arXiv:2302.07919v1 [cs.CV] 15 Feb 2023

COVID-VTS：短视频平台上的事实提取与验证

Fuxiao Liu **Yaser Yacoob** **Abhinav Shrivastava**

马里兰大学帕克分校 {fl3es, yaser, abhinav}@umd.edu

摘要

我们推出一个新的基准数据集 COVID-VTS，用于核查涉及短时长视频的多模态信息，其中包含来自现实世界和机器生成、聚焦于新冠肺炎的信息。我们提出了 TwtrDetective，这是一种有效模型，融入跨媒体一致性检查，以检测不同模态中的词元级恶意篡改，并生成解释。由于训练数据稀缺，我们还开发了一种高效且可扩展的方法，通过事件操纵或对抗性匹配自动生成误导性视频帖子。我们研究了多种最先进的模型，并证明了 TwtrDetective 的优越性。

1 引言

社交媒体上虚假信息的泛滥对社会构成了严重威胁，尤其是在 COVID-19 疫情期间。因此，有必要开发自动事实核查工具，以核实在线传播的主张。

近期的事实核查工作（Thorne 等人，2018；Wadden 等人，2020；Augenstein 等人，2019；Fung 等人，2021；Bekoulis 等人，2021）通过开发各种基准数据集以及最先进的神经网络架构，研究了自动错误信息检测，涉及维基百科页面、表格、新闻文章和科学文章等来源。然而，由于缺乏公开可用的基准数据集，短视频平台（如 TikTok、Twitter 和 Instagram）上的事实核查仍然面临挑战。

核实短视频平台上主张的事实正确性带来了新的挑战。首先，导致短视频帖文出现的多模态错误信息，比仅使用文本内容更具误导性（Micallef 等人，2022），因为这些主张通常只会篡改源视频中事实信息的细微元素。此外，包含Shang 等人，2021）的短视频多样的场景转换、人类活动和跨模态信息，大大增加了视频内容的复杂性和模糊性。此外，当前方法通过单模态分析探究 GAN 生成视频的真实性（Güera 和 Delp，2018）。然而，这些方法无法直接应用于短视频平台，因为平台上的视频内容往往会遭到蓄意篡改。

为应对这些挑战，我们提出了COVID-VTS，这是一个新的基准数据集，包含可信的主张及其对应的高质量视频，收集自Twitter视频帖子。COVID-VTS 包含不同的不一致分类体系，这种不一致来自不同的模态。示例见图1。此外，我们提出了一种新方法，通过操纵事件元素或进行对抗匹配，自动生成虚假视频帖子（Luoet al.，2021a）。我们的事件操纵工具的优势在于，只需编辑事实信息中的一个小组件，就能控制语义极化，并追踪操纵对象。我们实施质量控制，删除不合格的生成结果，并解决语言偏差问题。

我们提出了TwtrDetective，一种多模态事实核查模型，其输入由一条主张及其配对视频组成，目标是预测二者的一致性。TwtrDetective利用*Event Alert* 模块，从主张中精确提取细粒度的事实细节，同时利用*Pairwise Consistency Aggregation* 模块，有效衡量各模态之间的一致性。TwtrDetective 还能够指出不一致的模态（例如，见图3）。

实验结果表明，TwtrDetective 在两个数据集上的检测准确率均高于基线方法。我们的主要贡献总结如下：

• 我们推出 COVID-VTS，这是一个面向短视频平台的事实核查数据集，包含 1 万组视频文本对，涵盖多样化场景，且更易于获取



图1：COVID-VTS 中不同不一致分类体系的示例，由我们的自动操作工具生成。红框标示了不一致模态。我们的任务是独立判断视频-文本对是否一致，并指出哪种模态是伪造的。(1) 主张中的事件论元被修改；(2) 主张中的事件触发词遭到篡改；(3) 短视频通过对抗匹配进行筛选。

多种模态，以及来自现实世界和机器生成的可信声明。

• 我们提出了一种有效的方法，可自动生成大规模、可验证且可信的内容，以及误导性视频帖文，而不必依赖人工标注员。

• 我们提出了 TwtrDetective，这是一个面向短视频平台的新型可解释事实核查框架。在两个具有挑战性的数据集上，其表现均优于基线方法。

我们的代码已公开发布，网址为 https://github.com/FuxiaoLiu/Twitter-Video-dataset。

2 相关工作

虚假信息的传播推动了自动事实核查领域的研究不断发展。研究者推出了许多从维基百科和事实核查网站收集的大规模数据集，包括 FEVER（Thorne 等人，2018）、SciFact（Wadden 等人，2020）、UKP Snopes（Hanselowski 等人，2019）、MultiFC（Augenstein 等人，2019）。然而，这些数据集中的虚假主张均由人类手动生成，因此难以大规模部署。近期，研究者提出了一些合成数据集来解决这一局限。（Jiang 等人，2020）通过词语替换生成复杂的虚假主张。（Saakyan 等人，2021）利用掩码语言模型的词元填充能力替换显著词元。此外，（Fung 等人，2021）提出了一种新颖的数据合成方法，通过操纵多模态知识图谱中的知识元素来生成数据。为减轻机器生成主张中的语言学偏差，（Luo 等人，2021a）通过检索语义相似的现实世界句子，构造脱离上下文的字幕。与此类似，我们的数据集（Liu 等人，2020b）构建了用于视频-语言推理的 VIOLIN，而其中所有陈述均由专家撰写。（Wang 等人，2022）从 Twitter 收集了社交媒体视频帖子，但其中的虚假主张是通过随机交换构造的。相比之下，COVID-VTS 是首个面向短视频平台的新冠肺炎事实核查数据集，包含丰富的信息和多样的场景，涵盖了更多易于获取的模态，以及来自现实世界和机器生成的误导性声明。

