

# Video Misinformation Detection: A Systematic Review of Manipulation Tactics, Datasets and Algorithms

FLORIAN BRAUN, The Graduate University for Advanced Studies, SOKENDAI, Japan and National Institute of Informatics, Japan

TRUC HOANG, Independent Researcher, Vietnam

GIANLUCA DEMARTINI, The University of Queensland, Australia

ISAO ECHIZEN, National Institute of Informatics, Japan

TIMO SPINDE, National Institute of Informatics, Japan

Misinformation research has identified diverse strategies through which false or misleading content gains credibility, spreads, and influences audiences. Yet no systematic review has synthesized how computer science research on video misinformation detection defines and engages with different misinformation strategies. We therefore systematically review 51 of over 2200 retrieved computer science papers, organizing them by misinformation strategy, dataset and label design, and detection method. We find that while model architectures address specific misinformation strategies, prevailing binary real/fake labels obscure whether each strategy is actually detected, limiting the transparency and diagnostic value of current evaluations. These findings motivate strategy-aware video misinformation detection.

CCS Concepts: • **Computing methodologies** → **Computer vision tasks**; • **General and reference** → **Surveys and overviews**; • **Applied computing** → **Law, social and behavioral sciences**.

Additional Key Words and Phrases: Misinformation, Disinformation, Fake News, Video, Multimodal, Detection, Datasets, Evaluation, Systematic Literature Review, Prisma

## ACM Reference Format:

Florian Braun, Truc Hoang, Gianluca Demartini, Isao Echizen, and Timo Spinde. 2026. Video Misinformation Detection: A Systematic Review of Manipulation Tactics, Datasets and Algorithms. *J. ACM* 37, 4, Article 111 (August 2026), 35 pages. <https://doi.org/XXXXXX.XXXXXXX>

## 1 Introduction

While misinformation is not a new phenomenon, the widespread adoption of the internet and social media drastically increased the reach and impact of misinformation [41, 93], especially by packaging the misinformation into richer modalities, such as videos [27, 83]. For example, a supposed documentary video from 2020 called *Plandemic* spread misleading claims about the origins and transmission of Covid-19, attributing them to a government-led strategic campaign. The video reached over eight million views across several social media platforms, including Facebook, YouTube

---

This work was partially supported by JSPS KAKENHI Grant JP24H00732, by JST CREST Grants JPMJCR20D3 and JPMJCR2562 including AIP challenge program, and by JST K Program Grant JPMJKP24C2 Japan.

It was also funded by the German Federal Ministry of Education and Research (BMBF) through the DAAD (German Academic Exchange Service).  
Authors' Contact Information: Florian Braun, braun-florian@nii.ac.jp, The Graduate University for Advanced Studies, SOKENDAI, Hayama, Japan and National Institute of Informatics, Tokyo, Japan; Truc Hoang, Independent Researcher, Ho Chi Minh, Vietnam, hoangthuytruc@gmail.com; Gianluca Demartini, The University of Queensland, Brisbane, QLD, Australia, g.demartini@uq.edu.au; Isao Echizen, National Institute of Informatics, Tokyo, Japan, iechizen@nii.ac.jp; Timo Spinde, National Institute of Informatics, Tokyo, Japan, T.Spinde@media-bias-research.org.

---

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
Manuscript submitted to ACM

and Twitter, undermining information from authorities [23, 60, 82]. While *Plandemic* focused on health misinformation, video misinformation (i.e., false or inaccurate information [57] conveyed through video) is not restricted to this domain and has, among other areas, undermined open discourse [20] and public trust in science [95].

To combat video misinformation, fact-checkers such as Politifact<sup>1</sup> and Snopes<sup>2</sup> manually inspect posts. However, manual fact-checking cannot address the volume of content uploaded to social media platforms [22], and thus video misinformation detection must be automated to some degree. Social media platforms already deploy automated detection, but many do not disclose their methods, data, or evaluation results [69]. Thus, developing open and transparent detection is largely left to the research community, which has proposed many detection algorithms and datasets, each grounded in specific assumptions about how misinformation is conveyed or how misinformation videos differ from truthful videos [11, 53, 99]. Yet, no framework connects these assumptions to the detection strategies they motivate. As a result, it is difficult to determine whether an algorithm is suited to the misinformation it is applied to, or whether it is evaluated on the misinformation it was designed to detect. Micallef et al. [56] are the first to split the evaluation of misinformation detection approaches into misinformation types. Their results suggest that linking misinformation types to detection approaches clarifies how methods work and where they fail, but no systematic account of the field exists.

Therefore, our goal is to connect each detection method to the type of misinformation it is designed to detect. Thus, we conduct a systematic literature review of the field. We first investigate the conceptions of misinformation used in computer science video misinformation detection papers. We then analyze what datasets and algorithms exist, and how they model misinformation. To this end, we systematically screened over 2200 papers and reviewed 51 eligible papers published in A\*, A, or Q1 outlets between January 2015 and December 2025. Our contributions are threefold:

- C1 We propose a framework linking how videos convey misinformation with existing datasets and detection algorithms,
- C2 We provide the first systematic literature review focusing on video misinformation detection algorithms,
- C3 We identify shortcomings and gaps in the current literature, most significantly, how current evaluation practice lacks the granularity to diagnose what signals models actually learn to detect.

By linking how misinformation is conveyed with how it is detected, our review offers a reference point for researchers designing new detection algorithms, selecting or building datasets, and situating their work within the broader landscape of video misinformation detection. It further highlights where current approaches fall short: most detection paradigms address only a subset of ways that misinformation can be created, leaving ways without signal-level artifacts largely unaddressed; evaluation relies on binary in-the-wild datasets that cannot reveal whether a model detects the mechanism it claims to target; and this disconnect is not yet recognized as a limitation within the field.

## 2 Background

### 2.1 Concepts & Definitions

Because the key terms describing misinformation and related concepts are frequently used with different meanings [93], we first review the terms *Misinformation*, *Disinformation* and *Fake news* in Section 2.1, and specify how we use them throughout this review. To further clarify the scope of this review, we also briefly introduce *tampering detection*.

*Misinformation & Disinformation.* The terms misinformation and disinformation are mainly defined in two ways. Wardle and Derakhshan define misinformation as "information that is false, but not created with the intention of causing harm" [93, p. 20] and disinformation as "information that is false and deliberately created to harm [...]" [93, p.20]. While

<sup>1</sup><https://www.politifact.com/>, last accessed 01.03.2026

<sup>2</sup><https://www.snopes.com/>, last accessed 01.03.2026

misinformation and disinformation both involve false information, they are distinguished by the intent with which the information is conveyed. In contrast, the UNHCR [57] defines disinformation as a subset of misinformation. Similarly to Wardle and Derakhshan [93], they define disinformation as “deliberate and [...] malicious content”, but misinformation as “false or inaccurate information”, which does not assess the harm or intent of the misinformation [57].

Prior surveys on misinformation detection show that the definitional boundaries remain blurry in computer science as well. While Alam et al. [6] follow the misinformation definition proposed by Wardle and Derakhshan [93], they note that the factuality of content is usually studied separately from its harmfulness. Bu et al. [10] also confirm that factuality is studied without accounting for harmfulness. Consequently, they define the term “misinformation video” to describe “a video post that conveys false, inaccurate, or misleading content” [10], disregarding the harmfulness of the content.

To keep this review consistent with the analyzed papers in the computer science domain, we use *misinformation* to refer to any content that misleads the recipient, regardless of intent.

*Fake News.* Another related concept is *fake news*, which rose in popularity significantly during the 2016 US presidential elections [26, 85]. Tandoc et al. [85] summarizes fake news definitions from academic literature into six distinct categories, showing that the term has many different facets. Molina et al. [58] confirm that fake news is a term that is defined in many different ways across different fields. However, the term *fake news* is also used to discredit news outlets and people [9, 58, 93]. Because *fake news* is ill-defined and used with negative connotation, we avoid the term *fake news*.

*Tampering Detection.* A related concept to misinformation detection is the detection of tampered or generated content, which has been extensively covered in existing literature reviews [5, 76]. Fields such as deepfake detection [70], AI-generated content detection [97], copy-and-move detection [68], and splicing detection [40] fall under tampering detection. While often motivated as a way to fight disinformation, the algorithms check for structural alteration rather than the veracity of the content [10, 75]. Therefore, we do not cover tampering detection algorithms in this review.

## 2.2 Taxonomies & Frameworks

Because different fields approach misinformation from different perspectives and with different goals, misinformation is described through different taxonomies [2, 9, 10, 77]. Social scientists and political actors create taxonomies to describe misinformation theoretically [2, 42, 57, 81, 93]; fact-checkers to communicate the severity of misinformation to readers [34, 77]; and computer scientists to describe how misinformation is created from a technical perspective, or how it can be detected in multimedia content [4, 42, 81]. To connect how misinformation is conveyed with how it is detected, we first need to understand how prior work conceptualizes misinformation itself. We therefore provide an overview of how existing taxonomies classify misinformation. Because taxonomies from different fields can differ substantially depending on the purposes for which they were designed, we group them into three categories based on those purposes: (I) theoretical frameworks, (II) operational frameworks, and (III) technical frameworks. For each group, we extract the categories defined by the taxonomies and organize them into related concepts, as shown in Table 1.

*I. Theoretical frameworks.* Theoretical frameworks describe misinformation along dimensions such as how it is created, its degree of truthfulness, its intent and target, and its recurring themes [2, 42, 57, 81, 93]. Some taxonomies examine how misinformation is created, e.g., fabrications, false connections, false context, or clickbait. Others identify concepts that lie between truthful and false information, such as hoaxes (half-true information), rumors (ambiguous), or one-sided arguments (leaving out the whole picture) [4, 42, 81]. Frameworks also categorize by the intent behind

Table 1. Overview: Taxonomies and Frameworks on Misinformation

Group & Framework	Key Concepts & Definitions
<b>I. Theoretical Frameworks</b>	
<i>Conceptual Mechanics</i>	<b>Misinformation mechanic:</b> Fabricated (created/false), Manipulated (distorted), False Connection (headlines, visuals or captions don't support content) [2, 57, 93], Impostor (fake source) [57, 93], False Context (mismatched contextual information) [2, 57, 81, 93], Clickbait (exaggerated headlines) [4, 42], Polarizing (emotionally charged and one-sided) [58, 81], Misreporting (unintentionally false information) [58]. <b>Degree of truth:</b> Hoaxes (half-truths), Rumors (ambiguous) [4, 42], Biased (One-sided reporting) [42, 81].
<i>Intent</i>	<b>Misinformation goal:</b> Managing attitudes/values [4, 57], Flooding with contradictions, Attacking critical voices, Satire/Parody (no intent to misinform) [2, 57, 93], Persuasive Information (intent to change opinion) [58].
<i>Content</i>	<b>Information conveyed:</b> Conspiracy Theories [4, 42], Pseudoscience, Fake Reviews [42, 81], Ads disguised as news [57].
<b>II. Operational Frameworks</b>	
<i>Veracity Scales &amp; Labels</i>	<b>5-Point Scale:</b> True → Mostly True → Mixture / Half True → Mostly False → False [34, 77]. <b>Specific Labels:</b> "Pants on Fire" (false and ridiculous claim) [34]; Evidentiary (Unproven/Unfounded) & Error types (Miscaptioned, Fake, Wrong Attribution) [77].
<b>III. Technical Frameworks</b>	
<i>Content fabrication</i>	<b>Video Editing:</b> Missing Context (Misrepresentation, Isolation), Deceptive Editing (Omission, Splicing), Malicious Transformation (Doctoring, Fabrication) [86].
<i>Detection Levels (Computer Science)</i>	<b>Content-Centric:</b> Signal-level (artifacts) & Semantic-level (meaning) [10, 97]; News Content-based detection [4]. <b>Human-Centric:</b> Intent-level (motivation) [10], Perceptual (realism) & Human (emotions/behavior) levels [97], Social Context-based detection [4]. <b>Hybrid:</b> Approaches combining content and context analysis [4].
<i>Cross-modal relationship (Computer Science)</i>	<b>Cross-Modal Analysis:</b> Veracity of Text, Veracity of Video, Relationship of Text & Video, Veracity of Text & Video [56].

the misinformation, for instance Satire/Parody as false information without intent to harm [2, 57, 93]. Finally, some categories distinguish the content conveyed, such as conspiracy theories [4, 42], pseudoscience, or fake reviews [42, 81].

