共同看、听与知：深度伪造视频检测中的多模态策略

Chen Chen 新加坡南洋理工大学，新加坡
CHEN033@e.ntu.edu.sg

Dion Hoe-Lian Goh 新加坡南洋理工大学，新加坡
ASHLGoh@ntu.edu.sg

arXiv:2602.01284v1 [cs.MM] 1 Feb 2026

摘要

随着深度伪造视频越来越难以被人们识别，了解人们采用的策略是设计有效媒体素养干预的关键。我们对195名年龄在21至40岁之间的参与者开展了一项研究，请他们判断真实视频和深度伪造视频、评估自己的置信度，并报告他们在视觉、音频和知识策略中所依赖的线索。参与者对真实视频的判断比对深度伪造视频更准确，对真实内容的期望校准误差也更低。通过关联规则挖掘，我们识别出影响表现的线索组合。视觉外观、人声和直觉经常同时出现，并与成功识别相关，这凸显了多模态方法在人类检测中的重要性。研究结果揭示了哪些线索有助于或妨碍检测，并为设计能够引导人们有效运用线索的媒介素养工具提供了方向。基于这些见解，人们可以提升识别技能，增强抵御欺骗性数字媒体的能力。

CCS概念

• **人类 本中心计算 → HCI中的实证研究 中的。**

关键词

深度伪造检测、人类判断、策略使用、多模态线索、校准、媒介素养

1 引言

深度伪造视频是指高度逼真的人工智能生成视频，通过篡改或替换人脸来制作，已迅速成为一种具有挑战性的网络虚假信息形式 [10, 23]。随着这类视频在社交媒体上传播，它们可能误导受众、损害他人声誉，并削弱人们对数字内容的信任 [7, 22, 38]。尽管公众对深度伪造视频的认知不断提高，人们仍难以识别它们 [11, 17]。研究表明，人们常常会把深度伪造视频误认为真实视频 [17, 25]，而且许多人对自己判断的置信度，远高于其实际准确率所能支撑的程度 [17]。这种易受误导与过度自信并存的情况表明，在当今的数字媒体环境中，人类判断往往无法区分真实与伪造。

为应对这些挑战，学者们提出了干预措施，以提升人们识别深度伪造视频的能力。例如，研究发现，在评估前接触篡改内容可以提升辨别敏感度 [36]。其他研究探索了人机协作，结果表明，提供人工智能预测 [18] 或解释说明 [13, 18]，有助于人们更准确地识别深度伪造视频。尽管这些方法展现出潜力，但它们往往侧重于向参与者指定特定线索或提供指南，却没有先厘清人们自然会如何完成这项任务，以及哪些策略确实会影响表现。

因此，了解人们采用的识别策略至关重要。以往研究表明，人类与 AI 检测器处理这项任务的方式各不相同。计算机视觉模型能够识别细微的像素级伪影或异常，而这些往往难以被人眼察觉 [23]，；人们则依赖整体印象和情境知识 [12]。这些差异凸显了人类与机器在解决这一问题时各具优势，也表明对人类而言，检测可能取决于多种策略的结合，例如留意所见（视觉策略）、所闻（音频策略）和所知（知识策略）[11, 37]。每种策略都对应着特定线索。视觉策略涉及诸如眨眼不规律、光线不自然或唇部动作不匹配等线索。音频策略则依据音轨中的线索，例如语调、流利度或同步情况。知识策略依赖语境线索，例如了解某人可能会说什么、回想此前是否看过该视频，或通过事实核查交叉验证其中的主张 [11, 37]。这些类别涵盖了人类判断者可用的主要感知资源与外部资源，共同体现了人们处理和评估媒体内容的核心方式。

尽管先前研究已识别出人们用于判断深度伪造视频的几类策略，但这些发现大多停留在定性层面。对于人们在实际评估深度伪造视频时如何运用这些策略，我们仍缺乏清晰的定量认识。此外，过去的研究孤立地考察各种策略，因此难以确定哪些线索最重要、多种策略如何相互作用，以及在评估真实视频与深度伪造视频时，哪些策略会奏效或造成误导。理解这些过程不仅对推进理论至关重要，也有助于设计有效的媒体素养干预，引导用户更可靠地运用线索。

以往研究表明，年轻人经常使用在线媒体和社交网络平台，因此他们是虚假信息检测和媒介素养研究中特别值得关注的对象 [28]。因此，我们对195名年龄在21至40岁之间的年轻人开展了一项研究，请他们判断真实视频和深度伪造视频、报告自己的置信度，并描述影响其判断的线索。我们的分析着重探讨以下三个研究问题：

本文是已被 CHI 2026 接收的论文预印本。最终版本将在 ACM 数字图书馆发布。

研究问题1：人们判断真实视频与深度伪造视频时表现如何？他们的置信度校准得如何？

研究问题2：不同的检测策略（视觉、音频和知识策略）单独使用或结合使用时，对区分真实视频与深度伪造视频的准确率和校准有何贡献？

研究问题3：具体线索如何影响检测表现？在判断视频是真实还是深度伪造时，这些线索又会如何相互作用？

通过探讨这些问题，我们强调了“共同观看、聆听与认知”的重要性。研究结果表明，没有任何单一线索能够确保成功。相反，综合多种线索能够增强检测能力。

本文为媒介素养与虚假信息研究做出了三项贡献。第一，我们提供了有关人们如何判断真实视频和伪造视频的实证证据，包括其准确率和信心校准。第二，我们以定量方式展示了视觉、音频和知识策略及其组合与表现之间的关系。第三，通过考察各类策略中的线索，并绘制这些线索在网络中的相互作用，我们揭示了具体线索如何支持或阻碍对真实视频和伪造视频的检测。这项分析指出了哪些线索具有诊断性，哪些线索会产生误导，以及它们之间的相互作用如何影响表现。最后，本文还提出了媒体素养干预的设计启示。

2 相关工作

2.1 人类深度伪造检测表现

先前研究表明，在特定情境下，人类有时能够胜过人工智能检测模型。例如，判断名人视频 [12], 中的社会互动 [8],，或评估来自熟悉的文化或性别背景的个人 [19]。与此同时，如果没有结构化训练，总体检测准确率往往接近偶然水平 [9],，而且人们往往对自己的判断过度自信，这会增加其容易受到操纵影响的程度 [9, 16]。置信度并不总是与准确率一致，因为个人即使判断错误，也可能自信地将深度伪造内容判定为真实 [4, 17]。过度自信会加剧人们对虚假信息的易感性。例如，过度自信的个人更常访问不可信网站，也更难区分真实与虚假的主张 [20]。他们也较少通过外部来源核实信息 [26]。