传统事实核查方法（Zellers 等人，2019；Schuster 等人，2020；Atanasova 等人，2019）主要基于文本。如果证据来自视觉信息，这些方法就难以奏效。近期的多模态模型（Fung 等人，2021；Tan 等人，2020；Shang 等人，2021）能够根据不同模态所传达的信息检查一致性。相比之下，我们提出了一种面向短视频平台的事实核查模型，并提供多媒体解释，能够更准确地检测虚假信息。

3 COVID-VTS数据集构建

COVID-VTS 包含 10k 条格式规范的主张，并配有支持或反驳这些主张的视频。本节首先介绍如何从 Twitter 视频帖子中筛选出可信且一致的视频/主张对。其次，我们还介绍了自动构建格式规范的不一致视频—主张对的方法。

3.1 筛选真实且一致的视频帖子

为构建 COVID-VTS 数据集，我们使用 Twitter爬虫，借助 COVID相关关键词和话题标签收集了超过 10 万条英文视频帖文

时间范围为 2019年10月底至 2022年8月。示例见表 1。为减少潜在偏差，对于同一个视频链接，我们只保留一条帖子。为提取视频中的语音，我们使用 IBM Cloud 的语音转文本工具（Pitrelli 等人，2006）。此外，我们使用 SimpleOCR（Ko 和 Kim，2004）识别屏幕上的文字。初步筛选后，我们通过以下几个步骤筛选真实且内容一致的视频帖文：

已验证用户账号。根据 Twitter 的要求，只有真实、知名且活跃的账号才会获得蓝色认证徽章（包括政府官方账号、公司官方账号等）。经常发布骚扰、羞辱他人内容，或严重违反平台操纵与垃圾信息政策的账号，不符合获得蓝色徽章的资格。我们将已验证账号发布的真实帖子收录到 COVID-VTS 中，以提高数据质量。

主张必须可验证。事件结构对于揭示句子的事实信息至关重要，因为一旦事件元素发生变化，整体语义就可能截然相反。此外，COVID-VTS 中的主张应当是可验证的命题，其真实性由配对视频中的多模态证据决定。缺少事件结构，就无法验证其真实性。因此，我们删除包含个人观点或情绪的实例。例如，“为我的孩子们感到骄傲！*#GetVaccinated*”，其中没有可供验证的事实信息。

在实践中，我们首先利用在MECHANIC数据集上预训练的先进事件抽取框架DYGIE++(Wadden等人，2019)，移除不包含事件结构的主张。DYGIE生成的事件结构 ++ 包括事件触发词和事件论元。例如，在“官员们正争分夺秒地遏制中国以外地区的冠状病毒暴发”中，“遏制”是表达事件发生的触发词，“官员”和“冠状病毒暴发”则是在该事件中扮演不同角色的事件论元。

选出这些可验证的主张后，我们会人工检查它们是否与相关视频一致。我们还从未经核实的账号中选取可验证且一致的视频文本对，以增加数据集的规模。我们的假设是，经过“清理”步骤后，视频帖文是可信的。基于这一假设，主要任务是分析不同模态之间的不一致性。

关键词/话题标签
新冠肺炎，#新冠肺炎，*#COVID-19*，冠状病毒，#冠状病毒，戴口罩，大流行，重症监护室，疫苗，#疫苗，冠状病毒，隔离，#隔离，莫德纳，#新冠肺炎变异株，奥密克戎，加强针，信使*RNA*，辉瑞，#辉瑞，#世界卫生组织，#居家办公，#疫苗护照, #旅行禁令, #社交_距离, *N95*口罩

表1：数据收集过程中使用的新冠肺炎相关关键词和话题标签摘要。

	最常见的词元
事件触发词（原始主张）	抗争、感染、保护、预防，帮助、停止、隔离、遏制，确认、威胁、暴跌、人——*date, deal,* 原因
事件论元（原始主张）	冠状病毒、奥密克戎、辉瑞，新冠肺炎，德尔塔，莫德纳，加强针，口罩，封锁，抗议，社交距离 *, ban,* 疫苗
事件触发词（生成的主张）	产生、增加、创造、减少*velop*、移除、保护、避免，识别、支持、生成、规则，允许，建立
事件论元(生成的主张)	奥密克戎，变异株，死亡，*cancer*，医院、世界卫生组织、社区，医疗服务、药物、病毒，迁移、媒体、重症监护室

表2：原始主张和生成的主张中最常见的事件触发词和论元。

3.2 自动生成不一致的视频-主张对

本节将介绍自动生成不一致视频-主张对的步骤。我们的数据集包含以下三种不一致分类体系：（1）真实视频、真实语音和不一致主张；（2）真实视频、不一致语音和真实主张；（3）不一致视频、真实语音和真实主张。

不一致主张生成。受（Liu 等人，2020b）的启发，我们通过控制语义极化，并修改真实主张中少量事实信息来追踪操纵对象，从而自动生成虚假主张。在这种情况下，大部分陈述仍与视频内容相符。这一策略还有助于缓解（Fung 等人，2021；Zellers 等人，2019）所引入的语言偏差。在他们的数据集中，模型无需比较不同模态就能正确检测出虚假主张。这是因为所有虚假主张均由语言模型生成。另一项优势是，我们可以自动生成大规模训练集，而不必雇用人工标注员。

