COVID-VTS: Fact Extraction and Verification on Short Video Platforms

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

University of Maryland, College Park
{fl3es, yaser, abhinav}@umd.edu

Abstract

We introduce a new benchmark, COVID-VTS, for fact-checking multi-modal information involving short-duration videos with COVID19-focused information from both the real world and machine generation. We propose, TwtrDetective, an effective model incorporating cross-media consistency checking to detect token-level malicious tampering in different modalities, and generate explanations. Due to the scarcity of training data, we also develop an efficient and scalable approach to automatically generate misleading video posts by event manipulation or adversarial matching. We investigate several state-of-the-art models and demonstrate the superiority of TwtrDetective.

1 Introduction

The proliferation of misinformation in social media poses a serious threat to our society, especially during the COVID-19 pandemic. Therefore, it is necessary to develop automatic fact-checking tools to verify the claims propagated online.

Recent fact-checking work (Thorne et al., 2018; Wadden et al., 2020; Augenstein et al., 2019; Fung et al., 2021; Bekoulis et al., 2021) investigated automatic misinformation detection by developing various benchmark datasets as well as state-of-the-art neural network architectures involving sources such as Wikipedia pages, tables, news articles, and scientific articles. However, due to the lack of publicly available benchmarks, fact checking is still challenging on short video platforms, such as TikTok, Twitter, and Instagram.

Verifying the factual correctness of claims in short video platforms poses new challenges. Firstly, multi-modal misinformation that leads to short video posts is more misleading than using just textual content (Micallef et al., 2022), since the claims usually only tamper with subtle elements of the factual information from the source video. Furthermore, short videos (Shang et al., 2021) containing diverse scene shifts, human activities, and cross-modal information, greatly increase the complexity and ambiguity of the video content. Moreover, current methods explored the authenticity of GAN-generated video with unimodal analysis (Güera and Delp, 2018). Nevertheless, they cannot be directly applied to the short video platform, where the video content is often intentionally manipulated.

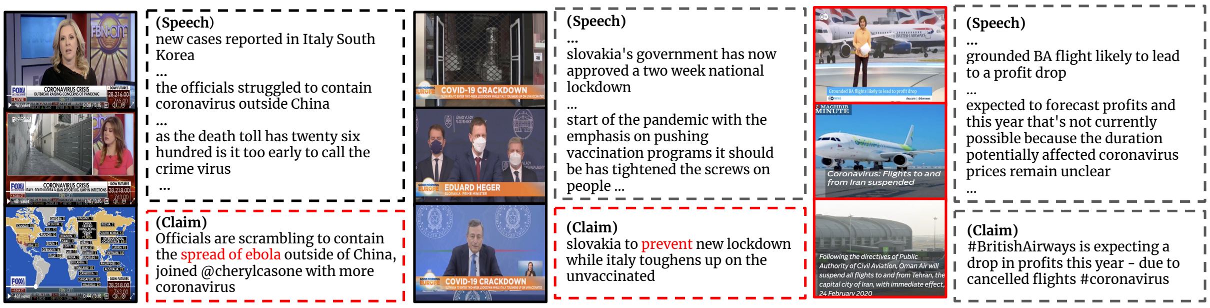
To tackle these challenges, we introduce COVID-VTS, a new benchmark dataset with trustworthy claims and corresponding good-quality videos, collected from Twitter video posts. COVID-VTS contains different inconsistent taxonomies and the inconsistency comes from different modalities. Examples are shown in Figure 1. We additionally introduce a novel approach to generate fake video posts automatically by manipulating event elements or adversarial matching (Luo et al., 2021a). The advantage of our event manipulation tool is that it's able to control the polarization of semantics and track the manipulation object by editing a small component of the factual information. We apply quality control to delete unqualified generations and address the linguistic bias.

We propose TwtrDetective, a multimodal fact-checking model, where the input consists of a claim with the paired video, and the goal is to predict the consistency. TwtrDetective takes advantage of the *Event Alert* module to precisely extract fine-grained factual details from the claim, as well as *Pairwise Consistency Aggregation* module to effectively measure the consistency between each modality. TwtrDetective is also able to point out the inconsistent modality (e.g., see Figure 3).

Experimental results show that TwtrDetective achieves higher detection accuracy over baselines on two datasets. Our main contributions are summarized as:

- We introduce COVID-VTS, a fact-checking dataset for short video platforms, consisting of 10k video text pairs with diverse scenes, more accessible

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 178–188
May 2-6, 2023 ©2023 Association for Computational Linguistics


CORONAVIRUS CRISIS
OUTBREAK RAISING CONCERNS OF PANDEMIC
CORONAVIRUS CRISIS
ITALY, SOUTH KOREA & IRAN REPORT BIG JUMP IN INFECTIONS
COVID-19 CRACKDOWN
EDUARD HEGER
SLOVAKIA PRIME MINISTER
COVID-19 CRACKDOWN
Grounded BA flights likely to lead to profit drop
MAGHRIB MINUTE
Coronavirus: Flights to and from Iran suspended
Following the directives of Public Authority of Civil Aviation, Oman Air will suspend all flights to and from Tehran, the capital city of Iran, with immediate effect, 24 February 2020

(Speech)
new cases reported in Italy South Korea
...
the officials struggled to contain coronavirus outside China
...
as the death toll has twenty six hundred is it too early to call the crime virus
...

