绕特定伪影为何重要所形成的连贯叙述，从而增强可解释性。我们的目标是鼓励研究人员和从业者将基于智能体的编排作为多媒体取证领域的下一种范式，在生成式人工智能时代增强真实性评估的严谨性、透明度和可信度。

人工智能生成媒体

如今，合成图像生成主要基于两类模型：生成对抗网络（GANs）和扩散模型（DMs），二者都能够合成高质量、受文本条件控制的图像。GANs 通过生成器与判别器之间的对抗博弈运作；随着双方相互较量、不断改进，生成器学会生成愈发逼真的图像。经典里程碑包括 BigGAN 和 ProGAN，随后发展出 StyleGAN 系列，它能够通过潜在编码精细控制合成过程；以及用于三维一致性人脸渲染的 EG3D。较新的 GANs，例如 StyleGAN-T、GALIP 和 GigaGAN，融合了文本条件控制与大规模训练，从而实现可控、高分辨率且快速的图像生成。

扩散模型则通过学习逆转加噪过程来生成图像：前向过程逐步向真实图像添加高斯噪声，使其受到扰动；而由神经网络实现的反向过程则逐步去噪，合成图像。它们在图像质量上可与 GANs 媲美甚至更胜一筹，训练也更加稳定，并且天然支持文生图。许多扩散架构采用带有文本编码器的 U-Net 主干网络（例如 GLIDE、DALL-E 2、eDiff-I），或在潜在空间中运行以降低成本（例如 Latent DM 和 Stable Diffusion）。InstructPix2Pix、Stable Diffusion XL、DeepFloydIF 和扩散变换器（DiTs）等变体，进一步提升了分辨率、可控性和架构多样性 [2]。

合成视频生成也发展得非常迅速，从早期的基于图像的方法逐渐转向专用的视频模型。早期方法往往难以确保运动稳定性和时间连贯性，而专用的视频模型能够更有效地捕捉时空结构。现代采用三维自编码器和 Transformer 的架构能够更好地与文本提示对齐，并生成质量更高、时间一致性更强的视频。最近，自回归方法也取得了令人印象深刻的视频生成性能 [3]。

在音频领域 [4]，，过去五年间，随着文本转语音和语音克隆技术的应用，语音生成技术取得了显著进展。前者利用声码器根据文本输入生成听起来自然的语音；后者则在保留底层语言内容的同时，修改原始语音信号以模仿目标说话人。除了传统的基于声码器的方法，近期的方案还利用大语言模型（LLMs）和神经音频编解码器，生成极为逼真的合成语音。

尽管合成图像、视频和音频十分逼真，但它们仍会留下取证痕迹。在语义（高）层面，早期的合成图像和视频生成器经常会产生明显的伪影，例如颜色怪异、光照不一致或几何结构错误；但较新的模型能够生成视觉上毫无瑕疵的图像（图1）。尽管如此，每个采集或生成流程都会留下特征性的底层伪影，类似于传统图像取证中使用的设备痕迹。在生成式人工智能中，可以通过统计分析和频域分析揭示架构特定模式，并将其用于检测或归因。常见的明显痕迹包括噪声残差功率谱中的异常峰值，以及在中高频范围内与真实数据不匹配的频谱分布。与视觉媒体一样，从 WaveNet 到现代神经编解码器等音频生成算法，也会留下独特的取证痕迹，例如声码器指纹和频谱幅度异常，可用于区分合成语音与真实人类音频。

多媒体取证领域的变迁

经典方法会利用相机内处理链或后续编辑操作引入的伪影，这些伪影会在所获取的图像上留下独特的痕迹 [1]。例如，不同的相机型号通常采用不同的去马赛克算法，因此，当伪造内容结合了来自多个设备的区域时，去马赛克模式中可能会出现不一致。同样，相机外编辑也会留下自身特有的痕迹，并且还可能改变或抑制传感器特定指纹模式，这两者都可以用来揭示篡改。