原始真实声明	生成的虚假声明
官员们正竭力控制冠状病毒疫情 在中国境外。	官员们正竭力控制 埃博拉的传播 疫情在中国境外。
莫德纳董事长接种加强针是阻止病毒的唯一方法。	莫德纳董事长感染病毒是阻止病毒的唯一方法病毒。
美联储主席鲍威尔警告称，奥密克戎变异株可能会冲击经济复苏。	美联储主席鲍威尔警告称，奥密克戎变异株可能助长经济复苏。
受中国危机波及的澳大利亚人终于回来了家在圣诞岛隔离*14*天后。	受中国危机波及的澳大利亚人失去了家园在在圣诞岛隔离*14*天。

表3：详细展示了我们高效自动化方法生成的示例。前两个示例采用事件论元操纵，最后两个采用事件触发词操纵。

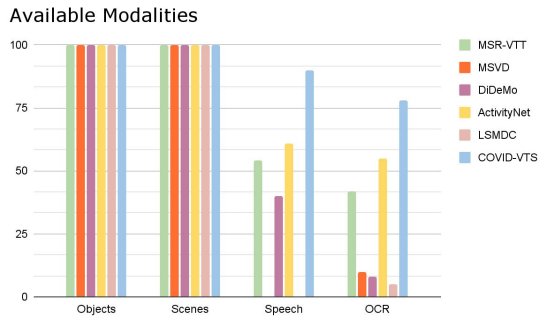


图2：直方图表明，与以往的数据集相比，COVID-VTS 更便于获取语音内容和屏幕文字。

只有事件触发词或事件论元会被选作 [口罩] 操纵词元，具体取决于不同的意图。我们遵循（Nguyen 等人，2020）中的流程，并使用 BERTWeet——一种基于与新冠肺炎疫情相关的推文预训练的模型——来预测可用于遮蔽事件元素的领域感知替代项。为避免引入额外噪声，如果候选项中的替换词元不在原始数据集的词汇表中，我们便将其删除，以减少潜在偏差。随后，我们将候选项及其对应的真实主张输入到基于 Mutli-NLI 数据集训练的 CrossEncoder 模型中（Williams 等人，2017），并选出带有矛盾标签的候选项。如果候选项对的矛盾分数最高，就将其指定为虚假主张。表3 展示了我们高效自动化方法生成的示例。前两个示例是事件论元操纵，后两个则是事件触发词操纵。

表 2 展示了原始主张和生成的主张中最常见的事件元素。实际上，我们发现这些高频替代项带来了额外的语言偏差。模型只需检测这些高频词元，就能将主张分类为虚假主张，而无需从其他模态中获取信息。

	COVID-VTS	VIOLIN
平均字幕长度	19.2	18.0
平均语音长度	69.2	76.4
平均视频长度	26.5秒	35.2秒
命名实体（句子）	90.8%	10.3%
来源	Twitter	电视节目

表 4：COVID-VTS 数据集概述。

因此，我们通过为每个替代项仅保留一个主张来减轻这种偏差。然后，从数据集中选取最相似的主张来替换不合格的主张。以这种方式构建的不一致配对更侧重于对配对进行全局理解。这种严格的设置使分析模型更难区分这些主张，因此需要进行深入推理才能识别虚假内容。

不一致语音生成。不同于（Tan 等人，2020）和（Fung 等人，2021）对文章的多个部分进行编辑，我们的目标仅是修改语音中的证据句，以减少语言偏差。首先，我们利用 SBERT 句子嵌入的余弦相似度，提取与真实主张最相似的句子。之后，我们修改证据句中的命名实体或事件元素。

不一致视频生成。对于第三种不一致对，我们保留原始语音和主张不变。不过，我们使用对抗匹配（Liu 等人，2020b）方法，将原视频替换为另一个与当前视频相似的真实视频。具体而言，我们使用预训练的 VideoCLIP（Xu 等人，2021）生成视频表示，以计算余弦相似度。

缺失模态。 为处理缺失语音文本的视频-主张对，我们从数据集中选取与当前主张具有共同实体的真实语音文本。这

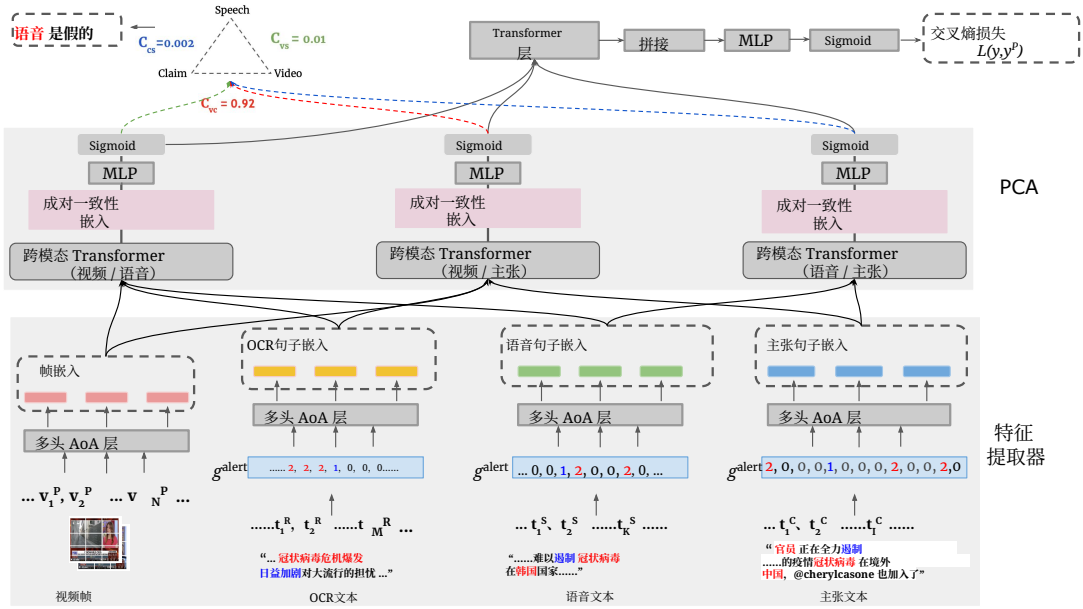


图 3：TwtrDetective 概览。它通过比较视频外观、语音内容、屏幕文本和主张来检测跨模态不一致性。 g^{alert} 表示事件提醒模块。在此示例中，TwtrDetective 将其预测为一个不一致对，并指出语音遭到篡改，因为 c^{vs} 和 c^{cs} 更小。

