Combating Online Misinformation Videos: Characterization, Detection, and Future Directions

Yuyan Bu
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
buyuyan22s@ict.ac.cn

Qiang Sheng
Institute of Computing Technology, Chinese Academy of Sciences
shengqiang18z@ict.ac.cn

Juan Cao
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
caojuan@ict.ac.cn

Peng Qi
Institute of Computing Technology, Chinese Academy of Sciences
pengqi.qp@gmail.com

Danding Wang
Institute of Computing Technology, Chinese Academy of Sciences
wangdanding@ict.ac.cn

Jintao Li
Institute of Computing Technology, Chinese Academy of Sciences
jtli@ict.ac.cn

ABSTRACT

With information consumption via online video streaming becoming increasingly popular, misinformation video poses a new threat to the health of the online information ecosystem. Though previous studies have made much progress in detecting misinformation in text and image formats, video-based misinformation brings new and unique challenges to automatic detection systems: 1) high information heterogeneity brought by various modalities, 2) blurred distinction between misleading video manipulation and nonmalicious artistic video editing, and 3) new patterns of misinformation propagation due to the dominant role of recommendation systems on online video platforms. To facilitate research on this challenging task, we conduct this survey to present advances in misinformation video detection. We first analyze and characterize the misinformation video from three levels including signal, semantic, and intent. Based on the characterization, we systematically review existing works for detection from features of various modalities to techniques for clue integration. We also introduce existing resources including representative datasets and useful tools. Besides summarizing existing studies, we discuss related areas and outline open issues and future directions to encourage and guide more research on misinformation video detection. The corresponding repository is at https://github.com/ICTMCG/Awesome-Misinfo-Video-Detection.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

KEYWORDS

misinformation video detection; multi-modal computing; survey

The authors are also with Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences.
Corresponding author: Juan Cao.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612426

ACM Reference Format:
Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating Online Misinformation Videos: Characterization, Detection, and Future Directions. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612426

1 INTRODUCTION

With the prevalence of online video platforms, information consumption via video streaming is becoming increasingly prominent. Popular video-sharing platforms like YouTube and TikTok have attracted billions of monthly active users [66]. Studies on news consumption show that about a quarter of U.S. adults under 30 regularly get news from these video-sharing platforms [37].

Unfortunately, the massive growth of video news consumption also boosts the rapid spread of misinformation videos, posing an increasingly serious challenge to the online information ecosystem. For instance, 124 TikTok misinformation videos about COVID-19 vaccines gathered over 20M views and 300K shares [47]. Fig. 1 shows a misinformation video post about the existence of microchips in COVID-19 vaccines. Compared with previously studied misinformation in text and image format mostly, video-based misinformation is more likely to mislead the audience and go viral. Research in the political domain shows that individuals are more likely to believe an event's occurrence when it is presented in video form [79]. Another experiment indicates that the video format makes news pieces perceived as more credible and more likely to be shared [68]. Such an effect makes the misinformation video dangerous and it may lead to further negative impacts for various stakeholders. For individuals, misinformation videos are more likely to mislead audiences to generate false memory and make misinformed decisions. For online platforms, the wide spread of misinformation videos may lead to reputation crises, users' inactivity, and regulatory checks. For watchdog bodies, misinformation videos may be weaponized to foment unrest and even undermine democratic institutions. Therefore, actions are urgently required to reduce the risk brought by misinformation videos.

As a countermeasure, online platforms have made efforts to mitigate the spread of misinformation videos. For instance, TikTok

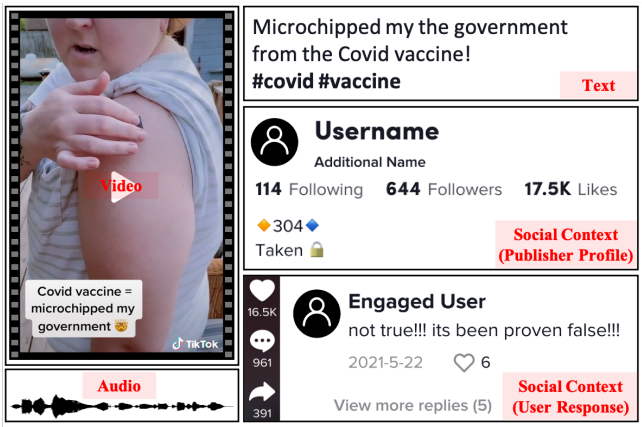


Figure 1: A misinformation video post on TikTok, along with the attached social context information, indicating that COVID-19 vaccines contain microchips. For privacy concerns, we replace the user avatars and names with placeholders.

introduces an enhanced in-app reporting feature to help curb the spread of COVID-19 misinformation [45]. However, the current solution relies much on human efforts from expert teams or active users, which is labor-intensive and time-consuming. It usually fails to perform real-time detection, and thus could not react rapidly when a new event emerges or previously debunked misinformation recurs [51]. Moreover, this solution may introduce uncertain individual biases and errors [24]. To tackle the above problems, developing techniques and systems for automatic misinformation video detection becomes a promising option.

Compared with text-based or text-image misinformation detection, video-based misinformation detection faces several unique challenges. First, the proliferation of heterogeneous information from diverse modalities brought more uncertainty and even noise to the final prediction. Second, nonmalicious video-editing behaviors blur the distinction between forged and real videos. Third, the recommendation-dominated content distribution of online video platforms reshapes the misinformation propagation from explicit behaviors like forwarding to implicit behaviors like re-uploading. These challenges necessitate new technical solutions for detecting video-based misinformation and also highlight the importance of conducting a careful, specific investigation into this problem.

Despite many valuable surveys conducted on broad misinformation detection, limited attention is given to video-based misinformation. Most of them regard the video as a kind of visual content as the image and discuss general multi-modal techniques for misinformation detection [2, 6, 11]. The above-mentioned uniqueness of video-based misinformation is not sufficiently considered. Other related surveys focus on a specific type of misinformation video, such as forged videos [4, 41], which provide detailed reviews but lack comprehensive analysis of the problem. Venkatagiri et al. [72] discussed the challenges and future research directions of studying misinformation on video-sharing platforms but did not provide an overview of specific detection methods. Considering the potential harms of misinformation videos, conducting a comprehensive survey on the detection problem is of urgent need.

To change the status quo and facilitate further exploration of this challenging problem by the research community, we present an overview of misinformation video detection in this survey. Our main contributions are as follows:

- **Comprehensive characterization:** We present a comprehensive analysis of characteristics of misinformation videos from three levels including signal, semantic, and intent;
- **Systematic technical overview:** We provide a systematical overview of existing multi-modal detection techniques and methods for misinformation in the video form with a principled way to group utilized clues and integration mechanisms;
- **Concrete future directions:** We discuss several open issues in this area and provide concrete future directions for both research and real-world application scenarios.

