FakeHunter：用于可解释视频取证的多模态逐步推理

Chen Chen[1*], Runze Li[2*], Zejun Zhang[3], Pukun Zhao[1], Fanqing Zhou[1], Longxiang Wang[4], Haojian Huang[5]

[1]Guangdong University of Finance and Economics [2]Westlake University [3]University of Southern California
[4]Chongqing University [5]The University of Hong Kong

Allen821@student.gdufe.edu.cn, lirunze@westlake.edu.cn, zejunzha@usc.edu, zhaopukun@student.gdufe.edu.cn, zhoyfanqing@gmail.com, longxiangwang@stu.cqu.edu.cn, haojianhuang@connect.hku.hk

Abstract

假猎手是一个融合了记忆引导检索、链式思维（观察－思考－行动）推理和工具增强验证的多模态深度伪造检测框架，能够提供准确且可解释的视频取证。它使用 CLIP 对视觉内容进行编码，并用 CLAP 对音频进行编码，生成联合的音视嵌入表示，从 FAISS 索引的记忆库中检索语义相似的真实示例以进行上下文定位。在检索到的上下文指导下，系统会迭代推理证据，定位操纵并解释它们。当信心不足时，它会自动调用专门工具——放大图像取证或梅尔频谱图检查——进行细粒度验证。基于 Qwen2.5- Omni-7B 构建，FakeHunter 生成结构化的 JSON 判决，指出什么被修改、其中发生的位置以及为什么它被判断为伪造。我们进一步介绍了 X-**虚假视频**，这是一个包含超过 5700 个操纵和真实视频（950+ 分钟）的数据集，标注了操纵类型、区域/实体、违反推理类别及自由形式的说明。在 X-AVFake 上，FakeHunter 达到了 34.75%的准确率——超过了 vanilla Qwen2.5-Omni-7B 16.87 个百分点和 MiniCPM-2.6 25.56 个百分点。消融实验显示，内存检索带来了 7.75 个百分点的增益，基于工具的检查将低置信度案例提升到了 46.50%。尽管其设计了多阶段流程，该流水线在单个 NVIDIA A800 上处理一段 10 分钟的片段需要 8 分钟（0.8× 实时），或在四个 GPU 上需要 2 分钟（0.2×），展示了其实用部署能力。

1 引言

近期生成模型和大型语言模型的进展大幅降低了创建逼真假音频、图像和视频的门槛 (Tolosana et al. 2020) 。这种广泛可访问性引发了越来越多的关注：深度伪造技术被用于传播政治错误信息 (Chesney and Citron 2019; Vaccari and Chadwick 2020)，进行声音克隆诈骗和冒充攻击 (Korshunov and Marcel 2018)，以及破坏对视频通话的信任 (Mirsky and Lee 2021; Agarwal and Farid 2020)。除了简单的面部操作，现代技术还实现了音频欺骗 (Tak et al. 2021) 、对象移除 (Mittal et al. 2023a) 以及全场景重合成 (Bar-Tal et al. 2024) 。这些趋势凸显了需要深度伪造检测系统不仅准确，而且多模态和可解释，能够分析视觉和听觉信号来识别什么被篡改了，篡改发生在何处，以及为什么被认为是假的。

尽管在深度伪造检测方面取得了进展，大多数现有方法仍存在范围和能力的限制。一些模型是单模态——仅基于视觉 (Rossler et al. 2019; Li et al. 2020c; Frank et al. 2020) 或音频 (Wang et al. 2023; Jung et al. 2022; Di Pierno et al. 2025) 输入运行——并且无法捕捉跨模式的一致性问题。其他模型结合了音频和视频流 (Yang et al. 2023; Nie et al. 2024; Oorloff et al. 2024)，但仍局限于二元分类，不能提供有关哪些内容被篡改或为何该内容被视为伪造的任何见解。

为了提高可解释性，最近的工作引入了具备解释功能的深度伪造检测器。TAENet (Du et al. 2024) 生成像素级热图以定位操纵区域，但缺乏文本推理能力。FakeShield (Xu et al. 2024) 识别片段级别的音视频不匹配，并基于视觉语义不一致产生人类可理解的解释。DD-VQA (Zhang et al. 2024) 利用通过视觉问答提示实现的链式思考推理，而 TruthLens (Kundu, Balachandran, and Roy-Chowdhury 2025) 则利用视觉和语言模型为面部图像检测生成文本理由。然而，所有这些方法都针对静态图像或单模态输入操作，这限制了它们对多模态或多动态操纵进行推理的能力。因此，这凸显出需要一个既能检测又能解释的统一多模态框架。

在本文中，我们介绍了**假猎手**，一个可解释的多模态深度伪造检测框架，它集成了记忆引导检索、链式思维（CoT）推理和工具增强分析。FakeHunter 通过预训练的 CLIP 和 CLAP 对视频和音频进行编码，从 FAISS 索引的记忆库中检索相似的真实样本，并使用它

*These authors contributed equally to this work.