(Claim)
Officials are scrambling to contain the spread of ebola outside of China, joined @cherylcasone with more coronavirus

(Speech)
...
slovakia's government has now approved a two week national lockdown
...
start of the pandemic with the emphasis on pushing vaccination programs it should be has tightened the screws on people ...

(Claim)
slovakia to prevent new lockdown while italy toughens up on the unvaccinated

(Speech)
...
grounded BA flight likely to lead to a profit drop
...
expected to forecast profits and this year that's not currently possible because the duration potentially affected coronavirus prices remain unclear
...

(Claim)
#BritishAirways is expecting a drop in profits this year - due to cancelled flights #coronavirus

Figure 1: Examples of different inconsistent taxonomies from COVID-VTS, which are generated by our automatic manipulation tool. The red box indicates the inconsistent modality. Our task is to independently decide whether the video-text pair is consistent and point out which modality is fake. (1) The event argument in the claim is modified; (2) The event trigger in the claim is tampered; (3) The short video is curated by adversarial matching.

modalities, and trustworthy claims from both the real world and machine generation.

- We propose an effective approach to automatically generate large-scale verifiable, trustworthy as well as misleading video posts rather than employing human annotators.
- We propose `TwtrDetective`, a new explainable fact-checking framework for the short video platform, showing superior results on two challenging datasets with respect to baselines.

Our code is publicly available at `https://github.com/FuxiaoLiu/Twitter-Video-dataset`.

2 Related Work

The spread of misinformation has led to a growing body of research in automatic fact-checking. Many large scale datasets collected from Wikipedia and fact-checking websites were introduced, including FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), UKP Snopes (Hanselowski et al., 2019), MultiFC (Augenstein et al., 2019). However, the fake claims in these datasets are manually generated by humans, making it expensive to deploy at scale. Recently, some synthetic datasets were proposed to address this limitation. (Jiang et al., 2020) generated complex fake claims using word substitutions. (Saakyan et al., 2021) took advantage of the token-infilling ability from Masked Language Model to replace the salient tokens. Also, (Fung et al., 2021) formulated a novel data synthesis method by manipulating knowledge elements within the multi-modal knowledge graph. In order to alleviate the linguistics bias within the machine-generated claims, (Luo et al., 2021a) constructed the out-of-context captions by retrieving the real-world sentence with the similar semantics. In a similar vein our dataset, (Liu et al., 2020b) constructed VIOLIN for the Video-Language Inference while all the statements are written by experts. (Wang et al., 2022) collected social media video posts from Twitter but the fake claims are constructed by random swap. In comparison, `COVID-VTS` is the first COVID-19 fact-checking dataset for short video platforms, containing rich information with diverse scenes, more accessible modalities as well as misleading claims from both the real world and machine generation.

Traditional fact-checking approaches (Zellers et al., 2019; Schuster et al., 2020; Atanasova et al., 2019) are mainly based on text. They fall short if the evidence stems from visual information. Recent multi-modal models (Fung et al., 2021; Tan et al., 2020; Shang et al., 2021) are equipped with the ability to check consistency according to the information conveyed across modalities. In contrast, we propose a fact-checking model for the short video platforms with multimedia explanations that achieves higher accuracy to detect misinformation.

3 COVID-VTS Dataset Construction

`COVID-VTS` contains 10k well-formed claims with the paired videos to support or refute the claims. In this section, we first describe the procedure to select the trustworthy and consistent video/claims pairs from Twitter video posts. Second, we also present our approach to automatically construct well-formed inconsistent video-claim pairs.

3.1 Filtering for Authentic and Consistent Video Posts

To assemble the `COVID-VTS` dataset, we used the Twitter scraper to collect over 100k English video posts using COVID related keywords and hashtags ranging from the end 2019.10 to 2022.8. Examples are shown in Table 1. We retain one post for a video link to reduce the potential bias. To retrieve the speech from videos, we leverage the Speech-to-Text tool from IBM Cloud (Pitrelli et al., 2006). In addition, we leverage SimpleOCR (Ko and Kim, 2004) to recognize text on screen. After the preliminary filtering, we have several steps to select the authentic and consistent video posts:

VERIFIED User Account. According to the requirements from Twitter, only the authentic, notable, and active accounts will receive the blue verified badge (including official government, official company, etc.). Accounts that routinely post content that harasses, shames, or engaged in severe violations of our platform manipulation and spam policy are ineligible for blue badge. We collect the authentic posts from the verified accounts into `COVID-VTS` to improve the data quality.

Claims Must be VERIFIABLE. Event structures are essential to reveal the factual information of a sentence, since the overall semantics might be opposite if event elements are changed. Moreover, claims in `COVID-VTS` are supposed to be verifiable propositions whose truthfulness is determined by multi-modal evidences from the paired videos. Without the event structures, its truthfulness isn't verifiable. Thus, we delete instances with personal opinions or emotions. For example, *'So proud of my boys! GetVaccinated'*, which has no factual information to be verified.