由真实视频、虚假语音和真实主张组成的新元组将被视为COVID-VTS中的不一致对。我们的工具生成的示例如图1 和图4所示。如果没有深入推理和跨模态理解，就很难将它们与真实对区分开来。

经过自动质量筛选和偏差控制的所有步骤，并进行人工验证后，我们将篡改样本与原始样本的数量设为相同。因此，最终数据集包含 1 万条格式良好的主张及其相关视频。

3.3 数据集分析

与现有的视频文本任务基准数据集（Xu 等人，2016；Chen 和 Dolan，2011年；Rohrbach 等人，2015；Anne Hendricks 等人，2017；Krishna 等人，2017）相比，COVID-VTS 有三个重要区别。首先，COVID-VTS 带来了新的挑战，场景更加多样、复杂，包括室内新闻发布会、新闻播报、抗议活动等户外活动、采访、屏幕录制以及幻灯片演示。与近期的视频文本数据集（Xu 等人，2016；Chen 和 Dolan，2011年；Rohrbach 等人，2015；Anne Hendricks 等人，2017；Krishna 等人，2017）相比，如图2所示，COVID-VTS 中可用的语音和屏幕文本更多。具体而言，87.5% 的视频包含语音，超过 77% 的视频包含屏幕文本。COVID-VTS 中的命名实体和视频更多（表4）。其中 90.8% 的句子COVID-VTS包含命名实体，而 VIOLIN（Liu 等人，2020b）中这一比例为 10.2%。这一巨大差距表明，COVID-VTS 不仅是支持视频与命名实体对齐研究的优秀资源，也为现有视频语言推理模型带来了新的挑战。最后，COVID-VTS 是短视频平台上的首个事实核查基准，其中不同模态的信息彼此不一致，这使事实核查任务更具挑战性。

4 事实核查模型

4.1 概述

如图3所示，我们的细粒度事实核验系统TwtrDetective能够通过整合不同模态之间的成对关系嵌入，评估整体事实一致性。TwtrDetective 还会指出哪种模态存在不一致，并提供解释以支持核验。

特征提取器。我们使用 CLIP（ViT-B/32）的视觉变换器（Radford 等人，2021）将每一帧编码为特征。具体而言，它会从帧中提取 N 互不重叠的图像块，并通过线性投影将每个图像块映射为一个一维词元 $\{v_1^p, \ldots, v_N^p\}$，其中$v_i \in \mathbb{R}^D$，其中 $D = 512$，N 表示每帧的图像块数量。其次，我们使用预训练的 RoBERTa（Liu 等人，2019b）生成

上下文文本表示，它使用字节对编码（Shibata 等人，1999）对句子进行分词。因此，语音内容中的每个句子 S 都表示为一个词元序列$\{t_1^s, \ldots, t_K^s\}$，其中 $t_i^s \in \mathbb{R}^D$和 $D = 768$ 以及屏幕文本 O为 $\{t_1^r, \ldots, t_M^r\}$，其中 $t_i^r \in \mathbb{R}^D$和 $D = 768$。最后，主张中的词元 C定义为 $\{t_1^c, \ldots, t_I^c\}$，其中 $t_i^c \in \mathbb{R}^D$和$D = 768$。

事件提醒模块 检验一致性的一种直观方法是使用视频-句子检索模型计算视频与文本之间的相似度（Liu 等人，2019a；Gabeur 等人，2020）。然而，如果真实信息被少量片段篡改，其性能会非常糟糕。这是因为现有模型主要在文档层面对文本进行编码，没有提供足够的信号来聚焦于事实要素，即事件触发词和事件论元。

为了提供明确的指导来学习文本的内部事件语义，我们首先采用 DY-GIE++ （Hope 等人，2020）来检测文本中的事件结构：如果词元是事件触发词，则赋值为 1；如果是事件论元，则赋值为 2；否则赋值为 0。随后，将这些索引输入可学习的嵌入表，以生成事件提醒 门控 g^{alert}，其中 $g^{alert} \in \mathbb{R}^D$ 和 $D = 512$。的一个关键属性是，它能够帮助我们的模型确定文本中词元的重要性。然后，通过学习参数 W^c， W^s， W^r，将主张词元 t_i^c、语音词元 t_i^s和光学字符识别词元 t_i^r投影到具有相同维度的共同语义子空间中。

$$t_i^{c'} = g^{alert} \odot \tanh(W^c t_i^c) \quad (1)$$
$$t_i^{s'} = g^{alert} \odot \tanh(W^s t_i^s) \quad (2)$$
$$t_i^{r'} = g^{alert} \odot \tanh(W^r t_i^r) \quad (3)$$

其中， $\odot$表示逐元素乘法操作， $tanh$ 表示激活函数。

多头 AoA 层。受（Liu 等人，2020a）提出的架构启发，我们通过堆叠多头注意力之上的注意力层，对 $t_i^{c'}$、 $t_i^{s'}$、 $t_i^{r'}$ 进行上下文建模。该方法利用“注意力之上的注意力”模块（Huang 等人，2019）来促进注意信息的生成。在对文本进行编码后，将[CLS] 词元的输出（命名为 s^c、 s^s 和 s^r ）分别用作主张、语音内容和屏幕文本的句子表示。

$$s^c = \text{MHAoA}^{\text{Mask}}(\{t_i^{c'}, \ldots, t_I^{c'}\})_{[cls]} \quad (4)$$
$$s^s = \text{MHAoA}^{\text{Mask}}(\{t_i^{s'}, \ldots, t_K^{s'}\})_{[cls]} \quad (5)$$
$$s^r = \text{MHAoA}^{\text{Mask}}(\{t_i^{r'}, \ldots, t_M^{r'}\})_{[cls]} \quad (6)$$