图1：人工智能生成图像的示例：部分图像具有近乎照片般逼真的质量（上），而另一些图像则带有明显的视觉伪影（下）。这些案例表明，有些生成器留下的痕迹主要细微且处于低层级，而另一些则在高层级引入明显的语义不一致。因此，仅依赖低层级或高层级伪影的检测器，难以稳健地应对合成图像的全部类型。

利用压缩伪影长期以来一直是图像取证的基石，例如利用块伪影网格或双重压缩痕迹。更通用的策略是分析整个基于相机的获取过程中引入的噪声伪影，例如检查局部噪声水平或高通噪声残差，以揭示不同区域之间的不一致。尽管这些痕迹极其细微，肉眼无法察觉，但经过适当增强和分析后，它们可以为评估数字内容的完整性提供有价值的证据。

这些方法功能强大，但高度依赖特定的统计模型；当底层假设不成立时，其性能就会下降。例如，它们对分辨率变化或额外压缩非常敏感，而这两种情况在社交媒体平台上都十分常见。 [5]在这种充满挑战的条件下，深度学习，尤其是非常深的架构，相较于传统方法能够显著提升性能。然而，由于训练和测试方案设计不当，数据驱动方法的效果可能被过分夸大。事实上，当训练数据与测试数据完全一致时，这些方法表现良好；但在分布偏移情况下，准确率则会大幅下降。检测由已知生成模型生成的图像相对容易；真正的挑战在于如何泛化到由未见过的生成器生成的合成内容。如今，这一问题尤为迫切，因为新的生成模型几乎在持续不断地推出。在此背景下，大型预训练视觉语言模型（VLM），例如CLIP，展现出对分布偏移显著的鲁棒性。它们在取证应用中的成功表明，在大型且多样化的数据集上进行预训练，有助于形成更具语义性且与模型无关的表征，从而构建出泛化能力更强的检测器，使其能够识别未见过的篡改和生成模型，即使训练时没有这些来源的样本也不例外。

此外，近期研究越来越多地利用多模态大模型，以获得更透明、可解释的取证决策 [6]。通过量身定制的提示学习，这些方法能够生成人类可读的解释，明确将细微的视觉伪影和不一致之处与最终分类结果联系起来，从而使决策过程更易于审查 [7], [8]。然而，尽管这类工具能够生成连贯的解释并检测语义异常，却无法对多个来源的取证证据进行联合推理，也无法通过经验不断改进。它们几乎只处理图像内容，而忽略元数据、来源信息和上下文线索等其他关键取证信息。因此，它们只能作为独立检测器运行，且其能力极其有限——

有限的上下文感知推理能力。

从单一检测器到编排器

当前的多媒体取证检测器——无论是用于识别人工智能生成媒体、图像修复、视频篡改，还是语音克隆——往往都被视为孤立的真相来源。每种方法通常都是针对特定的篡改或取证痕迹设计的，并在受控的实验设置中进行基准测试。这往往会导致系统**缺乏泛化能力**：某个检测器今天表现良好，明天却可能因生成技术不断演进、数据集发生漂移或对手利用其盲点而失效。过度依赖任何单一模型都会造成危险的不对称局面：攻击者不断演进策略，而防御者却只依靠一种工具。

然而，现代篡改已不再是单次编辑，而是由一系列变换构成的处理流程，会留下相互重叠、有时甚至彼此矛盾的痕迹。当操作序列未知时——这在现实调查中很常见——分析就应从孤立的检测器转向统一的推理系统。目标是将异质线索综合起来，对真实性进行连贯且不确定性感知的评估。

近期研究探索了检测器集成方法，通过多数投票、加权聚合或相关融合规则 [9]组合多个模型。这些集成方法可利用互补特征，表现优于单个模型，但也存在一个显著局限：不透明性。最终的置信度分数是对各检测器特定特征及临时采用的融合启发式方法进行黑箱组合的结果，可能包含难以预见或解释的非平凡偏差。因此，集成系统往往难以泛化至未见过的篡改类型或新兴生成模型，同时还会增加计算开销。