In practice, we first remove the claims without event structures, by using DYGIE++ (Wadden et al., 2019), a state-of-art event extraction framework, pretrained on MECHANIC dataset. The event structures generated by DYGIE++ include event triggers and event arguments. For example in *'Officials are scrambling to contain outbreaks of the coronavirus outside of China'*, *'contain'* is the trigger which express the occurrence of the event, *'officials'* and *'outbreaks of the coronavirus'* are event arguments which play different roles in this event.

After selecting these verifiable claims, we manually check whether they are consistent with the associated videos. We also select verifiable and consistent video text pairs from the unverified accounts in order to increase the size of our dataset. Our hypothesis is that the video posts are trustworthy after the "cleaning" steps. Given this hypothesis, the main task is to analyze the inconsistency between different modalities.

Keywords/Hashtags
covid, #covid, #covid19, corona, #coronavirus, maskup, pandemic, ICU, vaccine, #vaccine, coronavirus, quarantine, #quarantine, moderna, #covidvariant, omicron, booster, mRNA, phizer, #phizer, #who, #workfromhome, #vaccinepassports, #travelbans, #social_distance, n95

Table 1: Summary of the covid related keywords and hashtags in the data collection process.

	Most frequent Tokens
Event Trigger (Original Claims)	*fight, infect, protect, prevent, help, stop, quarantine, contain, confirm, threat, plunge, mandate, deal, cause*
Event Argument (Original Claims)	*coronavirus, omicron, pfizer, covid, delta, moderna, booster, mask, lockdown, protest, social distance, ban, vaccine*
Event Trigger (Generated Claims)	*produce, increase, create, develop, remove, protect, avoid, identify, support, generate, rule, allow, establish*
Event Argument (Generated Claims)	*omicron, variant, death, cancer, hospital, WHO, community, medical service, drug, viruses, migration, media, ICU*

Table 2: Most frequent event triggers and arguments in original claims and generated claims.

3.2 Automatic INCONSISTENT Video-Claim Pairs Generation

In this section, we describe the steps to automatically generate inconsistent video-claim pairs. Our dataset contains the following three inconsistent taxonomies: (1) Real video, Real speech and Inconsistent claim; (2) Real video, Inconsistent speech and real claim; (3) Inconsistent video, Real speech and Real claim.

Inconsistent Claim Generation. Inspired by (Liu et al., 2020b), we automatically generate fake-claims by controlling the polarization of semantics and tracking the manipulation object by modifying a small portion of the factual information in true claims. In this case, most of the statement remains true to the video content. This strategy can also alleviate the linguistic bias, which was introduced by (Fung et al., 2021; Zellers et al., 2019). In their datasets, the model correctly detects the fake claims without comparing different modalities. This is because all the fake claims are generated by language models. Another advantage is that we can generate the large-scale training set automatically rather than employing human annotators.

Original TRUE Claims	Generated FAKE Claims
Officials are scrambling to contain outbreaks of the coronavirus outside of China.	*Officials are scrambling to contain the spread of ebola outside of China.*
Moderna chairman getting vaccinated booster shots is the only way to stop the virus.	*Moderna chairman getting infected is the only way to stop the virus.*
Fed chair powell warns omicron variant could dent economic recovery.	*Fed chair powell warns omicron variant could facilitate economic recovery.*
Australians caught up in china's crisis have finally returned home after 14 days quarantined on christmas island.	*Australians caught up in china's crisis have lost home after 14 days quarantined on christmas island.*

Table 3: A detailed look into the examples generated by our efficient automatic approach. The first two examples are the event-argument manipulation and the final two are the event-trigger manipulation.

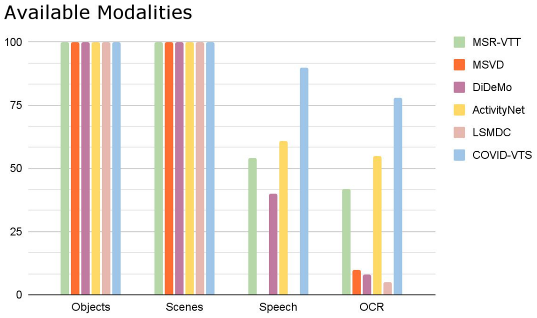


Figure 2: The histogram indicates COVID-VTS is more accessible to speech content and text on the screen than previous datasets.

	COVID-VTS	VIOLIN
Average Caption Length	19.2	18.0
Average Speech Length	69.2	76.4
Average Video Length	26.5s	35.2s
Named Entities (Sentence)	90.8%	10.3%
Source	Twitter	TV show

Table 4: Summary of `COVID-VTS` dataset.

Only event triggers or event arguments will be selected as the manipulated tokens `[mask]` depending on different intentions. We follow the procedures from (Nguyen et al., 2020) and use BERTWeet, pretrained on the Tweets related to the COVID-19 pandemic, to predict the domain-aware alternates of the event elements which can be subject to masking. In order not to introduce additional noise, we delete the candidates if their substituted tokens are not in the vocabulary of the original dataset to reduce the potential bias. Then, we feed the candidates and their corresponding true claims as input to the CrossEncoder model trained on Mutli-NLI dataset (Williams et al., 2017) and select the candidates with the CONTRADICTION label. If the candidate pair has the highest contradiction score, it will be assigned as the fake claim. Table 3 presents examples generated by our efficient automatic approach. The first two examples are the event-argument manipulation and the final two are the event-trigger manipulation.