图块特征也通过 W^v 投影到与文本相同的公共子空间中。为了学习每帧中各图块的重要性，我们将注入位置信息后的图块特征输入多头 AoA 层，以建模各图块之间的相关性。随后，我们利用全局平均池化输出每帧的表示 v^f。

$$v_i' = \tanh(W^v v_i) \quad (7)$$
$$v^f = Pool(\text{MHAoA}^{\text{Mask}}(\{v_1', \ldots, v_N'\})) \quad (8)$$

成对一致性聚合（PCA）。为了建模主张、语音和视频帧之间的一致性，我们应用跨模态 Transformer（Li 等人，2020）来学习成对关系。首先，我们将语音句子嵌入 $S = \{s_1^s \ldots, s_{L_s}^s\}$、视觉帧嵌入 $F = \{v_1^f \ldots, v_{L_f}^f\}$ 及其对应的光学字符识别句子嵌入 $O = \{s_1^r \ldots, s_{L_r}^r\}$ 作为输入。L_s、 L_f、 L_r 表示句子数量。我们还在输入序列的开头添加 [CLS] 词元。跨模态 Transformer 的输出是一系列经过语境化处理的嵌入。我们使用 [CLS] 词元的输出表示视频与语音之间的一致性关系。

$$R^{vs} = \text{Cross-Transformer}(F, O, S) \quad (9)$$
$$R^{vc} = \text{Cross-Transformer}(F, O, C) \quad (10)$$
$$R^{cs} = \text{Cross-Transformer}(C, S) \quad (11)$$

对于主张 $C = \{s_1^c \ldots, s_{L_c}^c\}$ 与语音 S、视频 F 之间的一致性度量，我们同样利用跨模态 Transformer 来整合不同模态，并使用 [CLS] 表示权重。主张可以被视为指针，用于从配对的视频和语音中精准检索相关证据，而这些证据对于推断配对内容的真实性起着关键作用。

解释。将成对一致性嵌入 R^{vs}、 R^{vc} 和 R^{cs} 输入多层感知机层和 Sigmoid 函数后，我们的模型会产生关系分数 c^{vs}，表示视频与语音之间的一致性分数； c^{vc} 表示视频与

模型	准确率	F1
CLIP4Clip (Luo 等人，2021b)	57.3	55.8
CLIP2Video(Fang 等人，2021)	59.4	56.8
VideoClip (Xu 等人，2021)	50.4	49.5
McCrae(McCrae 等人，2022)	62.7	62.3
MTS（Liu 等人，2020b）	56.3	55.4
MMT（Gabeur 等人，2020）	61.2	60.6
TwtrDetective（我们的方法）	**68.1**	**67.9**
人工核查	82.7	81.9

表5：与COVID-VTS 数据集上的基线方法相比的结果（%）。

模型	准确率
MTS（Liu 等人，2020b）	60.4
XML（Lei 等人，2020）	69.6
HERO（Li 等人，2020）	70.4
TwtrDetective（我们的方法）	**72.6**

表 6：与 VIOLIN 数据集上的基线方法进行比较的结果（%）。

主张与 c^{cs}之间的一致性分数。如果一个视频帖子被我们的模型分类为不一致，那么得分较低的两个链接所共有的模态将被标记为不一致模态。例如，在图3中，TwtrDetective 检测到语音不一致，因为 c^{vs} 和 c^{cs} 的得分较低。这一解释可用于支持判断。

目标函数。我们通过最小化图 3顶部所示的标准交叉熵来优化模型，其中 y 是真实标签，y^P 是将c^{vs}、 c^{vc} 和 c^{cs} 输入 Transformer 注意力层后得到的预测概率。

5 实验

在本节中，我们首先介绍实现细节，并将结果与竞争方法进行比较。最后，我们展示全面的实验结果并进行讨论。

5.1 实现细节

数据集。 我们在两个数据集上进行实验：(1) COVID-VTS 数据集，该数据集包含 1 万个视频-文本对，这些视频-文本对具有由我们的自动化工具生成的不同分类体系。其中一半样本为未篡改样本，另一半为篡改样本。(2) VI-OLIN 数据集（Liu 等人，2020b），这是一个视频-语言推理数据集，收集自电视节目和电影，包含 15,887 个由人类撰写的视频片段及主张。我们还使用我们的自动化

模型	准确率	F1分数
TwtrDetective	68.1	67.9
TwtrDetective（不含事件提醒模块）	64.4	64.5
TwtrDetective（不含主成分分析）	66.3	65.9

表7：通过移除事件提醒模块或成对一致性聚合模块，研究模型性能的消融研究。

工具生成不一致的语音和视频对，以确保不一致性来自不同模态。我们在两个数据集上将模型与基线方法进行比较。

模型训练。 我们的模型使用 PyTorch 实现。在实现中，公共子空间的维度为 512。模型使用 Adam 优化，学习率为 0.0005。此外，我们采用均匀采样策略提取帧，采样率为每秒 1 帧。在所有实验中，我们将 COVID-VTS 数据集划分为 80% 用于训练、10% 用于验证集、10% 用于测试。