们来为 Qwen2.5-VL 的逐步推理提供依据。该系统遵循观察 – 思考 – 行动序列来识别篡改行为，解释其发生的位置和原因，并在不确定性高时使用放大或频谱图工具来完善决策。

为了支持评估，我们构建了 X-**伪** AV，这是一个包含超过 5700 个视频和 950 多分钟内容的大规模多模态深度伪造数据集。每个样本涉及视觉对象移除或音频替换，并标注有实体/时间戳、操作类型、违反推理类别以及文本解释。该基准测试使在复杂音视场景中对准确性和可解释性的严格评估成为可能。**假猎手**达到了 34.75%的准确性，超过了最先进的基线。尽管其设计有多阶段，但整个管道在 4×A800 GPU 上以 0.3 倍实时运行，处理一段 10 分钟的视频大约需要 33 分钟。

我们的论文贡献总结如下：

- **可解释和多模态模型**。我们提出了**假猎手**，一个通过记忆引导的逐步推理来检测深度伪造并生成自然语言解释的统一模型。
- **多阶段推理管道**。该模型执行迭代的观察 – 思考 – 行动步骤，并通过解释验证以获得稳健且可解释的结果。
- **细粒度多模态数据集**。我们构建了 X-**虚假视频**，一个新的数据集，包含超过 5.7K 个视频（超过 950 分钟），该数据集的特点是针对视觉对象移除和音频替换进行了精细标注。

2 相关工作

2.1 深度伪造检测方法

近期在深度伪造检测方面的进展涵盖了音频、视觉和多模态领域。在音频方面，早期的方法依赖于手工特征和浅层分类器 (Qian, Chen, and Yu 2016)，而最近的方法使用原始波形 CNNs (Di Pierno et al. 2025; Tak et al. 2021; Wang et al. 2023)、频谱-时间图注意力 (Jung et al. 2022) 和非对比预训练的半监督 GNNs (Febrinanto et al. 2025)。在视觉上，检测器针对眨眼等人工痕迹 (Li, Chang, and Lyu 2018)、面部变形 (Rossler et al. 2019) 和频率异常 (Frank et al. 2020)。尽管对已知操作有效，单模态模型通常会在压缩或分布外攻击下表现不佳 (Li et al. 2020c)。为了提高鲁棒性，研究人员探索数据增强 (Agarwal and Farid 2020)、异常检测以及局部边界建模（例如，面部 X 光 (Li et al. 2020a)）。为了解决跨模态不一致的问题，近期的研究融合了音频 – 视频信号：AVFF (Oorloff et al. 2024) 学习联合特征，FRADE (Nie et al. 2024) 将音频线索注入视觉转换器，而 AVoiD-DF (Yang et al. 2023) 通过双流编解码器检测唇音错位。泛化进一步通过面部动作单元 (Bai et al. 2023) 、自监督的视听对齐 (Feng, Chen, and Owens 2023) 以及时间风格一致性 (Choi et al. 2024) 得到提升。

尽管取得了这些进展，大多数深度伪造检测器仍然局限于二分类——识别内容是否被篡改——而无法解释为什么它是假的或操纵发生在何处。为了提高可解释性，近期的研究利用大型语言模型（LLMs）和视觉-语言系统。FakeShield (Xu et al. 2024) 和 TruthLens (Kundu, Balachandran, and Roy-Chowdhury 2025) 生成基于视觉特征的自然语言理由。具有解释性的 VQA 框架如 DD-VQA (Zhang et al. 2024) 引入链式思维推理来证明决策。诸如 TAENet (Du et al. 2024) 等方法通过双解码器架构可视化操纵区域，而 Aghasanli 等人 (Aghasanli, Kangin, and Angelov 2023) 使用基于原型的可解释性检索相似的伪造作为支持证据。尽管前景看好，大多数具有意识的方法仍然局限于静态图像或单个模态，这促使我们开发一个多模态、推理驱动的检测和解释框架。

2.2 现有数据集

FaceForensics++ (Rossler et al. 2019) 是最早的大规模面部操纵基准之一，涵盖多种伪造技术。Celeb-DF (Li et al. 2020c) 通过减少视觉伪影来增强视频的真实感，而 DFDC (Dolhansky et al. 2019) 则扩大了主题多样性以进行现实世界评估。VideoSham (Mittal et al. 2023a) 将范围扩展到专业编辑的视频，不仅限于面部伪造。DeeperForensics-1.0 (Jiang et al. 2020) 提供了 1 万个高保真演员深度伪造视频，并加入扰动以模拟现实世界噪声，而 WildDeepfake (Zi et al. 2020) 则在不受限制的条件下收集野外视频。对于多模态评估，FakeAVCeleb (Khalid, Tariq, and Woo 2021) 将音频和视觉操纵融合以实现跨模态一致性分析。最近，ExDDV (Hondru et al. 2025) 引入了区域级标注和文本推理来解释视频深度伪造检测。尽管这些数据集显著推动了该领域的发展，但每个数据集在多样性、真实性和模式方面都存在局限性。我们的基准旨在通过提供更多样、更真实且多模态的深度伪造数据作为补充资源来填补这些空白。

3 数据集

为了在音频和视觉模态之间实现可解释的深度伪造检测，我们引入了 X-**伪** AV，这是一个包含双模态操作并配以基于自然语言推理的新基准。该数据集包括两种主要类型的篡改：(1) 可视对象移除和 (2) 音频内容替换。每个被操纵的样本都是通过任务级别的指令生成，并用细粒度标签标注了操作区域、类型和理由。总

共，X-**仿冒视音频**包含超过 5,700 个视频会话，涵盖了超过 950 分钟的内容。

标签	分类
A	Physical Laws
B	Time/Season
C	Location/Culture
D	Role/Profession
E	Causality/Order
F	Narrative Context