Table 2 shows the most frequent event elements in original claims and generated claims. In practice, we found that these frequent alternatives bring the additional linguistic bias. Models can learn to classify the claims as fake ones simply by detecting these frequent tokens without absorbing information from other modalities. Therefore, we alleviate this bias by only keeping one claim for each alternative. Then replace unqualified claims by selecting the most similar claims from the dataset. Inconsistent pairs constructed this way focus more on the global understanding of the pairs. This rigorous setting makes the claims more challenging to distinguish by the analysis model, and in-depth reasoning is required to identify the fake content.

Inconsistent Speech Generation. Unlike (Tan et al., 2020) and (Fung et al., 2021) editing multiple parts of the article, our target is to only modify the evidence sentences from the speech to reduce the linguistic bias. First, we use cosine similarity on SBERT sentence embeddings to extract the most similar sentences to the real claims. After that, we manipulate the named entities or event elements within the evidence sentences.

Inconsistent Video Generation. As for the third type of inconsistent pairs, we keep the speech and claims as the original ones. However, we replace the original videos with another real video that is similar to the current one by using the adversarial matching (Liu et al., 2020b) method. Specifically, we utilize pretrained VideoCLIP (Xu et al., 2021) to generate video representations to calculate the cosine similarity.

Missing Modalities. In order to handle the video-claim pairs missing the speech text, we select the real speech text from the dataset which has common entities with respect to the current claim. This

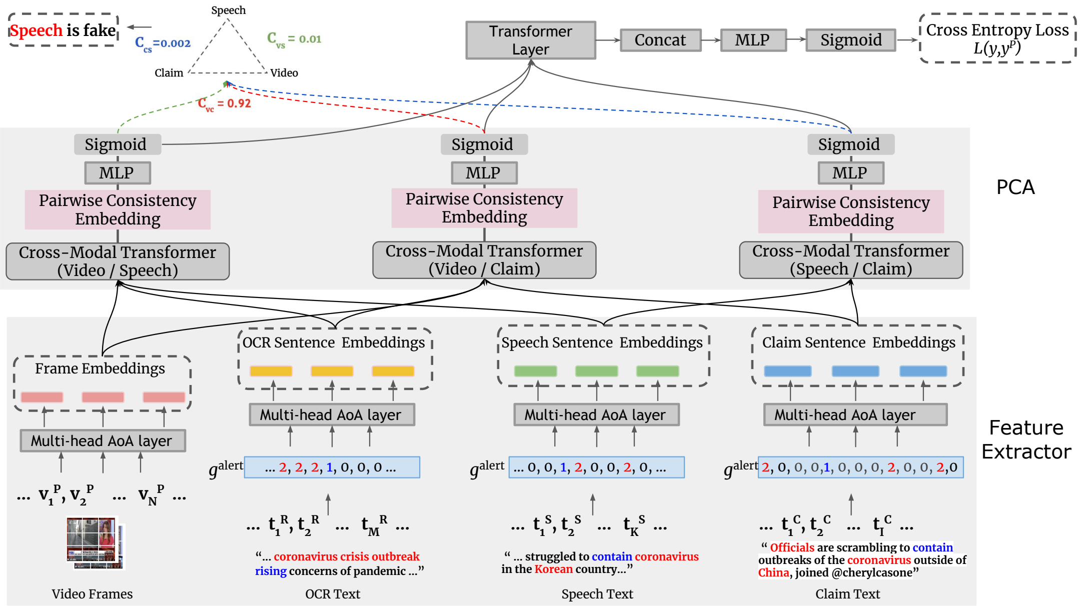

Speech is fake
Speech
C_{cs}=0.002
C_{vs} = 0.01
Claim
Video
C_{vc} = 0.92
Transformer Layer
Concat
MLP
Sigmoid
Cross Entropy Loss
$L(y,y^P)$
Sigmoid
MLP
Pairwise Consistency Embedding
Cross-Modal Transformer (Video / Speech)
Sigmoid
MLP
Pairwise Consistency Embedding
Cross-Modal Transformer (Video / Claim)
Sigmoid
MLP
Pairwise Consistency Embedding
Cross-Modal Transformer (Speech / Claim)
PCA
Frame Embeddings
OCR Sentence Embeddings
Speech Sentence Embeddings
Claim Sentence Embeddings
Multi-head AoA layer
Multi-head AoA layer
Multi-head AoA layer
Multi-head AoA layer
Feature Extractor
g^{alert} ... 2, 2, 2, 1, 0, 0, 0 ...
g^{alert} ... 0, 0, 1, 2, 0, 0, 2, 0, ...
g^{alert} 2, 0, 0, 0,1, 0, 0, 0, 2, 0, 0, 2,0
... v_1^P, v_2^P ... v_N^P ...
... t_1^R, t_2^R ... t_M^R ...
... t_1^S, t_2^S ... t_K^S ...
... t_1^C, t_2^C ... t_I^C ...
"... coronavirus crisis outbreak rising concerns of pandemic ..."
" ... struggled to contain coronavirus in the Korean country..."
" Officials are scrambling to contain outbreaks of the coronavirus outside of China, joined @cherylcasone"
Video Frames
OCR Text
Speech Text
Claim Text

Figure 3: Overview of `TwtrDetective`. It detects the cross-modal inconsistency by comparing with video appearance, speech content, screen text and claims. g^{alert} means the event alert module. In this example, `TwtrDetective` predicts it as a an inconsistent pair and points out speech is manipulated since c^{vs} and c^{cs} are smaller.

new tuple with the real video, fake speech and real claim will be regarded as the inconsistent pairs in `COVID-VTS`. Examples generated by our tool are shown in Figure 1 and Figure 4. Without the in-depth reasoning and cross-modal understanding, it's challenging to distinguish them from the real pairs.

