arXiv:2603.14992v1 [cs.AI] 16 Mar 2026

揭示短视频假新闻检测中的跨模态一致性

Chong Tian[1], Yu Wang[1],ChenxuYang[2],Junyi Guan[1],
Zheng Lin[2], Yuhan Liu[1],Xiuying Chen[1]*, Qirong Ho[1]*

[1]穆罕默德·本·扎耶德人工智能大学（MBZUAI）[2]中国科学院信息工程研究所（CAS）{chong.tian, xiuying.chen, qirong.ho}@mbzuai.ac.ae

摘要

短视频平台是新闻传播的重要渠道，但也为多模态错误信息提供了滋生土壤：各模态单独看似合理，跨模态关系却存在细微的不一致，例如视觉内容与图像说明文字不匹配。在两个基准数据集 FakeSV（中文）和 FakeTT（英文）上，我们观察到明显的不对称性：真实视频的文本–视觉一致性较高，但文本–音频一致性中等；而伪造视频则呈现相反模式。此外，单一的全局一致性分数构成一条可解释的轴线，虚假概率和预测误差沿其平滑变化。受这些观察启发，我们提出 **MAGIC**^{3} （模态对抗门控交互与一致性中心分类器），这是一种能够在多种粒度上明确建模并揭示跨三模态一致性信号的检测器。MAGIC^3 将显式的成对与全局一致性建模，与基于跨模态注意力得出的词元级和帧级一致性信号相结合；采用多风格大语言模型改写，以获得风格鲁棒的文本表示；并利用不确定性感知分类器进行选择性视觉语言模型路由。使用预提取特征时，MAGIC^3 在 FakeSV 和 FakeTT 上始终优于最强的非视觉语言模型基线。在准确率与视觉语言模型级准确率相当的同时，该两阶段系统实现了 18–27× 倍的吞吐量提升和 93% 的显存节省，提供了出色的成本–性能权衡。



图 1：**跨模态一致性模式示意图**。在真实新闻短视频中，文本、视觉内容和音频在语境上相互协调（一致）。在假新闻中，耸人听闻的说法（文本/音频）与实际视觉内容之间往往存在"语义鸿沟"。MAGIC^3 充当一致性透镜，用于量化这些多模态关系，以实现假新闻检测。

1 引言

TikTok、抖音、YouTube Shorts 和快手等短视频平台重新塑造了人们消费新闻和公共信息的方式（Bu 等人，2023年）。面对海量上传内容，检测系统不仅必须准确，还必须具备较高的计算效率。丰富的多模态信号和高度优化的推荐系统，使假新闻短视频得以在专业事实核查开展之前便吸引大量关注，加剧社会风险（Vosoughi等人，2018；Zhou和Zafarani，2020）。与纯文本假新闻相比（Devlin 等人，2019；Shu等人，2019；Tian等人，2025年），短新闻视频更难分析：操纵可能发生在任何模态、任何粒度上，而成功的伪造内容很少会在单一信息流中留下明显痕迹。相反，它们往往利用跨模态不一致性：每种模态单独看都合情合理，但联合传达的信息却具有误导性——如图1所示。

一种以一致性为中心的视角。我们采用面向任务的短视频新闻跨模态不一致性分类体系，涵盖实体/上下文、事件语义、情感、时间上的音视频错位以及风格/语用，并在附录中提供了详细定义和示例A。本文中，我们用"一致性"指代特征空间中学习得到的、面向任务的兼容性信号，而非相对于外部世界知识的事实正确性。我们没有再设计一个庞大的骨干网络，而是构建了一个

*通讯作者。

轻量且可解释的一致性透镜，可置于任意多模态特征提取器之上，并专为伪造新闻短视频量身打造。

关键发现。 基于这一以一致性为中心的视角，我们的分析揭示了四项关键发现：(i) 非对称一致性：真实视频的文本–视觉一致性较高，但文本–音频一致性中等；而虚假视频则呈现相反的模式；(ii)可解释的轴：全局一致性与预测难度高度相关，错误集中在中间数值；(iii)高效路由：结合一致性与不确定性，仅将 25% 的样本路由至重量级视觉语言模型，即可超越仅视觉语言模型准确率；(iv) 风格鲁棒性：多风格大语言模型改写可提升鲁棒性，且虚假视频的一致性方差更高。

我们的方法。 我们提出**MAGIC**^{3}，这是一种围绕显式一致性建模设计的特征级检测器。给定冻结的文本、视觉和音频特征（以及少量大语言模型改写文本），MAGIC^3 会：(i) 通过跨模态一致性门控（CMCG）计算成对及全局一致性分数；(ii) 根据跨模态注意力生成词元级和帧级一致性场；(iii) 在对抗感知改写融合（AARF）模块中构建风格鲁棒文本表示；(iv) 利用时间音频–视觉不一致性分数（TCMI）执行轻量级分层多模态融合；(v) 输出经校准的不确定性估计，以驱动两阶段流水线，仅在需要时调用计算开销较大的基于视觉语言模型的检测器。

贡献。 我们做出了三项贡献：(i) 据我们所知，这是首项针对假新闻短视频的研究，将三模态（文本–视觉–音频）一致性作为显式的多粒度输出：包括成对和全局标量分数，以及词元级和帧级一致性场，用以揭示文本–视觉与文本–音频之间的不对称性，并提供可解释的一致性–难度轴。(ii) MAGIC^3, 是一种检测器，基于这些显式一致性信号以及校准的不确定性估计构建，为短视频新闻提供可解释的“一致性透镜”。(iii) 一种由一致性和不确定性驱动的两阶段视觉语言模型路由策略，在 FakeSV 和 FakeTT 上达到最先进的准确率，同时将吞吐量提高 18–27× ，并显著改善成本–性能权衡。

2 相关工作

针对图像–文本对的多模态假新闻。先前研究通过联合建模图像说明文字和图像，将假新闻检测扩展到图像–文本帖子（Zhou 等人，2020；Müller-Budack 等人，2020；Li 等人，2025年）。EANN（Wang 等人，2018）等事件对抗方法学习事件不变的多模态特征，以提升泛化能力；SAFE（Zhou 等人，2020）等相似性感知方法则对图像–文本的一致或不一致关系进行编码，以识别篡改。这些方法启发我们关注显式的跨模态一致性，但其主要针对文章或帖子级内容，而非带有音频的短视频。

短新闻视频检测。现有的短视频假新闻检测器主要分为三类。基于特征的多模态模型采用强大的单模态编码器，并通过后期融合进行整合，例如 FakeSV Qi 等人，2023 年，但对交叉模态交互的利用较弱。深度融合架构，如 TikTec、FANVM、SVFEND、FakingRecipe 和 KDSGAT-FNVD（Shang 等人，2021；Choi 和 Ko，2021；Qi 等人，2023年；Bu 等人，2024年；Huang 等人，2025年）能够学习强大的联合表征，并隐式地推理一致性，但不会揭示不一致性出现的位置。较新的系统，如 CA-FVD（Wang 等人，2025b）、GLFE-SVFD（Wang 等人，2025a）和 SCMG-FND（Yang 等人，2025年）利用一致性感知训练、全局—局部特征增强或基于大语言模型的语义可信度分数，以更好地捕捉多模态关系；然而，其一致性线索仍然是粗粒度的视频级信号或内部特征。相比之下，我们聚焦于包含三种模态的短新闻视频，并将多粒度一致性——成对/全局分数、词元级/帧级一致性场以及时序视听不一致性——转化为显式输出，同时用于驱动不确定性感知的视觉语言模型路由。

3 方法

3.1 问题定义

在这项工作中，我们聚焦于短视频新闻，因为它是在线新闻消费和多模态虚假信息中的主导且具有实际重要性的模态。我们研究的两个基准数据集 FakeSV 和 FakeTT，均专门基于现实世界平台上的短视频构建，反映了用户在当今媒体生态系统中实际接触新闻的方式。因此，我们的建模选择、

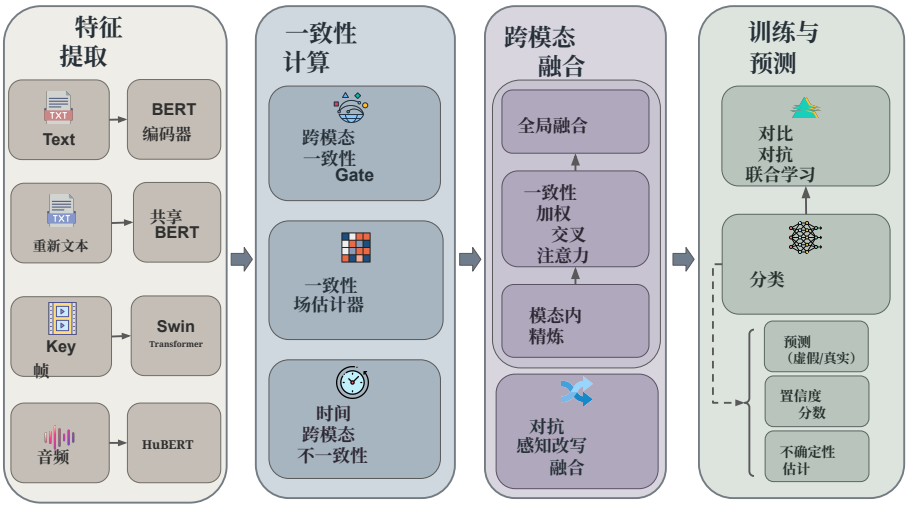


图2：MAGIC^3 概述。冻结编码器提供文本、视觉、音频和改写特征。跨模态一致性门控输出成对一致性得分和全局一致性得分；一致性场估计器将跨模态注意力转换为词元级和帧级一致性场；时序跨模态不一致性模块计算时序不一致性分数。对抗感知改写融合将原始文本与大语言模型改写融合为风格鲁棒表征；分层多模态变换器执行分层多模态融合；分类器则输出虚假概率和不确定性。该框架通过对比对抗联合学习进行训练。