评估指标。 我们计算准确率和 F1 分数，以区分不一致对和一致对。

基线方法。（1）CLIP4Clip（Luo 等人，2021b）（2）CLIP2Video（Fang 等人，2021）使用 CLIP（Rad- ford 等人，2021）对视觉帧和文本进行编码，并使用 Transformer 融合时间信息。（3）VideoCLIP（Xu 等人，2021）是一种先进的零样本视频与文本理解模型，它通过 Transformer 学习视频与文本之间的细粒度关联。（4）McCrae（McCrae 等人，2022）采用长短期记忆网络整合视频帖子中的视频、文本和命名实体信息。（5）MMT 是一种视频文本模型，它设计了一个多模态变换器，以联合编码视频中的不同模态。（7）HERO（Li 等人，2020）：一种基于 Transformer 的框架，具有用于计算局部与全局上下文的两种标准层级结构。

5.2 结果与讨论

对比实验。 如表 5 和表6 所示，与基线方法相比，我们的模型在 COVID-VTS 和 VIOLIN 数据集上均取得了最先进的结果。CLIP 变体模型表现不佳，说明它们无法通过分析句子级语义来检测词元级操纵。我们注意到，McCrae 和 MMT 在两项得分上均优于 MTS。这是因为 McCrae 和 MMT 从视频中提取了额外特征，例如目标检测和人脸检测，以辅助核验。然而，命名实体核



图4：Q 定性分析结果。红框标示不一致模态。TwtrDetective 预测 ts
前两个 除了第三个之外，示例均正确。

模型	准确率	F1
TwtrDetective	68.1	67.9
TwtrDetective（不含 $Pair_{[Claim,Video]}$）	66.6	66.3
TwtrDetective（不含 $Pair_{[Claim,Speech]}$）	67.2	67.0
TwtrDetective（不含 $Pair_{[Speech,Video]}$）	67.5	67.2

表8：分析PCA模块中不同模态对的重要性。

McCrae 提出的分类方法在我们的数据集上并未表现出改进，因为它无法发现事件触发词操纵。我们模型的主要优势在于，事件提醒模块能够从异构内容中提取细粒度的事实细节。成对一致性聚合模块能够精确衡量各模态之间的交互。这些观察结果也解释了为什么我们的模型优于 HERO，后者忽略了事件结构的引导。与 COVID-VTS 数据集上的结果相比，在 VIOLIN 数据集上的表现更好。这是因为我们的 COVID-VTS 数据集更具挑战性，包含更多命名实体和事件。

不同模态分析。我们进一步深入了解了 PCA 模块中不同模态对的重要性。表8 第一行中的模型在 PCA 中使用了全部三种模态对 $Pair_{[Claim,Video]}$、$Pair_{[Claim,Speech]}$ 和 $Pair_{[Speech,Video]}$。其他行则缺少其中一种模态对。表8表明，PCA模块能够有效提升多个模态之间的一致性检查性能。此外，$Pair_{[Claim,Video]}$ 的贡献大于另外两种模态对。这是因为在较长的语音文本中，证据句是唯一被操纵的部分，因此仅参考主张和视频很难检测出来。

消融研究。我们在图 5中研究了帧长度。可以看到，帧数从 1 增加到 6 时性能显著提升，这表明使用

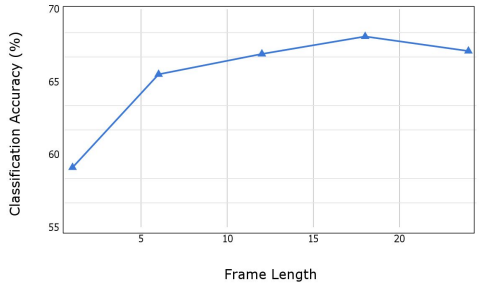


图5：消融研究，用于考察我们的模型在不同帧长度下的性能。

由多个帧组成的序列来编码视频，比只使用单帧更好。我们在实验中采样了 18 帧，兼具效率和效果。此外，表 7显示，我们的模型主要受益于事件提醒模块，该模块通过在词元级显式跟踪失真的事实片段，使分类准确率提升了 3.7%；主成分分析模块则通过聚合不同模态的特征，相比使用单流Transformer，带来了 1.8% 的提升。此外，模型在短主张（长度 < 15）上的准确率高于在长主张（长度 > 25）上的准确率。这是因为长主张比短主张包含更多事件触发词和论元，这使得事件提醒模块难以捕捉被篡改的词元。

定性分析。图4展示了我们模型的预测示例。正确的案例（前两个示例）展现了模型识别细微不一致词元的能力。由于第三个示例的原始配对缺少语音模态，我们加入了与主张包含更多共同实体的语音文本，作为不一致模态。TwtrDetective 未能正确预测第三条帖子，这表明它在视频与语音对齐方面表现不佳。原因可能在于视频和语音都描述了反对疫苗强制令的抗议活动

但发生在不同的国家。仅凭视频，很难判断地点是在澳大利亚、德国还是奥地利。此外，日期信息也难以核实。未来的研究方向可以是要求模型指出信息中哪些部分不一致或无法核实，并为这些情况定义不一致分类体系。

6 结论

本文发布了一个用于短视频平台事实核查的新基准数据集 COVID-VTS，其中包含 1 万个视频-主张对。我们开发了一种高效工具，可自动生成大量语义各异且可信的不一致对。此外，我们提出的事实核查模型 TwtrDetective 达到了最先进的检测准确率。我们希望这项工作能为多模态事实核查以及视频与语言领域的其他相关研究铺平道路。