After all the filtering steps of automatic quality and bias control with manual validation, we assign the same amount of manipulated samples with pristine ones. Thus, the resulted dataset consists of 10k well-formed claims with associated videos.

3.3 Dataset Analysis

`COVID-VTS` exhibits three important differences to current benchmark datasets for video-text tasks (Xu et al., 2016; Chen and Dolan, 2011; Rohrbach et al., 2015; Anne Hendricks et al., 2017; Krishna et al., 2017). First, `COVID-VTS` brings new challenges with more diverse and complex scenes, including indoor press conference, news broadcasting, outdoor activities like protests, interviews, and screen recordings as well a slide shows. In contrast to recent video-text datasets (Xu et al., 2016; Chen and Dolan, 2011; Rohrbach et al., 2015; Anne Hendricks et al., 2017; Krishna et al., 2017), `COVID-VTS` has more available speech and screen text in Figure 2. Specifically, 87.5% of the videos have speech and over 77% of the videos have the screen text. `COVID-VTS` has more named entities and videos (Table 4). 90.8% of the sentences in `COVID-VTS` have named entities while VIOLIN (Liu et al., 2020b) is 10.2%. This large gap indicates `COVID-VTS` not only is an excellent resource to support research in the alignment between videos and named entities, but also provides new challenges to existing video-language inference models. Finally, `COVID-VTS` is the first fact-checking benchmark on the short video platform with inconsistent information from different modalities, which makes the fact-checking task more challenging.

4 Fact-Checking Model

4.1 Overview

As shown in Figure 3, our fine-grained fact-checking system, `TwtrDetective`, is able to evaluate the overall factual consistency by integrating pairwise relation embeddings between different modalities. `TwtrDetective` also points out which modality is inconsistent and provides explanations to support the verification.

Feature Extractor. We use the vision transformer of CLIP (ViT-B/32) (Radford et al., 2021) to encode every frame into features. In particular, it extracts N non-overlapping image patches from the frame and perform linear projection to map every patch into 1D token $\{v_1^p, \ldots, v_N^p\}$, where $v_i \in \mathbb{R}^D$, where $D = 512$, N is the number of patches for each frame. Second, we apply pretrained RoBERTa (Liu et al., 2019b) to generate contextual text representations, which utilizes byte pair encoding (Shibata et al., 1999) to tokenize the sentences. Therefore, each sentence in the speech content S is represented as a sequence of tokens $\{t^s_1, \dots, t^s_K\}$, where $t^s_i \in \mathbb{R}^D$ and $D = 768$ and screen text O are $\{t^r_1, \dots, t^r_M\}$, where $t^r_i \in \mathbb{R}^D$ and $D = 768$. Finally tokens in the claim C are defined as $\{t^c_1, \dots, t^c_I\}$, where $t^c_i \in \mathbb{R}^D$ and $D = 768$.

Event Alert Module The intuitive method to check the consistency is to calculate the video-text similarity with video-sentence retrieval models (Liu et al., 2019a; Gabeur et al., 2020). However, the performance is dismal if the true information is tampered by small fragments. This is because existing models mainly encode the text at the document level without giving sufficient signal to focus on the factual elements, namely the event trigger and event argument.

To provide explicit guidance to learn the internal event semantic of the text, we first employ DYGIE++ (Hope et al., 2020) to detect the event structures within the text, assigning 1 if the token is the event trigger, 2 if it's the event argument and 0 otherwise. The indices are then fed into the learnable embedding table to generate *Event Alert* gate g^{alert}, where $g^{alert} \in \mathbb{R}^D$ and $D = 512$. A key property of g^{alert} is that it helps our model determine the salience of tokens in the text. Then, the claim token t^c_i, speech token t^s_i, OCR token t^r_i are projected into a common semantic subspace with the same dimension by learning parameters W^c, W^s, W^r.

$$t^{c\prime}_i = g^{alert} \odot \tanh(W^c t^c_i) \tag{1}$$

$$t^{s\prime}_i = g^{alert} \odot \tanh(W^s t^s_i) \tag{2}$$

$$t^{r\prime}_i = g^{alert} \odot \tanh(W^r t^r_i) \tag{3}$$

where $\odot$ represents the element-wise multiplication operation and $tanh$ is the activation function.