*II. Operational Frameworks.* Operational frameworks are often used by fact-checking agencies, such as Snopes [77] and Politifact [34]. Both Snopes and Politifact use a 5-point scale for denoting how true or false the claims are, rather than binary true/false labels. Both also employ further categories for edge cases such as ridiculous claims [34] or insufficient evidence [77]. Snopes furthermore denotes frequently occurring misinformation types, which are similar to the conceptual mechanics in theoretical frameworks [77]. These ratings usually summarize an in-depth article that describes how the video misinforms, with the goal of quickly communicating the gist of the review to the recipient.

*III. Technical Frameworks.* Lastly, technical frameworks are categorized by what part of the misinformation pipeline they address: some describe how misleading content is created, others bridge content characteristics with detection approaches, and others classify detection algorithms. An example of the first type is the video creation and manipulation framework by The Washington Post<sup>3</sup>, which describes how misinforming videos are fabricated [86]. While this taxonomy captures how misinforming videos are produced, it does not address how they can be detected.

To bridge the gap between misinformation and its detection, Micallef et al. [56] link content characteristics to detection performance by investigating how misinformation is conveyed through different modalities in social media posts. The resulting taxonomy categorizes misinformation by analyzing the veracity of the modalities individually, the

<sup>3</sup><https://www.washingtonpost.com/>

relationships between the modalities, and the veracity of the message conveyed by the combined modalities. Using their taxonomy, they can investigate on what misinformation specific detection algorithms succeed or fail. However, their analysis includes only a selection of text-based classifiers for post body, title, description, and speech transcript.

To our knowledge, the survey by Bu et al. [10] is the only literature review published in a high-ranking venue that focuses exclusively on video misinformation detection, although its paper selection was not systematic, leaving it susceptible to selection bias. Other existing reviews on multimodal misinformation detection do not focus on video content [1, 6, 65]. As a result, they cover only a few video misinformation algorithms, describe them generically, and classify detection algorithms solely by their technical properties without engaging with the content the algorithms operate on. Typically, existing frameworks organize misinformation detection algorithms by the level at which the model operates (signal vs. semantic), the human component analyzed (sender intent, receiver perception, social context), or hybrid combinations [4, 10, 97]. None of these frameworks links models to the types of misinformation they detect.

### 2.3 Summary & Positioning

As shown in Section 2.2, existing frameworks on misinformation capture only specific aspects. Theoretical frameworks describe how misinformation is conveyed, operational frameworks describe how severe the misinformation is, and technical frameworks describe how video material is altered to mislead or how video misinformation detection operates. Micallef et al. [56] come closest to bridging content and detection, but only distinguish through which modalities the misinformation is conveyed, leaving finer-grained analysis unaddressed. This disconnect also extends to evaluation: widely used datasets for video misinformation detection typically frame detection as binary classification [11, 63, 66], assigning a single true/false label regardless of how the misinformation is conveyed. Without such a distinction, it is unclear whether evaluation results reflect an algorithm's ability to detect the specific strategies and artifacts it targets.

Our literature review aims to close that gap by classifying misinformation detection algorithms not only by how they detect misinformation, but by what type of misinformation they address. To do so, we analyze the rationale behind each detection algorithm, from feature extraction to algorithm output. We then compare these types of misinformation to taxonomies from other fields, which lets us identify misinformation types that lack detection algorithms. Finally, to assess whether and which types of misinformation the currently available data reflect, we also review datasets.

## 3 Methodology

As shown in Section 2.3, no existing framework links how video misinformation detection algorithms work to how the misinformation they target is conveyed. To close this gap, we need to understand three things: how current detection algorithms work, how they describe the misinformation they target, and how they logically link the two in existing work. Because training and evaluation datasets directly encode which forms of misinformation an algorithm learns to recognize, we additionally review the datasets used.

### 3.1 Research Questions

Following our goal of linking how video misinformation is conveyed with how it is detected in computer science, we formulate four research questions:

**RQ1** What misinformation strategies are described in video misinformation detection literature, and how do they create misinformation?

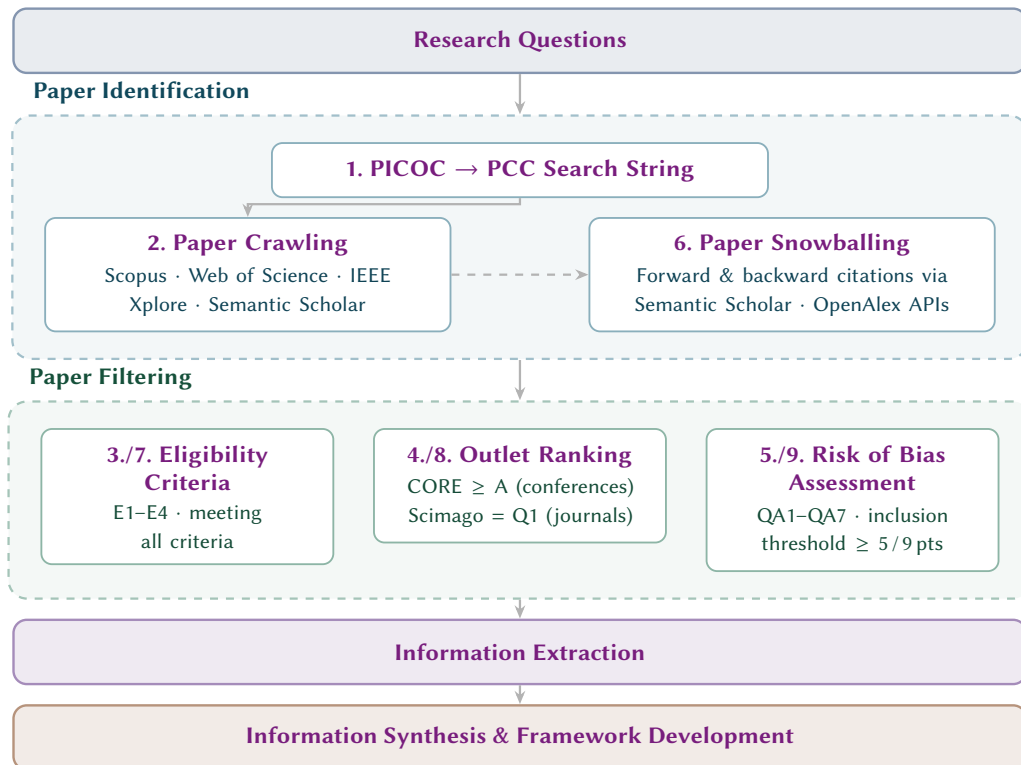


Fig. 1. Overview of the Methodology

- RQ2** What rationales do existing misinformation detection algorithms use to identify misinformation, and how do they implement the rationales?
- RQ3** What misinformation strategies are represented in existing datasets, and how are they used for training and evaluation?
- RQ4** How well do current detection approaches and datasets cover the identified misinformation strategies, and what gaps emerge?

Together, these questions trace the pipeline from how video misinformation deceives (RQ1), through how it is detected (RQ2) and what data is used to train the detection models (RQ3), to what gaps emerge when linking deception methods to their detection coverage (RQ4).

### 3.2 Study Selection

Because a goal of this review is to identify open gaps, the field must be representatively covered. Therefore, we followed a two-step paper identification approach to minimize oversights: In the first step, we queried bibliographic databases. In the second step, we conducted a snowball search, with the accepted papers from the first round as seed papers.

Table 2. PICOC scheme for identifying related papers. Italicized entries are not used in the query.

PICOC	Example	Synonyms
Problem	How misinformation is addressed in the literature	Misinformation, Fake news, Disinformation, Rumor
<i>Intervention</i>	<i>Content manipulation methods</i>	<i>Forgery, Cutting, Editing, Deepfake, Generation, Cheapfake</i>
Comparison	Algorithms, datasets	Algorithm, Detection, Classification, Dataset, Benchmark
<i>Outcome</i>	<i>Types of video misinformation</i>	<i>Fabrication, Manipulation, False Connection, False Context, Clickbait, Polarization, Impostor, Misreporting</i>
Context	Video, audiovisual	Video, Audiovisual, <i>Multimodal</i> , Movie, Recording, Clip, Footage

*Round 1: Structured Query.* For the first identification round, we used a structured query to search four databases. We first selected Scopus<sup>4</sup> and Web of Science<sup>5</sup> because their databases offer curated papers and complement each other [87]. However, we found that many of the papers listed in the survey by Bu et al. [10] were not found. Therefore, we also queried IEEE Xplore<sup>6</sup>, as many of the articles not found were published by IEEE, and Semantic Scholar<sup>7</sup>, because its paper database is very comprehensive [44] and allows for structured queries.

To set up the structured search query, we first defined a PICOC (Tab. 2) based on our research questions (Sec. 3.1). While we defined all five components to ensure comprehensive coverage of the research scope, we constructed the query using only Problem, Comparison, and Context. We excluded Intervention and Outcome because many detection algorithm papers do not explicitly state the content manipulation method or the type of misinformation they address, a gap this review aims to close. Including these terms would therefore systematically exclude relevant papers that lack explicit terminology for how misinformation is conveyed. We also made two adjustments to the initial search terms: Firstly, we excluded the term *Multimodal* from the context, because it is often used in papers on image-text misinformation, but yields no additional value when compared to the papers in Bu et al. [10]. Secondly, we added *Rumor* as one of the synonyms for misinformation, which is used synonymously by Papadopoulou et al. [64] and Bu et al. [10]. To build the query, we then combined synonyms with OR operators and components (i.e., P, C, C) with AND operators.

We restricted the publication period to studies published from 2015 onward, because one year later, the term *Fake News* gained prominence [26], and the Oxford Dictionary selected the term *Post-Truth* as “Word of the Year” [61].

*Round 2: Snowballing.* As mentioned in Section 2.1, the terms and concepts surrounding misinformation are often used ambiguously, which may lead to oversights when only structured queries are used to identify papers [59]. Therefore, we employed a second stage in which we identified papers using a single-hop citation graph. To find papers related to the papers identified by the query, we used Semantic Scholar and OpenAlex<sup>8</sup>, both of which provide free APIs to retrieve citations and references. Furthermore, they have a high forward citation coverage and a reasonable backward citation coverage [31]. Using the papers accepted in the first round as seed papers, we retrieved forward and backward citations and built a graph from the resulting connections. Next, we removed duplicate records by merging entries with identical DOIs. Finally, to focus the snowballing on papers with a strong topical connection to the reviewed literature, we removed all entries connected to only a single seed paper. This also reduced the candidate set to a manageable size.