表 1: 7 类推理违规

3.1 数据集生成

为了确保真实性和可解释性，我们遵循一个两阶段的生成管道：（1）识别有意义的操作目标并提供解释；（2）使用特定模态的编辑工具执行篡改。

3.1.1 操作目标识别。

我们使用 Qwen2.5-VL (Bai et al. 2025) 来分析输入的视频，并自动确定（1）要操作的实体是什么，以及（2）这种操作为何会引入语义或逻辑上的不一致。为了引导 Qwen 的推理过程，我们首先定义了一个包含七种推理违规类型的分类体系（标记为 A – F），如图 1所示。然后，我们构建了一个结构化的系统提示（展示在表 1 中），要求模型分析视频/音频并以标准化的 JSON 格式输出操作计划。此输出包括操作类型（视觉删除或音频替换）、目标实体、违反类别、自然语言解释以及时间锚点（即帧或时间戳）。此结构化的 JSON 作为基于事实的视听操作蓝图。

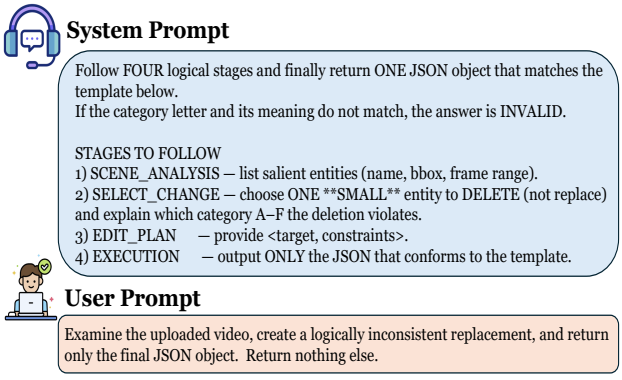

图 1: 提示 VLM 进行推理。我们提供了六个关于违反逻辑的选择，当 VLM 正确理解原始视频后，它将遵循指南选择合适的对象并给出违反逻辑的原因。

3.1.2 操作执行。

给定结构化的 JSON 计划，我们应用专门的编辑工具来操作视频或音频内容，具体取决于指定的模态和操作类型。

- **视频操控**：我们使用基于地面的 SAM 2 (Ren et al. 2024) 跟踪目标对象并生成掩码，然后应用 ProPainter (Zhou et al. 2023) 移除具有时空连续性的对象。
- **音频操纵**：我们使用视听模型 (Xing et al. 2024) 根据视觉线索替换音频片段，从而产生不匹配的说话人声音或不同步的声音。

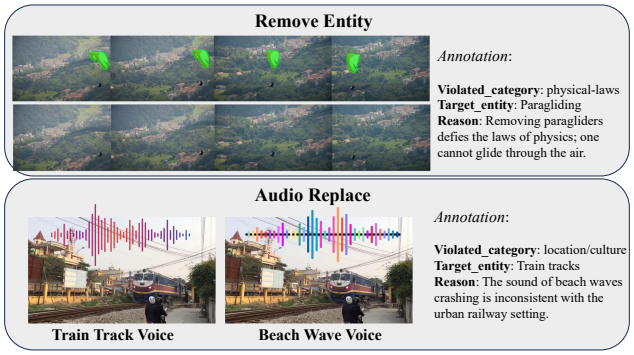

图 2: **假猎手**数据集。对于视觉内容，LLM 将首先选择合理的对象并给出删除违反规则的对象的原因；根据指导方针，我们将调用 SAM2 进行跟踪，并使用工具进行修复。对于音频内容，LLM 分析整个视频，并提供有关如何执行语义连贯的音频替换的高级指导。基于此指导，我们采用音频工具包生成与预期篡改策略相一致的操作音频片段。

3.2 数据集概述

如图 2所示，X-**虚假视频**中的每个样本包含一对原始和操纵过的视频或音频，并附有结构化的元数据以支持可解释的检测和推理监督。具体来说，我们存储：

- 视频/音频 ID：原始和篡改片段的文件 ID。
- 推理轨迹：一个详细描述操作类型、目标实体、违反类别、解释和时间锚点的 JSON 对象。
- 操作标注：像素级掩码（视觉）或时间戳（音频）。
- 解释标签：七个违规类别（A – F）之一及自然语言理由。

这种丰富的标注使在 X-**伪** AV 上训练的模型能够超越二分类，转而回答其中、什么和为什么内容是否被操纵。

4 方法

我们提出了**假猎人**，一个可解释的多模态深度伪造检测基准。如图 3所示，我们的框架采用了链式思考（CoT）推理策略，将决策过程分解为一个观察 – 思考 – 行动序列。通过结合记忆引导检索和工具增强的细粒度分析，**假猎手**逐步提高了检测鲁棒性和解释的质量。

4.1 CoT 推理流水线

为了支持可解释的深度伪造检测，我们采用了一种链式思维（CoT）推理策略，将决策过程分解为一个观察 – 思考 – 行动序列。这种结构化的方法使模型能够通过整合感知、比较和理由来分阶段对多模态内容进行推理。至关重要的是，这不仅允许系统检测深度伪造，还能够解释什么、其中和为什么操纵已经发生——从而支持透明和可解释的视频取证。

观察。 令 $(\mathcal{V}, \mathcal{A})$ 表示输入的视频和音频流，$\mathcal{M}$ 表示内存库。我们首先执行基于记忆的检索，从内存库 $\mathcal{M}$ 中使用 FAISS 检索出最语义相似的前 k 个样本 $\{\boldsymbol{x}^{(i)}\}_{i=1}^{k}$：

$$\mathcal{N}_k = \text{FAISS\_TopK}(\boldsymbol{x}, \mathcal{M}) \tag{1}$$

然后，我们利用检索集 $\mathcal{N}_k$ 作为基础上下文，并使用结构化描述提示来引导视觉语言大语言模型生成输入内容的初步高层次解释：

$$\text{Des}_0 = \text{VLM}(\mathcal{V}, \mathcal{A}, \mathcal{N}_k, \text{Prompt}_{\text{describe}}) \tag{2}$$