Multi-head AoA Layer. Motivated by the architecture presented in (Liu et al., 2020a), we contextualises $t^{c\prime}_i$, $t^{s\prime}_i$, $t^{r\prime}_i$ by using stacked Multi-Head Attention on Attention Layer, which takes advantage of the "Attention on Attention" module (Huang et al., 2019) to facilitate the generation of attended information. After encoding the text, the output of [CLS] tokens named as s^c, s^s and s^r are utilized as the sentence representations of claim, speech content and screen text correspondingly.

$$s^c = \text{MHAoA}^{\text{Mask}}(\{t^{c\prime}_i, \dots, t^{c\prime}_I\})_{[cls]} \tag{4}$$

$$s^s = \text{MHAoA}^{\text{Mask}}(\{t^{s\prime}_i, \dots, t^{s\prime}_K\})_{[cls]} \tag{5}$$

$$s^r = \text{MHAoA}^{\text{Mask}}(\{t^{r\prime}_i, \dots, t^{r\prime}_M\})_{[cls]} \tag{6}$$

Patch features are also projected into the common subspace with the text by W^v. In order to learn the salience of patches in each frame, we feed the patch features with injection of positional information into Multi-Head AoA Layer to model the correlation of each patch. After that, we leverage the global average pooling to output the representation of each frame v^f.

$$v'_i = \tanh(W^v v_i) \tag{7}$$

$$v^f = \mathit{Pool}(\text{MHAoA}^{\text{Mask}}(\{v'_1, \dots, v'_N\})) \tag{8}$$

Pairwise Consistency Aggregation (PCA). To model the consistency between the claim, speech and video frames, we apply the Cross-modal Transformer (Li et al., 2020) learn the pairwise relationship. First, we fed the speech sentence embeddings $S = \{s^s_1 \dots, s^s_{L_s}\}$, visual frame embeddings $F = \{v^f_1 \dots, v^f_{L_f}\}$ and its associated OCR sentence embeddings $O = \{s^r_1 \dots, s^r_{L_r}\}$ as the input. L_s, L_f, L_r present the number of sentences. We also add the [CLS] token in the first place of the input sequence. The outputs from Cross-modal Transformer is a sequence of contextualized embeddings. We use the output from the [CLS] token represent the consistency relationship between the video and speech.

$$R^{vs} = \text{Cross-Transformer}(F, O, S) \tag{9}$$

$$R^{vc} = \text{Cross-Transformer}(F, O, C) \tag{10}$$

$$R^{cs} = \text{Cross-Transformer}(C, S) \tag{11}$$

As for the consistency measurement between the claim $C = \{s^c_1 \dots, s^c_{L_c}\}$ and speech S, video F, we also utilize the Cross-modal Transformer to integrate different modalities and [CLS] to represent the weight. The claim can be regarded as the pointer to precisely retrieve relevant evidences from the paired video and speech, which play a key role to infer the truthfulness of the pairs.

Explanations. After feeding pairwise consistency embeddings R^{vs}, R^{vc} and R^{cs} into MLP layers and sigmoid functions, our model produces the relation scores c^{vs} which represents the consistency score between video and speech, c^{vc} which represents the consistency score between video and

Model	Accuracy	F1
CLIP4Clip (Luo et al., 2021b)	57.3	55.8
CLIP2Video (Fang et al., 2021)	59.4	56.8
VideoClip (Xu et al., 2021)	50.4	49.5
McCrae (McCrae et al., 2022)	62.7	62.3
MTS (Liu et al., 2020b)	56.3	55.4
MMT (Gabeur et al., 2020)	61.2	60.6
TwtrDetective (Ours)	**68.1**	**67.9**
Human Check	82.7	81.9

Table 5: Comparative results (%) with baselines on the COVID-VTS Dataset.

Model	Accuracy
MTS (Liu et al., 2020b)	60.4
XML (Lei et al., 2020)	69.6
HERO (Li et al., 2020)	70.4
TwtrDetective (Ours)	**72.6**

Table 6: Comparative results (%) with baselines on the VIOLIN Dataset.

claim and c^{cs}, which represents the consistency score between claim and speech. If a video post is classified as `inconsistent` by our model, the common modality of the two links with lower scores will be pointed as the `inconsistent` modality. For example in Figure 3, `TwtrDetective` detects that speech is inconsistent since c^{vs} and c^{cs} are smaller. It can be used the explanation to support the judgement.

Objective Function. We optimize our model by minimizing the standard cross-entropy as shown on the top of Figure 3, where y is the ground truth label and y^P is the prediction probability after putting c^{vs}, c^{vc} and c^{cs} into the transformer attention layer.

5 Experiments

In this section, we first introduce details of our implementation and compare the results to competing methods. Lastly, we present comprehensive experiment results and discussions.

5.1 Implementation Details

Datasets. We conduct experiments on two datasets: (1) COVID-VTS dataset, which consists of 10k video-text pairs with different taxonomies generated by our automatic tool. Half of the samples are pristine and half are manipulated. (2) VIOLIN dataset (Liu et al., 2020b), a Video-and-Language Inference, collected from TV shows and movies, which contains 15,887 video clips and claims written by human. We also generate inconsistent speech and video pairs with our automatic tool so as to make sure the inconsistency comes from different modalities. Our model is compared with baselines on both datasets.

Model	Accuracy	F1
TwtrDetective	68.1	67.9
TwtrDetective (w/o Event Alert Module)	64.4	64.5
TwtrDetective (w/o PCA)	66.3	65.9

Table 7: Ablation study to investigate our model's performance without the event alert module or the pairwise consistency aggregation module.