另一方面，一些研究发现，置信度与检测准确率之间存在微弱的正相关 [34]。从认知心理学的角度来看，关键在于校准，即置信度与实际正确性之间的一致程度。校准良好的判断与更高的决策质量相关 [2, 32]。游戏化训练等干预措施能够提升置信度和校准水平，帮助人们更有效地抵御虚假信息 [2, 33]。因此，置信度既是人们处理虚假信息的驱动因素，也是其结果，它会影响人们如何应对虚假内容，也会受到过往经历的重塑 [29]。

2.2 人类检测深度伪造视频中多种策略的使用

为了同时提升准确率和校准，理解影响人类真实性判断的因素至关重要。先前研究表明，人们会依赖三种主要策略来评估媒体内容是真实还是伪造：视觉策略、音频策略和基于知识的策略 [11, 28, 37]。视觉策略包括留意不自然的眨眼、扭曲的面部特征、光照不一致或过于光滑的皮肤纹理等伪影 [11]。音频线索则提供了另一项重要信息来源。人们可能会察觉到异常的语调、奇怪的停顿，或听起来略显人工的声音 [39]。唇部动作与语音不一致，例如说出的话与说话者的嘴型不符，也会进一步引起怀疑 [35]。最后，基于知识的策略让人们能够利用有关某个人或某个事件的先验信息。例如，如果视频中一位公众人物发表了不合常理或有悖其一贯表现的言论，人们可能会怀疑其真实性 [5]。

这些发现表明，人类判断往往会结合感知线索和语义线索，而非依赖单一信息来源。然而，既有研究大多孤立地对这些策略进行定性考察 [11, 37]。我们对人们在实践中如何结合多种策略，以及这些组合如何影响准确率和置信度，所知甚少。随着多模态深度伪造检测研究的发展，这一研究空白尤为重要，因为整合视听信号的算法表现始终优于仅依赖单一模态的算法 [27, 30]。

3 方法

3.1 参与者

我们通过便利抽样和滚雪球抽样招募了 200 名参与者。招募工作通过一则在社交媒体平台上发布的在线海报进行。每位参与者完成研究后均获得 10 新加坡元作为补偿。剔除不完整或无效的答复后，最终数据集包含 195 名参与者。每位参与者评估了四段视频，共形成 780 条有效的视频判断，并提供置信度评分和自我报告的策略使用情况。本研究已获大学机构伦理委员会批准。

在195名参与者中，女性111人（57%），男性84人（43%）。大多数参与者年龄介于21至30岁之间（171人，占87.7%），其余24名参与者（占12.3%）年龄介于31至40岁。在教育程度方面，略多于半数的参与者拥有学士学位（104人，占53.3%），其次为具有大专或同等学历者（75人，占38.5%），另有少数参与者拥有硕士学位（14人，占7.2%）。少数参与者报告称其教育程度属于其他级别（2人，占1.0%）。表1详细列出了参与者人口统计信息。

表1：参与者人口统计信息（N = 195）。

人口统计变量	类别	人数	%
性别	Male	84	43.0
	女性	111	57.0
Age	21–30岁	171	87.7
	31–40岁	24	12.3
教育程度	大专或同等学历	75	38.5
	学士学位	104	53.3
	硕士学位	14	7.2
	其他	2	1.0

3.2 刺激材料

参与者观看了同时包含视觉和音频成分的视频。每段视频都很短（通常少于60秒），内容是一位公众人物发表口头信息。我们的刺激材料集总共包含20段视频：其中10段为真实视频，10段为深度伪造视频，均从YouTube等公众可访问的来源收集。深度伪造刺激材料的选择旨在反映虚假信息中常见的操纵方式，包括换脸、唇形同步和面部动作重演，并涵盖政治和体育等一系列主题，以最大限度地提高多样性。附录B包含视频内容、时长以及两段真实视频和两段深度伪造视频的示例截图信息。

3.3 流程

在提供知情同意后，参与者通过在线平台单独完成本研究（附录D中提供了截图）。每位参与者观看了从上述包含20个视频的刺激材料集中随机抽取的四段短视频。对每位参与者而言，两段真实视频和两段深度伪造视频以随机顺序呈现，以尽量减少顺序效应。对于每段视频，参与者需要：（1）判断真实性，指出视频是真实还是伪造；（2）在五点李克特量表上评定其置信度（1 =完全没有信心， 5 = 非常有信心）；（3）使用结构化问卷指出影响其判断的策略。问卷包含三类线索：视觉（例如面部特征）、音频（例如语调）以及基于知识的线索（例如对当事人或事件的先验知识）。这些线索基于先前关于深度伪造识别的研究 [11, 37]。参与者可以在每种策略中选择多个线索。每位参与者重复此流程四次（每个视频一次）。向参与者呈现的完整线索条目列于附录C，每个条目均以带有简短描述的复选框选项形式显示。完成所有视频评估后，参与者还提供了年龄、性别、教育背景等人口统计信息。

3.4 指标

3.4.1 期望校准误差（*ECE*）。我们使用期望校准误差（ECE）[14, 24]评估校准情况。ECE 的计算方法是：根据报告的置信度分数，将预测划分为 M 个大小相等的区间，并对每个区间内准确率与置信度之间的绝对差值取平均。具体而言，

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \tag{1}$$

其中， N 是预测总数， B_m 是第 m 个置信度区间中的样本集合， $\mathrm{acc}(B_m)$ 是该区间的经验准确率， $\mathrm{conf}(B_m)$ 是报告的平均置信度。ECE 越高，表明校准失准越严重，即置信度水平与经验准确率不相符。

3.4.2 准确率。除了校准之外，我们还报告准确率，其定义很简单，即正确识别的项目所占的比例：

$$\mathrm{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}, \tag{2}$$

其中， $\hat{y}_i$ 是预测标签， y_i 是第 i 个项目的真实标签，而 $\mathbf{1}\{\cdot\}$ 是指示函数。准确率直观地衡量了总体检测表现，而期望校准误差（ECE）则量化了主观置信度与客观正确性之间的一致程度。

3.5 数据分析

我们的分析分为三个步骤。首先，我们比较了真实视频与深度伪造视频的总体表现。具体而言，我们使用配对样本t检验考察了准确率和期望校准误差（ECE）的差异。

其次，我们考察了自我报告的策略在塑造表现结果方面的作用。为此，我们进行了方差分析（ANOVA），比较依赖视觉、音频或基于知识的策略的参与者的准确率。随后进行事后检验，以确定哪些具体策略或策略组合造成了表现上的显著差异。