7 伦理声明

我们开发先进的一致性检查技术，旨在提升该领域检测假新闻的能力，并改善 Twitter 社区的健康状况。根据《Twitter开发者政策》，Twitter 支持为非商业研究目的，对垃圾信息、滥用行为或其他与平台健康相关的主题进行衡量和分析的研究。此外，我们收集的帖子均来自经过认证的真实账号。若某些账号发布骚扰、羞辱或侮辱任何个人或群体的内容，或违反《平台操纵和垃圾信息政策》，则不具备获得认证徽章的资格。为保护个人信息，我们只会使用说明文字和视频作为输入，不使用用户信息。我们还会对数据集进行筛选，只保留讨论公共新闻的帖子，而非涉及个人生活的帖子。个人信息包括但不限于用户姓名、健康状况、财务状况、种族或族裔出身、宗教或哲学信仰或理念、性取向、工会成员身份，以及被指控或实际实施的犯罪行为。需要特别说明的是，我们不会分享源视频文件，只会分享 URL 链接或从视频中提取的特征。这是出于许可政策的要求，也为了避免有人故意生成和传播虚假信息。因此，为供公众验证和审计，我们将公开模型代码，但不会公开本研究的输出，以便用于打击假新闻。

8 局限性

我们的模型与人类在准确率方面仍存在显著差距。我们希望 COVID-VTS 数据集能够推动社区开发更强大的模型。一个可能的方向是开发能够从语音中定位关键帧或关键句子的模型，以推断一致性检查的难度。

致谢

本研究得到 DARPA SemaFor (HR001119S0085) 项目的资助。我们感谢本文匿名审稿人提出的意见。

参考文献

Lisa Anne Hendricks、Oliver Wang、Eli Shechtman、Josef Sivic、Trevor Darrell 和 Bryan Russell。2017。用自然语言定位视频中的时刻。载于 *IEEE*国际计算机视觉会议论文集，第 5803–5812 页。

Pepa Atanasova、Preslav Nakov、Lluís Màrquez、Al- berto Barrón-Cedeño、Georgi Karadzhov、Tsve- tomila Mihaylova、Mitra Mohtarami 和 James Glass。2019。利用语境和话语信息进行自动事实核查。数据与信息质量期刊 *(JDIQ)*，11(3):1–27。

Isabelle Augenstein、Christina Lioma、Dongsheng Wang、Lucas Chaves Lima、Casper Hansen、Christian Hansen 和 Jakob Grue Simonsen。2019。Multifc：一个用于基于证据核查主张的真实世界多领域数据集。*arXiv* 预印本 *arXiv:1909.03242*。

Giannis Bekoulis、Christina Papagiannopoulou 和 Nikos Deligiannis。2021 年。事实提取与核验综述。*ACM* 计算调查（*CSUR*），55(1)：1–35。

David Chen 和 William B Dolan。2011 年。为释义评估收集高度平行的数据。载于计算语言学协会第 *49* 届年会论文集：人类语言技术，第 190–200 页。

Han Fang、Pengfei Xiong、Luhui Xu 和 Yu Chen。2021。CLIP2Video：通过图像剪辑掌握视频文本检索。*arXiv* 预印本 *arXiv:2106.11097*。

Yi Fung、Christopher Thomas、Revanth Gangi Reddy、Sandeep Polisetty、Heng Ji、Shih-Fu Chang、Kath- leen McKeown、Mohit Bansal 和 Avirup Sil。2021。Infosurgeon：用于假新闻检测的跨媒体细粒度信息一致性检查。载于第*59*届年会论文集：

计算语言学协会与第*11*届自然语言处理国际联合会议（第*1*卷：长篇论文），第1683–1698页。

Valentin Gabeur、Chen Sun、Karteek Alahari 和 Cordelia Schmid。2020。《用于视频检索的多模态变换器》。载于欧洲计算机视觉会议，第 214–229 页。施普林格。

David Güera 和 Edward J Delp。2018。使用循环神经网络进行深度伪造视频检测。载于*2018* 年第 *15* 届 *IEEE* 国际高级视频与基于信号的监控会议（*AVSS*），第 1–6 页。IEEE。

Andreas Hanselowski、Christian Stab、Claudia Schulz、Zile Li 和 Iryna Gurevych。2019。一个用于自动事实核查不同任务的丰富标注语料库。*arXiv* 预印本 *arXiv:1911.01214*。

TomHope、Aida Amini、David Wadden、Madeleine van Zuylen、Sravanthi Parasa、Eric Horvitz、Daniel Weld、Roy Schwartz 和 Hannaneh Hajishirzi。2020。从新冠肺炎论文中提取机制知识库。*arXiv* 预印本 *arXiv:2010.03824*。

Lun Huang、Wenmin Wang、Jie Chen 和 Xiao-Yong Wei。2019。用于图像描述生成的注意力之上的注意力。载于*IEEE/CVF*国际计算机视觉会议论文集，第4634–4643页。

Yichen Jiang、Shikha Bordia、Zheng Zhong、Charles Dognin、ManeeshSingh 和Mohit Bansal。2020。Hover：一种用于多跳事实提取和主张验证的数据集。*arXiv*预印本 *arXiv:2011.03088*。

Mi-Ae Ko 和 Young-Mo Kim。2004。一种基于强透视视图的简单光学字符识别方法。载于第*33*届应用图像模式识别研讨会（*AIPR'04*），第235–240页。IEEE。

Ranjay Krishna、Kenji Hata、Frederic Ren、李飞飞和 Juan Carlos Niebles。2017。视频事件的密集描述生成。载于*IEEE*国际计算机视觉会议论文集，第706–715页。

Jie Lei、Licheng Yu、Tamara L Berg 和 Mohit Bansal。2020。TVR：用于视频字幕片段检索的大规模数据集。载于欧洲计算机视觉会议，第 447–463 页。施普林格。