### 3.3 Eligibility Criteria

To identify relevant papers, we filtered the identified papers according to the following exclusion criteria (Tab. 3):

<sup>4</sup><https://www.scopus.com/>

<sup>5</sup><https://www.webofscience.com/>

<sup>6</sup><https://ieeexplore.ieee.org/>

<sup>7</sup><https://www.semanticscholar.org/>

<sup>8</sup><https://openalex.org/>

Table 3. Eligibility Criteria

ID	Inclusion Criterion	Exclusion Criterion
<b>E1. Form</b>		
1a	Article written in English	Article in a language other than English
1b	Full paper published in a peer-reviewed journal or conference proceedings	Any other publication type (workshop paper, short paper, demo, abstract, presentation notes etc.)
1c	Paper has not been withdrawn	Paper has been retracted or withdrawn
<b>E2. Relevancy</b>		
2a	Study investigates video content or information related to videos (e.g., title, description, comments)	Study does not analyze video or any information connected to videos
2b	Study addresses misinformation	Study does not address misinformation or any related concept
<b>E3. Contribution Type</b>		
3a	Proposes or evaluates a dataset or a detection algorithm	Does not contribute scientific findings to detection datasets or algorithms
3b	Aims to improve the detection performance of misinformation detection algorithms	Primarily aims at improving non-detection aspects, such as computational efficiency or adversarial robustness.
<b>E4. Detection Methodology</b>		
4a	Proposes a novel <i>passive</i> detection algorithm (post-hoc analysis) or dataset	Proposes an <i>active</i> detection algorithm (e.g., provenance watermarking)

**E1) Form:** We exclude all papers that are not full papers at a conference or journal, not written in English, or withdrawn,

**E2) Relevancy:** We exclude all papers that are not centered on video-based misinformation,

**E3) Contribution type:** We exclude all papers that do not propose a novel dataset, algorithm, or framework, and

**E4) Detection Methodology:** We exclude all papers that are not passive detection methods.

We focus on passive detection because active methods (e.g., watermarking) do not inspect how misinformation is conveyed. We then scanned the titles and abstracts for the eligibility criteria, and excluded papers meeting any exclusion criterion. Furthermore, if the inclusion or exclusion was unclear from the title or abstract, we scanned the full text.

### 3.4 Conference ranking

To further narrow in on high-impact papers, we collected conference and journal ranks for all eligible papers and rejected all papers published in conferences ranked below A by CORE ranking 2023<sup>9</sup> or in journals ranked below Q1 by Scimago<sup>10</sup>. To verify that the threshold does not exclude relevant work, we compared the filtered set against the papers covered by Bu et al. [10] and confirmed that no papers included in their survey were removed by this criterion.

### 3.5 Study Risk of Bias Assessment

To assess the risk of bias in our retrieved sources, we conducted a quality screening [14]. Each paper was scored on a 3-point scale (No / Partially / Fully) against the following weighted quality criteria, derived from the research questions:

**QA1** Does the paper primarily investigate video? (1/0.5/0)

**QA2** Does the paper primarily study misinformation? (1/0.5/0)

**QA3** Does the paper propose or build upon a dataset of misinformation videos? (2/1/0)

**QA4** Does the paper include an algorithm for detecting video misinformation? (1/0.5/0)

**QA5** Are the features derived from the audiovisual content? (1/0.5/0)

<sup>9</sup><https://portal.core.edu.au/conf-ranks/>

<sup>10</sup><https://www.scimagojr.com/>

**QA6** Does the paper state how misinformation is created? (2/1/0)

**QA7** Does the paper include keywords for the type of misinformation? (e.g., out-of-context, cheapfakes) (1/0.5/0)

The quality criteria are organized around three concerns that map to the research questions. QA1 and QA2 revisit the eligibility criteria on a graded scale: papers that meet the binary inclusion threshold but treat video or misinformation as secondary to their main contribution receive partial scores. QA3 and QA4 assess the depth of a paper’s contribution to detection: QA3 targets dataset engagement (RQ3), while QA4 targets algorithmic contribution (RQ2). QA5 evaluates whether the features used for detection are derived from the audiovisual content rather than from metadata alone (RQ2). QA6 and QA7 assess how explicitly a paper engages with the mechanisms behind misinformation creation (RQ1, RQ4).

We assign double weight to QA3 and QA6 because they reflect the two pillars of the review’s scope: empirical grounding in video misinformation data and engagement with specific deception mechanisms. The inclusion threshold is set at 5 out of 9 points. Therefore, passing the inclusion threshold requires alignment across most criteria. For example, a paper scoring almost full marks on all single-weighted items but zero on both double-weighted items would not pass.

### 3.6 Information Extraction

After selecting the relevant literature, we developed a standardized data extraction protocol to systematically gather information from the selected papers to answer the research questions. Several of the extraction steps required annotating how information is conveyed to the viewer. Modality labels alone were too coarse for this: audio, for instance, can carry speech, paralinguistic cues, or background noise, which are processed differently by detection algorithms even though they share a modality. We therefore split visual and auditory modalities by the level that conveys the information: linguistic levels, which cover captions, descriptions, on-screen text, and speech content, and non-linguistic levels, which cover visual content, non-speech audio such as paralinguistic cues, or background noise.

For the first research question, we extracted definitions, descriptions, and examples of video misinformation from full papers, which we refer to as manipulation strategies. For RQ2, we extracted the rationale for selecting specific feature extraction methods or architectures, or what the authors argue should be detectable if content is misinforming, which we call the detection paradigms. Next, we extracted the preprocessing, feature extraction, model architecture, and training paradigm, as well as any other information that describes how the approach detects misinformation, or how the authors operationalize that detection, which we call *implementation methods*. Furthermore, we extracted the logical connections that authors draw between manipulation strategies, detection paradigms, and implementation methods.

For the third research question, we extracted the dataset labels, e.g., binary or reasoning, and the methods used to annotate them. Additionally, we extracted the data’s origin (e.g., video-sharing platform) and, if applicable, how it was generated or manipulated. We also extracted what data the datasets provide, for example, video, audio, captions, descriptions, or comments. The data extracted for RQ4 includes which datasets are used by which algorithms, and the limitations that papers report about their proposed datasets and algorithms. The data extraction for RQ2, RQ3, and RQ4 was conducted by two researchers and cross-verified by the first researcher, who also extracted the data for RQ1.

### 3.7 Information Synthesis & Framework Development

To synthesize the information for RQ1, we used a hybrid inductive-deductive scheme to ensure that the labels resemble theoretical descriptions from the social sciences, while being grounded in video misinformation detection papers. Specifically, we annotated the mechanism used to misinform while referencing theoretical frameworks, but kept a more technical description of the authenticity of the modalities, the connection between linguistic and non-linguistic

content, and the modalities through which the misinformation is conveyed. After the codes stabilized, we grouped them semantically into strategies and categories. Furthermore, we did not collapse categories that few papers described but were theoretically distinct from other categories to find gaps in the literature. We also supplemented the groups with information on how the content is created or manipulated, bridging the gap between theory and detection methods.

For the second research question, the goal was to find patterns in how the information is processed (implementation methods) and why it is processed that way (detection paradigms). To extract the detection paradigms, we inductively coded the stated rationale behind the algorithm’s design choices and grouped the codes into categories based on the type of signal the algorithms analyze. For the implementation methods, we inductively coded preprocessing, feature extraction, model architecture, and training paradigms which we then grouped by how they interact with the content.

For RQ3, we inductively coded the annotation methods used and grouped them into broadly similar strategies. We first assumed that any dataset did not contain a specific type, and refined this assumption from the bottom up by combining three indicators: the dataset’s labels, the annotation method, and the data’s origin or generation.

Finally, for RQ4, we traced argumentative chains for the detection paradigms throughout each paper and mapped them back to the manipulation strategies synthesized in RQ1, where the paper clearly referenced a strategy’s definition or example. This mapping identified which manipulation strategies each algorithm was designed to detect. Algorithms that did not provide this connection were regarded as unspecified classifiers but were still included in the analysis. We also traced argumentative chains between the detection paradigms and the implementation methods. If the detection approach is stated without a clear link to a manipulation strategy, or an implementation method is stated without a clear connection to a detection approach, we noted no connection. Detection paradigms that are justified by being *an indicator for misinformation* or comparable generic justifications were treated without a connection. Logical connections between, for example, one detection approach and two implementation methods are annotated as the same argument chain, whereas unrelated argument chains in the same paper are treated as separate. In the resulting Sankey plot, each connection between two nodes is treated as a separate link due to the visualization’s limitations. For example, a single argument chain linking decontextualization to multimodal inconsistency, and then to both modality alignment and crossmodal attention, produces three connections in the Sankey: one from decontextualization to multimodal inconsistency, one from multimodal inconsistency to modality alignment, and one from multimodal inconsistency to crossmodal attention. Finally, we inductively coded the reported limitations and grouped them by theme. The information synthesis was conducted by one researcher to ensure consistency across the interconnected coding decisions.

## 4 Results

### 4.1 Collected Studies

We illustrate our search strategy in Figure 2. Through the query, we collect 3526 studies, of which we remove 1425 duplicates. Next, we exclude 1944 studies from the abstract and title screening, and 102 studies published in conferences below A or journals below Q1 rank. After filtering out 16 papers through the quality screening, we use the remaining 39 papers as seeds for the snowballing citation graph. We identify 2429 papers which are linked to at least 2 relevant seed papers through snowballing, and remove 2181 duplicates. We then apply the same exclusion criteria to the title and abstract screening as in the first round, removing 206 papers. In contrast to the first round, we first exclude 4 papers based on the study’s risk of bias assessment. We then identify 28 papers that do not pass the outlet rank restrictions. From these papers, we still include 2 papers that are cited by at least 4 of the seed papers, making them important for the

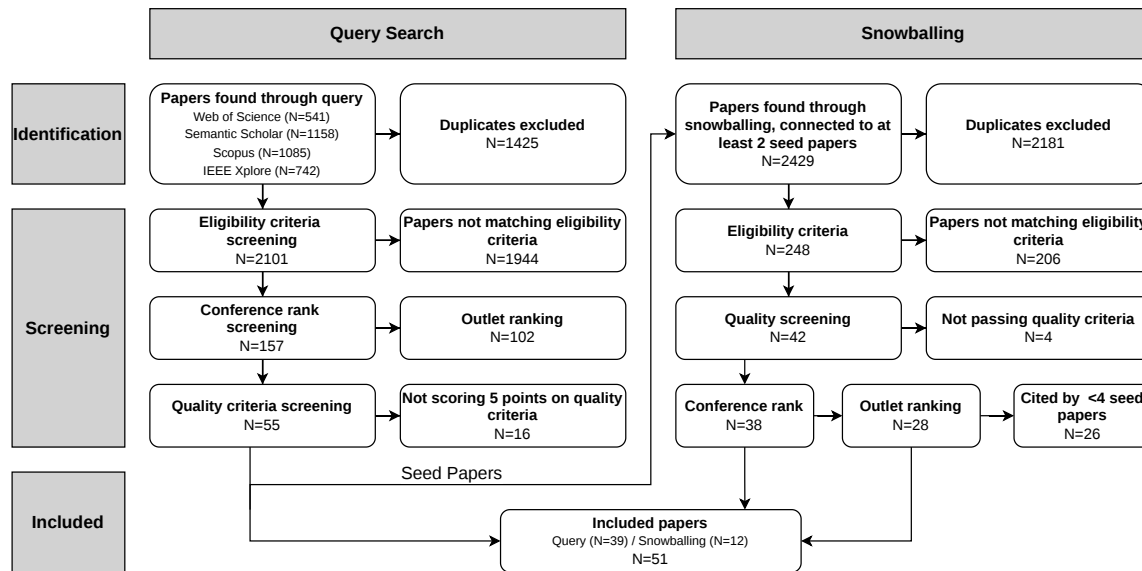


Fig. 2. PRISMA [62] Flowchart diagram of the systematic paper inclusion process.

field. Therefore, the snowballing leads to 12 additional papers being accepted. After completing all the above-mentioned steps for both the query and snowball results, we arrive at 51 papers selected for this review.

When excluding the papers, we found edge cases for criteria 3b and 4a (Tab. 3). For criterion 3b, some edge cases were rejected are [21, 46, 104], which detect tampering but do not clarify the connection between the tampering method and misinformation. Notably, the largest group of papers removed by this criterion are on deepfake detection, which assess the authenticity of content rather than its veracity, and do not evaluate on misinformation datasets. One accepted paper that addresses deepfakes is [3], which not only detect deepfakes, but also physical impersonations or swapping of audio as misinformation. Thus, they analyze whether what is said can be attributed to a person, rather than whether the content was generated. Papers that do connect synthetic media detection to a misinformation context are retained in the corpus. For criterion 4a, we exclude studies on adversarial robustness [71, 73] and model distillation [101], which do not propose new ways to detect misinformation, but build on existing approaches.

Notably, over half of the papers are from 2025 (Fig. 3), showing the timeliness of the review. To our knowledge, the presented corpus is the most comprehensive systematic review of video misinformation detection to date.