这些局限表明，我们需要从概念层面进行更深入的重新构想。多媒体取证不应只是简单地增加更多检测器，而应转向**人工智能取证编排器**——这类系统不仅融合输出，还会在统一框架内，对多种取证工具和专业子代理进行主动推理（图2）。类似的编排概念 [10], [11] 已应用于自动化交易和金融分析等领域，代理会动态管理由专业模型组成的投资组合，以实现复杂目标。在取证场景中，编排器发挥着类似作用：它决定调用哪些工具或子代理、以何种顺序调用，以及如何解读它们的输出。

取证编排器能够在恰当的时机启用恰当的组件，根据上下文信息评估每个检测器的可靠性，并校准异构线索所带来的不确定性。它的作用不止于汇总：当证据薄弱或相互矛盾时，它还可以选择不作判断，将不确定性从一种失效模式转化为防止过度自信结论的保障。编排也能实现更深入的推理。通过整合跨模态信息、协调跨多种变换的溯源重建，并针对不同用户画像（如调查人员、记者或法律决策者）定制解释，编排器能够实现简单的检测器融合无法达到的可解释性。这样一来，低层痕迹、语义不一致、元数据和上下文线索便可结合起来，形成连贯且以证据为依据的评估。

可信取证智能体

在这一设想中，多媒体取证系统是由专用工具构成的协调框架。其核心（图3）是一个人工智能驱动的协调器，它根据待分析媒体的特征，决定使用哪些工具、以何种顺序使用，以及各自的用途。提交内容后，协调器首先分析其语义和技术属性，例如元数据、格式、压缩历史以及任何可用的在线相关内容。借助网络搜索或来源分析（例如 C2PA[1]），它可以在使用有针对性的取证工具开展分析之前，重建媒体的生命周期。这些工具可能包括生物特征一致性检查、透视/阴影分析、指纹提取，和/或用于合成与真实内容分类及篡改定位的一些更通用算法。随后，协调器将这些证据综合成一份结构清晰、透明的报告，解释其结论背后的推理过程，包括为何对某些线索赋予更高权重（例如检测器训练领域、可靠性历史、模态一致性）。

鉴于取证工具会提供各种异构输出，例如检测分数、相似度分数和异常分数，人工智能智能体应妥善融合这些指标：这一过程的核心是对**不确定性进行严格量化**。系统不仅应说明其预测结果，还应说明自身的置信程度，以及