最后，为了超越个人策略层面，我们应用关联规则挖掘，识别三种策略中线索的频繁共现及其表现。在分析中，前件是参与者报告使用的具体线索，后项则是检测结果（识别正确或错误）。为确保模式具有意义，我们将支持度阈值设为 0.20，置信度阈值设为 0.60。支持度阈值为 0.20，意味着只有在至少 20% 的所有试次中出现的前件线索组合才会纳入考量，这有助于过滤罕见的线索集，避免得出可能过拟合的结论。置信度阈值为 0.60，则确保规则在至少 60% 的相关案例中能够正确预测检测结果。这些数值是通过对数据集进行预试探索得出的，也符合文献中的常见做法 [15]。由此，我们为深度伪造视频得出了 66 条关联规则，为真实视频得出了 128 条。随后，我们使用这些前件集构建子数据集，其中仅纳入前件中的所有线索均被选中的试次。对于每个线索集，我们都计算了平均准确率和期望校准误差。为直观呈现线索间关系，我们将包含多个线索的前件集拆解为所有可能的双线索组合。这些组合在网络图中表示为边，节点大小反映线索频率，边的粗细则反映平均表现（准确率或期望校准误差）。这种方法不仅能够捕捉单个线索的影响，还能展现线索在网络中的交互方式。

4 结果

4.1 人们识别深度伪造视频的能力如何？（研究问题1）

表2总结了参与者在总体、真实视频和深度伪造视频条件下的表现。总体准确率为 0.76。参与者识别真实视频（$M = .79$, $SD = .31$ $M = .72$ $SD = .33$ t 194 ）的表现显著优于识别深度伪造视频（, ）, （） = −2.08 $p = .038$, 。

然而，就置信度而言，参与者在判断真实视频（$M = 3.57$, $SD = .91$）和深度伪造视频（$M = 3.56$, $SD = .87$）时报告的置信度几乎相同。差异在统计上并不显著， $t(194) = -0.14$, $p = .89$。这表明，尽管

判断深度伪造视频时的准确率有所下降，但参与者的自信程度与判断真实内容时一样。

校准分析凸显了这种不匹配。深度伪造判断的期望校准误差（ECE）($M = .32, SD = .29$)高于真实视频的期望校准误差（$M = .24, SD = .26$），$t(194) = 2.96, p = .003$。换言之，参与者评估深度伪造视频时，其置信度与实际准确率的吻合程度较低。

图1展示了所有判断的混淆矩阵。参与者正确识别出282个深度伪造视频和308个真实视频，漏判了108个深度伪造视频，并将82个真实视频错误标记为伪造。与错误地将真实视频分类为伪造相比，参与者更容易将深度伪造视频误判为真实。

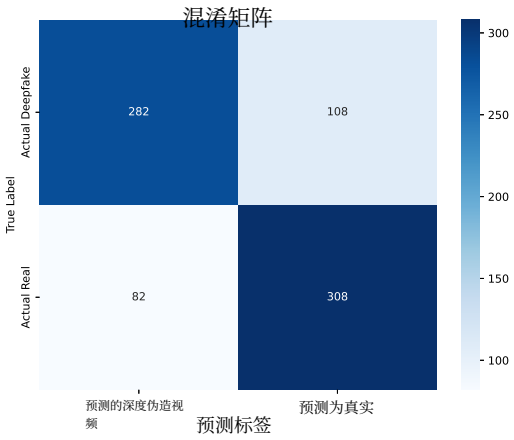


图1：混淆矩阵。

4.2 多种策略能否提升检测效果？（研究问题2）

表3展示了参与者报告使用的各种策略的检测表现。仅视觉（V）、仅音频（A）和仅知识（K）指参与者只选择了该策略中的线索。其次，视觉 + 音频（V+A）表示同时选择了视觉和音频线索，但未选择知识线索。最后，视觉+ 音频 + 知识（V+A+K）表示使用了所有三种策略类型的线索。深度伪造视频的试次中，约80%同时使用了所有三种策略；而真实视频中，这一比例为75%。

对于深度伪造视频，仅视觉和仅音频的使用取得了最高准确率，但样本量太小，结果不可靠。同时使用视觉和知识策略的参与者取得了第二高的准确率，且期望校准误差较低（0.95，0.18；$N = 39$），但样本量仍然有限。相比之下，对真实视频采用相同的策略组合，准确率较低，期望校准误差较高（0.50，0.21；$N = 18$）。使用多种策略（单一、双重或三重）并未显著影响深度伪造检测。策略分组未影响准确率，$F(3, 386) = 1.96$，$p = .120$；期望校准误差也未呈现显著的组间差异，$F(3, 241) = 1.42$，$p = .238$。这些发现表明即使参与者使用多种策略，深度伪造视频仍始终难以检测。

然而，对于真实视频，策略差异发挥了更重要的作用。单因素方差分析显示，策略对准确率有显著影响，$F(3, 386) = 3.41$，$p = .018$。采用多种策略的方法，尤其是结合知识来源与音频策略、视觉与音频策略，或同时采用全部三种策略的方法，表现优于仅使用视觉或仅使用音频的情况，部分情况下准确率接近0.90。这表明，将参与者看到或听到的内容与知识来源进行交叉核对，有助于他们更有效地识别真实内容。相比之下，仅依赖仅视觉策略会导致参与者在真实视频上的表现非常差（准确率为0.13），尽管该子组人数较少（$N = 8$）。

事后检验（表4）进一步对各组策略进行了统计比较。对于真实视频，总体方差分析显示各组在准确率上存在显著差异，但各组间的两两比较均未达到通常的显著性水平。最接近显著性的是第3组（三重策略）与第1组（单一策略）之间的差异，该差异接近显著（$p = .057$），这表明采用多种策略可能带来一定优势。此外，第2组（双重策略）的校准显著优于第1组和第3组，$p = .038$，这意味着他们的置信度与实际准确率更加吻合。

总的来看，研究问题2的结果显示出明显的不对称性。尽管策略组合并未提升深度伪造检测效果，但它们确实影响了真实视频的检测表现。多种策略，尤其是包含知识线索的策略，与更高的准确率和更好的校准相关。

4.3 哪些线索会帮助或误导人类？（研究问题3）

表5展示了具体的视觉、音频和知识线索如何影响参与者在准确率和期望校准误差方面的检测表现。关于线索定义的更多细节见附录 A。

视觉线索。 依赖视觉外观（例如面部异常）取得了相对较好的结果：深度伪造视频的准确率为 .84，真实视频为 .78。该差异具有统计显著性，$t(194) = 2.10$，$p = .036$，这表明，与真实视频相比，外观线索对于识别篡改内容尤其有帮助。相比之下，视觉环境线索（例如背景不一致）的总体准确率较低（.58 对 .55），不同视频类型之间没有显著差异。视觉制作质量线索的总体表现最差，不过深度伪造视频的准确率略高（.48 对 .41）。然而，依赖制作质量与显著更高的校准误差相关，$t(194) = 2.19$，$p = .030$，这表明关注制作质量的参与者往往过度自信。

音频线索。 声音线索（例如语调）提高了真实视频准确率（.81），相比深度伪造视频（.69），$t(194) = -3.58, p < .001$，并且也显示出更好的校准，$t(194) = 3.56, p < .001$。同样，语言线索（例如措辞）支持了真实视频准确率（.55 vs. .47, $p = .026$），并改善了校准，$t(194) = 4.48, p < .001$。相比之下，音质线索未能区分真实与伪造（.49 vs. .52, $t(194) = -0.86, p = .391$），但产生了更差的校准