## 4.2 RQ1: What misinformation strategies are described in video misinformation detection literature, and how do they create misinformation?

**4.2.1 Framework of deception mechanisms.** To identify the misinformation strategies described in the literature, we extract 103 definitions, explanations, or examples of misinformation. Of those, we exclude 16 that describe misinformation without specifying how it fools the recipient. For example, Zhang et al. [108] state that “creators of fake news [...] employ various techniques, such as malicious editing [...] and AI-generated content.”, but do not explain how editing or AI-generated content is used to misinform. Using the remaining 87 datapoints, we code how the recipient is deceived, i.e., the misinformation strategies. We identify 11 misinformation strategies and cluster them into four groups based on how

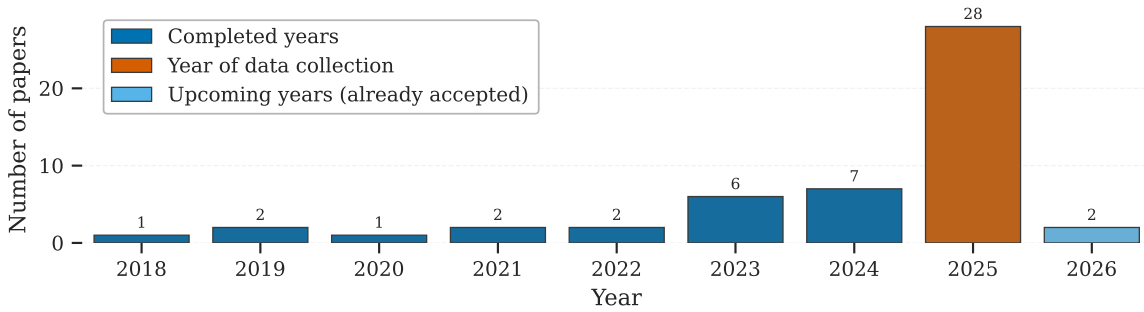


Fig. 3. Publishing dates of papers selected for review, per year.

they mislead: by providing false evidence, by providing deceptive arguments, by imitating an entity that recipients trust, or by presenting claims in a way that shifts the focus away from factuality. We show the framework of misinformation strategies in Table 4. Some strategies are supported by as few as one paper, which we attribute to the field’s early stage (Sec. 4.1) rather than irrelevance, and we retain every strategy that exhibits a distinct deception mechanism.

**4.2.2 Misinformation strategies.** *False evidence* describes misinformation that uses video content as supposed evidence for claims, for example, by providing a video that shows what is claimed but was generated [53, 64, 99]. The second category, *Deceptive argumentation*, makes claims believable through seemingly logical reasoning, e.g., by building an argument on factually wrong information [50, 66, 102]. In the third category, *Imitate trustworthy entity*, we capture all misinformation strategies that raise the perceived credibility by impersonating reliable sources, for example, by making it appear as though a claim was made by an expert [3]. The final category, *Cognitive bypass*, describes all types of misinformation strategies that circumvent critical analysis of the content, e.g., through emotionally charged framing [15, 74, 102].

*False evidence.* False evidence has four misinformation strategies: *Decontextualization*, *Malicious Editing*, *Generative*, and *Staged*. False evidence provides evidence for a claim, which is often communicated through a headline, description, on-screen captions or speech content [11, 29, 47, 49, 52, 63, 64, 88, 100], but can also be communicated through non-linguistic levels, for example, by editing footage to suggest that the moon is larger than it is [106]. For false evidence, the veracity of the claims does not determine whether a video is misleading, as many papers show false evidence paired with true [29, 51, 63, 64, 67, 99, 100] or false claims [11, 47, 49, 52, 75, 88, 90, 100]. Similarly, whether the video evidence is reused, edited, generated, or staged does not define whether content is misleading [11, 74, 75]. What all four strategies share is that the viewer assumes the video supports the claims, e.g., by presenting unaltered, authentic evidence. Some papers seemingly break this assumption by defining generated or manipulated content as misinforming [64, 99, 102], but they also carry an underlying assumption that the content authentically depicts a real event.

*Decontextualization* reuses existing content to create misinformation. It spreads unaltered footage with misleading claims, for example, in a caption [47, 63, 64, 100]. Commonly, videos taken at a different location [63, 64, 100] or depicting an unrelated event [47, 48, 63, 64] are used as supposed evidence. For example, a headline may claim that a bomb attack happened at the Brussels airport, but the video was captured at Moscow airport a few years prior [64]. As such, decontextualization relies primarily on visuals that look plausible to the average recipient. However, since the video is not tailored to the claim, a mismatch between the video’s content and the accompanying claims exists.

Table 4. Framework of Misinformation strategies

Method	Definition & Examples	Papers
<b>I. False Evidence</b>		
Decontextualization	Fabricate supporting evidence for a claim by suggesting that an unrelated video is related.	[8, 11, 15, 32, 39, 47, 48, 49, 53, 63, 64, 66, 99, 100, 105, 106]
Malicious Editing	Fabricate supporting evidence for a claim by altering the video content.	[3, 8, 11, 15, 29, 32, 39, 47, 49, 51, 52, 63, 64, 66, 67, 88, 90, 102, 106, 108]
Generative	Fabricate supporting evidence for a claim by generating video content.	[13, 29, 48, 63, 64, 96, 99, 100, 108]
Staged	Fabricate supporting evidence for a claim by staging video content.	[63, 64, 74, 75]
<b>II. Deceptive Argumentation</b>		
Wrong Facts	Arguing with factually wrong information.	[7, 11, 12, 13, 32, 37, 75, 92, 102]
One-sided Facts	Arguing based on cherry-picked facts.	[37]
Correct Facts, Wrong Conclusion	Reaching a false claim from accurate data.	[37, 38, 50]
<b>III. Imitate Trustworthy Entity</b>		
Dubbing	Suggest that a claim was made by a credible person by replacing the speech.	[3]
Impersonation	Suggest that a claim was made by a credible person by imitating them.	[3]
<b>IV. Cognitive bypass</b>		
Emotional Portrayal	Emotionalize the content.	[11, 74]
Illusion of Proof	Providing a video for a claim to suggest that the claim is backed up by evidence.	[15, 84, 99, 100, 102]

*Malicious editing* goes one step further: Rather than reusing footage as is, the content is altered to support the claim [8, 49, 64, 88]. Edits can be spatial [32, 66] or temporal [90], and the claims can be stated explicitly through the linguistic content, or implicitly through the visual content, as in the enlarged moon example stated above [32]. Because the evidence has been tailored to the claims, the content may exactly match the claims if the edits are executed well. However, manipulating the content introduces tampering artifacts, which can be exploited for detection.

While malicious editing alters existing content, *Generative* methods synthesize the supposed evidence [13, 39, 48, 64]. Therefore, one modality is at least partially generated, for example, when faces are swapped [3]. Like malicious editing, the content is created to support the claims, and the evidence may therefore completely align with them. Crucially, this type of misinformation relies on the viewer believing that what they see is authentic evidence and not generated content. However, because the content is generated rather than captured, synthesis artifacts are present.

Similar to generative evidence, *Staged* evidence fully fabricates evidence, but instead of using generation algorithms, the evidence is physically enacted and captured as normal footage would be. Like generated content, staged evidence can fully match the claims and conceals that it was fabricated [63, 64, 75]. For example, videos may show people with metallic objects stuck to their arms and claim that this is due to a magnetic vaccine [74]. This deception can be particularly effective because the video is genuinely captured in the physical world with no digital artifacts. Any clues come directly from the physical world, such as inconsistencies in the scene or hints of acting.

*Deceptive argumentation.* Deceptive argumentation supports false claims with supposed facts and logical reasoning. Its three strategies, *Wrong facts*, *One-sided facts*, and *Correct facts, wrong conclusion*, range from fabrication, through selective presentation of truthful information, to logically flawed reasoning over correct facts.

*Wrong facts* present fabricated or factually incorrect claims as true [11, 13, 32, 37, 75, 92, 102]. For example, a video may claim that vaccines harm people, and then base an argument on the claim [102]. This type of misinformation is often addressed in papers on health misinformation [37, 75, 92]. The false claims are usually conveyed through linguistic content [11, 32, 75]. However, the misinformation can also be conveyed through linguistic and non-linguistic information jointly, e.g., when linguistic content leads to a false interpretation of visual content, which would not mislead on its own [74]. This category often coincides with false evidence [11, 13, 75] to increase the credibility of the false claims. Because the core deception relies on factual claims that may align with content in other modalities and may not leave detectable traces in the content itself, detection requires verifying the claims.

Rather than stating false claims, *One-sided facts* present factually accurate information, but selected in a way that supports only one side of an argument [37]. One-sided facts are effective because the information does not conflict with external information; rather, it requires awareness of what is missing, which makes detecting this type of misinformation specifically challenging. Only one paper defines this strategy, in the domain of health misinformation [37].

In contrast to wrong facts or one-sided facts, *Correct facts, wrong conclusion* does not convey the misinformation through the facts themselves, but by constructing a chain of reasoning on top of the correct facts that is misleading. Common patterns include drawing conclusions that go beyond what the evidence supports [37, 50] and inferring causation from correlation [37, 38]. Because the individual facts are correct, fact-checking is insufficient. Rather than checking the presented facts, detecting this type of misinformation requires reasoning about the argument’s logical structure. So far, this type of misinformation is only defined for linguistic content [37, 38, 50].

*Imitate Trustworthy Entity*. While deceptive argumentation misleads directly through the claims, *Imitate trustworthy entity* exploits the credibility of an entity [3]. By making it appear as if a credible source, such as a public figure or news outlet, produced or endorsed the claims, they appear more credible and are challenged with less scrutiny. We identify two strategies that differ in how impersonation is achieved: digitally (by dubbing) or physically (by impersonation). This category is currently only supported by one paper in the corpus [3], which limits the analysis.

*Dubbing* uses digital manipulation or generation to make a person appear to say or do something they did not. For example, false statements about the Covid-19 pandemic and climate change may be spliced with footage from a political leader’s speech [3]. The employed techniques range from replacing the audio track, through lip-sync methods that adapt mouth movements to dubbed audio, to face-swap deepfakes that project a different face into the video. Because the content is digitally manipulated, it carries artifacts similar to false evidence. However, rather than assessing whether the content matches the claims, the task is to determine whether what is said can be attributed to the shown entity.

Mirroring the relationship between generated and staged false evidence, *Impersonation* is similar to dubbing, but the content was created in the physical world and captured with a camera, for example, by actors or look-alikes who imitate a person. For example, a look-alike may act like a renowned talk show host [3]. Because the content is not necessarily digitally altered, no digital artifacts are present. Detection, therefore, exploits biometric inconsistencies or behavioral patterns that distinguish the real entity from the imitator [3].

*Cognitive Bypass*. False evidence, Deceptive argumentation, and Imitate trustworthy entity manipulate what is shown, what is said, and who says it. In contrast, cognitive bypass methods target how recipients process information to circumvent critical analysis. We identify two strategies: *Emotional portrayal* makes the content believable by appealing to the recipient’s feelings, and *Illusion of proof* gives the false impression that evidence exists.

*Emotional portrayal* uses affective cues to influence how a recipient engages with the content, either to boost engagement [11] or to emphasize specific points [74]. Because of the emotional response, this strategy can reduce

Table 5. Detection paradigms identified in the reviewed algorithms, grouped by category.