2 MISINFORMATION VIDEO CHARACTERIZATION

In this section, we characterize the misinformation video by giving the definition and analyzing it from three levels. Following Zhou and Zafarani [86], we define the misinformation video as:

Definition 2.1 (Misinformation Video). A video post that conveys false, inaccurate, or misleading information.

Note that a video post may include not only the video itself. On text-dominated social platforms like Facebook and Twitter, there could be a text paragraph attached; and on video-dominated platforms mentioned before, a title or a short text description is generally included. Moreover, the concept of *misinformation video* is closely related to but different from *fake video* because the former is characterized by the intention of spreading false or misleading information while the latter typically refers to the video that has been manipulated or produced by Photoshop-like tools and generative AI techniques. A fake video can be certainly used to create a misinformation video, but not all video faking is malicious (e.g., for movie production). Conversely, a video that is not manipulated but accompanied by false or misleading text will be considered a misinformation video. Therefore, detecting misinformation videos requires the appropriate combination of multi-perspective clues.

To find helpful detection clues, we characterize the misinformation video according to how it is produced. From the surface to the inside, the analysis is presented from three levels, including signal, semantic, and intent.

2.1 Signal Level

Misinformation videos often contain manipulated or generated video and audio content in which the forgery procedure often leads to traces in underlying digital signals. Forgery methods that produce such traces can be classified into two groups: Editing and Generation. Editing refers to visual alterations on existing data of video and audio modality. Typical editing actions include screen cropping, frame splicing, wave cutting, tempo changing, etc., which could be done using editing software [8]. Generation actions, by contrast, are done by neural networks which are trained to directly generate complete vivid videos (mostly with instructions from human actions [12] or texts [64]). The generated videos may contain forged human faces or voices to mislead the audience.

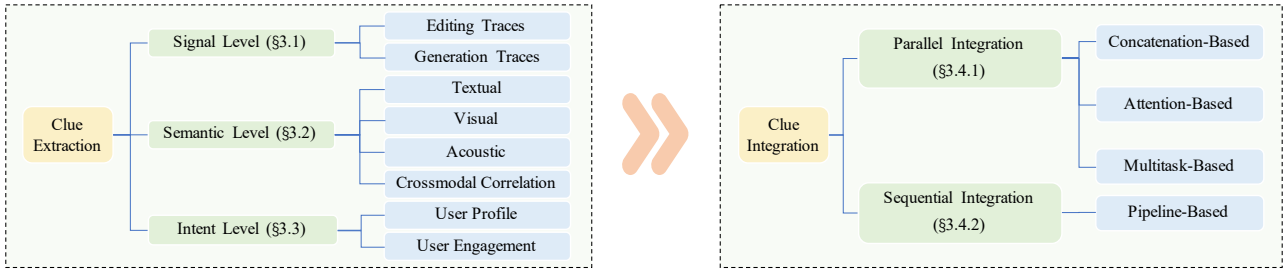


Figure 2: Overview of misinformation video detection techniques.

2.2 Semantic Level

The falsehood is conveyed through incorrect semantic changes against the truth. For a misinformation video, such changes may occur in one specific modality or across multiple ones. In the former case, manipulated elements (e.g., an exaggerated claim in the text description) are reflected by a single modality. In the latter case, which is more common in multi-modal situations, misinformation might be conveyed by wrong semantic associations among non-forged contents of different modalities. For instance, a creator may upload a real video of an event that happened before but add a real text description of a newly emerging event.

2.3 Intent Level

The creation of misinformation is often motivated by underlying intents, such as political influence, financial gain, and propaganda effects [77]. To achieve the underlying intent, misinformation videos generally pursue wide and fast spread. This leads to unique patterns of expression, propagation, and user feedback, which are different from real ones. For example, Qi et al. [51] find that compared with real news videos, fake news videos have more significant emotional preferences, involve more user engagement, and are more likely to be questioned in comments.

3 MISINFORMATION VIDEO DETECTION

The misinformation video detection problem is generally formulated as a binary classification task: Let $\mathcal{V}$ and $\mathcal{S}$ denote a *Video Post* and the attached *Social Context*, respectively. $\mathcal{V}$ consists of attributes such as title/description t, video v, and audio a. Social Context S consists of two major components: User Profile p and User Engagement e. User Profile p includes a set of features to describe the uploader account, such as the location indicated by IP address, self-written and verified introduction, and follower count. User Engagement e includes comments as well as statistics such as the number of views, likes, and stars. Fig. 1 illustrates an example of $\mathcal{V}$ from TikTok. The task of misinformation video detection is to predict whether the video post $\mathcal{V}$ contains misinformation given all the accessible features $\mathcal{E} = \{\mathcal{V}, \mathcal{S}\}$, i.e., $\mathcal{F} : \mathcal{E} \mapsto \{0, 1\}$. Existing studies usually take common metrics in classification tasks including accuracy, precision, recall, and F1 score for performance evaluation. Following the characterization in Sec. 2, we introduce a series of relevant works for detecting misinformation videos from clues at different levels to integration patterns. Fig. 2 gives an overview of this section.

3.1 Signal Level

Since misinformation videos are often created using forgery techniques, the detection of video forgery traces would provide a significant clue for detecting a misinformation video. Related works, which are always described as multimedia forensics, have received significant attention over the past decades. We introduce some commonly used techniques for the detection of editing traces and generation traces respectively. Given that there have been surveys on multimedia forensic techniques, here we will not go much into detail about the individual studies.

3.1.1 Editing Traces. As aforementioned, editing refers to alterations to the visual or audio content by multimedia editing software (mostly requiring manual operations). Existing detection methods for editing traces can be mainly categorized into two groups: active detection and passive detection.

In active detection methods, digital watermarks or digital signatures are pre-embedded and extracted to detect the trace of secondary editing. For instance, Tarhouni et al. [69] propose a blind and semi-fragile video watermarking scheme for detection. Combined watermarking (frames and audio watermarking) is used for detecting manipulation in both channels. In industry, the Coalition for Content Provenance and Authenticity develops a technical specification[1] based on digital signature techniques for certifying the provenance and authenticity of media content. Compared to passive detection, active detection methods could generally provide quicker responses and more accurate judgments.

However, most of the videos are not pre-embedded with such additional information in practice. The passive detection methods highlight their advantages as they use the characteristics of the digital video itself to detect tampering traces. For video frames, tampering detection methods leverage inter-frame and intra-frame information. The former detects the abnormal change in frame sequences, such as insertion [85], deletion [1], duplication [19], and shuffling [65] of frames. The latter detects the alteration of objects that could be reflected in a single frame, such as region duplication [5, 9] and splicing [31, 56]. Besides, artifacts acquired during the process of compression, noise artifacts left by the digital video camera, and inconsistencies (e.g., lighting, brightness, shadows) are also utilized as clues for detection [59]. For the audio content, statistical features inspired by observed properties are leveraged for forgery detection. Among them, Electric Network Frequency is

[1] https://c2pa.org/specifications/specifications/1.2/

widely used for forensics, thanks to its properties of random fluctuation around the nominal value and intra-grid consistency [55].