表2：真实视频与深度伪造视频的表现比较。数值以参与者层面的均值 ± 标准差报告。统计检验采用配对样本 *t*检验。

测量指标	真实（M ± 标准差）	深度伪造（M ± 标准差）	Δ (真实–伪造)	*t*-测试
总体				
准确率		0.76 ± 0.24	—	—
置信度		3.57 ± 0.77	—	—
ECE		0.27 ± 0.19	—	—
真实与深度伪造				
准确率	0.79 ± 0.31	0.72 ± 0.33	0.07	$t(194) = -2.08,\ p = .038$
置信度	3.57 ± 0.91	3.56 ± 0.87	0.01	$t(194) = -0.14,\ p = .89$
ECE	0.24 ± 0.26	0.32 ± 0.29	−0.08	$t(194) = 2.96,\ p = .003$

表3：按策略类型及组合划分的准确率与校准误差（期望校准误差），并按深度伪造与真实视频拆分。数值为试次级均值。单因素方差分析检验了各分组策略之间的差异。

策略使用	深度伪造			Real		
	Acc (M)	ECE	*N*	Acc (M)	ECE	*N*
无（无策略）	0.33	0.89	3	0.50	0.27	6
仅视觉	1.00	0.17	6	0.13	0.48	8
仅音频	1.00	0.13	2	0.67	0.22	9
仅知识	0.67	0.20	9	0.92	0.27	12
视觉 +仅音频	0.44	0.42	9	0.83	0.17	18
视觉 +仅知识	0.95	0.18	39	0.50	0.21	18
音频 + 仅知识	0.60	0.20	10	0.88	0.23	25
第1组（单项：A/V/K）	0.82	0.12	17	0.62	0.14	29
第2组（双项：V+A / V+K / A+K）	0.81	0.06	58	0.75	0.14	61
第3组（V+A+K）	0.71	0.18	312	0.82	0.12	294
总体	0.72	0.14	390	0.79	0.13	390
方差分析（深度伪造视频）：准确率 $F(3, 386) = 1.96,\ p = .120$；期望校准误差 $F(3, 241) = 1.42,\ p = .238$						
方差分析（真实视频）：准确率 $F(3, 386) = 3.41,\ p = .018$；期望校准误差 $F(3, 232) = 2.43,\ p = .066$						

表4：真实视频策略组的Tukey HSD事后检验结果。数值表示均值差。注：还对其他组间比较进行了检验，但所有结果的 *p*-值均大于 0.10。

测量指标	比较	均值差	*p*值	Sig.
准确率	第1组 – 第3组	0.199	0.057	**边缘显著**
ECE	第2组 – 第1组和第3组	0.281	0.038	**Weak**

对于深度伪造视频，$t(194) = 3.83,\ p = .0002$。这些发现表明，虽然人声和语言信号提高了准确率，但也有助于使置信度与真实视频的实际表现更加一致。

知识线索。 外部知识来源（例如事实核查）对深度伪造视频的准确率略有提升（.47 对 .39），但未带来校准方面的益处。对人物或事件的先验知识（.61 对 .62，差异不显著）在不同条件下相对稳定，且差异不显著。基于直觉的判断使用频繁，且准确率相对较高（深度伪造视频为 .70，真实视频为 .59），差异显著，$t(194) = 3.24,\ p = .001$。这表明，本能有时能帮助参与者抵御操纵，但也可能使他们对真实视频的判断产生偏差。

4.4 线索结合时会发生什么？（研究问题3）

图 2 展示了按平均准确率和期望校准误差（ECE）排名的前 10 种策略线索组合，并按深度伪造视频和真实视频分类。根据准确率，从每种视频类型中选取前 10 种组合。结果显示出一致的模式：与深度伪造视频相比，真实视频在大多数线索组合下的准确率更高、期望校准误差更低。例如，结合先验知识、直觉和环境线索，或结合外观、直觉和环境线索的组合，表现出较强的校准能力

在真实视频上，这些组合的准确率均高于 0.8，期望校准误差也显著低于深度伪造视频。虽然某些组合（如（知识 + 直觉 + 外观）或（直觉 + 知识））在真实视频和深度伪造视频上的表现差距较小。

图3和图4展示了深度伪造视频（左）和真实视频（右）的策略线索网络，呈现了这些线索在准确率和期望校准误差中的作用。

对于深度伪造视频，少数局部簇尤为突出。例如，视觉外观与环境、先验知识、人声和直觉之间的联系更为紧密。这表明，参与者在评估深度伪造视频时，有时会结合表层异常、情境知识和直觉判断，并取得更高的准确率。对于真实视频，中心枢纽包含更多线索，例如音频-声音、音频-语言、视觉外观、先验知识和直觉。

深度伪造视频的校准网络（反转期望校准误差）显示，基于知识的线索（如直觉和来源）之间的连线相对较粗。这表明，在某些情况下，这些线索有助于更好地校准置信度与准确率。相比之下，涉及视觉制作质量或音质的连线仍较细。这表明，参与者依赖这些信号时，校准效果较弱。对于真实视频，音频人声、音频语言、视觉外观以及与先验知识相关的线索之间可见较粗的连线。这表明，参与者综合运用这些线索时，校准效果稳健。

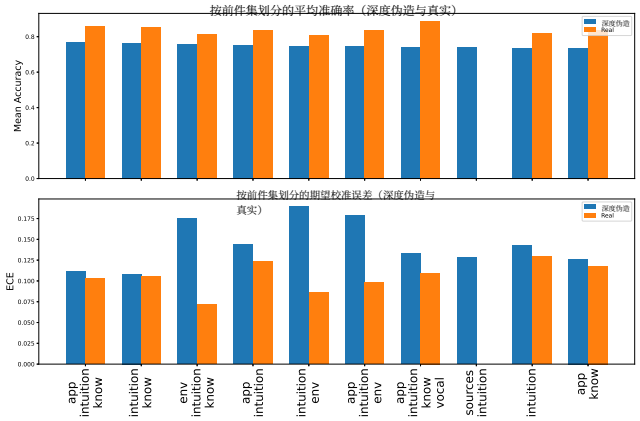


图 2：针对深度伪造视频与真实视频，按策略线索组合排列的准确率前 10 名。该图比较了特定多线索组合促成检测表现的频率。柱形图分别显示了真实视频和深度伪造视频中各线索组合的平均准确率和期望校准误差。