想法。 给定初始描述 Des_0 和检索到的记忆上下文 $\mathcal{N}_k$，我们提示 Qwen2.5-Omni-7B 形成初步判断：

$$\text{Verdict} = \text{VLM}(\text{Des}_0,\ \mathcal{N}_k,\ \text{Prompt}_{\text{classify}}) \tag{3}$$

模型输出包含四个字段：

$$\text{Verdict} = \{\text{label},\ \text{type},\ \text{reason},\ \text{confidence}\},$$

其中标签表示二元分类结果（真实或假的）；类型指定操作模式（例如，音频替换，可视化删除）；原因提供基于多模态上下文的自然语言解释；置信度表示模型自我评估的预测确定性，范围缩放至 $[0, 1]$。

如果置信度分数低于预定义的阈值 τ，我们将调用工具增强验证模块来进行额外的细粒度分析。

行动。 在最后阶段，模型整合所有可用信息，包括描述性总结 Des_0、检索到的示例 $\mathcal{N}_k$ 和可选工具分析 $\mathcal{E}_{\text{tools}}$，形成最终决定。我们重新提示 Qwen2.5-Omni-7B 生成最终裁定：

$$\text{FinalVerdict} = \text{VLM}(\text{Des}_0, \mathcal{N}_k, \mathcal{E}_{\text{tools}}, \text{Prompt}_{\text{final}}) \tag{4}$$

最终决策包括以下字段：

$$\text{FinalVerdict} = \{\,\text{label},\ \text{type},\ \text{region/timestamp},\ \text{explanation}\,\}$$

其中，标签指定真实性，类型表示操作类型，区域/时间戳表示操作发生的位置，而解释是对内容被认为是假的原因的简洁且易于理解的说明。

4.2 内存引导检索

4.2.1 特征编码

我们从每个输入视频中提取特定模式的特征，以形成一个联合的音视表示：

- **视觉特征**。我们均匀采样 T 关键帧 $\{I_t\}_{t=1}^{T}$ 并使用预训练的 CLIP 图像编码器对每一帧进行编码 (Radford et al. 2021)，得到图像特征 $f_t^{\text{img}} \in \mathbb{R}^{d_1}$。
- **音频特征**。对应的音频轨道被分割成 T 个块，$\{A_t\}_{t=1}^{T}$ 每个关键帧对齐。我们应用 CLAP 音频编码器 (Elizalde et al. 2023) 来获取 $f_t^{\text{aud}} \in \mathbb{R}^{d_2}$。

我们将两种模态连接起来以获得融合的多模态特征：

$$f_t = [f_t^{\text{img}} \,\|\, f_t^{\text{aud}}] \in \mathbb{R}^d.$$

我们将每段的嵌入向量取平均以获得视频级别的表示：

$$\boldsymbol{f}_v = \frac{1}{T}\sum_{t=1}^{T} f_t \in \mathbb{R}^d.$$

4.2.2 内存银行构建与检索

为了更可靠地检测细微的操作，我们构建了一个真实视频的记忆库 $\mathcal{M}$，并在推理时检索语义相似的样本。这些真实的参考作为基础上下文，使模型能够将输入与典型的音视频模式进行对比，并更有效地识别异常。

我们在训练集嵌入 $\{\boldsymbol{f}_v\}$ 上执行 K-均值聚类，以选择 $K = 300$ 个簇中心作为代表性记忆锚点。这些原型使用 FAISS (Johnson, Douze, and Jégou 2019) 进行索引，以支持推理时基于相似性的高效检索。

在测试时，给定一个输入视频表示 $\boldsymbol{f}_q$，我们从内存库中检索其最接近的 k 个邻居：

$$\mathcal{N}_k(\boldsymbol{f}_q) = \underset{\boldsymbol{f}\in\mathcal{M}}{\text{Top-}k}\left(\frac{\boldsymbol{f}_q^{\top}\boldsymbol{f}}{\|\boldsymbol{f}_q\| \cdot \|\boldsymbol{f}\|}\right).$$