3.1.2 Generation Traces. Videos generated by neural networks (e.g., generative adversarial networks) are also known as "deepfake videos". Among them, deepfake videos containing vivid, generated human faces have been used for impersonating celebrities and brought negative impacts. The past few years have witnessed significant progress in detecting visual deepfakes. With reference to how deepfake videos are created [41], detection methods leverage clues generated at different steps including model fingerprints [84], biological signals [36], and temporal consistency [25, 81]. With recent advances in Text-To-Speech and Voice Conversion algorithms, generating fake audio that is indistinguishable for humans become easier and attracted increasing attention. ASVspoof challenges [80] were organized for accelerating progress in deepfake audio detection. Most existing works for fake audio detection adopt hand-crafted acoustic features like Mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficient (LFCC), and constant-Q cepstral coefficients (CQCC) and apply classifiers such as Gaussian mixture model and light convolutional neural network to make predictions. Recent work also attempts to leverage pre-trained wav2vec-style models for speech representation extraction and build an end-to-end framework [35].

Joint audio-visual deepfake detection is to handle the case that the visual or/and auditory modalities have been manipulated. Detection methods focus on learning intrinsic (a)synchronization between video and audio frames in real and fake videos [42, 87]. Moreover, different modalities can also supplement each other for final judgments [32].

Considering that the action of video compression is widely used in the default upload setting of online video platforms, handling compressed deepfake videos becomes an important issue. The compression operation erases certain generation traces and thus makes videos more undetectable. Moreover, the quantization and inverse quantization operations during compression and decompression bring additional quantization noise and distortion. To tackle these problems, Hu et al. [28] propose a two-stream method by analyzing the frame-level and temporality-level of compressed deepfake videos for detection.

3.2 Semantic Level

Though signal-level clues could provide strong evidence, they are not decisive because of the wide use of portable editing tools. Even if editing or generation traces are detected, it does not necessarily mean that the video conveys misinformation. The existence of semantically unchanged edited videos has blurred the boundary between fake and real videos. For instance, a video that has been edited for the sake of brevity or clarity may still be truthful and informative. Conversely, a video that is technically untampered can be employed in a deceptive manner. Different from clues at the signal level, semantic-level clues offer a new perspective for identifying misinformative content. Thus, many recent works focus on leveraging multi-modal semantic clues from not only the video content but also descriptive textual information. In this section, we discuss the features exploited by semantic-based methods from a multi-modal perspective.

Textual Feature. The video content is always served with descriptive textual information such as video description, and title. Apart from these directly accessible texts, subtitles and transcriptions extracted from the video also present useful information. Early works most extract statistical features from these texts for classification. Papadopoulou et al. [49] first exploit linguistic features of the title, which contain basic attributes like text length, as well as designed indicators like whether a title contains specified characters (e.g., the question/exclamation mark and 1st/2nd/3rd person pronoun), the number of special words (e.g., slang words and sentiment-indicative words) and the readability score. Other works also consider the existence of specific expressions like clickbait phrases and violent words for detection[33, 48]. Corpus-aware features, such as n-grams, TF-IDF, lexical richness, and LIWC lexicon overlapping, are leveraged by Hou et al. [26] and Serrano et al. [57]. In addition to hand-crafted features, continuous representation generated using deep learning has been increasingly adopted. Jagtap et al. [30] employs GloVe [50] and Word2Vec [40] to generate subtitle embeddings. Shang et al. [58] and Choi and Ko [13] train bidirectional recurrent neural networks as text encoders to encode the semantics of textual information including the title, description, and transcription. Recent advances in pre-trained language models (e.g., BERT [17]) also drive the latest multi-modal detection models to obtain contextualized representation [13, 15, 38, 51, 73]. Moreover, factual elements like event triggers and event augments are utilized to provide explicit guidance to learn the internal event semantics in [34].

Visual Feature. The visual content is usually represented at the frame level or clip level. The former presents static visual features while the latter presents additional temporal features. Shang et al. [58] extract frames through uniform sampling and input the resized sampled frames into the advanced object detection network Fast R-CNN for visual features of object regions. Corresponding caption representation is used to guide the integration of object regions to help generate the frame visual representation. Choi and Ko [13] extract frames according to their similarity to the thumbnail. Pre-trained VGG-19 is utilized to extract visual features from the video frames. McCrae et al. [38] break video into 32-frames-long clips with each clip beginning at a keyframe. The keyframes are detected through the FFmpeg scene detection filter. For each clip, features related to human faces, objects, and activities are extracted through pre-trained FaceNet, ResNet50, and S3D networks, respectively. Wang et al. [73] break the video into clips with fixed duration directly and uses S3D to extract visual features likewise. Qi et al. [51] represent visual content both at the frame level and clip level. The pre-trained VGG19 model and pre-trained C3D model are used to extract frame features and clip features respectively. Liu et al. [34] encode frames into features using the pre-trained vision transformer ViT [18] used in CLIP [54].

Acoustic Feature. As a unique modality compared to text-image misinformation, the audio modality including speech, environment sound, and background music [51], plays an essential role in expressing information in videos. As for detection, in addition to the transcription mentioned above, current works search for useful clues from acoustic characteristics. Hou et al. [26] firstly import emotional acoustic features to the detection model, where predefined feature sets widely used for emotion recognition of raw

speech are exploited. Shang et al. [58] design an acoustic-aware speech encoder by introducing MFCC features. Qi et al. [51] use the pre-trained VGGish to extract the audio features for classification.

Cross-modal Correlation. Mismatches between modalities, such as video-text and video-audio, are often caused by video repurposing for the misleading aim, which would lead to important changes in cross-modal correlation. Liu et al. [34] apply the cross-modal transformer to learn the consistent relationship between video and speech, video and text, and speech and text, respectively. The pairwise consistency scores are then aggregated for final judgments. McCrae et al. [38] leverage an ensemble method based on textual analysis of the caption, automatic audio transcription, semantic video analysis, object detection, named entity consistency, and face verification for mismatch identification. Wang et al. [73] propose two methods based on contrastive learning and masked language modeling for joint representation learning to identify semantic inconsistencies. In [13], topic distribution differences between modalities are utilized to robustify the detection model.

The increasing diversity of semantic features has inspired the evolution of misinformation video detection from mining unimodal patterns to modeling multi-modal interactions. However, despite the progress, existing works still have some limitations. For instance, most of them extract features from the perspective of identifying misinformative patterns, which may result in overlooking factual information and failing to utilize clues that require external knowledge for reasoning and judgment. Despite the pioneering efforts of [53] in incorporating relevant factual information by rectifying the wrong predictions of previously detected news videos with reliable debunking videos, there remains an under-exploration of multi-modal verification. Moreover, logical constraints, such as inter-modality consistency and entailment, have not been sufficiently utilized, unlike in text-image modeling [52, 67]. These limitations suggest further improvements to enhance the accuracy and robustness of misinformation video detection methods.