Approach	Description	Papers
<b>Psychological &amp; Rhetorical</b>		
<i>Emotion / Sentiment</i>	Assumes that misinformation differs from truthful content in emotional tone, exploiting affective cues across modalities.	[7, 8, 11, 13, 28, 32, 37, 38, 39, 47, 50, 63, 64, 66, 74, 75, 90, 106, 110]
<i>Style</i>	Identifies surface-level linguistic patterns in titles or descriptions, such as punctuation, pronoun usage, or writing style.	[8, 37, 50, 63, 64]
<i>Intent</i>	Infers the communicative goal or purpose behind the content as an indicator of deceptive motivation.	[15, 17, 28, 66]
<i>Author Emotion / Stance</i>	Detects the position taken by the author or across modalities.	[15, 19, 32]
<b>Multimodal &amp; Media Signals</b>		
<i>Multimodal Inconsistency</i>	Assumes that modalities should be consistent; mismatches between visual, audio, and textual signals indicate manipulation or decontextualization.	[3, 11, 15, 24, 28, 29, 30, 32, 49, 51, 52, 53, 55, 66, 74, 88, 90, 96, 99, 100, 102, 105, 108, 109, 110]
<i>Fabrication / Manip. Artifact</i>	Analyzes individual modalities for forensic traces of content creation or editing, such as visual tampering or temporal irregularities.	[3, 11, 13, 15, 17, 51, 52, 66, 88, 90, 100, 102, 106, 108]
<b>External Validation</b>		
<i>Fact-checking</i>	Verifies claims by comparing them against external sources such as knowledge bases, fact-check databases, or retrieved evidence.	[12, 18, 39, 51, 67, 96, 100, 106, 107]
<i>Reasoning</i>	Reasons about the claims, adding context and a textual analysis	[17, 35, 38, 52, 88, 102, 108, 110]
<b>Social &amp; Credibility</b>		
<i>Social Reaction</i>	Uses audience responses (comments, likes, shares) as signals about content veracity.	[8, 18, 19, 29, 37, 48, 50, 63, 64, 66, 72, 103, 105, 106, 109, 110]
<i>Source Credibility</i>	Estimates veracity from properties of the content creator, such as account history, verification status, or reputation.	[8, 38, 48, 50, 64, 66]
<i>Propagation / Diffusion</i>	Assumes that misinformation spreads differently than truthful content, analyzing sharing patterns and temporal diffusion.	[43, 48]

critical evaluation of the claims [58]. In the current literature, this strategy is only defined for auditive, non-linguistic content [11, 74], for example, by adding emotionally charged background music to a video [11]. Detection therefore relies on identifying affective signals, such as sentiment or emotional tone, that are used to make the claims more persuasive than if they were stated in an unemotional framing.

*Illusion of proof* exploits a different cognitive trait: by attaching a video to the stated claims, recipients may trust the information more because they assume that the video shows evidence for the claims, even when the video content is unrelated [15, 84, 99, 100, 102]. In contrast to false evidence, the content does not support the claims. For example, a title may say that a public figure gave a speech to college freshmen, but in the video, no freshmen are visible [84]. The claims are typically conveyed through the linguistic levels [15, 84, 99, 100, 102], which may add information that cannot be inferred from the video itself [15] or use clickbait phrasing to attract attention [84]. Because the claims and video content are unrelated, detection can exploit the semantic mismatch between them.

### 4.3 RQ2: What rationales do existing misinformation detection algorithms use to identify misinformation, and how do they implement the rationales?

To understand how detection algorithms address the misinformation strategies identified in Section 4.2, we analyze the algorithms at two levels: detection paradigms (what trait is analyzed) and implementation methods (how it is analyzed).

*4.3.1 Detection Paradigms.* We find 11 detection paradigms, which cluster into four categories based on what they analyze and if they use external information: Psychological & Rhetorical paradigms target cognitive and linguistic cues in the content; Multimodal & Media Signal exploits cross-modal relationships and forensic traces; External Validation compares claims against external knowledge; and Social & Credibility leverages audience behavior and source reputation.

*Psychological and Rhetorical* paradigms use cues from how the content is communicated and operate on the content itself. The most common approach, Emotion & Sentiment analysis, assumes that misinformation presents content in a more emotional way than truthful information [32, 38, 66]. Notably, this assumption is not uncontested: Papadopoulou et al. [64] report no correlation between sentiment and misinformation in their data. Style analysis paradigms assume a direct correlation between misinformation and stylistic features, such as the number of personal pronouns [50, 63]. Moving from how to why misinformation is communicated, some paradigms analyze the author’s intent, assuming that misinformation is produced with a persuasive or deceptive purpose that leaves traces [17, 28, 66]. Finally, stance detection analyzes the author’s position toward the claims. Choi and Ko [19] observe that the difference in stance between title and user comments is usually large for misinformation. In contrast, Chen et al. [15] and Han et al. [32] assume that the author’s stance carries contextual information that complements other signals.

*Multimodal and Media Signal* paradigms focus on signals within or between modalities that indicate manipulation or misrepresentation. While this category contains only two paradigms, over half of the analyzed algorithms include at least one of them. Multimodal inconsistency, the most widely adopted paradigm in the entire corpus, uses the assumption that truthful content is consistent across modalities, and that misinformation introduces detectable mismatches between what is shown, said, and written [28, 29, 30]. The targeted inconsistencies vary across papers. The most common comparison is between claims and visual content, with papers assuming that the misinformation claims are imposed upon a video that does not support them in full detail [11, 15, 49]. Agarwal et al. [3] assume that speech patterns and expression patterns do not match with the speaker’s identity if the speech is dubbed or the speaker is impersonated. A few papers extend the comparison beyond the content itself, checking consistency between the claims and external knowledge sources [32, 51], which combines multimodal inconsistency and external validation analysis.

Fabrication and manipulation artifact paradigms, in contrast, look for forensic traces within individual modalities. They share the assumption that when a video is misinforming, it has been manipulated or generated, leaving traces in the content [13, 88, 90, 108]. Notably, most papers frame misinformation as re-edited or spliced material rather than fully generated content [13, 51, 52, 106]. Xu et al. [100] explicitly use artifacts from AI-generation as a detection signal, but do not evaluate on real-world examples. While Agarwal et al. [3] frame their analysis as detecting behavioral mismatches between impersonator and target, the detection functionally focuses on a creation artifact: a forensic trace of how the video was made, observable within a single modality. We therefore classify it alongside other artifact-based paradigms.

That misinformation can be detected through artifacts is not uncontested. Yang et al. [102] do not distinguish between misleading and benign splicing in their annotations, and attribute a drop in their grounding performance to this limitation. Agarwal et al. [3] further acknowledge that simple artifact analysis may not keep pace as generative methods improve. Out of the 14 papers using Fabrication & Manipulation artifacts, 11 use both inter-modal and intra-modal traces, suggesting that fabrication artifacts alone are insufficient as a standalone misinformation detection paradigm.

*External Validation* paradigms assume that the veracity of claims can be validated or falsified through external information, rather than relying solely on the content. Fact-checking paradigms assume that false claims can be identified by verifying them against external information. The data used for the detection, however, varies: Some paradigms use facts specific to the video’s topic [12, 39, 51, 96, 100, 106]. However, some paradigms also use videos from the same event with known labels, such as debunk or known false videos [18, 67, 107].

Conversely, reasoning assumes that VLMs or LLMs can enrich the analysis with related information about events or people by drawing on their instilled knowledge. Some paradigms assume that the model has internalized enough world knowledge to reason about the plausibility of claims [35, 88, 102, 108]. Some paradigms also assume that the model can derive deeper semantic understanding of the content than conventional feature extractors [17, 38, 52, 110].

*Social and Credibility* paradigms use meta information from the platforms on which the video was shared. As such, they shift the focus from the video to the social ecosystem. They draw on three signal types: audience reactions to the content, the uploader’s credibility, and the patterns of the video spreading across the platform.

Social reaction is the most frequently adopted paradigm in this category, appearing in 16 papers, and draws on community feedback such as comments, likes, and shares. The proposed paradigms share the assumption that misinformation is received differently than truthful content, either in the engagement of the users [37, 48, 72], or through comments that point out that the video is misleading [18, 63]. Most papers adopt this paradigm by citing other papers or even without justification. However, some papers investigate whether social reaction signals behave as the paradigm assumes, with mixed results. Qi et al. [66] report supportive within-corpus statistics, including that 18% of fake videos receive doubtful comments such as "Really?" or "Fake!" compared to only 4% of real videos. Papadopoulou et al. [64] find that fake videos accumulate fewer comments over time on YouTube, though the same pattern does not hold on Facebook. In contrast, Choi and Ko [19] report that title and description features remain more effective than comment-based features for fake video detection, and Zong et al. [109] show that comment features can introduce spurious correlations between common keywords and labels, actively misleading detection. The evidence for this paradigm is thus mixed, with effectiveness appearing to depend on the platform and the specific features used.

Source credibility assesses whether the uploader of the video can be trusted as a proxy for the content being misleading or not. This paradigm is grounded in empirical evidence [8, 38, 48, 50, 64, 66]. Specifically, Hu et al. [38] find that 81% of real news publishers in their corpus are verified, compared to only 11% of fake news publishers. Qi et al. [66] further show that fake publishers exhibit different engagement-per-follower patterns. Papadopoulou et al. [64] find that fake videos tend to be posted by younger accounts. Other papers draw on external sources: Li et al. [50] and Li et al. [48] cite prior studies on related detection tasks, and Boididou et al. [8] ground the feature choice in studies on journalistic practice. No paper in the corpus reports evidence against the paradigm.

Finally, Kim et al. [43] and Li et al. [48] exploit that misinformation and truthful videos spread differently on social media. Kim et al. [43] assume that videos recommended to watch next show different patterns for misleading and genuine content. Li et al. [48] further analyze event-based cascades, under the assumption that genuine content shows different patterns than misleading content in how the video is spread, adapted, and reuploaded.

Across the reviewed corpus, 32 papers combine multiple detection paradigms rather than relying on a single signal. The most common pairings are Multimodal Inconsistency and Fabrication/Manipulation Artifact with 11 papers, Emotion/Sentiment and Social Reaction with 8 papers, and Multimodal Inconsistency and Emotion/Sentiment with 6 papers. Among all paradigms, external validation is the most recently adopted: only 4 papers that employ it were published before 2025. While papers may act on the same assumption on how misinformation can be detected, how these assumptions are operationalized varies, which we explain in the following section.

**4.3.2 Implementation Methods.** We separate the implementation methods into five categories: Multimodal Interaction, Manipulation & Media Forensics, Linguistic & Text Analysis, Metadata & Context, and Social & Network Modeling (Tab. 6). The first three operate on video content by combining information across modalities, finding forensic traces, or analyzing the linguistic levels. The last two categories draw on external knowledge sources or platform-level signals.

Table 6. Implementation methods employed by the reviewed algorithms, grouped by category.