检索到的集合 $\mathcal{N}_k$ 被整合到推理过程中作为语境支持，使能够进行比较推理以检测不一致性和生成解释。

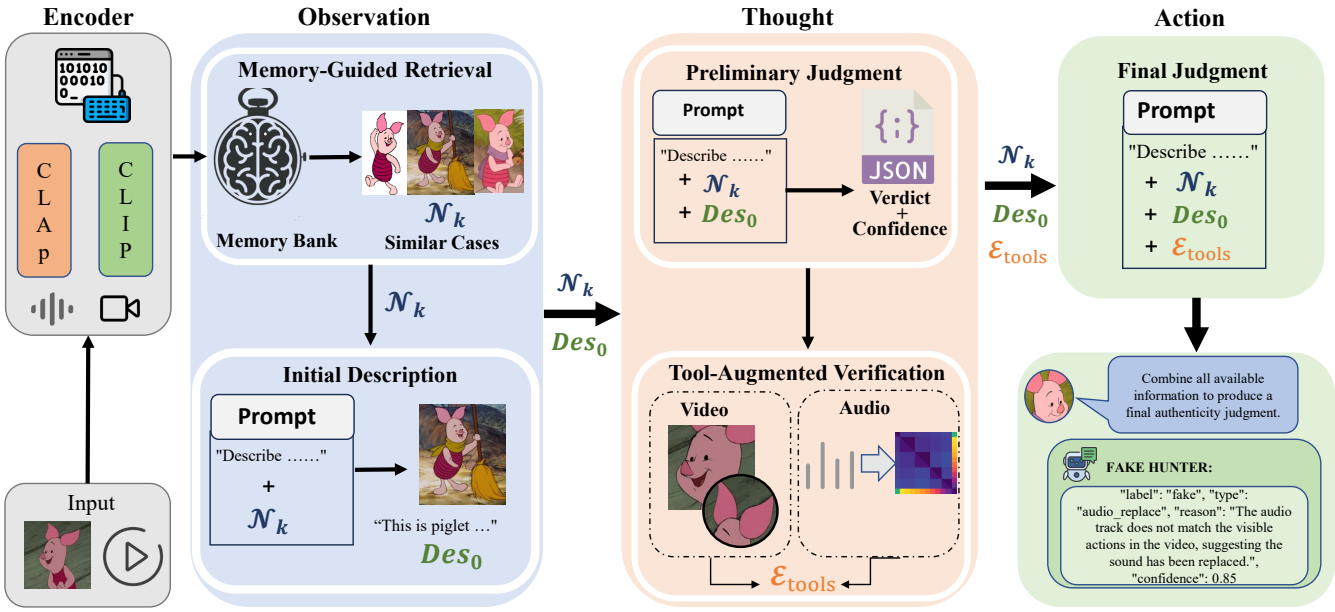

图 3: **假猎手**管道。该框架遵循观察 – 思考 – 行动过程。在观察阶段，对视听特征进行编码，并用于从记忆库中检索语义相似的真实示例，指导初始内容描述的生成。在思考阶段，模型根据描述和检索到的上下文做出初步判断；如果置信度低，则触发视觉和音频分析工具进行详细检查。在行动阶段，整合所有可用证据以产生最终的真实性判定，并附带可解释的理由，包括操纵类型、位置和推理。

4.3 工具增强验证

为了在低置信度场景中提高鲁棒性，我们引入了基于工具的验证模块来进行针对性、局部化的检查。这些工具仅在模型置信度低于预定义阈值 $\gamma \in [0,1]$ 低于预定义阈值 $\tau = 0.80$ 时被激活：

$$\text{Trigger}_{\text{tools}} = \mathbb{1}[\gamma < \tau] \tag{5}$$

4.3.1 视觉放大分析。

设 $\mathcal{I}_0$ 为来自视频 $\mathcal{V}$ 的参考帧，令 $\mathcal{R} \subset \mathcal{I}_0$ 表示一个疑似包含视觉伪影或篡改证据的兴趣区域（ROI）。该区域随后被传递给一个视觉-语言模型（VLM）进行局部多模态推理：

$$\text{VisualReport} = \text{Analyze\_image}(\mathcal{R}) \tag{6}$$

该模型返回一份描述性报告，突出显示诸如边缘缝合、空间失真或不一致的光照等局部视觉异常。

4.3.2 音频频谱图分析（如适用）。

如果存在音频，我们将嫌疑区域与音频段 $\mathcal{A}_t$ 进行时间对齐，并计算其梅尔频谱图：

$$\mathcal{S}_t = \text{MelSpec}(\mathcal{A}_t) \in \mathbb{R}^{F \times T} \tag{7}$$

其中 F 和 T 分别表示频率_bins 和时间步长。得到的频谱图 $\mathcal{S}_t$，编码时间域和频率域特征，然后由一个 VLM 进行分析：

$$\text{AudioReport} = \text{Analyze\_image}(\mathcal{S}_t) \tag{8}$$

这些模块的输出共同形成一个包含多模态理由和定位线索的结构化证据元组：

$$\mathcal{E}_{\text{tools}} = \{\text{VisualReport},\ \text{AudioReport}\} \tag{9}$$

组合证据被整合到 CoT 推理的最后阶段，以精炼真实性判断并生成详细、可解释的说明。

5 实验

5.1 实验设置

实现细节。所有实验均在配备 4 个 ×NVIDIA A800 GPU（80GB）的 Linux 服务器上使用 PyTorch 进行。以下是关键配置总结：

- **大语言模型推理：**我们使用带有 bfloat16 精度和启用 Flash Attention 2 的 Qwen2.5-Omni-7B 进行高效且优化内存的推理。该模型支持最大上下文长度为 32,768 个标记，能够在视频和音频输入之间进行长范围连贯的多模态推理。
- **视频预处理：**每个输入视频以每秒 1 帧的频率采样，并限制为 128 帧或 30 秒的音频。我们使用介质视觉分辨率设置来权衡效率和精度。

数据集	年	模态性	应用	操纵	# 攻击	# 实数	# 假的	解释
MTVFD	2016	V	Video Manipulation	User Generated	1	30	30	x
UADFV	2018	V	Face	Deep Learning	3	49	49	x
FaceForensics++	2019	V	Face	Deep Learning	4	1000	4000	x
CelebDF	2020	V	Face	Deep Learning	3	5907	5639	x
WildDeepFake	2021	V	Face	Deep Learning	4	3805	3509	x
Psynd	2022	A	Speech	Deep Learning	1	30	2,371	x
VideoSham	2022	A+V	Video Manipulation	User Generated	6	380	380	x
FakeBench	2024	I	Image Manipulation	Deep Learning	6	3000	3000	✓
VANE-Bench	2024	V	Video Manipulation	Deep Learning	5	1000	2000	✓
X-**虚假视频** (Ours)	2025	A+V	Video Manipulation	Deep Learning	2	5700	5700	✓