Linjie Li、Yen-Chun Chen、Yu Cheng、Zhe Gan、Licheng Yu 和 Jingjing Liu。2020。HERO：用于视频+语言全方位表征预训练的分层编码器。*arXiv* 预印本 *arXiv:2005.00200*。

Fuxiao Liu、Yinghan Wang、Tianlu Wang 和 Vicente Ordonez。2020a。视觉新闻：新闻图像描述生成的基准数据集与挑战。*arXiv* 预印本 *arXiv:2010.03743*。

Jingzhou Liu、Wenhu Chen、Yu Cheng、Zhe Gan、Licheng Yu、Yiming Yang 和 Jingjing Liu。2020b。VIOLIN：用于视频与语言推理的大规模数据集。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10900–10910 页。Yang Liu、Samuel Albanie、Arsha Nagrani 和 Andrew Zisserman。2019a。善用已有资源：利用协作专家表征进行视频检索。*arXiv* 预印本 *arXiv:1907.13487*。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019b。RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Grace Luo、Trevor Darrell 和 Anna Rohrbach。2021a。Newsclippings：自动生成脱离语境的多模态媒体。*arXiv* 预印本 *arXiv:2104.05893*。

Huaishao Luo、Lei Ji、Ming Zhong、Yang Chen、Wen Lei、Nan Duan 和 Tianrui Li。2021b。Clip4clip：一项关于 CLIP 端到端视频片段检索的实证研究。*arXiv* 预印本 *arXiv:2104.08860*。

Scott McCrae、Kehan Wang 和 Avideh Zakhor。2022。社交媒体新闻帖子中的多模态语义不一致检测。载于国际多媒体建模会议，第 331–343 页。施普林格。

Nicholas Micallef、Marcelo Sandoval-Castañeda、Adi Cohen、Mustaque Ahamad、Srijan Kumar 和 Nasir Memon。2022。跨平台多模态错误信息：文本帖子和视频的分类体系、特征与检测。载于国际 *AAAI* 网络与社交媒体会议论文集，第 16 卷，第 651–662 页。

Dat Quoc Nguyen、Thanh Vu 和 Anh Tuan Nguyen。2020。BERTweet：一种用于英语推文的预训练语言模型。*arXiv*预印本 *arXiv:2005.10200*。

John F Pitrelli、Raimo Bakis、Ellen M Eide、Raul Fer- nandez、Wael Hamza 和 MichaelA Picheny。2006。用于美式英语的IBM表现力文本转语音合成系统。*IEEE*音频、语音与语言处理汇刊，14(4):1099–1108。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第8748–8763页。PMLR。

Anna Rohrbach、Marcus Rohrbach、Niket Tandon 和 Bernt Schiele。2015。《电影描述数据集》。载于*IEEE* 计算机视觉与模式识别会议论文集，第 3202–3212 页。

Arkadiy Saakyan、Tuhin Chakrabarty 和 Smaranda Muresan。2021。Covid-fact：新冠肺炎疫情现实世界主张的事实抽取与核验。*arXiv* 预印本 *arXiv:2106.03794*。

Tal Schuster、Roei Schuster、Darsh J Shah 和 Regina Barzilay。2020。《文体计量学在检测机器生成的假新闻方面的局限性》。计算语言学，46(2)：499–510。

Lanyu Shang、Ziyi Kou、Yang Zhang 和 Dong Wang。2021。《面向 TikTok 上新冠肺炎短视频的多模态错误信息检测器》。载于*2021 IEEE* 国际大数据会议（*Big Data*），第 899–908 页。IEEE。

Yusuxke Shibata、Takuya Kida、Shuichi Fukamachi、Masayuki Takeda、Ayumi Shinohara、Takeshi Shinohara 和 Setsuo Arikawa。1999。《字节对编码：一种加速模式匹配的文本压缩方案》。

Reuben Tan、Bryan A Plummer 和 Kate Saenko。2020。检测跨模态不一致性以减少神经网络生成的假新闻危害。*arXiv* 预印本 *arXiv:2009.07698*。James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018。FEVER：用于事实提取与核验的大规模数据集。*arXiv* 预印本 *arXiv:1803.05355*。David Wadden、Shanchuan Lin、Kyle Lo、Lucy Lu Wang、Madeleine van Zuylen、Arman Cohan 和 Hannaneh Hajishirzi。2020。事实还是虚构：核验科学主张。*arXiv* 预印本 *arXiv:2004.14974*。David Wadden、Ulme Wennberg、Yi Luan 和 Hannaneh Hajishirzi。2019。利用上下文化跨度表示进行实体、关系和事件抽取。*arXiv* 预印本 *arXiv:1909.03546*。Kehan Wang、David Chan、Seth Z Zhao、John Canny 和 Avideh Zakhor。2022。社交媒体视频帖文中的虚假信息检测。*arXiv* 预印本 *arXiv:2202.07706*。Adina Williams、Nikita Nangia 和 Samuel R Bowman。2017。通过推理进行句子理解的广覆盖挑战语料库。*arXiv* 预印本 *arXiv:1704.05426*。Hu Xu、Gargi Ghosh、Po-Yao Huang、Dmytro Okhonko、Armen Aghajanyan、Florian Metze、Luke Zettlemoyer 和 Christoph Feichtenhofer。2021。VideoCLIP：用于零样本视频文本理解的对比式预训练。*arXiv* 预印本 *arXiv:2109.14084*。Jun Xu、Tao Mei、Ting Yao 和 Yong Rui。2016。MSR-VTT：用于连接视频与语言的大型视频描述数据集。载于*IEEE* 计算机视觉与模式识别会议论文集，第 5288–5296 页。

Rowan Zellers、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。2019。抵御神经网络生成的假新闻。神经信息处理系统进展，32。