研究发现和效率主张，应结合这一具有实际重要性但范围经过有意界定的领域来理解。给定一段短新闻视频 $\mathcal{V}$ 及其关联文本 $\mathcal{T}$（标题、字幕和描述），任务是预测其为伪造（$y = 1$）还是真实（$y = 0$）。

3.2 特征提取

遵循先前的研究（Qi 等人，2023年；Huang 等人，2025年），我们采用固定预训练编码器来提取特征级表征。具体而言，我们使用 BERT（Devlin 等人，2019）对文本标题和描述进行编码，使用 Swin Transformer（Liu 等人，2021）对视觉关键帧进行编码，并使用 ExHuBERT（Amiriparian 等人，2024年）提取音频情感嵌入。这种解耦设计使 MAGIC^3 能够利用最先进的表征，而无需承担端到端微调的负担。为增强鲁棒性，我们还为每段文本生成三种大语言模型改写（中性、正式、煽情），并使用相同的文本编码器对其进行编码。

3.3 一致性计算

该模块计算多粒度一致性信号，以揭示不同抽象层级上的跨模态差异。我们假设，虚假新闻视频会以多种方式呈现不一致：有些整体上缺乏连贯性（例如，字幕与视觉内容无关），而另一些则包含细微的局部篡改（例如，某个短语与某一视频帧相矛盾）或时间伪影。为捕捉这种多样性，我们设计了三种互补机制：（1）用于高层筛选的全局门控；（2）用于局部定位的细粒度一致性场；（3）用于检测同步伪影的时间不一致性分数。这种多视角方法确保 MAGIC^3 既能敏锐识别低成本伪造内容，也能识别复杂的篡改。

跨模态一致性门控 (CMCG)。如图所示2，该模块将各模态的特征汇聚为向量 $\mathbf{h}_{\text{text}}$、$\mathbf{h}_{\text{vis}}$和 $\mathbf{h}_{\text{aud}}$。随后，它通过门控机制计算每一对的标量一致性得分(m_i, m_j)：

$$c_{ij} = \sigma(\mathbf{w}_{ij}^{\top}[\bar{\mathbf{h}}_{m_i}; \bar{\mathbf{h}}_{m_j}] + b_{ij}). \tag{1}$$

这些成对分数组成一个向量 $\mathbf{c} = [c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}]$，它简洁地概括了模态间一致性。最后，将这些分数压缩为单个全局一致性标量：

$$c_{\text{global}} = \sigma(\text{MLP}_c(\mathbf{c})). \tag{2}$$

这个全局分数 c_{global} 提供了对视频整体连贯性的高层次衡量，并调制跨模态注意力。全文中，我们用“一致性”指代在冻结编码器特征空间中计算的、面向任务的学习型兼容性代理指标，从而使一致性透镜保持轻量，并与庞大的主干网络解耦。

一致性场估计器 (CFE)。它将跨模态注意力矩阵转换为逐词元和逐帧的“一致性场”，从而定位视频中哪些地方 存在不一致。给定来自跨模态 Transformer 的交叉注意力矩阵 $\mathbf{A}_{m_i \leftarrow m_j} \in \mathbb{R}^{L_{m_i} \times L_{m_j}}$，我们定义：

$$F_{ij}^{(m_i)}(t) = \max_k \mathbf{A}_{m_i \leftarrow m_j}(t,k), \quad t = 1, \ldots, L_{m_i}. \tag{3}$$

直观地说，$F^{(m_i)}{}_{ij}(t) \in [0,1]$ 反映了模态 t 中的词元 m_i 在模态m_j中找到支持证据的强度。关注度高且集中的词元得分较高；关注分散或较弱的词元得分较低。这些字段支持热力图可视化和局部分析，FakeSV 和 FakeTT 的一些示例可视化如附录6中的图B所示。

时序跨模态不一致性(TCMI)。与此同时，我们沿时间维度对齐音频和视觉特征，以捕捉粗粒度的不匹配。我们将两种模态重采样至长度 T'，并计算每个时间步的距离：

$$d_t = \left\| \tilde{\mathbf{h}}_{\text{vis},t} - \text{Proj}_{\text{av}}(\tilde{\mathbf{h}}_{\text{aud},t}) \right\|_2, \tag{4}$$

将其汇总为 $\mathbf{s}^{\text{temp}}$[均值(d)；方差(d)；最大 = 值(d)]，并映射为：

$$c_{\text{temp}} = 1 - \sigma\big(\text{MLP}_{\text{temp}}(\mathbf{s}_{\text{temp}})\big). \tag{5}$$

这捕捉了受深度伪造检测器启发的粗粒度音频–视觉不匹配（Gao 等人，2024年）。

3.4 跨模态融合

对抗感知改写融合(AARF)。多视角对抗感知改写融合模块将大语言模型改写视为语义保持但风格有所改变的视角（例如中性、正式和煽情风格）。设$\mathbf{h}^{\text{orig}}$ = Pool($\mathbf{H}_{\text{text}}$)为原始文本，$\mathbf{h}_{\text{rew}}^{(v)}$ 为第 v 次改写文本。我们计算门控权重：

$$q^{(v)} = \sigma\big(\text{MLP}_q[\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(v)}]\big), \tag{6}$$

其中，$q^{(v)}$ 用于估计改写质量。门控多层感知机根据所有视图生成融合 logits，并强制原始文本的权重至少为 $\alpha_{\min}$：

$$\mathbf{h}_{\text{text}}^{\text{fuse}} = \alpha_0 \, \mathbf{h}_{\text{orig}} + \sum_{v=1}^{V} \alpha_v \, \text{Proj}\big(\mathbf{h}_{\text{rew}}^{(v)}\big), \tag{7}$$

其中 $\alpha_0 \geq \alpha_{\min}$ 和 $\sum_v \alpha_v = 1$，Proj(·) 是一种学习得到的线性投影，用于将重写文本嵌入对齐到原始文本嵌入的维度。对比损失会拉近来自同一视频的重写文本的表示，并推远来自不同视频的表示，从而促使模型学习具有风格不变性的嵌入。

分层多模态变换器(HMT)。该组件分三个阶段运行。(i)模态内细化：对于每种模态 $m \in$ {文本、视觉、音频}，我们使用带有自注意力的模态专属 Transformer 编码器 $\mathbf{H}_m^{(A)}$ = TransEnc$_m$($\mathbf{H}_m$)，对该模态内的词元或帧序列进行细化。(ii)一致性加权交叉注意力：随后，我们加入交叉注意力模块，使一种模态的查询关注另一种模态的键/值，并使用由成对一致性向量 $\mathbf{c}$ 得出的权重聚合生成的消息。此阶段的交叉注意力矩阵也供 CFE 用于构建词元级和帧级一致性场。(iii)全局聚合：最后，我们对所有模态（以及融合文本）的细化表示进行池化，并将其输入一个小型全局融合层，以获得单一的视频级表示 $\mathbf{h}^{\text{global}}$，再将其传递给分类器。

3.5 训练与预测

我们使用对比-对抗联合学习（CAJL）目标训练 MAGIC^3，以联合优化分类准确率，以及用于预测和路由的一致性与不确定性信号的可靠性。主要监督损失采用标准二元交叉熵：

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N} \sum_i \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big). \tag{8}$$

为提高鲁棒性和对齐效果，我们采用 InfoNCE 风格的对比损失进行模态内和跨模态对齐（$\mathcal{L}_{\text{intra}}$，$\mathcal{L}_{\text{cross}}$）。对于一个包含 N 对（$\mathbf{u}_i$，$\mathbf{v}_i$）的批次，损失为：

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N} \sum_i \log \frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau)}. \tag{9}$$

为提升预测的鲁棒性，我们还在训练过程中向隐藏表示中引入随机噪声扰动，并最小化干净特征与受扰动特征上的预测之间的 KL散度，将其定义为对抗一致性正则项 $\mathcal{L}_{\text{adv}}$。最终的复合目标将这些项与一个一致性

正则化项 $\mathcal{L}_{\text{reg}}$ 结合起来，以对齐全局和局部分数：

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \quad (10)$$

其中 λ 为权衡超参数。所有术语的更多细节见附录 C。

分类器。 最终分类器会生成三个不同的信号：(1) 预测：二元分类概率 $\hat{y}_{fake}$，表示视频为伪造视频的可能性。(2) 置信度分数：根据 Softmax 输出得出，conf $=\max(\hat{y}, 1-\hat{y})$，表示模型对其所选类别的置信程度。(3) 不确定性估计：由预测熵$u_{\text{ent}} = -\sum p\log p$和学习得到的辅助标量 $\hat{u}$（经训练以回归误差裕度）组合而成的复合指标。这些信号为后续路由决策提供了可靠依据，使系统能够区分有把握但错误的预测与真正的模糊情况。