表 2: 深度伪造和操纵检测的数据集比较。

- **特征嵌入：**我们从 CLIP（视觉）和 CLAP（音频）中提取 512 维的嵌入，并将它们拼接成一个用于下游推理的统一的 1024 维多模态特征向量。
- **记忆检索：**我们从包含最多 10,000 个参考样本的 FAISS 索引内存库中检索最接近的前 $k = 5$ 个邻居。应用了 0.7 的相似度阈值进行过滤。
- **推理配置：**每个输入最多经过 3 轮推理处理。如果 Qwen2.5-Omni-7B 的预测信心低于 0.8，则触发工具增强验证。

基线。我们评估了 FakeHunter 和 X-AVFake 相对于强基线的表现。对于数据集比较，我们用 X-AVFake 与 9 个广泛使用的数据集进行基准测试：MTVFD (Al-Sanjary, Ahmed, and Sulong 2016)、UADFV (Yang, Li, and Lyu 2019)、FaceForensics++ (Rossler et al. 2019)、CelebDF (Li et al. 2020b)、WildDeepFake (Zi et al. 2020)、Psynd (Zhang and Sim 2022)、VideoSham (Mittal et al. 2023b)、FakeBench (Li et al. 2024) 以及 VANE-Bench (Gani et al. 2025)。

为了比较，我们考虑两组基线：

- **基于 LLM 的推理器：** Qwen2.5-全能-7B 和 MiniCPM-o-2_6。
- **深度伪造检测模型：**我们包括了强模态特定的基线，例如用于音频视觉伪造分类的 FTCN (Oorloff et al. 2024) 和用于音频欺骗检测的 AASIST (Jung et al. 2022)。
- **深度伪造推理模型：**由于缺乏公开可用的基于推理的深度伪造检测器，在这一类别中没有进行直接比较。

数据集。我们仅在 X-AVFake 上评估我们的方法，因为没有现有的数据集为视觉和音频操作提供细粒度的标注以及相应的推理标签。

5.2 与竞争数据集的比较

表 2汇总了 X-AVFake 和九个具有代表性的深度伪造或操纵数据集在视觉、音频和多模态领域之间的比较。早期的数据集，如 FaceForensics++(Rossler et al. 2019)、CelebDF(Li et al. 2020b) 和 UADFV(Yang, Li, and Lyu 2019) 主要集中在面部视觉伪造上，但缺乏音频内容和细粒度的推理标注。类似地，像 Psynd(Zhang and Sim 2022) 这样的特定于音频的数据集是单模态的，并且在操纵类型方面受限。

VideoSham (Mittal et al. 2023b) 和 VANE-Bench (Gani et al. 2025) 包括多模态操作，但没有提供逐步解释或明确的可解释性推理类别。FakeBench (Li et al. 2024) 是基于图像并且包含文本解释，但是缺乏时间动态性和跨模态复杂性。

相比之下，X-AVFake 在同一基准中引入了音频和视频操作，并配以结构化的元数据，如违反的推理类别（A – F）、被操纵区域/时间戳以及自然语言解释。它是唯一支持可解释多模态深度伪造检测的数据集，拥有超过 5,700 个高质量样本和均衡的真实/虚假配对样本，为检测和可解释性研究提供了独特的测试平台。

5.3 与最先进的方法比较

我们将 FakeHunter 与一组具有代表性的基线方法进行比较，涵盖单模态和多模态的深度伪造检测方法。竞争方法分类如下：

- **单模检测模型：**对于仅视觉检测使用 FTCN (Oorloff et al. 2024)，对于仅音频检测使用 AASIST (Jung et al. 2022)。

模型	总体准确性	音频子集	可视子集
FTCN (video only)	—	—	0.00%
AASIST (audio only)	—	18.69%	—
Qwen2.5-Omni-7B	18.68%	12.27%	24.83%
MiniCPM-o-2_6	9.19%	17.88%	0.78%
FakeHunter with Qwen2.5-Omni-7B	34.75%	23.00%	46.50%
FakeHunter with MiniCPM-o-2_6	27.00%	25.50%	28.50%

表 3: 准确性跨模态对比。多模态模型在音频和视觉子集上进行评估。

- **多模态大语言模型：** Qwen2.5-Omni-7B 和 MiniCPM-o-2_6，它们可以接受音频和视觉输入。

单模态基线仅在其支持的模态上进行评估，而 FakeHunter 和基于大语言模型的方法则在完整的多模态数据集和特定模态的数据子集上进行评估。具体来说，我们报告：

- **音频准确性** 超过 $\mathcal{D}_A$（操纵音频的视频），
- **视觉准确性** 超过 $\mathcal{D}_V$（视觉操控视频），
- **总体准确率** 超过 $\mathcal{D}_A \cup \mathcal{D}_V$：

$$\mathrm{Acc}_{\mathrm{Overall}} = \frac{|\mathrm{Correct}_A| + |\mathrm{Correct}_V|}{|\mathcal{D}_A| + |\mathcal{D}_V|}.$$

表 4展示了结果。FTCN 和 AASIST，限于单一模态，表现出的性能不如多模态方法。Qwen2.5-Omni 和 MiniCPM-o-2_6 展现了适度的基础能力，但在多模态对齐方面存在困难。

FakeHunter 在所有指标中均取得了优异的结果：

- **多模态优势：假猎手** (Qwen2.5) 在 $\mathcal{D}_{\mathrm{all}}$ 上达到了 34.75%，几乎是 Qwen2.5-Omni (18.68%) 和 MiniCPM-o-2_6 (9.19%) 的两倍。
- **增强的视觉检测：** 放大视觉检查使**假猎手**（Qwen2.5）在 $\mathcal{D}_{\mathrm{visual}}$ 上达到 46.50% 的准确率。
- **框架增益：** 虽然 MiniCPM-o-2_6 单独表现不佳（例如，在视觉任务上仅为 0.78%），但我们的框架通过检索、推理和基于工具的优化将其提升至 28.50%。