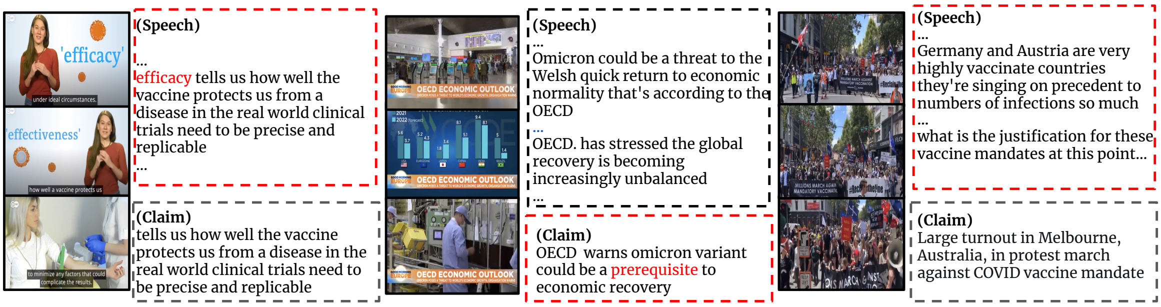
Model Training. Our model is implemented using Pytorch. In our implementation, the dimensions for the common subspace is 512. Models are optimized using Adam with learning rate as 0.0005. In addition, we adopt a uniform sampling strategy to extract the frames and the sampling rate is 1 frame per second. In all the experiments, we split the COVID-VTS dataset into 80% for training, 10% for validation and 10% for testing.

Evaluation Metric. We compute the accuracy and F1 score at distinguishing inconsistent pairs from consistent ones.

Baselines. (1) CLIP4Clip (Luo et al., 2021b) (2) CLIP2Video (Fang et al., 2021) uses CLIP (Radford et al., 2021) to encode visual frames and text and use the transformer to fuse temporal information. (3) VideoClip (Xu et al., 2021) is a state-of-art zero-shot video and text understanding model, which learns fine-grained association between video and text in a transformer. (4) McCrae (McCrae et al., 2022) employs Long LSTM to integrate the video, text and named entity information from the video posts. (5) MMT, a video-text model that designs a multi-modal transformer to jointly encode the different modalities in video. (7) HERO (Li et al., 2020): a transformer-based framework with two standard hierarchies for local and global context computation.

5.2 Results and Discussion

Comparison Experiments. As Table 5 and Table 6 show, our model achieves SOTA results compared with baselines on the both `COVID-VTS` and VIOLIN dataset. CLIP-variant models fail to perform well, revealing they do not detect token-level manipulation by analyzing the semantics of the sentence level. We notice that McCrae and MMT outperform MTS on both scores. This is because McCrae and MMT extract additional features from videos to help verification like objective detection and face detection. However, named entity verification



(Speech)
...
efficacy tells us how well the vaccine protects us from a disease in the real world clinical trials need to be precise and replicable
...

(Claim)
tells us how well the vaccine protects us from a disease in the real world clinical trials need to be precise and replicable

(Speech)
...
Omicron could be a threat to the Welsh quick return to economic normality that's according to the OECD
...
OECD. has stressed the global recovery is becoming increasingly unbalanced
...

(Claim)
OECD warns omicron variant could be a prerequisite to economic recovery

(Speech)
...
Germany and Austria are very highly vaccinate countries they're singing on precedent to numbers of infections so much
...
what is the justification for these vaccine mandates at this point...

(Claim)
Large turnout in Melbourne, Australia, in protest march against COVID vaccine mandate

'efficacy'
under ideal circumstances.
'effectiveness'
how well a vaccine protects us
to minimize any factors that could complicate the results.
OECD ECONOMIC OUTLOOK
OECD ECONOMIC OUTLOOK
OECD ECONOMIC OUTLOOK

Figure 4: Qualitative analysis results. The red box indicates the inconsistent modality. `TwtrDetective` predicts the first two examples correctly except the third one.

Model	Accuracy	F1
TwtrDetective	68.1	67.9
TwtrDetective (w/o $Pair_{[Claim,Video]}$)	66.6	66.3
TwtrDetective (w/o $Pair_{[Claim,Speech]}$)	67.2	67.0
TwtrDetective (w/o $Pair_{[Speech,Video]}$)	67.5	67.2

Table 8: Analysis on the importance of different modality pairs in the PCA module.

proposed by McCrae does not show improvement on our since because it's incapable of finding the event-trigger manipulation. The main advantages of our model is that our event alert module is able to extract fine-grained factual details from the heterogeneous content. The pairwise consistency aggregation module is able to measure the interaction between each modality precisely. These observations also explain why our model outperforms HERO, which ignores the guide from the event structures. The performance on the VIOLIN dataset is better compared to the results on `COVID-VTS` dataset. This is because our `COVID-VTS dataset` is more challenging with more named entities and events.