两阶段路由。 在推理过程中，我们采用路由函数 $\mathcal{R}(x)$来决定是信任 MAGIC^3 的预测，还是升级至重量级视觉语言模型。不同于简单的置信度阈值判定，我们利用多维输出制定有效的路由策略（详见附录D.4）：(1) 基于不确定性：将预测熵较高的样本（$u_{\text{ent}} > \tau$）路由出去，优先处理模型在信息论意义上不确定的情况。(2) 基于难度：使用复合难度分数 $d = \text{Norm}(u_{\text{ent}}) + (1 - c_{\text{global}}) + (1 - \text{conf})$ 将同时具有不确定、不一致且置信度低等特征的样本路由出去。过滤掉 MAGIC ∼75能够正确处理的“简单”案例（占 ^{3} %），我们便能以极低的成本达到视觉语言模型级准确率。

4 实验

4.1 实验设置

我们在 FakeSV 和 FakeTT 基准数据集上评估 MAGIC^3 在短视频假新闻检测任务中的表现(Qi 等人，2023；Huang 等人，2025）。我们遵循 Huang 等人 (2025) 的按时间顺序划分协议，将视频按 70%/15%/15% 的比例划分为训练集、验证集和测试集。数据集统计数据汇总于表 10，更多详情见附录 E。

文本、视觉和音频特征按照第3.2 节所述，使用固定预训练编码器提取。所有可训练参数均位于 MAGIC^3 检测器头（CMCG、CFE、AARF、HMT+TCMI 和分类器）中。我们使用 AdamW，学习率为 5×10^{-5}，批次大小为 32，并采用余弦衰减策略，预热比例为 10%，隐藏层维度为 256，注意力头数量为 8。大语言模型改写使用 DeepSeek-V3.2 离线生成，采用三种风格提示（中性、正式、煽情）；详细提示列于附录F。损失权重和超参数在验证集上进行调优（附录11 中的表G）。视频/音频/文本特征预先提取后，检测器头在单张 RTX 5090 GPU 上、批次大小为 128 时，吞吐量约为 ∼680 个特征级样本/秒（约 1.5 毫秒/样本，且使用缓存特征时峰值 GPU 显存为 ∼10.67GB）；端到端延迟主要由解码和特征提取决定，可通过缓存和异步预处理进行摊销。附录H 中也进一步讨论了效率与部署。

4.2 基线方法

我们将 MAGIC^3 与三类基线方法进行比较。(i) 零样本视觉语言模型：GPT-4o-mini、GPT-4.1-mini、Qwen2.5-VL（Bai 等，2025）、InternVL2.5（Chen 等，2024）和 InternVL2.5-MPO（Wang 等，2024b），通过直接提示这些模型执行检测任务来进行评估。(ii) 监督式多模态方法：HCFC-Hou（Bu 等，2024）、HCFC-Medina（Serrano 等，2020）（手工特征分类器）、TikTec（Shang 等，2021）、FANVM（Choi 和 Ko，2021）、SVFEND（Qi 等，2023）、FakingRecipe（Bu 等，2024）和 KDSGAT-FNVD（Huang 等，2025）。(iii) 基于视觉语言模型的检测器：FakeSV-VLM（Wang 等，2025c），这是一种专为假新闻视频检测设计的微调视觉语言模型。

4.3 主要结果

表1 报告了在 FakeTT 和 FakeSV 上的结果。我们将 MAGIC^3 与第4.2节详述的基线进行比较。对于每个数据集，我们报告准确率、宏平均F1以及各类别精确率/召回率。在 FakeSV 上，MAGIC^3 达到 86.71% 的准确率和 86.84 的宏平均F1，分别比 KDSGAT-FNVD 高出 1.98 和 2.16 个百分点，并且在真实和伪造类别上的精确率/召回率均衡。在 FakeTT 上，MAGIC^3 达到 84.95% 的准确率和 82.28 的宏平均F1，分别比 KDSGAT-FNVD 高出 3.01 和 3.34 个百分点。零样本视觉语言模型的表现远远落后：在 FakeSV 上，表现最好的视觉语言模型比 MAGIC^{3}低约 16 个准确率百分点；在 FakeTT 上，差距超过 20 个百分点，这凸显了监督学习和显式一致性建模的价值。这些优势在宏平均F1和各类别精确率/召回率方面同样明显，尤其是在 FakeTT 的真实类别上

方法	FakeSV						FakeTT					
	Acc	宏平均 F1	Real		Fake		Acc	宏平均 F1	Real		Fake	
			Pre	Rec	Pre	Rec			Pre	Rec	Pre	Rec
零样本视觉语言模型												
GPT-4o-mini	68.08	68.05	65.35	73.97	74.41	64.93	61.54	61.20	81.62	52.00	47.28	79.80
GPT-4.1-mini	70.30	70.25	68.12	73.53	73.10	68.21	49.16	48.54	77.42	48.00	37.50	71.40
Qwen2.5-VL	64.21	60.79	62.50	60.54	66.60	62.50	45.82	45.31	72.10	45.00	41.28	65.84
InternVL2.5	64.39	57.89	63.80	57.50	73.24	63.50	46.82	45.29	77.60	49.00	52.24	69.46
InternVL2.5-MPO	65.13	61.07	63.75	59.16	69.17	65.08	43.14	40.84	74.00	47.00	49.80	65.46
监督基线												
HCFC-Hou	74.91	73.61	77.72	60.08	73.46	86.51	73.24	72.00	87.04	70.50	56.93	78.79
HCFC-Medina	76.38	75.83	74.77	69.75	77.50	81.58	62.54	62.23	84.92	53.50	46.24	80.81
TikTec	73.43	73.26	68.08	74.37	78.37	72.70	66.22	65.08	82.35	63.00	49.32	72.73
FANVM	79.52	78.81	80.98	69.75	78.64	87.17	71.57	70.21	85.28	69.50	55.15	75.76
SVFEND	80.88	80.17	74.53	83.61	85.82	77.63	77.14	75.63	87.91	76.33	62.33	78.79
FakingRecipe	82.87	83.52	**91.35**	71.00	80.67	**94.74**	79.26	77.87	89.66	78.00	64.80	81.82
KDSGAT-FNVD	84.73	84.68	82.23	83.61	87.00	85.86	81.94	78.94	92.68	76.00	64.44	87.88
Ours												
FakeSV-VLM	89.60	89.40	89.20	90.10	90.50	88.00	88.90	87.60	88.20	86.50	86.60	89.10
MAGIC^3（我们的方法）	86.71	86.84	86.52	83.61	87.50	89.80	84.95	82.28	**94.12**	80.00	68.99	89.90
MAGIC^3+视觉语言模型（我们的方法）	**90.93**	**90.41**	90.85	**90.95**	**91.10**	88.90	**89.52**	**88.14**	89.70	**88.50**	**87.20**	**90.10**
端到端效率												
FakeSV-VLM	18.67 个样本/秒 @24.68 GB						30.25 个样本/秒 @22.76 GB					
MAGIC^3（我们的方法）	**680.43 个样本/秒 @1.67 GB**						**693.76 个样本/秒 @1.67 GB**					
MAGIC^3+视觉语言模型（我们的方法）	509.71 样本/秒 @24.68 GB						530.73 样本/秒 @22.76 GB					

表1：FakeSV 和 FakeTT 上的主要结果。最佳数值以**粗体**标示，次佳数值加下划线。MAGIC^3+将 25% 的样本路由至视觉语言模型（图5），同时取得相当或更高的准确率，带来**最佳成本—性能权衡**：准确率达到视觉语言模型水平或更高，吞吐量最高提升 18–27× 倍于仅使用视觉语言模型。

其中 MAGIC^3 的精确率最高。

与重量级视觉语言模型（FakeSV-VLM）结合，构成两阶段流水线时，MAGIC^3+VLM 进一步提升了性能，甚至在 FakeSV（90.93% 对 89.60%）和 FakeTT（89.52% 对 88.90%）上均超过专用的仅视觉语言模型检测器 FakeSV-VLM，同时仍明显优于所有非视觉语言模型基线（表 1）。关于路由策略及其卓越性价比优势的详细分析见第5.3节。

4.4 消融研究

为评估各组件的贡献，我们在两个数据集上进行了消融实验（表2）。移除核心一致性模块（CMCG、CFE、TCMI）会持续导致性能下降，这证实了显式一致性建模的有效性。AARF 的风格鲁棒融合对性能的影响在 FakeTT 上尤为显著，说明处理风格偏移的必要性。CFE 和 TCMI 还有助于提升可解释性，推动了场域可视化和时间不一致性检测。

5 分析与发现

除主要结果外，我们还研究了一致性信号，以深入揭示真实视频与虚假视频之间的差异。

5.1 非对称跨模态一致性

我们首先考察真实视频与虚假视频在聚合层面的跨模态一致性。使用 CMCG，我们为每个视频计算四个标量分数：文本–视觉、文本–音频、视觉–音频和全局一致性。图3 展示了 FakeSV（左）和 FakeTT（右）数据集上的分布。在这两个数据集中，真实视频的文本–视觉一致性较高，文本–音频一致性中等；而虚假视频则呈现相反的情况：文本–音频一致性非常高，但文本–视觉一致性骤降。真实视频和虚假视频的视觉–音频一致性均保持较高水平。真实视频的全局一致性平均而言更高，但真正起区分作用的是这种不对称性。