3.3 Intent Level

Misinformation videos are often created and shared with deliberate intents, e.g., for financial and political gains [63] or self-expression [20]. Compared with clues at the signal and semantic levels, clues at the intent level are usually more robust to elaborately produced misinformation videos that avoid being detected, because the underlying essential intent and its effects on the behaviors of social media users are less likely to be changed or manipulated at scale. Starting from the motivations of those who create and spread misinformation, some effective features for detecting misinformation videos are leveraged by researchers.

Social contexts refer to user social engagements and profiles when information spreads on platforms with social media characteristics. As mentioned in Sec. 2.3, unique social contexts might reflect the spreading intent of misinformation creators and thus provide useful features [51]. Current works mostly make use of user comments and statistics on user engagement. The comments are usually exploited by extracting hand-crafted features [49] or generating general representation vectors through deep models [13, 48, 51]. Some works go deeper in mining comments. For example, Serrano et al. [57] learn a feature of comment conspiracy and Choi and Ko [14] give an eye to the domain knowledge. User engagement statistics such as the number of likes, comments, and views are generally directly concatenated with other features before being put into the classifier [26]. Some works also use statistical numbers as importance weights to help generate embedding. Choi and Ko [14] generate video comment embeddings by calculating the weighted sum of embeddings of each comment using their numbers of likes. The publisher profile provides auxiliary information about source credibility in post-level detection. Papadopoulou et al. [49] and Li et al. [33] leverage a series of features, such as the number of views of the publisher channel, number of published videos, and follower-following ratio. Qi et al. [51] also point out that user profiling features like geolocation information and whether the account is verified or not can be useful to the detection, and exploit the textual publisher description in their model.

Note that social context is not the only choice to capture intent-level clues. In text-based and text-image studies, researchers have started preliminary trials to mine the intents by directly analyzing the content itself [16, 21].

3.4 Clue Integration

In previous subsections, we categorized the various features utilized in existing works for detecting misinformation videos at the signal, semantic, and intent levels. In practice, existing methods generally combine multiple features from different modalities to make a judgment. In this section, we group the methods for integrating clues into two types (i.e., parallel and sequential integration) and introduce the current advances.

3.4.1 Parallel Integration. In parallel integration, all clues from different modalities contribute to the final decision-making process, although their participation may not be equal. The integration process can be performed at both the feature level (where features from different modalities are fused before being input into the classifier) and the decision level (where predictions are produced independently by different branches and combined using strategies like voting for final decisions). In existing works, feature fusion is used more frequently than decision integration. Here we discuss three types of feature fusion techniques used in existing works.

Concatenation-Based: The majority of the existing works on multi-modal misinformation detection embed each modality into a representation vector and then concatenate them as a multi-modal representation. The generated representation can be utilized for classification tasks directly or input into a deep network (e.g., the convolutional neural network) for deeper fusion and classification [33]. Linear combination is another simple but effective way to combine feature vectors of different modalities [13]. The fusion process that combines features from different modalities can be done at the video level or at the frame/clip level [38].

Attention-Based: The attention mechanism is a more effective approach for utilizing embeddings of different modalities, as it jointly exploits the multi-modal feature by focusing on specific parts and allows dynamic fusion for sequential data. Shang et al. [58] use a co-attention module that simultaneously learns the pairwise relation between each pair of a video frame and spoken word to fuse the visual and speech information. Wang et al. [73] model the joint distribution of video and text by using a variant of masked

Table 1: Summary of misinformation video detection methods. E/G: Editing/Generation Traces. T: Textual. V: Visual. A: Acoustic. CMC: Cross-modal Correlation. UE: User Engagements. UP: User Profile. C: Concatenation-Based. ATT: Attention-Based. MT: Multitask-Based. PL: Pipeline-Based. Methods for fake video detection are not included due to the space limit.

Method		Dataset			Clue Extraction							Clue Integration			
		Source	Amount	Released?	E/G	T	V	A	CMC	UE	UP	C	ATT	MT	PL
Liu et al. [34]	2023	Twitter	10,000	N		•	•		•				•		
Qi et al. [51, 53]	2023	Douyin, Kuaishou	5,538	Y		•	•	•	•	•	•		•		
Ganti [22]	2022	Not Specified	-	N	•	•									•
McCrae et al. [38]	2022	Facebook, YouTube	4,651	N		•	•		•	•		•			
Wang et al. [74]	2022	Twitter	943,667	N		•	•	•	•				•	•	
Wang et al. [73]	2022	Twitter	160,000	N		•	•		•				•	•	
Li et al. [33]	2022	Bilibili	700	N		•				•	•	•			
Choi and Ko [14]	2022	YouTube	2,912	N		•	•			•		•			
Choi and Ko [13]	2021	YouTube	4,622	N		•	•		•	•		•		•	
Shang et al. [58]	2021	YouTube	891	N		•	•	•					•		
Jagtap et al. [30]	2021	YouTube	2125	Y		•									
Serrano et al. [57]	2020	YouTube	180	N		•				•		•			
Hou et al. [26]	2019	YouTube	250	N		•		•		•		•			
Palod et al. [48]	2019	YouTube	546	Y		•				•		•			
Papadopoulou et al. [49]	2017	YouTube, Twitter, Facebook	5,006	Y		•				•	•	•			

language modeling. A transformer is trained to predict each text token given its text context and the video. Qi et al. [51] and Wang et al. [74] utilize a cross-modal transformer to model the mutual interaction between different modalities.

Multitask-Based: Another utilized fusion architecture is based on multitask learning. Under this architecture, auxiliary networks are applied to learn individual or multi-modal representations, spaces, or parameters better and improve the classification performance [2]. For example, Choi and Ko [13] use a topic-adversarial classification to guide the model to learn topic-agnostic features for good generalization. Wang et al. [73] use contrastive learning to build the joint representation space of video and text.

In practice, no single method demonstrates overwhelming superiority. Concatenation-based fusion preserves all available information but it treats all features equally without explicitly considering their relative importance. Attention-based fusion enables the model to focus on more informative parts, but it necessitates more complex mechanism design and computing resources. Multitask-based fusion enhances the model's generalization capabilities but requires a larger amount of labeled data as well as more computing resources. The choice of fusion method should depend on the specific requirements of the task, available resources, and the trade-off between simplicity, computational demands, and performance improvement.

3.4.2 Sequential Integration. In sequential integration, clues from different modalities are combined in a step-wise manner with each modality contributing incrementally to the final decision. This way of integration can promote overall efficiency and effectiveness, especially when some modalities provide redundant or overlapping information. We exemplified how sequential integration can be utilized in detecting misinformation videos with the two-pronged method proposed by Ganti [22]. This method operates by finding the original video corresponding to the given suspicious video and measuring the similarity between them. Initially, the method uses the reverse image search to retrieve the original video and calculates the similarity between frames of two videos to judge if the given video is a face-swapped deepfake video. If so, the video will be judged as a misinformative one; otherwise, the method analyzes the semantic similarity of video captions to detect the shifts in the meaning and intent behind the two videos. High semantic similarity leads to a real video judgment. To avoid the impact of sentiment changes on semantic similarity, the method compares video captions' sentiments for videos obtaining low semantic similarities. If the sentiments are similar, the method confirms that the semantic difference is not influenced by sentiments and judges the given video as a misinformative one.