Analysis on Different Modalities. We gain further insights into the importance of different modality pairs in PCA module. Our model in the first row of Table 8 uses all three pairs $Pair_{[Claim,Video]}$, $Pair_{[Claim,Speech]}$ and $Pair_{[Speech,Video]}$ in PCA. Other rows miss one of these pairs. Table 8 indicates the PCA module is effective to improve the consistency checking performance between multiple modalities. In addition, $Pair_{[Claim,Video]}$ contributes more than the other two pairs. This is because the evidence sentences are the only manipulated parts in the long speech text, making it challenging to be detected by referring to claims and videos.

Ablation Study. We investigate frame length in Figure 5, we can see a significant increase between 1 and 6 frames which shows it is better to encode the videos with a sequence of multiple frames instead of one single frame. We sampled 18 frames for our experiment, which is both efficient and effective. Furthermore in Table 7, our model mainly benefits from *Event Alert* module which provides 3.7% boost in classification accuracy by explicitly tracking distorted factual pieces at the token level, *PCA* module contributes 1.8% improvement compared to using one-stream transformer to aggregate features from different modalities. Additionally, our model achieves better accuracy on short claims (length < 15) than on long claims (length > 25). This is because long claims have more event triggers and arguments than short claims, which makes it challenging for our event alert module to capture the manipulated tokens.

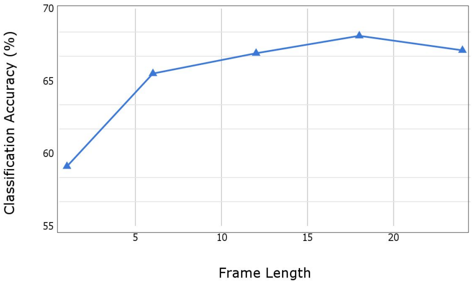


Figure 5: Ablation study to investigate our model's performance with different frame length.

Qualitative Analysis. Figure 4 presents prediction examples from our model. The correct cases (first two examples) demonstrate the model's ability to recognize the tiny inconsistent tokens. Since the original pair of third example miss the speech modality, we add the speech text which shares more common entities with the claim as the inconsistent modality. `TwtrDetective` fails to predict third post, suggesting that it does not work well on the video and speech alignment. The reasons behind might because both the video and speech describe the protest against the vaccine mandate but in different countries. Only from the video, it's hard to detect whether it's in Australia, Germany or Austria. In addition, the date information is also challenging to check. The future direction could be asking the model to point out which part of the information is inconsistent or unverifiable and define the inconsistency taxonomy for them.

6 Conclusions

In this paper, we release a new benchmark, `COVID-VTS`, for fact-checking in short video platforms, consisting of 10k video-claim pairs. We develop an efficient tool to automatically generate large-scale trustworthy inconsistent pairs with different semantic meanings. Furthermore, our proposed fact-checking model `TwtrDetective` achieves state-of-the-art detection accuracy. We hope this work paves the way for future studies in multi-modal fact-checking as well as other related research areas in video and language.

7 Ethical Statement

Our goal in developing state-of-art consistency checking technique is to enhance the field's ability to detect fake news and improve the Twitter community health. According to Twitter Developer Policy, Twitter supports the research that measures and analyzes topics like spam, abuse, or other platform health-related topics for non-commercial research purposes. In addition, the posts we collected are from verified and authentic accounts. Certain accounts are ineligible for the verified badge if they post content that harasses, shames, or insults any individual or group, or violate the Platform manipulation and spam policy. To protect the personal information, we will only use the captions and videos as input without the user information. We also work to filter the dataset and only keep the posts discussing public news instead of personal life. Personal information but not limited to, user's name, health, financial status, racial or ethnic origin, religious or philosophical affiliation or beliefs, sexual orientation, trade union membership, alleged or actual commission of crime. It's crucial to mention that we will not share our source video file but the URL links or the extracted features from the videos. This is to due to the licence policy and avoid anyone to deliberately generate and spread misinformation. As such, we will release the model code but not the output in our work for public verification and auditing so it can be used to combat fake news.

8 Limitations

There is a significant gap between our model and human performance on the accuracy. We hope `COVID-VTS` dataset will encourage the community to develop stronger models in the future. One possible direction is to develop models to localize key frames or key sentences from the speech to deduce the difficulty of consistency check.

Acknowledgements

This work was supported by DARPA SemaFor (HR001119S0085) program. We are thankful for the feedback from anonymous reviewers of this paper.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Computing Surveys (CSUR)*, 55(1):1–35.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pages 1683–1698.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer.

David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2020. Extracting a knowledge base of mechanisms from covid-19 papers. *arXiv preprint arXiv:2010.03824*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Mi-Ae Ko and Young-Mo Kim. 2004. A simple ocr method from strong perspective view. In *33rd Applied Imagery Pattern Recognition Workshop (AIPR'04)*, pages 235–240. IEEE.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020a. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020b. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019a. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021a. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021b. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Scott McCrae, Kehan Wang, and Avideh Zakhor. 2022. Multi-modal semantic inconsistency detection in social media news posts. In *International Conference on Multimedia Modeling*, pages 331–343. Springer.

Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 651–662.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. 2006. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 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, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. IEEE.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Reuben Tan, Bryan A Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Kehan Wang, David Chan, Seth Z Zhao, John Canny, and Avideh Zakhor. 2022. Misinformation detection in social media video posts. *arXiv preprint arXiv:2202.07706*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.