跨模态一致性不对称的根本原因 这种不对称性与“低成本伪造内容”的制作方式一致（Paris 和 Donovan，2019）。从定性角度来看，我们观察到，假新闻制作者往往会先选用高

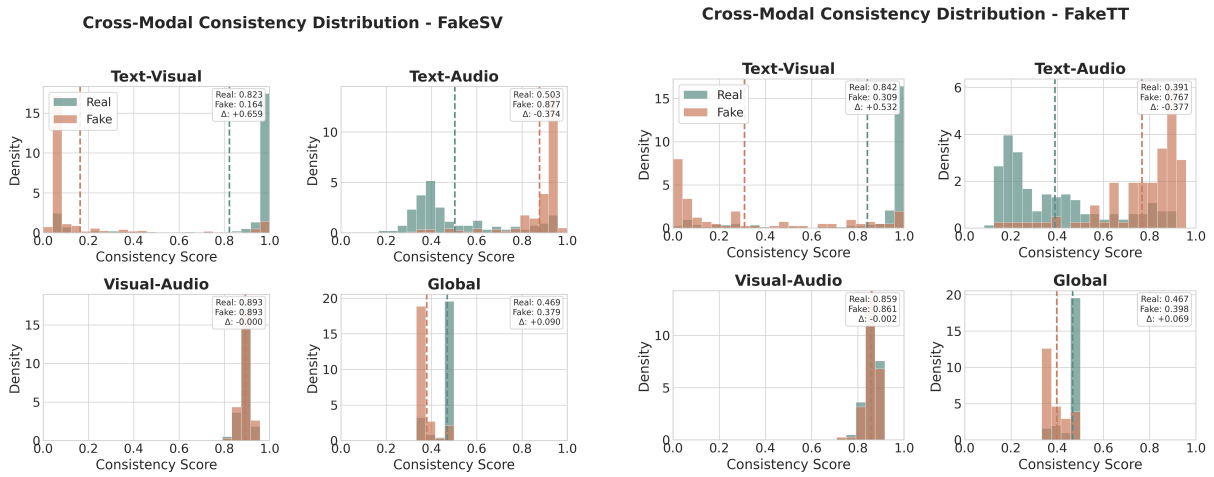


图 3：真实视频与伪造视频的文本–视觉、文本–音频、视觉–音频及全局一致性分数的跨模态一致性分布。真实新闻呈现较高的文本–视觉一致性和中等的文本–音频一致性；假新闻则呈现相反模式，而视觉–音频一致性仍然较高。左：FakeSV；右：FakeTT。

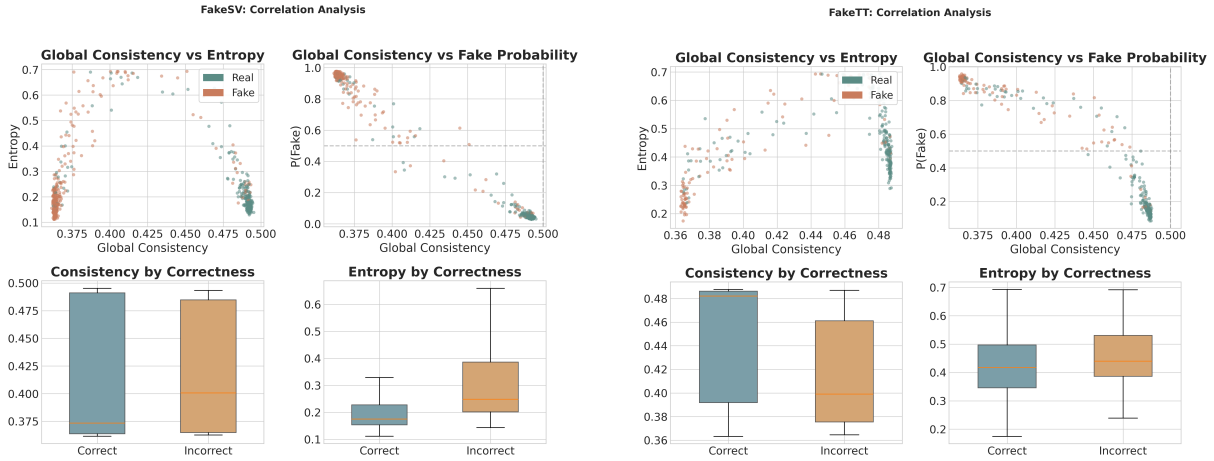


图 4：一致性–预测分析。全局一致性与虚假概率呈强负相关；预测误差集中在中等一致性区间，而 c_{global} 与熵呈非线性关系，不确定性在中等一致性时达到峰值。左：FakeSV；右：FakeTT。

质量高、视觉连贯的库存素材或盗用片段，这些素材自然能保持视觉流与背景音频之间的高度一致。随后，制作者会重新利用这些素材，添加伪造字幕和配音，以传达虚假叙事。生成的文本旨在引发情感共鸣，并与音频的情感基调紧密契合，同时逐渐偏离实际的视觉证据，从而形成 MAGIC[3] 所检测到的语义鸿沟。相比之下，真实新闻视频，尤其是 FakeTT 中由专业媒体发布的视频，更注重事实依据，其配音会贴切地描述视觉事件，因此文本–视觉一致性更强。更多统计数据见附录 D.1。

5.2 全局一致性的可解释性

我们考察全局一致性分数c_{global} 与预测结果的关系。图 4 绘出了 c_{global}与 FakeSV（左）和 FakeTT（右）上的预测伪造概率之间的关系。c_{global} 与模型的伪造概率呈强烈且大致负相关；预测误差聚集在c_{global}的中间区间，表明该分数提供了一个可解释的低维难度轴。

全局一致性的可解释性 由于c_{global} 是显式输入特征，因此这种相关性是意料之中的。然而，其价值在于实际应用价值。与不透明的端到端概率不同，c_{global} 为人工审核员提供了对模型决策的直观解释：一段视频被标记出来，不只是因为它“看起来像伪造的”，而是具体因为其模态无法相互协调。

	变体	Acc	F1	AUC
	FakeSV			
	MAGIC^3 （完整）	**86.71**	**86.84**	92.23
一致性	无CMCG	84.90	84.85	91.77
	无CFE	85.69	85.61	92.30
	无TCMI	85.75	85.89	91.94
融合	无AARF	84.49	84.71	91.43
	无HMT	85.86	85.79	**92.48**
训练	无对比	85.75	85.89	92.15
	无对抗	86.56	86.53	92.23
	FakeTT			
	MAGIC^3 (完整)	**84.95**	**82.28**	**89.08**
一致性	无CMCG	81.94	78.68	87.44
	无CFE	78.83	80.13	87.88
	无TCMI	78.64	77.22	86.92
融合	无AARF	80.91	78.90	88.30
	无HMT	82.91	81.48	88.47
训练	无对比	82.93	80.47	88.50
	无对抗	78.37	77.78	88.37

表2：消融研究。一致性模块（CMCG、CFE、TCMI）和 AARF 至关重要。

确实是因为其模态无法相互协调。我们在附录D.2中表明，不确定性恰好在中等一致性区域达到峰值，这表明c_{global} 可作为筛选器，用于识别需要人工或视觉语言模型干预的模糊案例。

5.3 不确定性感知的两阶段视觉语言模型路由

现在，我们研究如何利用一致性和不确定性，通过重量级视觉语言模型检测器来协调两阶段流水线。图5 总结了两个数据集上的不确定性分布、校准曲线和路由行为。我们采用 FakeSV-VLM（Wang 等人，2025年c）作为第二阶段检测器。第一行显示，正确预测的熵通常低于错误预测，且置信度校准得较为合理。第二行则展示了用于识别难以处理、需要升级至视觉语言模型处理的样本的路由策略。我们将路由阈值视为超参数，并在验证集上进行调优，使两个数据集上约 25% 的样本进入路由。

路由样本确实更难（仅使用检测器时准确率较低），而两阶段系统的整体准确率有所提升。如表1所示，在 FakeSV 上，将 25% 的样本路由至视觉语言模型（MAGIC^3+VLM）可将整体准确率从 86.71% 提升至 90.93%，比 FakeSV-VLM 的 89.60% 高出 1.33 个百分点。在 FakeTT 上，25% 的路由比例可将准确率从 84.95% 提升至 89.52%，比仅使用视觉语言模型的基线 88.90% 高出 0.62 个百分点。重要的是，与纯视觉语言模型流水线相比，这种混合方法的吞吐量提高了 18–27× 倍，并大幅节省显存。MAGIC^3 有效过滤掉了大部分简单样本（∼75%），让成本高昂的视觉语言模型只需处理最难的样本，从而实现最佳的成本–性能权衡。