Method	Description	Papers
<b>Multimodal Interaction</b>		
<i>Modality Alignment</i>	Projects multiple modalities into a shared embedding space to compare or fuse their representations.	[3, 13, 15, 19, 24, 28, 45, 47, 48, 52, 53, 84, 88, 90, 96, 99, 100, 108, 109]
<i>Crossmodal Attention</i>	Uses attention mechanisms across modalities to capture fine-grained inter-modal dependencies or compare modalities.	[11, 13, 29, 30, 45, 49, 52, 53, 55, 66, 74, 75, 88, 90, 96, 105, 106, 107, 108, 109, 110]
<i>Linguistic Guided Vision</i>	Guides visual feature extraction using linguistic cues, focusing the vision encoder on text-relevant regions.	[3, 13, 32, 52, 74, 75, 96, 100, 102, 108]
<i>Audio-Visual Sync</i>	Detects temporal or semantic desynchronization between audio and visual streams.	[3]
<b>Manipulation &amp; Media Forensics</b>		
<i>Frame-level Anomaly</i>	Detects visual tampering artifacts within individual frames.	[11, 66, 88, 102, 108]
<i>Temporal Anomaly</i>	Identifies irregularities in the temporal signature of video or audio streams.	[3, 11, 13, 66, 90, 102, 108]
<i>Deepfake Detector</i>	Employs specialized modules for detecting synthetically generated content.	[100]
<b>Linguistic / Text Analysis</b>		
<i>Sentiment &amp; Emotion</i>	Extracts sentiment or emotional features.	[7, 8, 11, 13, 28, 32, 37, 38, 39, 47, 50, 63, 64, 66, 75, 90, 110]
<i>Stylometric Features</i>	Measures surface linguistic patterns such as punctuation frequency, pronoun use, or vocabulary complexity.	[8, 37, 50, 63, 64]
<i>Intent Features</i>	Encodes signals of communicative purpose, e.g., general audio features interpreted as markers of intent.	[66]
<i>Stance Detection</i>	Classifies the position or attitude expressed by the author, often comparing across modalities.	[19, 32]
<b>Metadata &amp; Context</b>		
<i>External info retrieval</i>	Retrieves factual information from structured or unstructured knowledge sources for claim verification.	[12, 32, 39, 43, 96, 100]
<i>Fact-Check Database</i>	Compares claims against existing fact-check verdicts.	[32, 51, 67, 106, 107]
<i>Parametric Knowledge</i>	External information recalled from model parameters	[13, 18, 32, 35, 38, 52, 102]
<i>LLM/VLM Reasoning</i>	Uses language or vision-language models to reason over the content, e.g., assessing semantic entailment between modalities or analyzing model-generated reasoning chains as detection signals.	[32, 35, 38, 52, 88, 108, 109]
<b>Social &amp; Network Modeling</b>		
<i>Comment / Reaction</i>	Uses comments, likes, or shares as direct features or for contrastive analysis against the content.	[18, 19, 29, 37, 45, 50, 63, 64, 66, 72, 103, 105, 106, 109, 110]
<i>User profiling</i>	Inspects metadata from the uploader account.	[8, 38, 39, 48, 50, 64, 66, 109]
<i>Temporal Modeling</i>	Models the sharing history of content as a graph.	[48]

*Multimodal Interaction Methods*, the most common group, combine information from various modalities through alignment, crossmodal attention, linguistic guided vision, or by examining the synchronization between audio and video. Modality alignment transforms two or more modalities into a joint representation, either to bridge features [28, 45, 47, 99, 100] or to check for consistency [3, 24, 53, 88, 108]. To transform the information into a joint representation, three approaches are used. Firstly, by using contrastive or shared-space losses, the embeddings of matching modality pairs are pulled together, while mismatched modality pairs are pushed apart [28, 53, 96]. Secondly, some papers use pretrained joint encoders, such as VideoCLIP [98] or Imagebind [25], usually as a feature extraction method [84, 99, 100]. Thirdly, pretrained LLM/VLMs are used to check the consistency between modalities [13, 52, 108].

Crossmodal attention frequently co-occurs with modality alignment but operates at the feature rather than representation level. It is primarily used to fuse modalities [11, 29, 30, 49], but also to incorporate external information [52, 96] or compute consistency signals [12, 88].

Linguistic guided vision uses text to extract specific information from visual content [13, 32, 52]. We identify two ways to guide visual extraction, which use different guiding information. The first group uses text extracted from the content, either by fusing the modalities [3, 74, 75] or by selecting relevant frames for further processing [13]. The second group extracts relevant information using VLMs with content-agnostic prompts [32, 52, 96, 100, 102, 108].

Audio-visual sync is only proposed by Agarwal et al. [3]. It temporally aligns speech with the speaker’s expressions to find irregularities. As such, it fuses information from the same timeframe, but requires a frontal view of the speaker.

*Manipulation and Media Forensics* focuses on detecting traces introduced by content editing or generation. The identified methods operate at different granularities within the video signal. In the reviewed corpus, manipulation and media forensics methods appear exclusively in combination with other method categories.

Frame-level anomaly detection searches for visual traces. While some algorithms look for traces like compression [66] or superimposed CGI artifacts [11, 102], others do not mention targeting specific tampering types [88, 108].

Temporal anomaly detection analyzes content over time. Most studies focus on video content, looking for artifacts in how video frames are combined or edited over time [11, 66, 102, 108]. Furthermore, Cao et al. [13] check whether the audio has been manipulated, and Agarwal et al. [3] compare audio with visual information to find out whether the audio was replaced. Notably, over half of the papers that use temporal anomaly also use frame-level anomaly analysis.

Only Xu et al. [100] use a deepfake detection algorithm to differentiate between generated and authentic content.

*Linguistic and Text Analysis* targets how content is expressed and positioned, often analyzing text or spoken content.

Sentiment & Emotion capture whether claims are expressed in emotionally charged or neutral language. Implementations use emotion and sentiment features from various modalities, such as text [8, 32], audio [37, 39], text and audio [7, 11, 13, 28], or multiple modalities jointly [47, 110]. Text sentiment or emotion features are often extracted using RoBERTa variants [7, 11, 38, 90]; audio sentiment features are often extracted using VGGish [38, 66] or HuBERT [11, 39, 75, 90]. Notably, Kumari et al. [47] use sentiment as an additional prediction target rather than a feature, and Zong et al. [110] prompt an MLLM to extract the emotion from all modalities simultaneously.

Style analysis uses surface linguistic patterns as features, for example, by analyzing the number of personal pronouns and exclamation marks, or the wording of the claims [50, 63, 64]. Throughout the literature, stylistic features are hand-engineered rather than learned, of which we find four categories: syntactic cues such as pronoun usage, part-of-speech counts, and sentence structure [8, 37, 50, 64]; orthographic cues such as capitalization and punctuation [8, 63, 64]; readability and lexical-richness scores [8, 37]; and informal-register markers such as slang, modal particles, or clickbait phrases [8, 50, 63, 64]. Hou et al. [37] go beyond sentence or word analysis by using LIWC and n-gram features.

The remaining methods are seldom used. Stance detection models the author’s position, either as disagreement between video titles and viewer comments via topic modeling [19], or as LLM-generated author stance used as auxiliary context [32]. Qi et al. [66] additionally extract publisher intent as a learned audio representation via VGGish.

*Metadata and Contextual methods* use external information to analyze the content. External knowledge integration methods augment inference with information related to the video, such as external events or interpretive information. The source from which this knowledge is drawn varies between approaches: Some algorithms use a search query to gather related information from online search engines. The retrieved data can be in many modalities, such as related text [12, 100], images [100], or related videos [32]. Instead of a query, Kim et al. [43] use YouTube’s recommendation algorithm to find similar videos. Other algorithms retrieve related information about entities and events from external

knowledge graphs [39, 96] Finally, some algorithms use fact-check databases, which can either contain samples of previous verdicts [32, 51, 67, 107], or examples that debunk the misinformation [32, 67, 106]

Parametric knowledge, in contrast, builds on information stored in model weights, rather than drawing on examples during inference. Common approaches use LLMs [13, 32, 38] or multimodal variants [35, 52, 102]. In contrast, Choi and Ko [18] store embeddings that reflect the domain during training, and use them during inference to extract domain-relevant information. As such, parametric knowledge does not reference databases but relies on learned knowledge.

Finally, some approaches use language model reasoning as features. Both text-only [32, 38, 109] and multimodal models [35, 52, 88, 108] are used for this paradigm. The reasoning can be generated based on different information: Some papers combine related external knowledge sources in their reasoning, summarizing the information [32, 35, 88, 109]. Others use the VLMs/LLMs directly on the content to extract more sophisticated features, either by direct prompting [38, 108, 109], or by pitting a question and an answer generation model against each other [52].

*Social and Network Modeling* methods shift the analysis from the content to the social ecosystem in which it circulates. Comment and reaction features are the most adopted, extracting signals from audience engagement such as comment text, like/dislike ratios, and sharing behavior. The techniques differ in how they extract information.

Some algorithms extract handcrafted features from individual comments, such as sentiment polarity, text length, pronoun use, or punctuation and capitalization patterns [50, 64], while others use aggregates across all comments, such as the total comment count, the ratio of comments flagging the video as fake, or the frequency of sentiment-bearing words [37, 63, 72]. A larger group encodes the comments with pretrained language models, typically BERT, and feeds the embeddings into a detection model [18, 29, 45, 66, 106, 110]. Some additionally weight the embeddings by engagement signals such as like counts [66, 106]. Beyond encoding, a few algorithms embed the comments into specialized processing components: Choi and Ko [19] compare the comments' content to the title or description to detect disagreement between the two; Zeng et al. [105] apply a causal reasoning module that removes spurious correlations between textual content, e.g., emotional words, and the detection labels; and Zong et al. [109] combine comment encodings with title, transcript, and user information, and use an LLM to reason about whether the content is misinforming. Finally, Yin et al. [103] are the only ones to encode time-sync comments, which they link with normal comments into a temporal graph.

User profiling exploits uploader information to detect misinformation. The features draw on three signal types, often in combination: self-declared information such as usernames, biographies, or posting history [8, 38, 39, 48, 66, 109]; community-generated signals such as follower counts, likes, and view counts [8, 38, 39, 48, 50, 64]; and platform-granted markers, most notably verification status [8, 39, 66]. These features are processed in two ways: as numerical features fed into a classifier [8, 50, 64], or encoded through a pretrained language model [38, 39, 66, 109]. Only Li et al. [48] embed user attributes as node features in a heterogeneous graph rather than treating them as independent inputs.

Temporal modeling is the only method in social and network modeling that captures how social signals evolve over time. So far, only Li et al. [48] employ temporal modeling by using a Hawkes process to model how topics rise and fall in popularity and how videos relate temporally, specifically targeting platforms where reposting cannot be traced directly.

#### 4.4 RQ3: What misinformation strategies are represented in existing datasets, and how are they used for training and evaluation?

*Dataset annotations.* In total, 27 out of 51 reviewed papers propose own datasets (Tab. 7). From the proposed datasets, 22 collect in-the-wild data. Since in-the-wild datasets collect misinformation samples without knowing the deception strategy, the distribution of strategies within the dataset is unknown unless annotated. YouTube was the most popular data source until 2023, when short-form platforms TikTok, Douyin, and Kuaishou gained greater attention (Fig. 4).

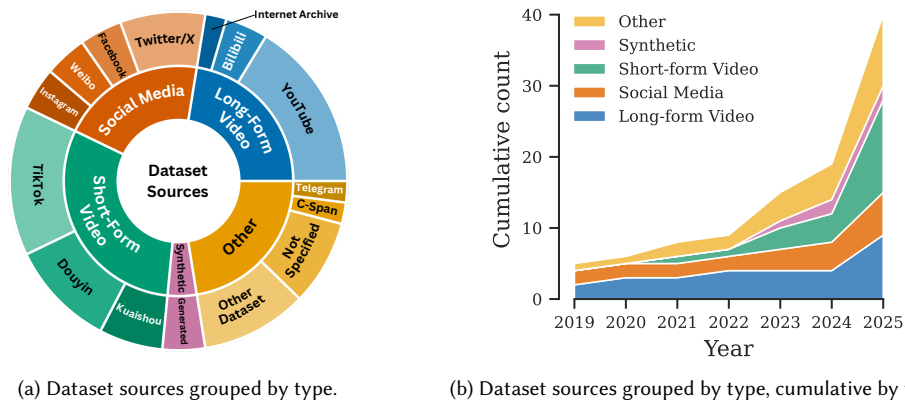


Fig. 4. Overview of the video sources used by each dataset.

For annotating the videos, common strategies are to use non-experts [19, 63, 84], who oftentimes use fact-checking articles as a reference [38, 74, 75, 102]. Expert annotations are rare and do not reference fact-checking articles [37, 50] except for Zeng et al. [106]. Labels are also directly extracted from fact-checking articles [29, 72] or sourced from trusted outlets of truthful and misleading information [12, 43, 47, 99]. Furthermore, some datasets generate labels by training models [92] or prompting LLMs [17], but more often generate misinforming samples [3, 30, 53]. Many datasets also collect samples by crawling for videos related to the corpus [3, 7, 11, 29, 64, 66, 103, 106, 110].

Annotations further vary in form. 19 datasets provide binary veracity labels (truthful vs. misinforming), three of them with a third category for debunk videos [29, 38, 66]. Further veracity labels are ternary, which add an *uncertain* category to binary [47, 63], and 6-point scale, which follows Politifact’s scheme [7]. Three datasets annotate different misinformation strategies [17, 99, 106], of which only Zeng et al. [106] provides human annotations. Finally, Sung et al. [84] and Yang et al. [102] are the only datasets that provide reasoning or grounding annotations, respectively.