[1]https://c2pa.org/ C2PA 是一项开放标准，供出版商、创作者和消费者确认数字内容的来源及其编辑记录。

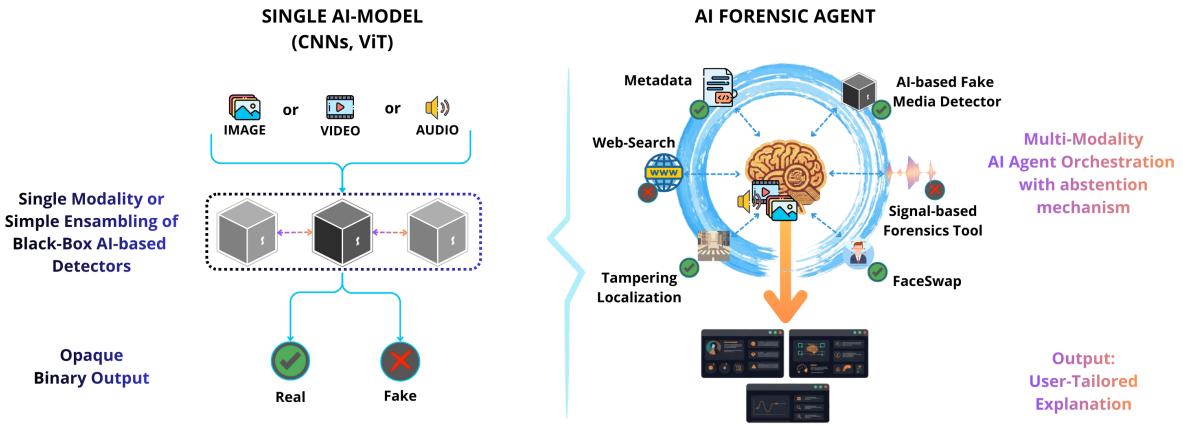


图 2：重新构想多媒体取证：从单一检测器系统转向人工智能取证协调器。左图：传统集成方法将一组固定模型组合起来，生成单一且往往不透明的置信度分数。右图：人工智能取证代理范式会动态选择、分析并融合多种工具，例如低层取证痕迹、人脸生物特征、元数据分析和基于网络的溯源信息检索。这种基于代理的架构能够实现明确的推理、经过校准的不确定性评估，以及对新检测器的自适应整合，从而使取证分析更加稳健、透明。

在证据不足时是否应弃答。在当前的多媒体取证实践中，不确定性往往只是流于表面：置信度分数很少经过校准，而且在分布偏移下的表现也基本不为人所知。

医学影像、天气预报、金融风险建模和自动驾驶等相邻的安全关键领域，提供了更成熟的方法，相关社区可以借鉴。温度缩放、深度集成、贝叶斯神经网络和适当评分规则等技术，展示了如何将概率预测与经验误差率联系起来，并系统评估其校准和区分能力 [12]。当检测器意见不一或不确定性较高时，取证智能体应采用**选择性预测**或**拒绝选项**机制，选择弃权，而不是提出误导性结论。诸如保形预测之类的框架通过为保留的案例提供误差界，来形式化这一行为。这些方法能够持续减少过度自信导致的错误，并提升用户信任；这些特性对于取证分析同样至关重要，因为误报可能损害声誉，或影响新闻报道和法律结果。

另一个挑战在于，如何汇总基于不同线索运行的检测器所面临的不确定性：低层级指纹、语义不一致、元数据痕迹或跨模态相关性。融合过程应考虑每种工具的可靠性，纳入相关性因素，并惩罚过度自信的错误。概率图模型、集成贝叶斯推断和不确定性感知多模态融合提供了颇具前景的方向。尽管布里尔分数等适当评分规则在多媒体取证领域尚未成为标准，但它们为评估汇总概率如何反映现实世界不确定性提供了一种有原则的方法。

上下文感知对于做出可靠决策也至关重要。新闻编辑部、社交媒体平台或执法调查等不同的现实场景，需要采用不同的评估指标、不同的证据权重分配方式，以及不同的决策阈值。为此，协调器可以利用记忆机制，存储有关既往案例、模型行为和特定领域约束的信息。这样，系统便能调整自身行为，避免重蹈已知失效模式（例如，曾导致误报或漏检的错误）。

从问责角度来看，系统应清晰记录每项内容的分析过程：使用了哪些模型及其版本、采用了哪些参数、按什么顺序进行分析，以及由哪位分析人员或哪个系统组件执行分析。这有助于建立稳健的“数字证据保管链”，从而能够复现验证流程。框架还应能够以非技术利益相关方（法官、律师、记者）可以理解的方式，说明所用方法、不确定性和局限性；同时，应允许在有更好的工具或新证据可用时更新结果。