5 讨论

5.1 真实视频与深度伪造视频检测表现的不对称性

我们的研究结果凸显了人类检测能力的潜力与局限。参与者的准确率高于随机水平，这与先前的研究一致；先前研究表明，人类有时能够较好地识别涉及名人的篡改媒体 [12]。

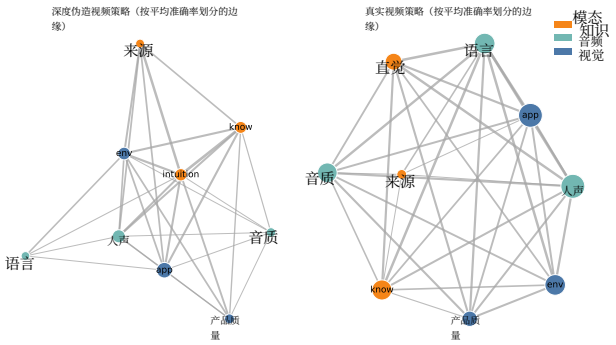


图3：深度伪造视频和真实视频的策略线索准确率网络可视化。节点大小表示线索的使用频率。边的粗细表示 av-

线索组合时的平均准确率。颜色表示模态（蓝色 = 视觉、青绿色 = 音频、橙色 =知识）。节点标签显示线索名称（app-应用程序外观、env-环境、prodqual-制作质量；soundqual-音质；sources-信息来源、know-对人物/视频内容/视频背景的先验知识）。此图突出显示了哪些线索组合有助于成功检测各类视频。

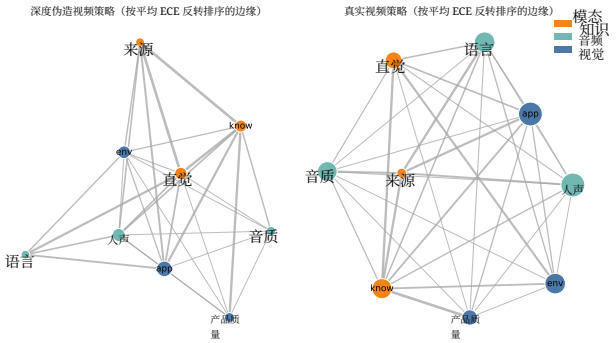


图4：深度伪造视频和真实视频的策略线索校准（反转ECE）网络可视化。节点大小表示线索的使用频率。边的粗细表示置信度与准确率之间的一致性（ECE越低，一致性越好）。颜色表示模态（蓝色 = 视觉、青绿色 =音频、橙色 =知识）。此图展示了哪些线索有助于提高不同视频类型下判断的校准程度。

然而，不同视频类型的检测表现并不均衡。参与者对真实视频的识别比对深度伪造视频更可靠。同时，不同视频类型下的置信度评分保持稳定。以往研究也发现了深度伪造视频检测中的过度自信模式 [17]。在以往研究的基础上，我们的研究采用校准指标（ECE），而不仅仅考察相关性，结果显示，参与者对深度伪造视频的校准较差（ECE较高），而对真实视频的校准较好（ECE较低）。

这种不对称性可能有两个原因。首先，以往研究记录了一种“真实性偏差”：除非能看到明显伪影，否则人们会默认内容具有真实性

表5：深度伪造与真实视频中策略线索的准确率、期望校准误差和频率。数值为试次级均值。配对样本 *t*-检验比较深度伪造与真实条件。粗体表示 $p < .05$

策略线索	深度伪造			Real			*t*-检验	
	Acc (M)	ECE	*N*	Acc (M)	ECE	*N*	准确率 *t*, *p*	期望校准误差 *t*, *p*
视觉外观	0.84	0.14	327	0.78	0.12	304	**2.10, .036**	*1.83, .068*
视觉环境	0.58	0.18	228	0.55	0.09	214	1.01, .312	1.22, .225
视觉制作质量	0.48	0.19	189	0.41	0.12	158	**2.24, .026**	**2.19, .030**
音频人声	0.69	0.18	271	0.81	0.15	314	**-3.58, .0004**	**3.56, .0004**
音频语言	0.47	0.15	184	0.55	0.14	215	**-2.22, .026**	**4.48, <.001**
音频音质	0.49	0.19	191	0.52	0.13	203	-0.86, .391	**3.83, .0002**
知识（来源）	0.47	0.12	185	0.40	0.11	155	**2.17, .030**	1.08, .281
知识（先验知识）	0.61	0.13	237	0.62	0.13	243	-0.44, .659	**2.82, .005**
知识（直觉）	0.70	0.14	274	0.59	0.13	231	**3.24, .001**	**2.84, .005**

。其次，由于我们的刺激材料中包含知名人物，参与者在认出熟悉的面孔时，可能会高估自己的检测能力。以往研究表明，使用不熟悉的目标时，置信度与准确率的相关性更高 [34]。

5.2 多策略整合的益处与限制

通过分别分析各种策略及其组合，我们发现人类检测本质上是多模态的，因为大多数参与者都报告称使用了全部三种策略类型。这与认知心理学的研究发现一致：判断会综合感知线索和基于知识的线索 [38]。

然而，不同策略组合的有效性存在差异。对于深度伪造视频，单一、双重和三重策略组之间未观察到显著的表现差异。相比之下，真实视频判断受益于多模态整合，其中视觉、音频和知识策略的组合提高了准确率并改善了校准。这种差异可能反映了深度伪造视频的新颖性与复杂性。对这种不对称性的一种可能解释是，真实视频往往提供连贯的跨模态证据，其中视觉外观、语调、语言流畅度和语境线索自然一致。当这些线索相互强化时，整合多种线索会增强总体判断信号并提升表现。相比之下，深度伪造视频可能包含局部上令人信服的元素，例如逼真的面部纹理或流畅语音，但仍表现出细微的跨模态不一致，这些不一致过于微弱或不为参与者所熟悉，因而难以可靠解读 [6]。当线索相互冲突或显得模糊时，增加更多模态可能会引入额外噪声，而非带来清晰度。此外，在深度伪造视频与真实视频上使用多种策略的表现差异，也可能反映了深度伪造视频的新颖性与复杂性。参与者可能对先进的操纵技术不够熟悉，从而限制了策略组合的有效性。这些发现凸显了教导用户不仅要注意不同策略，还要理解何时以及如何有效运用这些策略的重要性。

5.3 线索层面的模式

线索层面分析进一步揭示了各模态特有的优势与不足。对于深度伪造视频，视觉外观（例如面部不规则）和直觉线索相对有效，音频与声音线索也有所贡献。这一模式与以下事实一致：深度伪造操纵往往针对眼睛或嘴巴等面部区域，从而可能产生可见的不规则之处。相比之下，对于真实视频，音频线索发挥了更重要的作用，其次是视觉外观和先验知识。这可能反映了当前生成模型不易模仿真实语音模式和自然的语言流畅度这一事实 [13]。