Table 1 summarizes misinformation video detection methods and specifies the data sources, exploited clues, and clue integration techniques for each method. Most of the methods exploit multiple clues and use concatenation and attention for clue integration. Including more modalities and better modeling inter-modality relationships are promising for improving misinformation video detection.

4 RESOURCES

Resources are often the limiting factors for conducting research on this task. Here we introduce relevant datasets and tools and highlight their features and application contexts.

4.1 Datasets

Due to the difficulty of video crawling (mostly based on carefully selected keywords) and human annotation, many existing datasets are small-scale and topic-specific. We have compiled information about these datasets in Appendix B, including the data source, topic

domain, size, accessibility, and state-of-the-art (SOTA) methods (along with their performances) on these datasets. Unfortunately, most of the datasets are not released. Here we detail four large and publicly available datasets:

- **FVC**[2]**.** The initial FVC comprises videos from a variety of event categories (e.g., politics and sports), and contains 200 fake and 180 real videos. Using the initially collected videos as seeds and searching on three platforms (YouTube, Facebook, and Twitter), researchers extend FVC to a multi-lingual dataset containing 3,957 fake and 2,458 real videos, with textual news content and user comments attached.
- **YouTubeAudit**[3]**.** This dataset contains 2,943 videos that were published in 2020 on YouTube and cover five popular misinformative topics [29]. Each sample is labeled as promoting, debunking, and neutral to misinformation. It also provides social contexts like metadata (e.g., video URL, title, duration), statistics of user engagements, and user profiles (e.g., gender and age).
- **FakeSV**[4]**.** This dataset contains 5,538 Chinese short videos (1,827 fake, 1,827 real, and 1,884 debunking videos) crawled from the short video platforms Douyin and Kuaishou. It covers 738 news events from 2019 to 2022. It also provides social contexts including user responses and publisher profiles.
- **COVID-VTS**[5]**.** This dataset contains 10k COVID-related video-text pairs and is specifically for research on inter-modality consistency. Half of the samples are pristine and others are generated by partially modifying or replacing factual information in pristine samples.

4.2 Tools

Tools for specific utilities are valuable for verifying suspicious videos because they provide important auxiliary information that could be hardly learned by a data-driven model. We list three representative publicly available tools for different application contexts:

- **DeepFake Detector**: An AI service to judge whether a given video contains deepfake manipulated faces, developed within WeVerify Project[6]. It uses the URL of a suspicious image or video as the input and returns the deepfake probability score. It could help detect generation traces discussed in Sec. 3.1.2.
- **Reverse Image Search**: Web search services with an image as a query. By submitting extracted keyframes, we could check if there exist similar videos elsewhere before. This could help detect the manipulation by comparison [22] and identify the original source, as discussed in Secs. 3.1 and 3.2. Many general search engines (e.g., Google and Baidu) provide such services. Task-specific tools include TinEye, ImageRaider, and Duplichecker[7].
- **Video Verification Plugin**: A Google Chrome extension to verify videos provided by the InVID European Project[8]. It provides a toolbox to obtain contextual information from YouTube or Facebook, extract keyframes for reverse search, show metadata, and perform forensic analysis, which is useful for the detection at the signal and semantic levels (Secs. 3.1 and 3.2).

[2]https://mklab.iti.gr/results/fake-video-corpus/
[3]https://social-comp.github.io/YouTubeAudit-data/
[4]https://github.com/ICTMCG/FakeSV
[5]https://github.com/FuxiaoLiu/Twitter-Video-dataset
[6]https://weverify.eu/tools/deepfake-detector/
[7]https://tineye.com/, https://infringement.report/api/raider-reverse-image-search/, & https://www.duplichecker.com/
[8]https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

Table 2: Misinformation video detection vs. related areas.

Misinformation Detection vs. Deception Detection	Deceptive techniques could be used to convey fabricated information, but non-intentionally created misinformation may not correspond to deceptive behaviors.
Misinformation Detection vs. Harmful Content Detection	Harmful content is more likely to cause mental harm to specific populations or figures, while misinformation is meant to mislead readers.
Misinformation Detection vs. Clickbait Detection	Clickbait could be seen as a manifestation of misinformation materials.

5 RELATED AREAS

In this section, we discuss three areas related to misinformation video detection and clarify the relationships between misinformation video detection and these areas (Table 2).

Deception Detection: It aims at identifying the existence of deceptive behaviors, which is crucial for personal and public safety. In earlier research, both verbal and nonverbal cues play important roles in deception detection [7]. Verbal cues mainly refer to the linguistic characteristics of the statement while non-verbal cues include neurological, visual, and vocal indicators.

Harmful Content Detection: Harmful content generally renders as doxing, identity attack, identity misrepresentation, insult, sexual aggression, and the threat of violence [10]. Detecting video-based harmful content often relies on capturing indicative features in multiple modalities, such as linguistic features of audio transcription, video sentiment, and flagged harmful objects [6].

Clickbait Detection: Clickbait is a term commonly used to describe eye-catching and teaser headlines (thumbnails) in online media [63]. Clickbait video detection is to determine whether a video is faithfully representing the event it refers to. Content-based methods focus on analyzing the semantic gaps between the initially presented information (e.g., title and video thumbnail) and that expressed by the whole video, while others exploited creator profiles and audience feedback [71].

6 OPEN ISSUES AND FUTURE DIRECTIONS

Though existing works have demonstrated significant advances in detecting misinformation videos, there are still many issues that barrier their application to real-world systems. Here we present two open issues and two future directions along with concrete tasks to advance the landscape of practical detection systems.

6.1 Open Issues

6.1.1 Transferability. Transferability reflects how well a detection system tackles data distribution shift which is common and inevitable in real-world applications. Despite being a hot research topic, this issue remains largely underexplored in misinformation video detection, which is a crucial barrier for detection methods

to be put into practice. Here we provide three transfer-related subproblems in different aspects:

1) Multi-platform detection. The differences in contents and user groups among platforms shape different social contexts and provide extra clues, which has been found helpful in detection [39, 51]. This is important for this area given that cross-platform sharing (re-uploading) is common for videos. However, the principle of tackling multi-platform distribution gaps remains unclear.

2) Multi-domain detection. Misinformation texts in different news domains have different word use and propagation patterns, leading to data shifts [44, 60, 89]. Therefore, the investigation and mitigation of the domain gap for video misinformation is a promising direction.