表 1 总结了所提出的愿景。总体而言，这些原则为多媒体取证智能体指明了一条发展路径，使其既具备强大的技术能力，又能在运营中恪守责任。在适当的

表1：从独立检测器到编排器：对比当今的多媒体取证实践与新兴的 AI 取证智能体范式。

方面	当今实践	拟议愿景
检测器的作用	检测器作为彼此隔离的真相来源运行，每个都专注于特定的篡改方式或模态。	专用 工具 协调编排 动态地基于上下文、可用证据和调查目标。
结果融合	薄弱或临时性的融合（例如手动阈值， 多数投票），往往缺乏原则性的不确定性处理。	概率推理与结构化融合跨 模态， 纳入 不确定性，可靠性和互补信号。
适应性	对新的生成技术较为脆弱；存在较高的快速过时风险。	模块化、持续自适应编排参考既往案例，并适应不断演变的操纵手法模式。
溯源	可见性有限；训练数据、决策路径和工具交互很少能够追溯。	Full 溯源链 详细说明 使用了哪些工具按什么顺序使用，以及每种工具如何促成了最终评估。
不确定性处理	二元决策（真实/伪造）或校准不佳的分数；对置信度缺乏透明度。	校准概率，并在证据不足或相互矛盾时基于原则弃答，当证据不足或相互矛盾时。
用户交互	千篇一律的输出，可能晦涩难懂非专家。	根据用户的专业程度量身定制解释（例如，调查人员、记者、法律专业人士），并提供清晰的理由。
范围	单模态 关注或依赖 单独检测器。	多模态 编排集成 图像，视频，音频，元数据 ，以及跨模态推理-理。

治理机制下，经协调的系统能够在维护人类权威、透明度和伦理保障的同时，增强调查和验证工作流程。

迈向可解释性

可解释性对于建立人们对多媒体取证智能体的信任至关重要。这些系统必须提供透明且可解释的输出，帮助用户了解不仅得出了什么结论，还要了解为何该结论有充分依据。研究人员、工程师、记者和平台运营者应能追溯评估背后的推理，并判断支撑证据是否充分。信任源于清晰性、可复现性，以及系统推断与底层取证线索之间始终如一的联系。

当前方法及其局限性 可解释人工智能（XAI）的最新进展提供了有用的原则，但将其应用于多媒体取证仍面临重大挑战。在目标识别等计算机视觉领域，显著性图和基于概念的解释有助于阐明模型的推理。然而，取证痕迹往往十分细微、经过蓄意遮掩，并且可能涉及多种模态。因此，难以将现有的可解释人工智能方法应用于相关解释

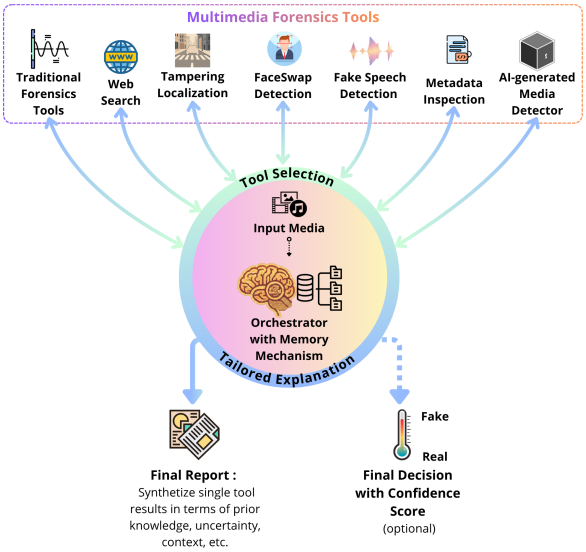


图3：人工智能取证智能体的概念架构。中央编排器选择并组合多种检测工具——从低层次的取证分析到元数据和溯源核查——并将这些工具的输出综合为一份透明的最终报告，其中包含校准置信度分数。

噪声级别不一致、压缩伪影、元数据异常和跨模态相关性等方面，而这些因素对于取证评估都可能至关重要。

如图所示，多媒体取证中的可解释人工智能仍处于起步阶段。4在2024年之前，该领域主要聚焦于事后可视化技术，通过利用梯度加权类激活映射（Grad-CAM）、LIME 和 SHAP 生成的热力图等计算机视觉方法，突出深度学习模型的决策过程 [13]。虽然这些技术能够识别视觉伪影，例如融合边界或颜色异常，但当取证痕迹并非局部存在，而是分布于整个媒体中时（例如完全由人工智能生成内容的情况），它们无法提供明确的证据。此外，采用不同方法获得的事后可视化结果往往会突出不同区域，导致相互矛盾的指示，使非专业用户难以解读。此外，基于物理或生理信号的方法（例如角膜镜面高光或不规则瞳孔形状）受严格的环境假设限制，因此在不受约束的场景中适用性有限 [1]。