5.4 风格扰动下的一致性稳定性

最后，我们探讨多风格大语言模型改写的影响。对抗感知改写融合将原始说明文字与中性、正式和煽情风格的改写文本融合，并采用风格感知对比损失进行训练。表2 报告了一项消融实验：移除 AARF（即仅使用原始文本），同时保持其他组件不变。两个数据集上的性能均有所下降，表明接触风格变化能够提升鲁棒性。从概念上说，这与风格感知神经网络假新闻检测的相关研究相呼应，例如 Style-News（Wangetal., 2024a），该研究将风格化生成用作对抗手段，以对检测器进行压力测试。我们的设置有两点不同：我们将重写文本视为由通用大语言模型生成的语义保持型扰动（而非由特定任务生成器生成），并明确测量跨模态一致性在风格变化下的表现（统计细节见附录D.3）。我们的假设是，真实新闻在风格改写下仍能保持一致，而假新闻则表现出更大的方差，尤其是文本–视觉一致性方面。图7从定性角度支持了这一假设，其中伪造样本在风格扰动下呈现出更大的变化。附录 D.3 报告了详细的方差统计数据。

6 结论

我们提出了 MAGIC^3, , 一种以一致性为核心的假新闻短视频检测器，可显式呈现多个粒度上的跨三模态一致性：成对分数和全局分数、词元级一致性场、时序视听不一致性分数，以及不确定性估计。在 FakeSV 和 FakeTT 上，MAGIC^3 在冻结特征上取得了最先进的监督性能，并实现了更优的成本–性能权衡：与视觉语言模型路由结合时，其准确率超过仅视觉语言模型准确率，吞吐量则高出 18–27× 。未来工作应整合外部知识检索，以开展事实级验证。

局限性

首先，我们的方法依赖预先提取的多模态特征和由大语言模型离线生成的重写文本，这会带来额外的预处理成本，并可能对大语言模型的选择和提示设计较为敏感；然而，这是经过审慎考虑的设计选择，它将繁重的骨干网络计算与检测器头解耦，从而实现高吞吐量、低内存使用量，并能灵活集成不同的预训练编码器；我们的消融实验结果也进一步表明，即使不使用大语言模型改写，核心的一致性建模仍然有效。其次，当前的 TCMI 模块侧重于粗粒度的时序视听不一致性，并未显式建模细粒度的时序逻辑或因果结构；不过，该模块有意设计为静态一致性信号的轻量级补充，旨在检测重新利用或低成本篡改视频中常见的显著同步失配。第三，MAGIC^3 是一个特征级检测器，不直接处理原始视频输入；尽管这限制了其在设备端严格端到端部署的能力，但使该框架能够保持骨干网络无关性，并适用于大规模平台环境中的部署，因为在这些场景下，特征通常会被缓存或跨任务共享。此外，基于一致性的检测器在应对精心炮制的错误信息时仍可能遇到困难，因为这类信息的所有模态内部一致，却在事实层面不正确；解决此类问题通常需要外部知识和核验资源。我们认为，以一致性为核心的方法应作为一种互补的检测维度，而非独立的事实核查方案。最后，尽管所提出的一致性场和不确定性估计提升了模型透明度，并支持细粒度分析，但仍需开展更多以人为中心的评估，以充分了解从业者如何解读并依据这些信号采取行动；我们将此视为未来研究的重要方向。

伦理声明

本研究旨在减轻短视频新闻中多媒体错误信息造成的危害。然而，这也存在双重用途风险：了解检测机制可能为试图规避这些机制的对手提供信息。因此，我们侧重于通用架构原则，而非针对特定攻击的启发式方法，并鼓励采取负责任的部署实践，将自动检测与人工监督相结合。

所有使用的数据集均可公开获取；我们遵循其原始许可协议，并尊重现有的隐私保护措施。在现实场景部署中，应特别注意监测假阳性，确保不同主题和语言之间的公平性，并在自动化决策影响用户时提供适当的申诉渠道。

参考文献

Shahin Amiriparian、Filip Packa´n、Maurice Gerczuk 和 Björn W Schuller。2024 年。Exhubert：通过块扩展以及在 37 个情绪数据集上进行微调来增强 hubert。*arXiv* 预印本 *arXiv:2406.10275*。

Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、LianghaoDeng、Wei Ding、Chang Gao、Chunjiang Ge 等 1 人。2025年。Qwen3-vl 技术报告。*arXiv* 预印本 *arXiv:2511.21631*。

Yuyan Bu、Qiang Sheng、Juan Cao、Peng Qi、Danding Wang 和 Jintao Li。2023年。应对网络虚假信息视频：特征分析、检测与未来方向。载于第*31*届 *ACM* 多媒体国际会议论文集，第8770–8780页。

Yuyan Bu、Qiang Sheng、Juan Cao、Peng Qi、Danding Wang 和 Jintao Li。2024年。FakingRecipe：从创作过程的角度检测短视频平台上的假新闻。载于第*32*届 *ACM* 多媒体国际会议论文集，第1351–1360页。

Zhe Chen、Weiyun Wang、Yue Cao、Yangzhou Liu、Zhangwei Gao、Erfei Cui、Jinguo Zhu、Shenglong Ye、Hao Tian、Zhaoyang Liu 等。2024年。通过模型、数据和测试时扩展，拓展开源多模态模型的性能边界。*arXiv*预印本 *arXiv:2412.05271*。

Hyewon Choi 和 Youngjoong Ko。2021。使用主题建模和对抗神经网络进行假新闻视频检测。载于第*30*届*ACM*信息与知识管理国际会议论文集，第2950–2954页。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深度双向变换器预训练。载于*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第*1*卷（长篇和短篇论文），第4171–4186页。

Yuan Gao、Xuelong Wang、Yu Zhang、Ping Zeng 和 Yingjie Ma。2024年。音视频深度伪造检测中的时间特征预测。《Electronics》，13(17):3433。

Xuejian Huang、Tinghuai Ma、Hao Tang 和 Huan Rong。2025年。用于假新闻视频检测的知识增强型动态场景图注意力网络。《IEEE 多媒体汇刊》。

Xianghua Li、Jiao Qiao、Shu Yin、Lianwei Wu、Chao Gao、Zhen Wang 和 Xuelong Li。2025年。《多模态假新闻检测综述：跨模态交互视角》。*IEEE* 计算智能新兴主题汇刊。

Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。2021。Swin Transformer：使用移位窗口的分层视觉 Transformer。收录于*IEEE/CVF* 计算机视觉国际会议论文集，第 10012–10022 页。

Eric Müller-Budack、Jonas Theiner、Sebastian Diering、Maximilian Idahl 和 Ralph Ewerth。2020。利用跨模态实体一致性度量对现实世界新闻进行多模态分析。收录于*2020* 年国际多媒体检索会议论文集，第 16–25 页。

Britt Paris 和 Joan Donovan。2019。深度伪造与廉价伪造。美利坚合众国：*Data & Society*，1。

Peng Qi、Yuyan Bu、Juan Cao、Wei Ji、Ruihao Shui、Junbin Xiao、Danding Wang 和 Tat-Seng Chua。2023年。FakeSV：一种具有丰富社会语境的多模态基准，用于短视频平台上的假新闻检测。载于*AAAI* 人工智能大会论文集，第37卷，第14444–14452页。

Juan Carlos Medina Serrano、Orestis Papakyriakopoulos 和 Simon Hegelich。2020。利用基于自然语言处理的特征提取检测 YouTube 上的 COVID-19 错误信息视频。载于 *ACL 2020* 首届新冠疫情自然语言处理研讨会论文集。

Lanyu Shang、Ziyi Kou、Yang Zhang 和 DongWang。2021。一种用于检测 TikTok 上 COVID-19 短视频中错误信息的多模态错误信息检测器。载于*2021 IEEE* 大数据国际会议（大数据），第899–908页。IEEE。

Kai Shu、Suhang Wang 和 Huan Liu。2019年。《超越新闻内容：假新闻检测中的社会语境作用》。载于第十二届 *ACM* 国际网络搜索与数据挖掘会议论文集，第 312–320 页。

Chong Tian、Qirong Ho 和 Xiuying Chen。2025年。《用于不断演进的假新闻生成与检测的符号对抗学习框架》。载于*2025*年自然语言处理实证方法会议论文集，第 12307–12321 页。

Soroush Vosoughi、Deb Roy 和 Sinan Aral。2018。《在线真假新闻的传播》。《科学》，359(6380):1146–1151。

Haoran Wang、Yan Yang 和 Yingli Zhong。2025a。用于短视频假新闻检测的全局与局部特征增强。载于《智能计算国际会议》，第435–446页。Springer。

Junxi Wang、Jize Liu、Na Zhang 和 Yaxiong Wang。2025b。《短视频平台上的一致性感知虚假视频检测》。载于智能计算国际会议，第 200–210 页。Springer。

Junxi Wang、Yaxiong Wang、Lechao Cheng 和 Zhun Zhong。2025c。《FakeSV-VLM：通过渐进式专家混合适配器驾驭视觉语言模型，检测短视频虚假新闻》。*arXiv* 预印本 *arXiv:2508.19639*。

Wei-Yao Wang、Yu-Chieh Chang 和 Wen-Chih Peng。2024a。Style-news：融合风格化新闻生成与对抗验证的神经网络虚假新闻检测。*arXiv*预印本 *arXiv:2401.15509*。

Weiyun Wang、Zhe Chen、Wenhai Wang、Yue Cao、Yangzhou Liu、Zhangwei Gao、Jinguo Zhu、Xizhou Zhu、Lewei Lu、Yu Qiao 等。2024b。通过混合偏好优化增强多模态大语言模型的推理能力。*arXiv*预印本 *arXiv:2411.10442*。