这些见解对媒介素养干预措施的设计具有直接启示。现有项目往往会规定具体线索或提供指南 [7, 36], , 而辅助工具则会突出人工智能检测出的伪影 [1, 3]。我们的结果表明，应超越“找出故障”的方法，转而开展培训，帮助人们认识到自己的置信度何时可能并不可靠。例如，视觉外观有助于发现表层伪影，但当故障不明显或不存在时，也可能产生误导。因此，媒介素养项目不仅可以帮助用户了解如何识别操纵的视觉指标，还可以帮助他们理解，依赖某些视觉线索何时可能导致错误自信。此外，与深度伪造视频相比，音频策略（例如语调、语言连贯性、音质）对于判断视频是否真实更具诊断性。干预措施可以强调这种不对称性，教导用户：音频线索虽能确认真实性，但在发现操纵方面效果较差，还可能导致校准错误。此外，知识策略（例如先验知识、外部来源、直觉）对于检测真实内容和伪造内容都至关重要。然而，仅依赖直觉可能导致过度自信。干预措施可以强调事实核查和交叉核对外部信息，以此更有效地补充感知线索。

5.4 检测挑战

随着生成式人工智能的发展，深度伪造视频在感知上可能会变得与真实视频难以区分。我们的结果已经显示出这一趋势：多模态策略有助于处理真实视频，但在识别深度伪造视频方面几乎没有带来益处。即使存在这些局限，

人类能否区分真实内容与伪造内容仍然十分重要。尽管AI检测器可以支持验证，但它们往往落后于新的生成技术，而且并不总是向公众开放。例如，商业工具需要付费，而开源模型则要求用户具备技术专长 [40]。这凸显了持续培训人们的必要性，使其能够识别自身的置信度何时可能被误置，以及何时需要进行额外验证。

在涉及政治问责、新闻核查或人际信任的情境中，人类判断仍将至关重要。这是因为深度伪造虚假信息可能造成严重危害，例如损害个人和组织的声誉、破坏政治进程、助长金融欺诈，以及促成非自愿色情内容的传播 [38]。在这些情况下，人类判断充当第一道防线，尤其是在无法使用 AI 工具或需要快速决策时。此外，受过训练的个人可能会识别出 AI 模型所忽略的线索。在人机协作方法中，将人工评估与 AI检测器相结合，可能会提升总体检测效果，并增强我们应对深度伪造视频的能力。

然而，随着深度伪造视频变得更加逼真，需要采取更广泛的防护措施，包括来源溯源或水印标准、平台层面对合成媒体的披露，以及惩罚恶意制作和传播深度伪造视频的法规 [21, 31]。例如，英国近期出台的《2023年在线安全法》要求平台承担责任，中国的《深度合成规定》则要求强制标注和算法审查，这些法规表明各国政府正开始应对深度伪造视频带来的风险 [31]。

6 结论

本文通过分析人们在视觉、音频和知识

e-

基础策略方面的准确率、置信度和校准情况，考察了人们如何检测深度伪造视频。我们发现，真实视频受益于线索的多模态整合，而深度伪造视频则会干扰这些策略，使参与者的准确率降低，校准也更不准确。进一步将策略拆解为线索并绘制其共现网络后，我们确定了哪些组合有助于或阻碍检测，并突出显示了支持有效判断的核心线索。总之，这些发现加深了我们对人类深度伪造检测的理解，并为设计媒介素养干预提出了具体方向，使其能够以人们自然运用的策略为基础。

本研究存在一些局限。参与者仅限于年轻人，因此研究结果在不同年龄群体中的普适性有限。刺激材料集中包含的深度伪造视频数量有限，可能无法涵盖现实中遇到的各种操纵方式。未来研究可以纳入更广泛的人群和更多样的视频操纵方式，以拓展本研究。此外，参与者从预先设定的清单中选择线索。虽然这种设计便于进行定量分析，但可能促使参与者报告他们在自然情况下未必会使用的线索。未来研究可以采用不加提示的方法，以捕捉更贴近真实情境的策略。

参考文献

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Ben- netot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins 等。2020。可解释人工智能（XAI）：概念、分类体系、机遇与迈向负责任人工智能的挑战。载于《信息融合》58（2020），82–115。[2] Melisa Basol、Jon Roozenbeek 和 Sander Van der Linden。2020。关于坏消息的好消息：游戏化免疫训练能够增强人们抵御假新闻的信心和认知免疫力。《认知》3，1（2020），2。[3] Aidan Boyd、Patrick Tinsley、Kevin Bowyer 和 Adam Czajka。2023。人工智能指导在人类审查合成生成的人脸中的价值。载于《AAAI人工智能会议论文集》，第37卷。5930–5938。[4] Sergi D Bray、Shane D Johnson 和 Bennett Kleinberg。2023。测试人类检测人脸“深度伪造”图像的能力。《Journal of Cybersecurity》9，1（2023），tyad011。[5] Chen Chen 和 Dion Hoe-Lian Goh。2025。了解用户对用于教授深度伪造视频检测的对话式人工智能的接受度：一种扩展技术接受模型方法。载于《国际人机交互会议》。Springer，165–175。[6] Chen Chen、Dion Hoe-Lian Goh、Hao Ran Qiu 和 Celene Neo。2025。Z世代对抗深度伪造视频：识别策略调查。《信息科学与技术协会论文集》62，1（2025），892–896。[7] Bobby Chesney 和 Danielle Citron。2019。深度伪造：隐私、民主和国家安全面临的迫在眉睫的挑战。《加利福尼亚法律评论》107（2019），1753。[8] Athirasree Das、KS Angel Viji 和 Linda Sebastian。2022。使用深度学习技术检测深度伪造视频的调查。载于《2022年第二届下一代智能系统国际会议（ICNGIS）》。IEEE，1–4。[9] Alexander Diel、Tania Lalgi、Isabel Carolin Schröter、Karl F MacDorman、Martin Teufel 和 Alexander Bäuerle。2024。人类检测深度伪造的表现：对56篇论文的系统综述和元分析。《人类行为报告中的计算机》16（2024），100538。[10] Christopher Doss、Jared Mondschein、Dule Shu、Tal Wolfson、Denise Kopecky、Valerie A Fitton-Kane、Lance Bush 和 Conrad Tucker。2023。深度伪造与科学知识传播。《Scientific Reports》13，1（2023），13429。[11] Dion Hoe-Lian Goh。2024。“他看起来非常真实”：用于识别深度伪造的媒体、知识和基于搜索的策略。《信息科学与技术协会杂志》75，6（2024），643–654。[12] Matthew Groh、Ziv Epstein、Chaz Firestone 和 Rosalind Picard。2022。由人类群体、机器和机器辅助群体检测深度伪造。《美国国家科学院院刊》119，1（2022），e2110013119。[13] Matthew Groh、Aruna Sankaranarayanan、Nikhil Singh、Dong Young Kim、Andrew Lippman 和 Rosalind Picard。2024。人类通过文字稿、音频和视频检测政治言论深度伪造。《Nature Communications》15，1（2024），7629。[14] Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。2017。现代神经网络的校准。载于《国际机器学习会议》。PMLR，1321–1330。[15] Liusheng Huang、Huaping Chen、Xun Wang 和 Guoliang Chen。2000。一种快速挖掘关联规则的算法。《Journal of Computer Science and Technology》15，6（2000），619–624。[16] Achhardeep Kaur、Azadeh Noori Hoshyar、Vidya Saikrishna、Selena Firmin 和 Feng Xia。2024。深度伪造视频检测：挑战与机遇。《Artificial Intelligence Review》57，6（2024），159。[17] Nils C Köbis、Barbora Doležalová 和 Ivan Soraperra。2021。屡次受骗：人们无法检测深度伪造，却自以为能做到。《iScience》24，11（2021）。[18] Vivian Lai 和 Chenhao Tan。2019。关于人类对机器学习模型解释和预测的预测：欺骗检测案例研究。载于《公平性、问责与透明度会议论文集》。29–38。[19] Juniper Lovato、Jonathan St-Onge、Randall Harp、Gabriela Salazar Lopez、Sean P Rogers、Ijaz Ul Haq、Laurent Hébert-Dufresne 和 Jeremiah Onaolapo。2024。多样化的错误信息：人类偏见对网络中深度伪造检测的影响。《npj Complexity》1，1（2024），5。[20] Benjamin A Lyons、Jacob M Montgomery、Andrew M Guess、Brendan Nyhan 和 Jason Reifler。2021。对新闻判断的过度自信与虚假新闻易感性相关。《美国国家科学院院刊》118，23（2021），e2019527118。[21] Siwei Lyu。2024。深度伪造的威胁：减轻人工智能生成内容的负面影响。《Organizational Cybersecurity Journal: Practice, Process and People》，在线优先（2024）。[22] S Maddocks。2020。“一场意在让我噤声的深度伪造色情阴谋”：探讨色情深度伪造与“政治”深度伪造之间的延续性。《Porn Studies》7（4），415–423。[23] Yisroel Mirsky 和 Wenke Lee。2021。深度伪造的制作与检测：一项调查。《ACM Computing Surveys（CSUR）》54，1（2021），1–41。[24] Mahdi Pakdaman Naeini、Gregory Cooper 和 Milos Hauskrecht。2015。使用贝叶斯分箱获得校准良好的概率。载于《AAAI