Despite many datasets proposed, few are reused (Tab. 7): FakeSV [66] and FakeTT [11] are reused the most, with the former sourcing from Douyin and Kuaishou, and the latter from TikTok. Both come from the same research group and use the same annotation strategy. The two oldest datasets, FVC [64] and VAVD [63] follow in popularity. M3A [99] and MYVC [19] are only reused by the authors in a follow-up paper, and TikCron is proposed in two papers by the same first three authors [38, 92], with the same number of videos retrieved but labeled differently. 21 datasets are not reused.

Similar to how detection algorithms define misinformation strategies (Sec. 4.3), 12 out of 27 datasets include data from certain strategies (Tab. 8). Three approaches to annotate misinformation strategies are used: direct annotation from real-world data, restricting the corpus to one misinformation strategy, or generating type-specific misinformation.

To annotate the data, Yang et al. [102] ask human annotators where decontextualization, malicious editing or wrong facts occur. Sung et al. [84] instead ask annotators how the content is stating wrong facts or going beyond what the video suggests. Chen et al. [17] use VLM agents to annotate misinformation subcategories, which, however, are not clear to us (e.g., misinformation is a subcategory of misinformation). Datasets with specific strategy annotations often filter the strategies. Li et al. [50] only annotate videos where the factual claims are correct, but the conclusion based on the facts is wrong. Wang et al. [92] restrict their data to misinformation that conveys factually wrong information, and Biçer et al. [7] to individuals stating factually false information, resulting in three misinformation strategies. Datasets that annotate specific misinformation strategies often synthetically create data rather than labels. To create decontextualized evidence,

Table 7. Overview of the 27 datasets proposed by reviewed papers, sorted by publication year. **Data Origin:** W = in-the-wild, S = synthetic, W+S = both. **Annotation:** NE = non-expert annotations, EX = expert annotations, FC = fact-checking articles, OS = online search / video crawling, GL = generated labels (automated), TO = trusted outlet videos, LLM = LLM-generated labels. **Reuse** = number of other reviewed papers using this dataset.

Paper	Dataset Name	Source	Origin	Labels	Annotation	# Samples	Reuse
[63]	VAVD	YouTube	W	Ternary	NE	546	4
[64]	FVC	YouTube, Twitter, Facebook	W	Binary	EX, OS	6,392	10
[37]	—	YouTube	W	Binary	EX	250	0
[19]	MYVC	YouTube	W	Binary	NE	1,805	1
[74]	—	TikTok	W	Binary	FC, NE	891	0
[50]	—	Bilibili	W	Binary	EX	700	0
[3]	—	Other datasets	W+S	Binary	GL, OS	177.9 hours	0
[47]	FakeClips	YouTube	W	Ternary	NE, TO	5,454	0
[53]	TwitterDetective	Twitter	W+S	Binary	GL	10,000	0
[66]	FakeSV	Douyin, Kuaishou	W	Binary, Debunk	FC, OS, NE	5,538	22
[84]	VMH	Facebook	W	Reasoning	NE	N/A	0
[11]	FakeTT	TikTok	W	Binary	FC, OS, NE	1,991	13
[17]	SafeWatchBench	Other datasets, Generated	W+S	Manip. Type	LLM	1,442,500	0
[30]	3M	Weibo, Douyin, Kuaishou	W+S	Binary	GL	17,352	0
[72]	—	YouTube, Twitter, Facebook, Instagram, TikTok, Telegram	W	Binary	FC	756	0
[99]	M3A	Instagram, Generated	W+S	Manip. Type	TO, GL	2,146,146	1
[7]	—	YouTube	W	6-point scale	FC, OS, NE	201	0
[12]	—	Douyin	W	Binary	TO	12,080	0
[29]	—	TikTok	W	Binary, Debunk	OS, FC	361	0
[38]	TikCron	Douyin	W	Binary, Debunk	FC, NE	4,169	0 <sup>†</sup>
[92]	TikCron	Douyin	W	Binary	FC, NE, GL	42,201	0 <sup>†</sup>
[43]	FYKE	YouTube	W	Binary	FC, TO	4,000	0
[75]	—	TikTok	W	Binary	FC, NE	2,190	0
[102]	GroundLie360	Twitter, YouTube, TikTok	W	Grounding	FC, NE	2,044	0
[103]	TSC-VRD	Bilibili	W	Binary	FC, OS, NE	645	0
[106]	—	Douyin, Kuaishou, Weibo	W	Manip. Type	FC, OS, EX	2,051	0
[110]	CH-SV	Kuaishou	W	Binary	OS, FC, NE	6,728	0

<sup>†</sup> TikCron is proposed by both papers; due to the difference in size, we count it as two datasets with 0 external reuse.

modalities are swapped between samples, either randomly [3, 30] or more strategically combined by identifying similar yet different samples [53, 99]. Another way to create decontextualized evidence is to replace entities in the linguistic content [53, 99], or to generate false video descriptions using VLMs [99]. For impersonation and false evidence examples, Agarwal et al. [3] use face-swap algorithms. Finally, Xu et al. [99] generate false evidence videos from misleading text prompts. As such, only false evidence and impersonation data are currently synthetically created.

In-the-wild datasets contain different misinformation strategies, but many do not annotate them, and instead use veracity labels [3, 7, 30, 64]. For misinformation strategies with annotations, coverage varies substantially (Tab. 8). False evidence is the most-covered category (DEC: 3, ME: 3, GEN: 2), though no dataset annotates staged content. Among deceptive argumentation strategies, 5 datasets annotate wrong facts and 3 datasets correct facts wrong conclusion, while one-sided facts has no coverage. No dataset annotates dubbing or impersonation. In cognitive bypass, 3 datasets annotate illusion of proof, but emotional portrayal has no coverage. In total, 5 of 11 strategies lack annotated labels.

*Dataset usage.* Moving from datasets to algorithms, we find that 42 algorithms predict general binary labels, and 2 papers predict veracity scores with either ternary [47] or 6-point [7] labels. As such, they do not differentiate between misinformation strategies. Liu et al. [53] predicts whether any of speech, video, or claims is inconsistent with the other two as a proxy for misinformation detection. 4 papers generate reasoning for why a video is misinforming or not [15,

Table 8. Misinformation strategies explicitly provided by datasets, sorted by publication year. Only datasets that annotate at least one strategy are shown; the remaining 15 datasets provide binary or untyped veracity labels only (see Table 7 for the full list).  $\Delta$  = unsure,  $(\checkmark)$  = included but collapsed to binary label. **Acronyms:** **DEC** = Decontextualization, **ME** = Malicious Editing, **GEN** = Generative, **ST** = Staged, **WF** = Wrong Facts, **OSF** = One-sided Facts, **CFWC** = Correct Facts, Wrong Conclusion, **DUB** = Dubbing, **IMP** = Impersonation, **EP** = Emotional Portrayal, **IOP** = Illusion of Proof.

Paper	False Evidence				Deceptive Argumentation			Imitate Trustworthy Entity		Cognitive Bypass	
	DEC	ME	GEN	ST	WF	OSF	CFWC	DUB	IMP	EP	IOP
[64]	( $\checkmark$ )	( $\checkmark$ )	( $\checkmark$ )	( $\checkmark$ )							
[50]							( $\checkmark$ )				
[3]		( $\checkmark$ )		( $\checkmark$ )	( $\checkmark$ )			( $\checkmark$ )	( $\checkmark$ )		
[53]	( $\checkmark$ )	( $\checkmark$ )			( $\checkmark$ )		( $\checkmark$ )				
[84]							( $\checkmark$ )				( $\checkmark$ )
[17]	$\Delta$			$\Delta$							
[30]	( $\checkmark$ )										
[99]	( $\checkmark$ )		( $\checkmark$ )		( $\checkmark$ )						( $\checkmark$ )
[7]					( $\checkmark$ )	( $\checkmark$ )	( $\checkmark$ )				
[92]					( $\checkmark$ )						
[102]	( $\checkmark$ )	( $\checkmark$ )			( $\checkmark$ )						( $\checkmark$ )
[106]		( $\checkmark$ )	( $\checkmark$ )		( $\checkmark$ )						
$\Sigma \checkmark$	3 (+2)	3 (+2)	2 (+1)	0 (+2)	5 (+2)	0 (+1)	3 (+1)	0 (+1)	0 (+1)	0 (+0)	3 (+0)

17, 35, 100]. Focusing more deeply on where the misinformation comes from, Yang et al. [102] predict what part of the video is misinforming. Finally, Zeng et al. [106] predict misinformation types, matching their proposed dataset’s labels.

Notably, Xu et al. [99] train their model on binary labels, even though their dataset contains labels for different misinformation strategies. However, they evaluate per-strategy performance using the strategy labels. Their evaluation results show significant differences in performance across misinformation types, ranging from 1.0 AUC for generated multimodal content to 0.5 AUC for changing the name of the entity shown with the Languagebind backbone.

#### 4.5 RQ4: How well do current detection approaches and datasets cover the identified misinformation strategies, and what gaps emerge?

**4.5.1 Connection between misinformation strategies and their detection.** From 51 papers analyzed, 37 define a misinformation strategy, 25 of which logically connect a misinformation strategy to the proposed algorithm. The other 14 papers do not define any specific misinformation strategy, but 12 connect detection paradigms to implementation methods.

In total, 45 logical links connect different misinformation strategies to 6 of 11 detection paradigms. Most connections follow expected patterns: false evidence strategies connect primarily to multimodal inconsistency and fabrication / manipulation artifact detection, while deceptive argumentation strategies connect to fact-checking and reasoning. Multimodal & Media signals are the most connected approaches, with 32 connections. External validation is also quite prominent, with 10 connections. Psychological and Social & Credibility detection paradigms are only linked to a misinformation type very infrequently, with 2 and 1 connections respectively.

Six detection paradigms have weak or no connection to misinformation strategies: Social Reaction, Source Credibility, Stance, Style and Intent have no connection, and Emotion & Sentiment features, the second most used detection paradigm, is only connected to a misinformation strategy in two papers out of 19. In total, almost half of the detection paradigms are employed without connection to specific misinformation strategies. However, they are well-connected to implementation methods, following predictable group-level patterns: Psychological & Rhetorical methods map almost exclusively to Linguistic / Text Analysis, and Social & Credibility to Social & Network Modeling. Outliers from these mapping patterns are Kumari et al. [47], who align multiple modalities to detect misinformation alongside emotion

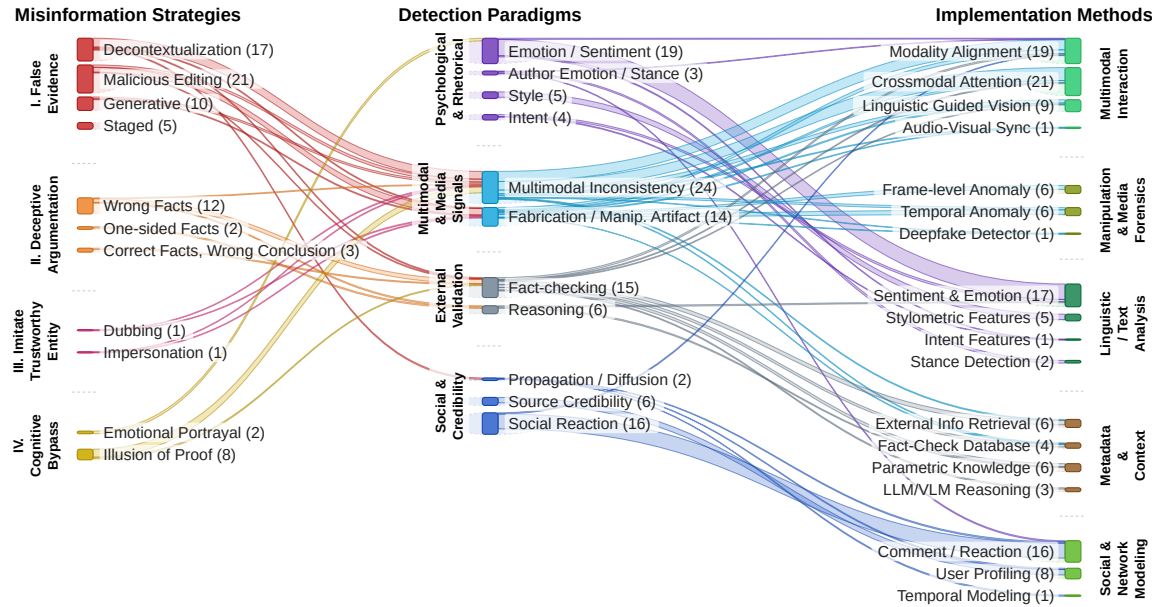


Fig. 5. Connection between misinformation strategies, detection paradigms and implementation methods. The height of the nodes is proportional to the number of papers; faded bands represent the proportion of papers that do not have a connection to the nodes before or after. The width of the connected bands represents the relative amount of connections.

recognition tasks, Choi and Ko [19] who compare comments to the author’s stance extracted from the other modalities of the video, and Zong et al. [110], who prompt a VLM to extract emotions from the video content.