3) Temporal generalization. Distribution shift over time is unavoidable for online video data. Effective features on past data might perform poorly online [27, 43, 88]. How to find stable features and rapidly adapt to current video data require further exploration.

6.1.2 Explainability. Most existing methods focus on improving accuracy and neglect the importance of providing an explanation. Without explanations aligned with human expectations, human users could hardly learn about the strengths and weaknesses of a detection system and decide when to trust it. This issue should be tackled in two aspects:

1) Distinguishing fine-grained types of misinformation. In addition to binary classification, the model should further predict a concrete type of detected misinformation samples (e.g., video misuse). This requires a new taxonomy and fine-grained annotation on datasets.

2) Multi-modal clue attribution. The detection model should attribute its output to the multi-modal clues extracted. Due to the complicated characteristics and fine-grained types of misinformation videos, a single piece of misinformation video, on the one hand, may only have a few clues. On the other hand, normal real videos can also contain some aforementioned features, e.g. editing traces, and propaganda intentions. It is important to disentangle the clue integration process to justify the final output by clue attribution, whether the video is real or fake.

6.2 Future Directions

6.2.1 Clue Integration & Reasoning. The diversity of involved modalities in a video post requires the detection model to have a higher clue integration and reasoning ability than that for text- and image-based detection. In most cases, the final judgment of misinformation depends on neither a single modality nor all modalities, and finding out effective combinations is non-trivial. For example, for a previously recorded accident video that is repurposed to be the scene of a new accident and added background music, what is crucial for judgment is the mismatch between video and text, rather than that between video and audio.

However, clue integration in this area is typically accomplished by directly aligning and fusing all representation vectors obtained from different modalities, which makes it hard for models to learn to reason among modalities. We believe that enabling reasoning among modalities will be important for better clue integration and more flexible detection. The possible directions include:

1) Inter-modality relationship modeling. Following tasks requiring reasoning ability like visual question answering [23], one can build graphs to guide interaction among modalities.

2) Problem decomposition. By transforming the detection as a mixture of several subproblems, one can use Chain/Tree of Thoughts [78, 83] to prompt large language models (e.g., GPT-4 [46]) to reason. This is useful for video claim verification when combined with external knowledge sources like online encyclopedias, fact-checking reports and relevant news (similar to that in text and image format [3, 61, 62, 70]).

6.2.2 Recommendation-Detection Coordination. The coordination between recommendation-based video distribution and misinformation video detection is crucial for practical systems, whose ultimate goal is to keep recommending videos that are of interest to users while avoiding misinforming them. To achieve this, detection systems are expected to contain different models and strategies to exploit rich side information from recommender systems as well as make recommendations more credible. Here we provide three concrete coordination scenarios:

1) User-interest-aware detection. The viewing history of the videos reflects not only users' interests but also how susceptible they are to specific topics (e.g., elections). Therefore, we could prioritize these recommended videos and detect misinformation with awareness of topics (a similar case for text fake news is [75]).

2) User-feedback-aware detection. Feedback from the crowd to the platform might be valuable indicators of suspicious videos. A recent example is to use users' reports of misinformation as weak supervision in text-based fake news detection [76]. Using more user feedback derived from recommender systems like expressions of dislike due to factuality issues will be a promising direction.

3) Credibility-aware recommendation. Considering information credibility in recommender systems can mitigate the exposure of misinformation videos and make the recommendation more accountable. A possible solution is to include misinformation video detection as an auxiliary task or use a well-trained detector as a critic to provide feedback.

7 CONCLUSION

We surveyed the existing literature on misinformation video detection and provided an extensive review of the advanced detection solutions, including clues at the signal, semantic, and intent levels and clue integration techniques. We also summarized publicly available datasets and tools and discussed related areas. Furthermore, we presented critical open issues and future directions for both research and real-world applications. Also, we open-sourced a GitHub repository that will be updated to include future advances in this area. We hope this survey could shed light on further research for defending against misinformation videos.

ACKNOWLEDGMENTS

This research is supported in part by the National Key Research and Development Program of China (2022YFC3302102), the National Natural Science Foundation of China (62203425), the Project of Chinese Academy of Sciences (E141020), the China Postdoctoral Science Foundation (2022TQ0344), and the International Postdoctoral Exchange Fellowship Program by Office of China Postdoc Council (YJ20220198).

REFERENCES

[1] Javad Abbasi Aghamaleki and Alireza Behrad. 2017. Malicious Inter-Frame Video Tampering Detection in MPEG Videos Using Time and Spatial Domain Analysis of Quantization Effects. *Multimedia Tools and Applications* 76 (2017), 20691–20717.

[2] Sara Abdali. 2022. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. *arXiv:2203.13883* (2022).

[3] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14940–14949.

[4] Ronak Agrawal and Dilip Kumar Sharma. 2021. A Survey on Video-Based Fake News Detection Techniques. In *2021 8th International Conference on Computing for Sustainable Global Development*. 663–669.

[5] Omar Ismael Al-Sanjary, Ahmed Abdullah Ahmed, Adam Amril Bin Jaharadak, Musab AM Ali, and Hewa Majeed Zangana. 2018. Detection Clone An Object Movement Using An Optical Flow Approach. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics*. 388–394.

[6] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. 6625–6643.

[7] Haya Alaskar, Zohra Sbaï, Wasiq Khan, Abir Hussain, and Arwa Alrawais. 2022. Intelligent Techniques for Deception Detection: A Survey and Critical Study. *Soft Computing* (2022), 1–20.

[8] Shivangi Aneja, Cise Midoglu, Duc-Tien Dang-Nguyen, Michael Alexander Riegler, Paal Halvorsen, Matthias Nießner, Balu Adsumilli, and Chris Bregler. 2021. MMSys' 21 Grand Challenge on Detecting Cheapfakes. *arXiv:2107.05297* (2021).

[9] Neema Antony and Binet Rose Devassy. 2018. Implementation of Image/Video Copy- Move Forgery Detection Using Brute-Force Matching. In *2018 2nd International Conference on Trends in Electronics and Informatics*. 1085–1090.

[10] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the 4th Workshop on Online Abuse and Harms*. 125–137.

[11] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the Role of Visual Content in Fake News Detection. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 141–161.

[12] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 2019. Everybody Dance Now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5933–5942.

[13] Hyewon Choi and Youngjoong Ko. 2021. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2950–2954.

[14] Hyewon Choi and Youngjoong Ko. 2022. Effective Fake News Video Detection Using Domain Knowledge and Multimodal Data Fusion on YouTube. *Pattern Recognition Letters* 154 (2022), 44–52.

[15] Christos Christodoulou, Nikos Salamanos, Pantelitsa Leonidou, Michail Papadakis, and Michael Sirivianos. 2023. Identifying Misinformation on YouTube through Transcript Contextual Analysis with Transformer Models. *arXiv:2307.12155* (2023).