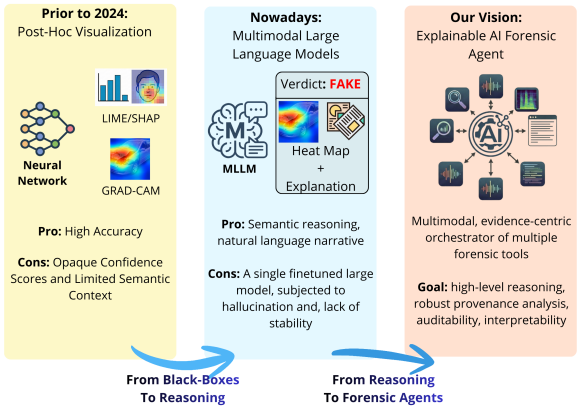


图 4：数字取证领域可解释人工智能的发展演变。左：早期的“黑箱”神经网络依赖事后可视化。中：现代多模态大语言模型（MLLM）赋予的新兴语义推理能力。右：本文提出的可解释人工智能取证智能体，它整合以证据为重点的工具，以支持改进推理、可审计性和可解释性。

多模态大语言模型（MLLMs）近期作为可解释取证的潜在组成部分引发了广泛关注，但其当前能力仍不足以在这一领域可靠应用。新兴方法正探索通过微调多模态大语言模型，使其执行从二元检测到细粒度伪造定位等取证任务，从而为更高级的语义推理开辟道路 [14]。虽然它们能够生成高层次描述并标记粗略的视觉不一致之处，但这类分析仍局限于语义缺陷。

我们对可解释取证人工智能智能体的愿景 为突破这些局限，未来多媒体取证领域的可解释性研究应采用混合型证据中心方法，明确将模型决策与可测量的领域特定取证线索联系起来。下一代取证智能体不应仅依赖事后可视化，而应指出检测到了哪种信号、如何测量该信号，以及该信号为何表明存在篡改。基于可量化伪影的解释——例如频率不一致、传感器模式扰动、光照几何、运动场偏差或元数据与像素不匹配——更有可能经受住对抗条件的考验，并支持人工核验。

在我们提出的人工智能智能体编排器构想中，多模态大语言模型可以发挥积极作用，通过人类可理解的解释实现可解释性。它们能够将专用检测器的输出整理成连贯的叙述，并帮助用户理解某些伪影为何重要：将低层次的取证证据转化为**更高层次的推理**。在这方面，可以利用思维链推理生成解释，明确说明检测器输出如何与最终决策相关联。这有助于清楚展示每条线索如何影响最终决策，并突出说明决策何时受到特定类型证据的驱动，或受到歧义的限制。在人工智能智能体编排器中，这类思维链式解释充当低层次取证分析与人类决策者之间的接口层，将原始完整性分数转化为透明的推理分析。更进一步，可以考虑多智能体环境，其中每个智能体独立提出答案，随后共同讨论各自的回答和推理，以达成共同结论。这种迭代交互有助于构建事实准确性更高、幻觉更少的答案 [15]。

可解释性也受益于**稳健的溯源分析**。了解一段媒体内容的来源、它经历了哪些转换，以及它经由哪些平台传播，有助于提供背景信息，从而增强真实性评估的可解释性。将被动取证技术与数字水印或基于 C2PA 的溯源等主动方法相结合，有助于厘清观察到的

痕迹是否源自良性的编辑流程。

最后，可解释性应与**隐私保护**并存。多媒体验证经常涉及敏感材料，解释不应披露不必要的个人细节。安全处理、访问限制，以及在适当情况下使用特征级表示，有助于确保通过透明度获得的信任不会因隐私风险而受到损害。