人工智能会议，第29卷。[25] Sophie J Nightingale 和 Hany Farid。2022。人工智能合成面孔与真实面孔难以区分，且更值得信任。《美国国家科学院院刊》119，8（2022），e2120481119。[26] Irmak Olcaysoy Okten、Tianshu Huang 和 Gabriele Oettingen。2023。更新错误信念：错误确信与恰当确信的作用。《心理学通报与评论》30，2（2023），712–721。[27] Trevine Oorloff、Surya Koppisetti、Nicolò Bonettini、Divyaraj Solanki、Ben Col- man、Yaser Yacoob、Ali Shahriyari 和 Gaurav Bharaj。2024。AVFF：用于视频深度伪造检测的视听特征融合。载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》。27102–27112。[28] Ana Pérez-Escoda、Luis Miguel Pedrero-Esteban、Juana Rubio-Romero 和 Carlos Jiménez-Narros。2021。社交网络上的假新闻触达年轻人：不信任给媒介素养带来的挑战。《出版物》9，2（2021），24。[29] David N Rapp 和 Mandy M Withall。2024。置信度作为元认知因素，既是错误信息经历的成因，也是其结果。《心理学当前观点》55（2024），101735。[30] Muhammad Anas Raza 和 Khalid Mahmood Malik。2023。Multimodaltrace：使用视听表征学习进行深度伪造检测。载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》。993–1000。[31] Felipe Romero-Moreno。2025。生成式人工智能中的深度伪造检测：保护人权的法律框架提案。《计算机法律与安全评论》58（2025），106162。[32] Nabil F Saleh、JON Roozenbeek、Fadi A Makki、William P McClanahan 和 Sander Van Der Linden。2024。主动免疫训练能增强抵御极端主义劝服技巧的态度抵抗力：一种预防暴力极端主义的新方法。《行为公共政策》8，3（2024），548–571。[33] Nikita A Salovich 和 David N Rapp。2021。受到错误信息影响却浑然不觉？元认知与不准确信息的影响。《实验心理学杂志：学习、记忆与认知》47，4（2021），608。[34] Klaire So moray 和 Dan J Miller。2023。提供检测策略以提升人类深度伪造检测能力：一项实验研究。《人类行为中的计算机》149（2023），107917。[35] Klaire Somoray、Dan J Miller 和 Mary Holmes。2025。深度伪造检测中的人类表现：系统综述。《人类行为与新兴技术》2025，1（2025），1833228。[36] Rashid Tahir、Brishna Batool、Hira Jamshed、Mahnoor Jameel、Mubashir Anwar、Faizan Ahmed、Muhammad Adeel Zaffar 和 Muhammad Fareed Zaffar。2021。眼见为实：探索深度伪造视频中的感知差异。载于《*2021*年*CHI*计算机系统中的人为因素会议论文集》。1–16。[37] Zhong Tang、Dion Hoe-Lian Goh、Chei Sian Lee 和 Yihao Yang。2024。了解老年人识别深度伪造的策略。载于《国际人机交互会议》。Springer，236–244。[38] Cristian Vaccari 和 Andrew Chadwick。2020。深度伪造与虚假信息：探索合成政治视频对欺骗、不确定性和对新闻的信任的影响。社交媒体+与社会6，1（2020），2056305120903408。[39] Kevin Warren、Tyler Tucker、Anna Crowder、Daniel Olszewski、Allison Lu、Caroline Fedele、Magdalena Pasternak、Seth Layton、Kevin Butler、Carrie Gates 等。2024。“最好是电脑，不然我就是笨蛋”：对人类作为音频深度伪造检测器的大规模评估。载于《*2024*年*ACM SIGSAC*计算机与通信安全会议论文集》。2696–2710。[40] Y Kelly Wu、Saniat Javid Sohrawardi、Candice R Gerstner 和 Matthew Wright。2025。了解并赋能情报分析员：用于深度伪造检测工具的以用户为中心的设计。载于《*2025*年*CHI*计算机系统中的人为因素会议论文集》。1–26。