[16] Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited Media Understanding Frames: Reasoning About the Intent and Implications of Visual Misinformation. 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)*. 2026–2039.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[19] Sondos M Fadl, Qi Han, and Qiong Li. 2018. Authentication of Surveillance Videos: Detecting Frame Duplication Based on Residual Frame. *Journal of Forensic Sciences* 63, 4 (2018), 1099–1109.

[20] Ryoya Furukawa, Daiki Ito, Yuta Takata, Hiroshi Kumagai, Masaki Kamizono, Yoshiaki Shiraishi, and Masakatu Morii. 2022. Fake News Detection via Biased User Profiles in Social Networking Sites. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 136–145.

[21] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3108–3127.

[22] Dhanvi Ganti. 2022. A Novel Method for Detecting Misinformation in Videos, Utilizing Reverse Image Search, Semantic Analysis, and Sentiment Comparison of Metadata. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4128499.

[23] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12746–12756.

[24] Ankur Gupta, Neeraj Kumar, Purnendu Prabhat, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Combating Fake News: Stakeholder Interventions and Potential Solutions. *IEEE Access* 10 (2022), 78268–78289.

[25] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5039–5049.

[26] Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *2019 International Conference on Multimodal Interaction*. 235–243.

[27] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 116–125.

[28] Juan Hu, Xin Liao, Wei Wang, and Zheng Qin. 2021. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1089–1102.

[29] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[30] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. 2021. Misinformation Detection on YouTube Using Video Captions. *arXiv:2107.00941* (2021).

[31] Pamela Johnston, Eyad Elyan, and Chrisina Jayne. 2020. Video Tampering Localization Using Features Learned from Authentic Content. *Neural Computing and Applications* 32 (2020), 12243–12257.

[32] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors. In *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*. 7–15.

[33] Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A CNN-based Misleading Video Detection Model. *Scientific Reports* 12, 1 (2022), 1–9.

[34] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 178–188.

[35] Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. 2022. Fake Audio Detection Based On Unsupervised Pretraining Models. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. 9231–9235.

[36] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops*. 83–92.

[37] Katerina Eva Matsa. 2022. More Americans are getting news on TikTok, bucking the trend on other social media sites. https://www.pewresearch.org/fact-tank/2022/10/21/more-americans-are-getting-news-on-tiktok-bucking-the-trend-on-other-social-media-sites/.

[38] Scott McCrae, Kehan Wang, and Avideh Zakhor. 2022. Multi-modal Semantic Inconsistency Detection in Social Media News Posts. In *MultiMedia Modeling: MMM 2022*. 331–343.

[39] 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*, Vol. 16. 651–662.

[40] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781* (2013).

[41] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys* 54, 1 (2021), 1–41.

[42] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2823–2832.

[43] Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about Time: Rethinking Evaluation on Rumor Detection Benchmarks using Chronological Splits. In *Findings of the Association for Computational Linguistics: EACL 2023*. 736–743.

[44] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-Domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.

[45] Arjun Narayan. 2020. Our efforts towards fighting misinformation in times of COVID-19. https://newsroom.tiktok.com/en-in/our-efforts-towards-fighting-misinformation-in-times-of-coronavirus.

[46] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774* (2023).

[47] Ciarán O'Connor. 2021. How TikTok Sounds Are Used to Fuel Anti-Vaccine Fears. https://www.isdglobal.org/digital_dispatches/how-tiktok-sounds-are-used-to-fuel-anti-vaccine-fears/.

[48] Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading Metadata Detection on YouTube. In *Advances in Information Retrieval: ECIR 2019*. 140–147.

[49] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. Web video verification using contextual cues. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*. 6–10.

[50] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[51] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14444–14452.

[52] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-Enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212—-1220.

[53] Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11947–11959.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. 8748–8763.

[55] Paulo Max Gil Innocencio Reis, João Paulo Carvalho Lustosa da Costa, Ricardo Kehrle Miranda, and Giovanni Del Galdo. 2016. ESPRIT-Hilbert-Based Audio Tampering Detection With SVM Classifier for Forensic Analysis via Electrical Network Frequency. *IEEE Transactions on Information Forensics and Security* 12, 4 (2016), 853–864.

[56] Mubbashar Saddique, Khurshid Asghar, Usama Ijaz Bajwa, Muhammad Hussain, and Zulfiqar Habib. 2019. Spatial Video Forgery Detection and Localization using Texture Analysis of Consecutive Frames. *Advances in Electrical and Computer Engineering* 19, 3 (2019), 97–108.

[57] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

[58] 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*. 899–908.

[59] Nitin Arvind Shelke and Singara Singh Kasana. 2021. A Comprehensive Survey on Passive Techniques for Digital Video Forgery Detection. *Multimedia Tools and Applications* 80 (2021), 6247–6310.

[60] Qiang Sheng, Juan Cao, H Russell Bernard, Kai Shu, Jintao Li, and Huan Liu. 2022. Characterizing Multi-Domain False News and Underlying User Effects on Chinese Weibo. *Information Processing & Management* 59, 4 (2022), 102959.

[61] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4543–4556.

[62] Qiang Sheng, Juan Cao, Xueyao Zhang, Xirong Li, and Lei Zhong. 2021. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. 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)*. 5468–5481.

[63] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[64] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*.

[65] K Sitara and BM Mehtre. 2017. A Comprehensive Approach for Exposing Inter-Frame Video Forgeries. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications*. 73–78.

[66] Matthew S Smith. 2022. TikTok vs. YouTube: Which One Is Right for You? https://www.lifewire.com/tiktok-vs-youtube-6665455.

[67] Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1412–1423.

[68] S Shyam Sundar, Maria D Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication* 26, 6 (2021), 301–319.

[69] Nesrine Tarhouni, Salma Masmoudi, Maha Charfeddine, and Chokri Ben Amar. 2022. Fake COVID-19 videos detector based on frames and audio watermarking. *Multimedia Systems* (2022), 1–15.

[70] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 809–819.

[71] Deepika Varshney and Dinesh Kumar Vishwakarma. 2021. A Unified Approach for Detection of Clickbait Videos on YouTube using Cognitive Evidences. *Applied Intelligence* 51 (2021), 4214–4235.

[72] Sukrit Venkatagiri, Joseph S Schafer, and Stephen Prochaska. 2023. The Challenges of Studying Misinformation on Video-Sharing Platforms During Crises and Mass-Convergence Events. *arXiv:2303.14309* (2023).

[73] Kehan Wang, David Chan, Seth Z Zhao, John Canny, and Avideh Zakhor. 2022. Misinformation Detection in Social Media Video Posts. *arXiv:2202.07706* (2022).

[74] Kehan Wang, Seth Z Zhao, David Chan, Avideh Zakhor, and John Canny. 2022. Multimodal Semantic Mismatch Detection in Social Media Posts. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing*. 1–6.

[75] Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. 2022. Veracity-aware and Event-driven Personalized News Recommendation for Fake News Mitigation. In *Proceedings of the ACM Web Conference 2022*. 3673–3684.