Yaqing Wang、Fenglong Ma、Zhiwei Jin、Ye Yuan、Guangxu Xun、Kishlay Jha、Lu Su 和 Jing Gao。2018。EANN：用于多模态虚假新闻检测的事件对抗神经网络。载于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第849–857页。

Yukun Yang、Xiwei Shi、Haoxu Li、Buwei Fan 和 Yijia Xu。2025年。通过整合语义可信度与多粒度对比学习进行短视频假新闻检测。*AppliedSciences*，15(23):12621。

X Zhou、J Wu 和 R Zafarani。2020。SAFE：基于相似性感知的多模态假新闻检测。arXiv。*arXivpreprint arXiv:2003.04981*。

Xinyi Zhou 和 Reza Zafarani。2020。假新闻综述：基础理论、检测方法与机遇。*ACM*计算调查*(CSUR)*，53(5):1–40。

A 跨模态不一致性分类体系（详情）

表3 详细列出了引言中提到的面向任务的不一致性分类法。我们为每个类别提供了简要说明和典型模式；MAGIC[3]对应的信号将在正文中讨论。

B 补充图表

参见图 5和图 7。

C 方法详情

本节提供了第3节中概述的各组件的数学细节。

C.1 跨模态一致性门控（CMCG）

我们首先计算每种模态的池化表示：

$$\bar{\mathbf{h}}_{\text{text}} = \text{Pool}(\mathbf{H}_{\text{text}}), \quad \bar{\mathbf{h}}_{\text{vis}} = \text{Pool}(\mathbf{H}_{\text{vis}}), \\ \bar{\mathbf{h}}_{\text{aud}} = \text{Pool}(\mathbf{H}_{\text{aud}}). \tag{11}$$

此处，Pool(·)表示均值池化或特殊的CLS标记。

对于每个模态对（m_i、m_j）∈{（文本、视觉）、（文本、音频）、（视觉、音频）}，CMCG计算标量一致性分数

$$c_{ij} = \sigma\Big(\mathbf{w}_{ij}^{\top}[\bar{\mathbf{h}}_{m_i}; \bar{\mathbf{h}}_{m_j}] + b_{ij}\Big), \tag{12}$$

其中，$[\cdot;\ \cdot]$ 表示拼接，σ 表示Sigmoid函数。由此得到一个三维一致性向量

$$\mathbf{c} = [c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}] \in \mathbb{R}^3. \tag{13}$$

随后，一个小型多层感知机生成全局一致性分数

$$c_{\text{global}} = \sigma(\text{MLP}_c(\mathbf{c})). \tag{14}$$

C.2 一致性场估计器（CFE）

假设我们有一个从模态 m_i 到m_j的交叉注意力矩阵，

$$\mathbf{A}_{m_i \leftarrow m_j} \in \mathbb{R}^{L_{m_i} \times L_{m_j}}, \tag{15}$$

由跨模态 Transformer 生成。第 t 行对应模态 m_i 中的一个词元或帧。我们将 m_i 相对于 m_j 的一致性场定义为

$$F_{ij}^{(m_i)}(t) = \max_k \mathbf{A}_{m_i \leftarrow m_j}(t,k), \quad t = 1, \ldots, L_{m_i}. \tag{16}$$

直观地说，$F^{(m_i)}{}_{ij}(t) \in [0,1]$ 反映了模态 m_i 中的词元 t 在多大程度上能在模态 m_j 中找到支持性证据。在训练过程中，我们加入一个轻微的正则化项，使全局 CMCG 分数与聚合后的一致性场保持一致：

$$\mathcal{L}_{\text{reg}} = \sum_{(i,j)} \big(c_{ij} - \text{mean}(F_{ij}^{(m_i)})\big)^2. \tag{17}$$

C.3 多视图对抗感知改写融合（AARF）

设$\mathbf{H}_{\text{text}}$ 为原始文本序列，$\mathbf{H}_{\text{rew}}^{(v)}$为第 v个改写。我们将它们池化为向量

$$\mathbf{h}_{\text{orig}} = \text{Pool}(\mathbf{H}_{\text{text}}), \tag{18}$$

$$\mathbf{h}_{\text{rew}}^{(v)} = \text{Pool}\big(\mathbf{H}_{\text{rew}}^{(v)}\big). \tag{19}$$

对于每个视图 v，我们估计一个标量改写质量

$$q^{(v)} = \sigma\big(\text{MLP}_q[\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(v)}]\big), \tag{20}$$

并计算门控 logit

$$\boldsymbol{g} = \text{MLP}_g\big([\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(1)}; \ldots; \mathbf{h}_{\text{rew}}^{(V)}; q^{(1)}, \ldots, q^{(V)}]\big). \tag{21}$$

我们得到初步权重$\boldsymbol{\alpha}^{\text{soft}}$ =softmax($\boldsymbol{g}$)，对原始文本施加最小权重 $\alpha_{\min}$，并进行重新归一化：

$$\tilde{\alpha}_0 = \max(\alpha_0^{\text{soft}}, \alpha_{\min}), \tag{22}$$

$$\tilde{\alpha}_v = \alpha_v^{\text{soft}}, \quad v \geq 1, \tag{23}$$

$$\alpha_v = \begin{cases} \tilde{\alpha}_0, & v = 0, \\ (1 - \tilde{\alpha}_0)\, \dfrac{\tilde{\alpha}_v}{\sum_{u=1}^{V} \tilde{\alpha}_u}, & v \geq 1. \end{cases} \tag{24}$$

融合表示为

$$\mathbf{h}_{\text{text}}^{\text{fuse}} = \alpha_0\, \mathbf{h}_{\text{orig}} + \sum_{v=1}^{V} \alpha_v\, \text{Proj}\big(\mathbf{h}_{\text{rew}}^{(v)}\big). \tag{25}$$

InfoNCE 风格损失促使同一视频的所有视图彼此接近，并使不同视频的视图彼此分离。

C.4 分层多模态Transformer（HMT）

A 层执行模态内细化：

$$\mathbf{H}_{\text{text}}^{(A)} = \text{TransEnc}_{\text{text}}(\mathbf{H}_{\text{text}}), \tag{26}$$

$$\mathbf{H}_{\text{vis}}^{(A)} = \text{TransEnc}_{\text{vis}}(\mathbf{H}_{\text{vis}}), \tag{27}$$

$$\mathbf{H}_{\text{aud}}^{(A)} = \text{TransEnc}_{\text{aud}}(\mathbf{H}_{\text{aud}}). \tag{28}$$

类别	描述及典型模式
实体与语境不一致	文本中提及的实体（人物、组织、地点）或语境（例如国家、城市、场所）与视频或音频中呈现的内容不一致。典型模式是字幕声称“数千人在 Paris 抗议”而视频展示的是另一座城市中一场规模较小的集会。
事件层面的语义不一致性	不同模态对核心事件存在分歧：发生了什么、规模如何，或有哪些参与者。例如，说明文字声称“爆发暴力冲突”，但视频展示的是一场和平集会；又或者音频描述的是另一起事件。
情感不一致	音频传达的情绪或情感与文本或视觉内容相矛盾，例如叠加在欢快背景音乐上的悲伤字幕，或高度戏剧化的视觉内容上平静的旁白。
时序视听不一致性	音频流与视觉流在时间上错位、取自不同素材，或经过重新排序。这包括配音视频、重新剪辑的素材或时序错位的音效。
风格/语用不一致性	一种模态中的修辞风格、框架或煽情程度与其他模态不一致，例如，相对平淡的视频配上煽情的字幕，或情绪强烈的文本搭配中性音频。

表3：短视频新闻中不一致现象的面向任务的分类体系。这些类别为 MAGIC[3] 的一致性中心架构提供了设计目标。

B层执行一致性加权交叉模态注意力；对于模态 m_i，m_j 我们计算

$$\tilde{\mathbf{H}}_{m_i \leftarrow m_j} = \mathrm{CrossAttn}\big(\mathbf{H}^{(A)}_{m_i}, \mathbf{H}^{(A)}_{m_j}\big). \tag{29}$$

为 CFE 保存注意力权重。随后，我们使用由 $\mathbf{c}$ 得出的权重聚合消息：

$$\boldsymbol{\beta}_m = \mathrm{softmax}\big(\mathrm{MLP}_m(\mathbf{c})\big), \tag{30}$$

$$\mathbf{H}^{(B)}_m = \sum_{j \neq m} \beta_{mj}\, \tilde{\mathbf{H}}_{m \leftarrow j}. \tag{31}$$

LayerCpools 对表示进行细化，并将其与 AARF 文本融合：

$$\mathbf{Z} = \begin{bmatrix} \mathrm{Pool}(\mathbf{H}^{(B)}_{\mathrm{text}}) \\ \mathrm{Pool}(\mathbf{H}^{(B)}_{\mathrm{vis}}) \\ \mathrm{Pool}(\mathbf{H}^{(B)}_{\mathrm{aud}}) \\ \mathbf{h}^{\mathrm{fuse}}_{\mathrm{text}} \end{bmatrix}, \tag{32}$$

$$\mathbf{h}_{\mathrm{global}} = \mathrm{TransEnc}_{\mathrm{global}}(\mathbf{Z})_{[\mathrm{CLS}]}. \tag{33}$$

C.5 对比-对抗联合学习（CAJL）

设 $y_i \in \{0,1\}$ 为标签，$\hat{}y_i$ 为预测概率。交叉熵损失为

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{N}\sum_i \Big(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i) \Big). \tag{34}$$