这些结果突出了我们记忆引导的 CoT 推理流水线的有效性，展示了它处理具有强解释性和泛化性的挑战性多模态深度伪造内容的能力。

5.4 剔除实验研究

为了评估 FakeHunter 中每个模块的贡献，我们进行了一项消融研究，比较了在 X-AVFake 基准测试（表 4）上的四种系统变体。我们隔离了记忆引导检索和工具增强验证的影响。

基础大语言模型性能。在没有我们框架的情况下，原始的 Qwen2.5-Omni-7B 达到了有限的准确性（整体为 18.68%），在音频方面的结果（12.27%）低于视觉方面（24.83%）。这突显了仅通过端到端的 LLM 推理检测多模态不一致性的困难。

工具增强验证的效果。添加我们的工具模块（例如，放大或频谱图检查）可以提高音频深度伪造检测。与仅使用内存的变体（21.00%）相比，在保留工具的同时移除内存得到 27.00%的准确率——这表明局部、细粒度分析能够恢复被全局推理忽略的细微线索。

记忆引导检索的影响。内存仅变体在视觉样本上表现良好（40.50%），但在音频上的表现较差（1.50%），这表明视觉定位更多受益于检索对比，而音频不一致性需要更深入的检查。

完整模型。我们的完整 FakeHunter（包含记忆和工具）表现出色，整体达到了 34.75%，在各种模态下均有良好的结果（音频为 23.00%，视觉为 46.50%）。这证实了上下文检索和精细检查对于稳健、可解释的多模态检测至关重要。

5.5 运行效率

完整的 FakeHunter 管道——包括特征提取、内存检索和多模态推理——在四块 NVIDIA A800 GPU 上以约 0.9× 实时的速度运行。尽管其设计有多阶段的模块化推理和基于工具的检查，推理仍然保持实用。这种效率来源于轻量级的 Qwen2.5-Omni 主干和 FAISS 加速的内存检索。

6 结论

假猎手是一种逐步多模态框架，集成了记忆检索、连贯思维推理和工具增强验证，用于准确的和可解释深度伪造检测。基于 Qwen2.5- Omni-7B，它共同分析视频

模型变体	整体准确性	音频子集	可视子集
Qwen2.5-Omni-7B (Raw)	18.68%	12.27%	24.83%
FakeHunter (Qwen) w/ Memory, w/o Tool	21.00%	1.50%	40.50%
FakeHunter (Qwen) w/o Memory, w/ Tool	27.00%	27.00%	27.00%
FakeHunter (Qwen) Full (w/ Memory & Tool)	34.75%	23.00%	46.50%

表 4: 不同消融设置下 FakeHunter 在 X-AVFake 基准上的性能比较。记忆引导检索和工具增强验证均有助于提高准确性。

和音频，并通过检索示例来锚定判断，在置信度较低时调用放大或频谱图工具。每个阶段都会发出结构化的证据，描述什么被操纵了，其中它发生了，以及为什么它是假的——弥合了原始预测与法医级解释之间的差距。

为了支持可解释的取证，我们发布了 X-AV **伪造**，一个带有操作类型、目标实体、推理类别和自由形式理由注释的 ∼5.7k 视频基准。涵盖视觉实体移除和音频替换，X-AVFake 实现了对复杂音视操作准确性和可解释性的整体评估。

实验表明，FakeHunter 达到了 34.75%的准确率，超过了 vanilla Qwen2.5-Omni-7B +16.9 pp 和 MiniCPM-2.6 +25.6 pp。内存检索带来了稳定的增益，而基于工具的检查对于关键低置信度片段提升了准确率至 46.5%。尽管采用了多阶段设计，该管道在 4 个 ×NVIDIA A800 GPU 上以 ∼0.3× 实时运行，处理一段 10 分钟的片段需要大约 33 分钟——展示了部署的可能性。

广泛影响。 可解释的检测对新闻业、内容审核和法律取证有益，因为它揭示了模型推理并允许进行法庭可接受的分析。其模块化设计允许集成未来的探测器或特定领域的工具而不必重新训练核心模型。

未来工作。 我们计划（1）通过场景重新合成和跨模态不匹配等高阶操作扩展 X-AVFake，（2）开发自适应工具选择以加快推理速度，以及（3）探索将推理痕迹与视听特征融合的端到端训练。我们将发布数据集、代码和模型，以促进可解释视频取证领域的透明且社区驱动的进步。

参考文献

Agarwal, S.; and Farid, H. 2020. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In CVPR Workshops.

Aghasanli, A.; Kangin, D.; and Angelov, P. 2023. Interpretable-through-prototypes deepfake detection for diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, 467–474.

Al-Sanjary, O. I.; Ahmed, A. A.; and Sulong, G. 2016. Development of a video tampering dataset for forensic investigation. Forensic science international, 266: 565–572.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Bai, W.; Liu, Y.; Zhang, Z.; Li, B.; and Hu, W. 2023. AUNet: Learning Relations Between Action Units for Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 24709–24719.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. In SIGGRAPH Asia 2024 Conference Papers, 1–11.

Chesney, R.; and Citron, D. 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff., 98: 147.

Choi, J.; Kim, T.; Jeong, Y.; Baek, S.; and Choi, J. 2024. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1133–1143.

Di Pierno, A.; Guarnera, L.; Allegra, D.; and Battiato, S. 2025. End-to-end Audio Deepfake Detection from RAW Waveforms: a RawNet-Based Approach with Cross-Dataset Evaluation. arXiv preprint arXiv:2504.20923.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The DeepFake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:1910.08854.