附录 A。策略类型与线索

策略 Type	Cue	描述
视觉	外观	面部特征（例如肤色不均、面部对称性、发型、眼睛渲染不佳）排序良好）身体和行为特征（例如口型同步问题、异常异常的身体动作、换脸、僵硬——（表情僵硬／毫无表情）
	环境	视频中物体的特征（例如，形状怪异的伪影、不自然的反光——、重叠物体）背景和环境细节（例如，场景设置、异常背景、水印或标志问题或字幕）色彩与光照不一致（例如，变色、光照不一致，对比度问题）
	制作	视频质量问题（例如，视频分辨率、失真、故障、像素化）制作问题（例如，剪辑问题、摄像机角度／拍摄问题、抖动和抖动)
音频	人声	人声特征（例如，音高和音调变化、声音自然度、缺乏情-绪)
	语言	可懂度和语言（例如，语言使用不匹配、流利度、发音、语言使用、流利度、发音、丧失可懂度）音质丧失）
	音频质量ity	音质问题（例如，清晰度问题）题，音量不均、失真、回声，故障）背景声音问题（例如，背景——混响/回声，整体噪声、机械噪声）
知识	信息来源	在线工具（例如使用搜索引擎如 Google 或社交媒体平台如 Facebook 和 YouTube）使用多个来源
	先验知识	对当事人的了解（例如，先验对视频中人物的了解）对视频内容的了解（例如，熟悉——对视频中的事件的熟悉程度）对视频内容的背景的了解视频内容
	直觉	直觉与情绪（例如，基于你自己的本能、观点或情绪）

表6：策略类型及其对应线索

附录B。视频信息

视频 Type	主题	描述	时长
深度伪造	娱乐	马克·扎克伯格说他控制着数十亿人的机密数据，因此拥有他们的未来。	00:18
		金·卡戴珊讲述她喜欢通过操纵粉丝来赚钱。	00:22
		汤姆·克鲁斯的日常生活。	00:41
		一段来自《闪灵》的片段。	00:51
	政治	马诺杰·蒂瓦里批评了一则……摆姿势的政党并鼓-励人们投票给他的政党。	00:48
		杰里米·科尔宾支持鲍里斯·约翰逊担任首相。	00:33
		一段宣布阿波罗11号任务失败的语音。	01:03
	教育程度	奥巴马提醒人们要对假新闻更加警觉。	00:27
		摩根·弗里曼问道：“眼见就一定为实吗？”	00:28
	体育	何塞·穆里尼奥评论足球。	01:27
Real	娱乐	马克·扎克伯格表示他能够判断人们的在线行为行为。	00:29
		金·卡戴珊主张她在考试中作弊。	00:15
		汤姆·赫兰德休息一下来自社交媒体。	00:51
		《闪灵》的另一段片段。	00:50
	政治	拜登批评 MAGA 共和党人——共和党人。	00:51
		特朗普敦促他的支持者在国会遭到袭击后冷静下来sional攻击。	00:51
		乌胡鲁总统哀悼——肯尼亚前总统姆瓦伊基巴基。	00:42
	教育程度	希拉里·克林顿讨论了假新闻的危害。	00:56
		艾伦警告人们提防伪造news.	00:29
	体育	何塞·穆里尼奥谈论爵士亚历克斯·弗格森在波尔图赢得欧洲冠军联赛战胜曼彻斯特联之后的反应。	01:17

表7：研究中使用的视频



图5：杰里米·科尔宾的深度伪造视频



图6：马克·扎克伯格的深度伪造视频

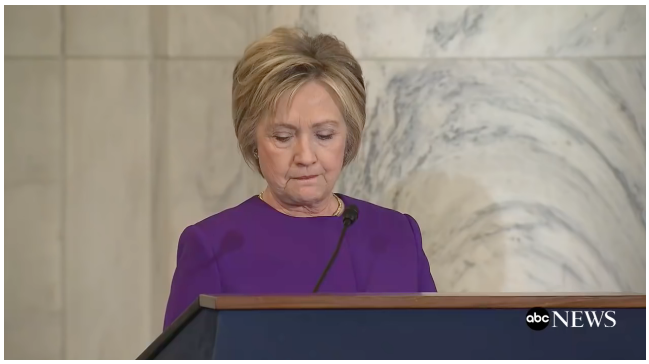


图7：希拉里·克林顿的真实视频



图8：特朗普的真实视频

附录C. 问卷条目

对4个视频中的每一个重复此操作

(1) **视频是真实的还是伪造的?**
- 真实
- 伪造

(2) **置信度评分**
你对上述选择有多大把握？请用 1 到 5 的等级表示你的置信度。

- 1
- 2
- 3
- 4
- 5

(3) **视觉特征**
根据你在视频中看到的内容，哪些视觉线索帮助你判断视频是真实还是伪造的?
• 面部特征（例如，肤色不均、面部对称性、发型、眼睛渲染不佳）• 身体与行为特征（例如，口型同步问题、异常身体姿势、换脸、僵硬或面无表情）• 视频中物体的特征（例如，形状奇怪的伪影、不自然的反射、重叠物体）• 背景与环境细节（例如，场景设置、不寻常的背景、水印、标志或字幕方面的问题）• 视频质量问题（例如，分辨率、失真、故障、像素化）• 制作问题（例如，剪辑问题、拍摄角度问题、画面抖动或抖动）
• 色彩与光照不一致（例如，变色、光照不一致、对比度问题）

• 其他属性（请具体说明）

（4）**音频特征**
根据您在视频中听到的内容，哪些音频线索帮助您判断该视频是真实的还是伪造的?
• 人声特征（例如音高或语调变化、声音自然度、缺乏情感）• 可懂度和语言（例如语言不匹配、流利度、发音、可懂度丧失）• 音质问题（例如清晰度问题、音量不均、失真、回声、故障）• 背景声音问题（例如混响、机械噪声、总体背景噪声）

• 其他属性（请具体说明）

（5）**基于知识的特征**
你使用自己的知识或外部信息时，哪些线索帮助你判断视频是真实还是伪造的?
• 在线工具（例如通过 Google、Facebook、YouTube 搜索）• 与他人沟通（例如在线或离线向家人或朋友求证）

• 使用多个来源 • 对当事人的了解（例如，对视频中个人的熟悉程度）• 对视频内容的了解（例如，对所呈现事件的熟悉程度）

• 对周边背景的了解（例如，更广泛的背景、意义一致性、政治立场）• 直觉与情绪（例如，本能、个人观点、情绪印象）

• 其他属性（请具体说明）

附录D. 研究界面

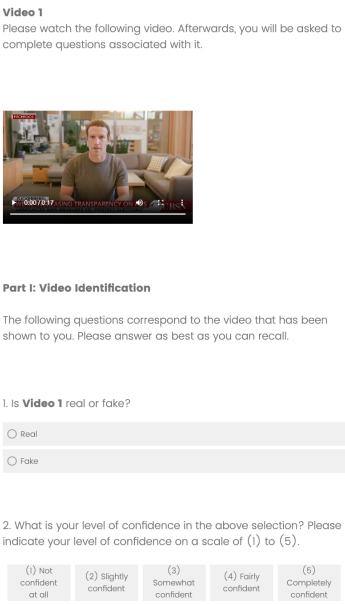


图9：研究界面