我们使用 InfoNCE 作为对比项。对于来自同一视频的表示对（$\mathbf{u}_i$,$\mathbf{v}_i$）以及来自其他视频的负样本，

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{N}\sum_i \log \frac{\exp(\mathrm{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(\mathrm{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau)}. \tag{35}$$

$\mathcal{L}_{\mathrm{intra}}$ 使用同一视频的不同视图（原始文本、重写文本、音频、视觉内容），而 $\mathcal{L}_{\mathrm{cross}}$ 则对齐融合文本与全局表示。

为实现对抗一致性，我们对中间表示施加微小扰动 δ_i，并强制保持预测一致性：

$$\mathcal{L}_{\mathrm{adv}} = \frac{1}{N}\sum_i \mathrm{KL}\big(p_\theta(y \mid \mathbf{h}_i) \| p_\theta(y \mid \mathbf{h}_i + \delta_i)\big). \tag{36}$$

我们还鼓励原始文本与融合文本之间的预测保持一致：

$$\mathcal{L}_{\mathrm{sem}} = \frac{1}{N}\sum_i \mathrm{KL}\big(p_\theta(y \mid \mathbf{h}_{\mathrm{orig},i}) \| p_\theta(y \mid \mathbf{h}^{\mathrm{fuse}}_{\mathrm{text},i})\big). \tag{37}$$

风格不变性项对所有文本视图的投影应用 InfoNCE：

$$\mathcal{L}_{\mathrm{style}} = \frac{1}{N}\sum_i \mathcal{L}_{\mathrm{NCE}}\big(\{\mathbf{z}^{(v)}_{\mathrm{text},i}\}_{v=0}^{V}\big). \tag{38}$$

最终损失为

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{intra}}\mathcal{L}_{\mathrm{intra}} + \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{cross}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{style}}\mathcal{L}_{\mathrm{style}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}. \tag{39}$$

超参数见表11。

D 发现的补充定量分析

本附录报告了支持发现1–4的定量统计数据。

D.1 聚合一致性统计数据

表5 列出了真实与伪造视频在四种一致性得分上的均值和 ± 标准差，

组件	功能	一致性级别	在研究发现与可解释性中的作用
一致性计算			
CMCG	计算成对一致性分数 $[c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}]$ 和全局分数 c_{global}	成对 → 全局	揭示了不对称的文本–视觉与文本–音频模式；定义了全局一致性轴（图3–4）.
CFE	从跨模态注意力中推导词元/帧级一致模态间一致性字段自交叉模态 attention	局部（词元/帧）	定位可疑短语或帧；支持热力图可视化（图6）。
TCMI	计算时序视听不一致性分数 c_{temp}	时间	检测粗略的视听同步时间同步不匹配很常见在再利用视频中。
跨模态融合			
AARF	将原始文本与中性／正式／煽情 LLM 重写文本通过门控权重进行融合	文本内部	构建具备风格鲁棒性的文本嵌入——用于研究风格扰动——的影响（发现4）。
HMT	三阶段 层次化 fusion：模态内细化，一致性加权 交叉-注意力，全局聚合	结构	融合多模态表征与一致性调制的注意力。
训练与预测			
CAJL	对比–对抗学习 结合 InfoNCE 联合 and 扰动损失	训练	对齐多模态表示并提升在每次扰动下的鲁棒性扰动。
分类器	输出虚假概率、置信度，以及不确定性，用于两阶段路由	预测	为……提供校准不确定性选择性视觉语言模型升级处理提供校准不确定性（发现3）。

表4：MAGIC³ 组件概览。组件按功能分组：一致性计算（CMCG、CFE、TCMI）、跨模态融合（AARF、HMT）以及训练/预测（CAJL、分类器）。这些组件共同提供了用于四项发现和两阶段视觉语言模型流程的信号。

以及双样本 t-检验。两个数据集上的文本–视觉与文本–音频不对称性都高度显著；不同类别之间的视觉–音频几乎完全相同。

数据集	得分	Real	Fake	p-数值
FakeSV	c_{tv}	0.82 ± 0.32	0.16 ± 0.27	$p \ll 0.001$
	c_{ta}	0.50 ± 0.20	0.88 ± 0.13	$p \ll 0.001$
	c_{va}	0.89 ± 0.02	0.89 ± 0.02	$p = 0.82$
	c_{global}	0.47 ± 0.04	0.38 ± 0.04	$p \ll 0.001$
FakeTT	c_{tv}	0.84 ± 0.29	0.31 ± 0.35	$p \ll 0.001$
	c_{ta}	0.39 ± 0.23	0.77 ± 0.18	$p \ll 0.001$
	c_{va}	0.86 ± 0.03	0.86 ± 0.03	$p = 0.49$
	c_{global}	0.47 ± 0.04	0.40 ± 0.04	$p \ll 0.001$

表5：发现1的统计数据：按类别划分的一致性得分（均值±标准差）及双样本 t-检验的 p值。在两个数据集中，真实视频与伪造视频的文本–视觉模式和文本–音频模式呈现相反趋势。

D.2 全局一致性与预测之间的相关性

表6 报告了c_{global} 与预测的伪造概率之间的斯皮尔曼/皮尔逊相关，以及线性拟合 R^2。表7 展示了五个c_{global} 分箱中的经验伪造率，体现了总体下降趋势。

数据集	斯皮尔曼相关系数。ρ	皮尔逊相关系数。r	R^2
FakeSV	-0.91（$p \ll 0.001$）	-0.99（$p \ll 0.001$）	0.99
FakeTT	-0.96($p \ll 0.001$)	-0.96($p \ll 0.001$)	0.91

表6：发现2： c_{global} 与预测的伪造概率之间的相关性。

c_{global} 个区间	FakeSV率	FakeTT率
最低（Q1）	0.93	0.83
Q2	0.89	0.55
Q3	0.76	0.17
Q4	0.19	0.03
最高（Q5）	0.05	0.07

表7：发现2：c_{global} 的五个分位数区间中的伪造率（分数越低， =越不一致）。

D.3 风格敏感性统计

表8 比较了真实视频与虚假视频的风格引起的方差（在原始文本和+三种改写文本之间）。在两个数据集上，伪造样本的文本–视觉、文本–音频和全局一致性方差均始终较高；视觉–音频方差则如预期般没有变化。

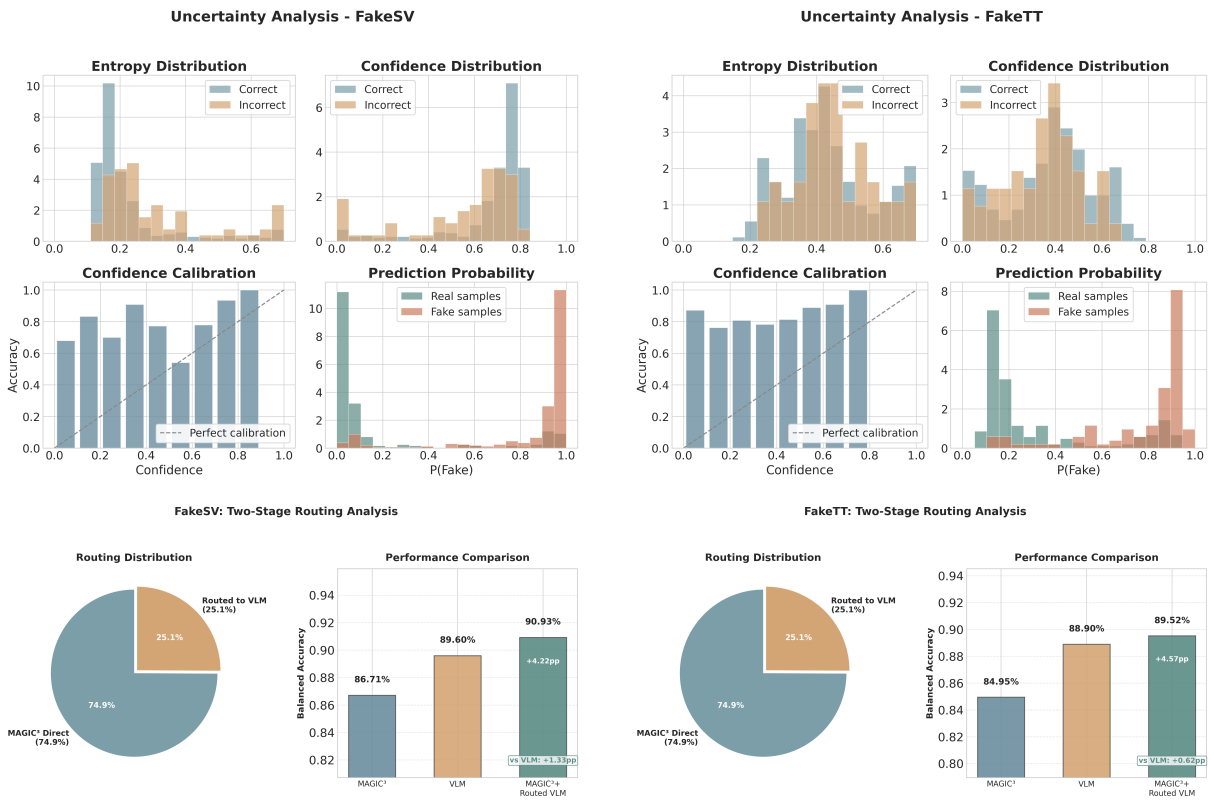


图5：MAGIC[3] 的不确定性与两阶段路由行为。上：熵/置信度分布及校准图。下：直接预测与路由样本的划分，以及使用基于视觉语言模型的第二阶段检测器时，路由与未路由子集的准确率。左：FakeSV；右：FakeTT。