结论

所有文明社会都依赖信任机制。生成式人工智能很容易被恶意行为者利用，通过传播错误信息来破坏**社会信任**。人工智能生成媒体近期的爆发式增长，可能助长一种普遍观念：我们所看到或听到的内容很可能是合成的或虚假的。反过来，这可能侵蚀社会凝聚力，并损害民主进程。为缓解这一问题，我们提出了能够提供可解释且可靠工具的人工智能取证编排器。为此，应解决以下研究挑战：

- 基于不确定性的分析：取证智能体跨模态编排多种检测器，并提供校准置信度分数；当不确定性较高时明确避免作出判断，从而遵循以下原则：**不要猜测，升级处理**。

- 溯源追踪：结合主动与被动方法的取证智能体，例如 C2PA、生成式人工智能的数字水印，以及合成与真实内容取证检测器。该框架必须应对极具挑战性的场景，包括追踪社交媒体平台上的分享链；平台特有的压缩、尺寸调整和滤波会削弱取证痕迹。

- 可解释性：能够提供可解释结果的取证智能体，涵盖低级和高级取证痕迹。这将确保我们的框架在法律上具备可采性，从而能够整合到调查与司法工作流程中。

总而言之，我们提议采用人工智能取证智能体，使其能够整合多种工具，涵盖从被动到主动的方法、从检测到归因，并对异构证据进行推理。所提出的框架必须提供校准置信度分数，并在可用线索薄弱或相互矛盾时避免作出明确判断，从而降低过度自信预测的风险。这要求进行领域感知推理，以提供可解释且有科学依据的论证。只有通过这种综合而透明的推理方式，取证智能体才能获得在现实运营环境中部署所必需的信任。

致谢

本研究由欧盟通过地平线欧洲计划资助，资助协议编号为 101092887（NGI Sargasso–Open Call OC5 | DeepshieldProject）。

参考文献

1. L. Verdoliva，“媒体取证与深度伪造：概述”，IEEE 信号处理精选主题期刊，第14卷，第5期，第910–932页，2020年。2. W. Peebles 和 S. Xie，“基于 Transformer 的可扩展扩散模型”，载于国际计算机视觉大会（ICCV），2023年。3. H. Deng、T. Pan、H. Diao、Z. Luo、Y. Cui、H. Lu、S. Shan、Y. Qi 和 X. Wang，“无需向量量化的自回归视频生成”，载于国际学习表征会议（ICLR），2025年。4. M. Li、Y. Ahmadiadli 和 X.-P. Zhang，“语音深度伪造检测综述”，ACM 计算调查，第57卷，第7期，第1–38页，2025年。5. D. Tariang、R. Corvi、D. Cozzolino、G. Poggi、K. Nagano 和 L. Verdoliva，“生成式人工智能时代的合成图像验证：哪些方法有效，哪些仍有待探索”，IEEE 安全与隐私，第22卷，第37–49页，2024年。6. Y. Zou、P. Li、Z. Li、H. Huang、X. Cui、X. Liu、C. Zhang 和 R. He，“人工智能生成媒体检测综述：从非多模态大语言模型到多模态大语言模型”，arXiv 预印本 arXiv:2502.05240v2，2025年。7. Z. Huang、J. Hu、X. Li、Y. He、X. Zhao、B. Peng、B. Wu、X. Huang 和 G. Cheng，“SIDA：利用大型多模态模型进行社交媒体图像深度伪造检测、定位与解释”，载于 IEEE 计算机视觉与模式识别会议（CVPR），2025年。8. P. Yu、J. Fei、H. Gao、X. Feng、Z. Xia 和 C.-H. Chang，“释放大型视觉语言模型在可泛化且可解释的深度伪造检测中的能力”，载于国际机器学习会议（ICML），2025年。9. S. H. Silva、M. Bethany、A. M. Votto、I. H. Scarff、N. Beebe 和 P. Najafirad，“深度伪造法证分析：一种可解释的弱监督模型层次集成方法”，《国际法医学：协同》，第4卷，文章编号100217，2022年。