Du, F.; Yu, M.; Li, B.; Chow, K. P.; Jiang, J.; Zhang, Y.; Liang, Y.; Li, M.; and Huang, W. 2024. TAENet: Two-branch Autoencoder Network for Interpretable Deepfake Detection. Forensic Science International: Digital Investigation, 50: 301808.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. IEEE.

Febrinanto, F. G.; Moore, K.; Thapa, C.; Ma, J.; Saikrishna, V.; and Xia, F. 2025. Vision Graph Non-Contrastive Learning for Audio Deepfake Detection with Limited Labels. arXiv preprint arXiv:2501.04942.

Feng, C.; Chen, Z.; and Owens, A. 2023. Self-Supervised Video Forensics by Audio-Visual Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10491–10503.

Frank, J.; Eisenhofer, T.; Knöth, J.; Rathgeb, C.; Busch, C.; and Damer, N. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In ACM Workshop on Information Hiding and Multimedia Security.

Gani, H.; Bharadwaj, R.; Naseer, M.; Khan, F. S.; and Khan, S. 2025. Vane-bench: Video anomaly evaluation benchmark for conversational lmms. In Findings of the Association for Computational Linguistics: NAACL 2025, 3123–3140.

Hondru, V.; Hogea, E.; Onchis, D.; and Ionescu, R. T. 2025. Exddv: A new dataset for explainable deepfake detection in video. arXiv preprint arXiv:2503.14421.

Jiang, L.; Wu, W.; Li, R.; Qian, C.; and Loy, C. C. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. arXiv preprint arXiv:2001.03024.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3): 535–547.

Jung, J.-w.; Heo, H.-S.; Tak, H.; Shim, H.-j.; Chung, J. S.; Lee, B.-J.; Yu, H.-J.; and Evans, N. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), 6367–6371. IEEE.

Khalid, H.; Tariq, S.; and Woo, S. S. 2021. FakeAVCeleb: A Novel Audio–Video Multimodal Deepfake Dataset. arXiv preprint arXiv:2108.05080.

Korshunov, P.; and Marcel, S. 2018. Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

Kundu, R.; Balachandran, A.; and Roy-Chowdhury, A. K. 2025. TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data. arXiv preprint arXiv:2503.15867.

Li, Y.; Bao, J.; Yang, X.; Dong, M.; Wen, F.; Liu, D.; and Lyu, S. 2020a. Face X-ray for More General Face Forgery Detection. In Conference on Computer Vision and Pattern Recognition (CVPR).

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. Exposing DeepFake Videos By Detecting Eye Blinking. In IEEE International Workshop on Information Forensics and Security (WIFS).

Li, Y.; Liu, X.; Wang, X.; Lee, B. S.; Wang, S.; Rocha, A.; and Lin, W. 2024. Fakebench: Probing explainable fake image detection via large multimodal models. arXiv preprint arXiv:2404.13306.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 3207–3216.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020c. Celeb-DF: A New Dataset for DeepFake Detection. In Conference on Computer Vision and Pattern Recognition (CVPR).

Mirsky, Y.; and Lee, W. 2021. The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR), 54(1): 1–41.

Mittal, T.; Sinha, R.; Swaminathan, V.; Collomosse, J.; and Manocha, D. 2023a. Video Manipulations Beyond Faces: A Dataset with Human-Machine Analysis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops.

Mittal, T.; Sinha, R.; Swaminathan, V.; Collomosse, J.; and Manocha, D. 2023b. Video manipulations beyond faces: A dataset with human-machine analysis. In Pro-

ceedings of the IEEE/CVF winter conference on applications of computer vision, 643–652.

Nie, F.; Ni, J.; Zhang, J.; Zhang, B.; and Zhang, W. 2024. FRADE: Forgery-aware Audio-distilled Multimodal Learning for Deepfake Detection. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), 6297–6306.

Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; and Bharaj, G. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 27092–27102.

Qian, Y.; Chen, N.; and Yu, K. 2016. Deep features for automatic spoofing detection. Speech Communication, 85: 43–52.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, 8748–8763. PmLR.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, 1–11.

Tak, H.; Patino, J.; Todisco, M.; Nautsch, A.; Evans, N.; and Larcher, A. 2021. End-to-end anti-spoofing with rawnet2. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6369–6373. IEEE.

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64: 131–148.

Vaccari, C.; and Chadwick, A. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society, 6(1): 2056305120903408.

Wang, C.; Yi, J.; Tao, J.; Zhang, C.; Zhang, S.; Fu, R.; and Chen, X. 2023. TO-Rawnet: improving RawNet with TCN and orthogonal regularization for fake audio detection. arXiv preprint arXiv:2305.13701.

Xing, Y.; He, Y.; Tian, Z.; Wang, X.; and Chen, Q. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7151–7161.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761.

Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; and Ren, K. 2023. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. IEEE Transactions on Information Forensics and Security, 18: 2015–2029.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), 8261–8265. IEEE.

Zhang, B.; and Sim, T. 2022. Localizing fake segments in speech. In 2022 26th International Conference on Pattern Recognition (ICPR), 3224–3230. IEEE.

Zhang, Y.; Colman, B.; Guo, X.; Shahriyari, A.; and Bharaj, G. 2024. Common sense reasoning for deepfake detection. In European Conference on Computer Vision, 399–415. Springer.

Zhou, S.; Li, C.; Chan, K. C.; and Loy, C. C. 2023. Propainter: Improving propagation and transformer for video inpainting. In Proceedings of the IEEE/CVF international conference on computer vision, 10477–10486.

Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM international conference on multimedia, 2382–2390.