数据集	得分	真实样本方差	伪造样本方差	p
FakeSV	c_{tv}	0.0336	0.0815	$p < 10^{-16}$
	c_{ta}	0.0109	0.0148	$p < 10^{-2}$
	c_{va}	0.0	0.0	n.s.
	c_{global}	6.19×10^{-4}	1.44×10^{-3}	$p < 10^{-15}$
FakeTT	c_{tv}	0.0194	0.0801	$p < 10^{-21}$
	c_{ta}	0.0081	0.0170	$p < 10^{-7}$
	c_{va}	0.0	0.0	n.s.
	c_{global}	3.13×10^{-4}	1.23×10^{-3}	$p < 10^{-20}$

表8：发现4：风格改写下（真实与伪造）一致性得分的方差。“n.s.”表示差异不显著（$p \geq 0.05$）。

D.4 两阶段路由分析详情

我们将路由阈值视为超参数，并在验证集上进行调优，使两个数据集上约 25% 的样本进入路由。表9 总结了所使用的路由策略。对于 FakeSV，我们采用基于熵的路由（熵高表示预测不确定）；对于 FakeTT，我们使用难度分数，该分数结合了熵、全局一致性和置信度：$d = \text{熵}_{\text{norm}} + (1 - c_{\text{global, norm}}) + (1 - \text{置信度}_{\text{norm}})$，并将难度分数最高的样本进行路由。

数据集	路由比例	MAGIC[3]+视觉语言模型准确率	路由策略
FakeSV	25.1%	90.93%	基于熵
FakeTT	25.1%	89.52%	难度分数

表9：发现3：用于构建两阶段系统的路由策略和路由比例。两种策略的表现均优于单独使用的视觉语言模型检测器（FakeSV：+1.33pp，FakeTT：+0.62pp）。

E 数据集与划分

表 10 总结了本文使用的两个基准数据集。我们遵循 Huang 等人 (2025年) 提供的官方时间顺序划分，将 70%/15%/15% 分别用于训练、验证和测试。

统计数据	FakeSV	FakeTT
时间范围	2017/10–2022/02	2019/05–2024/03
平均时长（秒）	39.88	47.69
# 虚假视频	1,810	1,172
# 真实视频	1,814	819
# 总计	3,624	1,991

表 10：FakeSV 和 FakeTT 的统计数据。所有视频均为现实世界平台上的短视频片段。

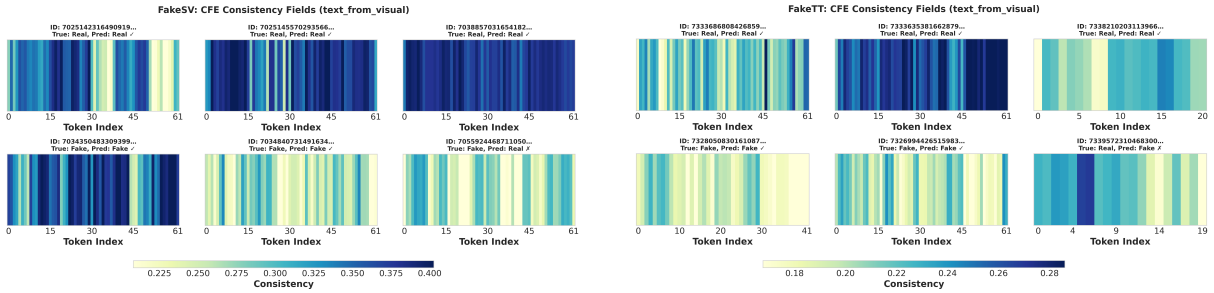


图6：CFE一致性场。颜色表示与视觉流的逐词元一致性；高分区域清晰且集中，表明局部内容得到支持，而分数分散或偏低则表明可能存在不一致。

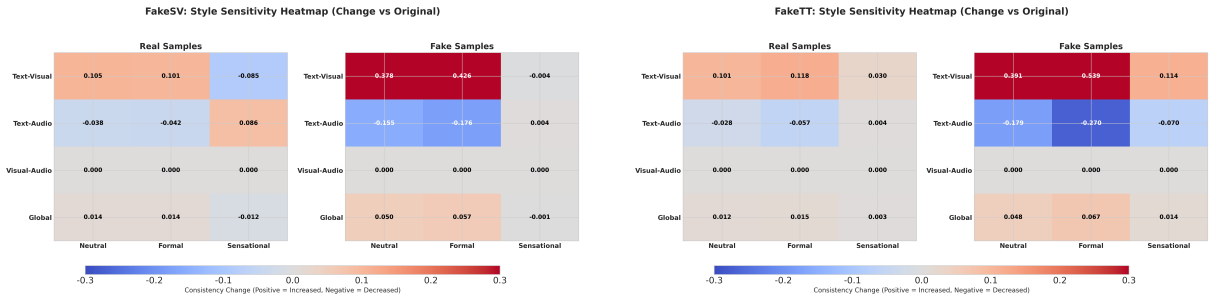


图7：风格敏感性热图（发现4）。行：一致性类型；列：风格改写与原始文本。伪造样本的变化幅度大于真实样本。

F 大语言模型改写提示

我们使用三条提示词来生成多视角改写文本。在实验中，我们使用 DeepSeek-V3.2；对于 FakeSV，我们使用这些提示词的中文翻译。

中性新闻风格。

> 你是一名专业的新闻编辑。请以中性、客观的新闻风格改写以下视频描述。请注重表达清晰和事实准确。不要添加新信息。原始文本：{input_text} 改写后的文本：

正式详尽风格。

> 您是一位资深记者。请以更正式、详尽的风格改写以下视频描述，明确呈现关键事实，并使结构清晰。避免使用煽情语言。原始文本：{input_文本}改写后的文本：

煽情风格。

> 你正在撰写一篇关于突发新闻的煽情社交媒体帖子。请改写以下视频描述，使其更具情感张力、更能吸引注意力，同时保持核心事实不变。原始文本：{input_text} 改写文本：

G 实现细节与超参数

骨干编码器使用 `bert-base-multilingual-uncased` 处理文本（最大长度为 256），使用 Swin Transformer（Liu 等人，2021）处理 K 个视频关键帧，并使用基于 ExHuBERT的情感嵌入处理音频（Amiriparian 等人，2024年）。这些编码器仅用于离线特征提取；在 MAGIC^3 训练期间，我们只使用预提取特征，不更新编码器参数。

训练使用 AdamW，学习率为 5×10^{-5}，批次大小为 32，采用 10% 预热的余弦衰减策略，Transformer 层的丢弃率为 0.3，并根据验证集宏F1进行早停。训练轮数最多为 50 轮，并保留验证集宏F1最佳的检查点。在主要实验中，我们设置 $\lambda_{\text{intra}} = \lambda_{\text{cross}} = 0.1$、$\lambda_{\text{adv}} = 0.5$、$\lambda_{\text{sem}} = 0.1$、$\lambda_{\text{style}} = 0.1$和 $\lambda_{\text{reg}} = 0.05$，以及 $\alpha_{\min} = 0.5$、$V = 3$和 $K_{\text{MC}} = 5$。

H 效率与部署讨论

由于所有编码器仅用于离线特征提取，且 MAGIC^3 直接处理预提取特征，因此对缓存特征进行推理时只需使用一个检测器头。可训练的

超参数	数值
学习率	5×10^{-5}
批次大小	32
训练轮数	50
预热比例	0.1
隐藏维度	256
注意力头（HMT）	8
丢弃率	0.3
λ_{intra}	0.1
$\lambda_{\text{交叉}}$	0.1
λ_{adv}	0.5
λ_{sem}	0.1
λ_{style}	0.1
λ_{reg}	0.05
$\alpha_{\min}$ (AARF)	0.5
V (每个视频的重写文本)	3

表11：MAGIC^{3}的关键超参数设置。

检测器头约有 185M 个参数（无需进行编码器微调），与拥有数十亿参数的视觉语言模型骨干网络相比十分轻量。在单张 RTX 5090 GPU 上，每个数据集的完整训练流程（包括验证和测试评估）约需 15 分钟。在同一 GPU 上，推理批次大小为 128 时，我们测得吞吐量约为 680 个样本/秒，每个样本耗时 $\sim$1.5 毫秒；处理缓存特征时，GPU 显存峰值为 $\sim$1.67 GB。同一 GPU 上训练期间的 GPU 显存峰值使用量约为 11GB。在端到端延迟中，视频解码和特征提取占主要部分；在实际系统中，可通过缓存和异步预处理来摊销这部分开销。

相比之下，基于视觉语言模型的检测器通常需要使用大型视觉—语言骨干网络进行端到端处理，因此训练成本更高（通常需要多GPU配置），大规模部署的成本也更高。MAGIC^3 仅在特征层级上运行，因此可以针对短视频虚假信息领域的新主题或细分领域快速重新训练，使其成为受硬件限制、需要快速适应的平台的实用选择。