[76] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak Supervision for Fake News Detection via Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 516–523.

[77] Claire Wardle. 2017. Fake News. It's Complicated. https://firstdraftnews.org/articles/fake-news-complicated/.

[78] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35.

[79] Chloe Wittenberg, Ben M Tappin, Adam J Berinsky, and David G Rand. 2021. The (Minimal) Persuasive Advantage of Political Video over Text. *Proceedings of the National Academy of Sciences* 118, 47 (2021), e2114388118.

[80] Junichi Yamagishi. 2022. Lessons Learned from ASVSpoof and Remaining Challenges. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 1–2.

[81] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8261–8265.

[82] Yuxing Yang, Junhao Zhao, Siyi Wang, Xiangyu Min, Pengchao Wang, and Haizhou Wang. 2023. Multimodal Short Video Rumor Detection System Based on Contrastive Learning. *arXiv:2304.08401* (2023).

[83] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. [n. d.]. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. ([n. d.]).

[84] Ning Yu, Larry S Davis, and Mario Fritz. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7556–7566.

[85] Markos Zampoglou, Foteini Markatopoulou, Gregoire Mercier, Despoina Touska, Evlampios Apostolidis, Symeon Papadopoulos, Roger Cozien, Ioannis Patras, Vasileios Mezaris, and Ioannis Kompatsiaris. 2019. Detecting Tampered Videos with Multimedia Forensics and Deep Learning. In *MultiMedia Modeling: MMM 2019*. 374–386.

[86] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys* 53, 5 (2020), 1–40.

[87] Yipin Zhou and Ser-Nam Lim. 2021. Joint Audio-Visual Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14800–14809.

[88] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the Future: Mitigating Entity Bias in Fake News Detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2120–2125.

[89] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-Guided Multi-View Multi-Domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering* (2022).

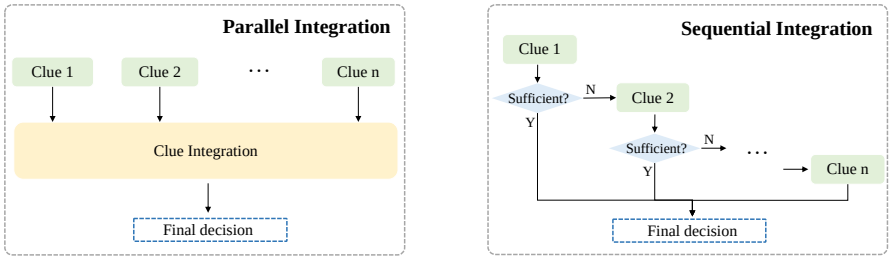
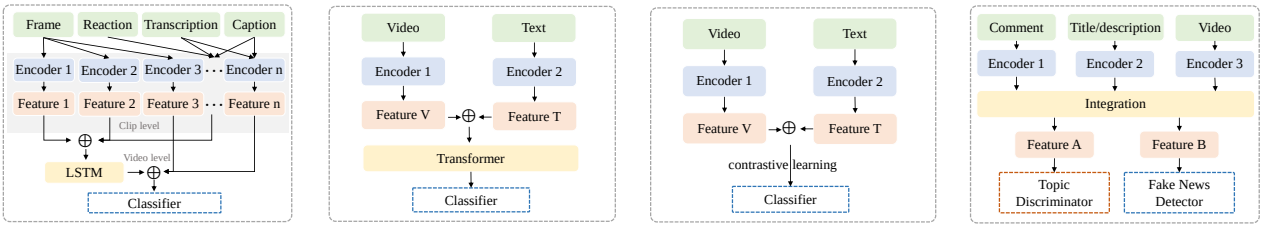


Figure 3: Comparison of the parallel and sequential clue integration in misinformation video detection.



(a) McCrae et al. [38] **(b) Wang et al. [73] (MLM)** **(c) Wang et al. [73] (CL)** **(d) Choi and Ko [13]**

Figure 4: Illustrations of representative misinformation video detection models.

Table 3: Overview of datasets for misinformation video detection. Note that Christodoulou et al. [15] conducted experiments on a subset of YouTubeAudit. SOTA: state-of-the-art.

Dataset	Sources	Domain	Size	Released?	SOTA Method	SOTA Performance (%) Accuracy	F1
FVC	YouTube, Twitter, Facebook	various	5,006	Y	FANVM [13]	––	85.84
VAVD	YouTube	various	546	Y	FANVM [13]	––	86.71
YouTubeAudit	YouTube	Conspiracy theories	2,943	Y	(subset) RoBERTa [15]	94.00	94.00
FakeSV	Douyin, Kuaishou	various	5,538	Y	SVFND+NEED [53]	84.62	84.61
COVID-VTS	Twitter	COVID-19	10,000	Y	TwtrDetective [34]	68.10	67.90
(Hou et al. 2019)	YouTube	Prostate cancer	250	N	(Hou et al.) [26]	74.41	––
(Serreno et al. 2020)	YouTube	COVID-19	180	N	(Serreno et al.) [57]	84.40	––
(Jagtap et al. 2021)	YouTube	various	2,125	N	(Jagtap et al.) [30]	92.00	78.00
(Shang et al. 2021)	YouTube	COVID-19	891	N	TikTec [58]	72.31	60.51
(Choi and Ko 2021)	YouTube	various	4,622	N	FANVM [13]	––	86.28
(Choi and Ko 2022)	YouTube	various	2,912	N	FVDM [14]	––	82.55
(Li et al. 2022)	Bilibili	Health	700	N	(Li et al.) [33]	90.00	89.00
(Wang et al. 2022a)	Twitter	various	160,000	N	(Wang et al.) [73]	71.07	––
(Wang et al. 2022b)	Twitter	various	943,667	N	(Wang et al.) [74]	85.43	––
(McCrae et al. 2022)	Facebook, YouTube	various	4,651	N	(McCrae et al.) [38]	60.50	––
(Yang et al. 2023)	Douyin, Kuaishou	various	8,213	N	(Yang et al) [82]	87.71	87.38

A ADDITIONAL ILLUSTRATIONS

For better clarification, we visualize and compare the processes of the two paradigms of clue integration, as presented in Fig. 3. Since most existing works adopt parallel integration, here we present four representative models to help readers gain an intuitive understanding of detection frameworks. Figs. 4a and 4b respectively demonstrate the concatenation-based fusion and the utilization of attention mechanisms to model cross-modal interactions. Figs. 4c and 4d respectively showcase the employment of auxiliary objectives like contrastive learning and topic classification to optimize the learning of multi-modal features.

B OVERVIEW OF DATASETS

In Sec. 4.1, we delve into the detailed description of four large publicly available datasets. However, it is worth noting that there are several small-scale and topic-specific datasets that have not been open-sourced but still contribute significantly to the research community. Table 3 summarizes all these datasets, including their data source, topic domain, size, accessibility, state-of-the-art(SOTA) methods, and the corresponding performance based on accuracy and F1 scores.