10. X. Li、Y. Zeng、X. Xing、J. Xu 和 X. Xu，“Hedgeagents：一种具备平衡意识的多智能体金融交易系统”，载于 ACM 网络会议2025年大会配套论文集，2025年。
11. P. Peigné、M. Kנותski、F. Sondej、M. David、J. Hoelscher-Obermaier、C. S. de Witt 和 E. Kran，《多智能体安全税：在多智能体系统中权衡安全性与协作能力》，载于 AAAI 人工智能会议论文集，2025年。
12. R. Rahaman 和 A. H. Thiery，《不确定性量化与深度集成》，载于神经信息处理系统年会（NeurIPS），2021年。 13. M. S. Momin、A. Sufian、D. Barman、M. Leo、C. Dis- tante 和 N. Damer，《跨不同模态的可解释深度伪造检测：方法与挑战综述》，《图像与视觉计算》，第 105738 页，2025年。 14. Y. Li、X. Liu、X. Wang、B. S. Lee、S. Wang、A. Rocha 和 W. Lin，《Fakebench：通过大型多模态模型探究可解释假图像检测》，IEEE《信息取证与安全汇刊》，第 20 卷，第 8730–8745 页，2025年。 15. Y. Du、S. Li、A. Torralba、J. B.Tenenbaum 和 I. Mordatch，《通过多智能体辩论提升语言模型的事实准确性与推理能力》，载于国际机器学习会议（ICML），2024年。

Giulia Boato 是意大利特伦托大学的正教授，也是 Truebees 的联合创始人。自 2012 年起，她一直领导媒体实验室的多媒体取证研究团队。她的研究兴趣包括图像与信号处理，尤其关注多媒体数据保护与认证、数据隐藏和数字取证，以及智能多维数据管理与分析。她已在国际会议和期刊上发表 170 多篇论文。联系邮箱：giu- lia.boato@unitn.it。

Andrea Montibeller 是 Truebees srl 的联合创始人兼首席人工智能官，也是意大利特伦托大学的助理教授。他于 2024年 获得特伦托大学信息与通信技术博士学位，此前于 2020年 获得理学硕士学位。他曾在德雷塞尔大学（2023年）和维戈大学（2018年）担任研究学者。其研究兴趣包括人工智能生成图像/视频伪造检测与定位，以及相机来源归属。联系邮箱：andrea@truebees.eu。

Edward J. Delp 出生于俄亥俄州辛辛那提。他是美国普渡大学电气与计算机工程查尔斯·威廉·哈里森杰出教授兼生物医学工程教授。他的研究兴趣包括图像分析、计算机视觉、机器学习、图像与视频压缩、多媒体安全、医学影像、多媒体系统、通信和信息论。Delp 博士是 IEEE 终身会士、ACM 会士、SPIE 会士、成像科学与技术学会（IS&T）会士、美国医学与生物工程学会会士，以及国家发明家科学院会士。请通过 ace@purdue.edu 联系他。

LuisaVerdoliva 是那不勒斯腓特烈二世大学电气工程与信息技术系的正教授，地址为意大利那不勒斯 80125。她的研究兴趣集中于图像与视频处理领域，主要贡献在多媒体取证方面。Verdoliva 获得了那不勒斯腓特烈二世大学信息工程博士学位。她是 IEEE《信息取证与安全汇刊》的主编，并获得了2025年科学前沿奖、2018年谷歌教师奖以及慕尼黑工业大学高等研究院汉斯·菲舍尔高级研究员奖。她是 IEEE 会士。联系方式：verdoliv@unina.it。

Daniele Miorandi 是 Truebees 的首席执行官、Afliant（原 U-Hopper）的人工智能负责人，也是 Trentino AI 的联合创始人。作为连续创业者，他目前的研究和业务兴趣包括可信人工智能、多媒体取证和隐私保护分析。他已发表约200篇关于物联网、大数据平台和应用人工智能的科学文章与论文。Miorandi 获得了意大利帕多瓦大学通信工程博士学位。联系方式：daniele@truebees.eu。