However, not all misinformation strategies connect to detection paradigms. Most notably, 5 papers define staged content but do not connect it to any detection paradigm. Correct facts wrong conclusion and one-sided facts each connect through a single paper, and emotional portrayal through two. Among frequently defined strategies, trace rates vary: decontextualization connects in 11 of 17, malicious editing in 10 of 21, and generative content in 3 of 10 papers.

**4.5.2 Connection between detection algorithms and datasets.** Of the 51 papers, 21 trace at least one argumentation chain from a misinformation strategy through a detection paradigm to an implementation method (Tab. 9). Of the 21 that trace complete argumentation chains, 15 train and evaluate exclusively on datasets with veracity labels. Out of the six remaining papers, five use datasets that annotate specific deception mechanisms, but the dataset annotations go beyond the paper’s argued chains. Thus, the algorithms do not provide a detection paradigm for every misinformation type in the dataset. Of these five, two generate synthetic misinformation with built-in type labels [53, 99], one uses reasoning labels from an expert-annotated dataset [15, 16], one collects type-specific data [3], and one provides human annotations informed by fact-checking articles [102]. Only Xu et al. [100] fully align their argued misinformation strategies with their dataset annotations. The algorithms trained on strategy-specific data evaluate solely on their own datasets, making them difficult to compare with other approaches or to assess their in-the-wild performance.

**4.5.3 Self-reported limitations.** To see whether the field recognizes the above-stated characteristics, we examine their self-reported limitations. For algorithms, the most frequently reported limitation is flaws of the detection method leading to reduced performance (16 papers), followed by limited data variety (6) and limited evaluation (5). For datasets,

Table 9. Alignment between argued misinformation strategies and dataset annotations for the 21 papers with complete argumentation chains. **Align**: whether the dataset annotates all (Full), some (Partial), or none (None) of the paper’s argued misinformation strategies. **Abbreviations**: DEC = Decontextualization, ME = Malicious Editing, GEN = Generative, ST = Staged, WF = Wrong Facts, OSF = One-sided Facts, CFWC = Correct Facts Wrong Conclusion, DUB = Dubbing, IMP = Impersonation, EP = Emotional Portrayal, IOP = Illusion of Proof. Datasets annotated with “–” provide only binary or untyped veracity labels.

Paper	Argued methods	Dataset (annotated types)	Label	Align
[11]	DEC, EP, ME, WF	FakeTT [11] (–); FakeSV [66] (–)	Binary	None
[48]	DEC, GEN, ME	FakeSV [66] (–); FakeTT [11] (–); <i>Weibo (image/text)</i> [91] (–)	Binary	None
[52]	IOP, ME, OSF	FakeSV [66] (–); FakeTT [11] (–); FVC [64] (–)	Binary	None
[49]	DEC, IOP, ME	FakeSV [66] (–); FakeTT [11] (–)	Binary	None
[74]	EP, ST, WF	Own dataset [74] (–)	Binary	None
[75]	ST, WF	Own dataset [75] (–)	Binary	None
[96]	GEN, WF	FakeSV [66] (–); <i>Weibo (image/text)</i> [91] (–)	Binary	None
[66]	DEC, ME	FakeSV [66] (–)	Binary	None
[88]	ME	FakeSV [66] (–); FakeTT [11] (–)	Binary	None
[90]	ME	FakeSV [66] (–); FakeTT [11] (–)	Binary	None
[51]	ME	FakeSV [66] (–); FakeTT [11] (–)	Binary	None
[105]	DEC	FakeSV [66] (–); FVC [64] (–)	Binary	None
[30]	IOP	3M [30] (–)	Binary	None
[12]	WF	Own dataset [12] (–)	Binary	None
[38]	CFWC	TikCron [38] (–)	Binary	None
[3]	DUB, IMP, ME	Own dataset [3] (DUB, IMP, ME, <b>GEN, WF</b> )	Binary	Partial
[102]	IOP, ME, WF	GroundLie360 [102] ( <b>DEC, ME, WF, IOP</b> )	Grounding	Partial
[99]	DEC, GEN, IOP	M3A [99] (DEC, GEN, IOP, <b>WF</b> )	Binary	Partial
[53]	DEC	TwtrDetective [53] (DEC, <b>CFWC, WF</b> )	Consistency	Partial
[15]	DEC, IOP, ME	FakeVE [16] (DEC, ME, IOP, <b>WF</b> )	Reasoning	Partial
[100]	DEC, GEN, IOP, WF	M3A [99] (DEC, GEN, IOP, WF)	Manip. type, Reasoning	Full

annotation quality (5), data diversity (4), and dataset size and bias (3 each) are the most common concerns. Notably, none of the self-reported limitations concern the granularity of the labels or a lack of transparency in the evaluation.

## 5 Discussion

### 5.1 Misinformation Strategy Framework

As shown in Section 2.3, existing frameworks have several strengths and weaknesses: Theoretical frameworks classify how misinformation deceives, but neither answer how videos are created or used to deceive, nor how the misinformation can be detected [2, 42, 57, 58, 81, 93]. Technical frameworks, in contrast, capture what techniques are used to create misinforming videos [86], or what algorithms are used to detect video misinformation [4, 10, 97], but do not capture why videos are misinforming. Our proposed framework bridges the gap between theoretical and technical frameworks by capturing how misinforming videos deceive the recipients. The framework was built inductively from how detection papers view, describe, and characterize misinforming video content, rather than from how they classify misinformation. Because the framework is content-oriented and built on the misinformation detection literature, it exhibits two unique properties: Firstly, video misinformation strategies are defined from the content rather than through the proxy of the detection algorithm. Secondly, because most detection approaches are motivated by properties of different misinformation strategies, it can be used to connect *how misinformation deceives* to *how misinformation can be detected*.

The proposed framework also deepens the understanding of misinformation detection: first, it describes what properties or artifacts are expected in which modalities for certain misinformation strategies, expanding on cross-modal relationship frameworks [56]. Compared to Bu et al. [10], the framework also structures why and how misinformation detection algorithms use specific properties of misinformation, rather than organizing models by whether they operate on signal, semantic or intent level. Because the framework links how videos misinform to how models detect misinformation, we can audit each step in the pipeline - model design, training data, evaluation - and examine whether the whole pipeline aligns with the misinformation strategy the paper states it detects.

## 5.2 Misinformation Strategies

Our framework identifies eleven video misinformation strategies, grouped into four categories. Because the coding process drew on theoretical misinformation frameworks as reference points, several strategies correspond to theoretical misinformation categories: decontextualization corresponds to false context [2, 57, 81, 93] as it is defined as the reuse of content in different contexts, malicious editing to manipulated content [2, 57, 93] because the editing creates the misinformation, and illusion of proof to false connection [2, 57, 81, 93] since they both denote unrelated information. Additionally, wrong facts corresponds to misreporting [58], as both describe factually incorrect claims, though misreporting is typically defined as unintentional, a distinction the detection literature does not make. Correct facts, wrong conclusion and emotional portrayal do not map onto theoretical categories, as theoretical frameworks tend to classify misinformation by content type rather than by the structure of the argument or the mode of persuasion.

Beyond classifying how misinformation misleads, the presented misinformation strategies differ in what they imply for detection: which modalities carry the misinformation, what artifacts are present, and what detection paradigms can be applied. Not all strategies leave traces in the content that detection algorithms can exploit, and the traces vary across categories. Strategies that alter or synthesize content, such as generative content, malicious editing, and dubbing, produce signal-level artifacts from the synthesis or manipulation process. Decontextualization and illusion of proof produce semantic mismatches between video and claims, but for different reasons: decontextualization presents the video as depicting the claimed event, while illusion of proof does not attempt to match, relying on the presence of a video to lend credibility instead. In contrast, staged content, impersonation, one-sided facts, and emotional portrayal leave no digital artifacts. Staged content may contain real-world inconsistencies in the recording, and impersonation can be detected through biometric or behavioral analysis [3], but one-sided facts produces no traces because the facts and the statement are correct, and emotional portrayal may be difficult to distinguish from legitimate editorial choices.

However, the proposed framework also makes distinctions that theoretical frameworks do not: Fabricated content [2, 57, 81, 93] is split into generative and staged false evidence, because generation creates digital artifacts while staging does not. Therefore, a detection model trained to find generation artifacts would miss staged content, even though both fall under the same theoretical category. Similarly, impostor content [57, 93] is split into dubbing and impersonation: dubbing can be detected by digital artifacts from audio replacement or lip-sync methods, while impersonation requires biometric or behavioral analysis [3]. In contrast, theoretical categories may consist of multiple strategies: polarizing content [58, 81] requires both one-sided facts and emotional portrayal by definition. In the first two cases, collapsing the misinformation strategies into a single category would obscure whether a detection model can actually address the full range of content within that category. In the latter case of polarizing content, splitting it into two misinformation strategies makes for a more atomic description of how misinformation is conveyed. More broadly, the eleven strategies are compositional rather than mutually exclusive: a single video can employ several strategies simultaneously, each operating through different modalities and producing different artifacts. For example, a video can be manipulated and

emotionalized, or wrong facts can be reinforced with generated evidence. The misinformation strategies are therefore combinable and not a typology in which each video is assigned to a single category.

Referencing theoretical frameworks also reveals categories that exist in the social science literature, but see little attention in the inspected corpus. While theoretical frameworks define goal-oriented misinformation categories such as managing attitudes, attacking critical voices, or flooding the information space with contradictions [4, 57], these categories are not addressed in the misinformation detection literature. Instead, all algorithms in the corpus model misinformation detection as a single event in time, which may be insufficient, for example, when detection models are flooded with the same incorrect information across multiple sources. Furthermore, theoretical frameworks define rumors as claims whose veracity cannot be determined [4, 42], but in the video misinformation detection literature, the term is oftentimes used synonymously with misinformation, evidenced by how our data collection had to include the term in the search queries to not miss relevant literature from Bu et al.'s [10] survey (Sec. 3.2). Some papers in the corpus acknowledge uncertainty [7, 47, 63], but those with rumor in the title do not [12, 29, 35, 103].

In the proposed framework, we observe an imbalance between the papers supporting each misinformation strategy (Tab. 4). When inspecting the artifacts produced by different misinformation strategies, we see that the depth of research conducted on them mostly correlates with the artifacts they generate. Strategies that are well-covered in the literature tend to have known signal-level artifacts, such as decontextualization, malicious editing, or generative content (Tab. 4). Staged content, one-sided facts, and emotional portrayal, in contrast, are much less researched, as few to no traces remain in the signal. One-sided facts is mostly unexplored, with only Hou et al. [37] mentioning it.

However, one-sided facts also illustrates that the boundaries to related research domains are blurry. It overlaps with media bias concepts such as selection bias and framing bias. All three describe cases in which the reported facts can be correct, but their selection, omission, or framing supports a particular conclusion [79]. The overlap suggests that adjacent research fields may already contain concepts [78], annotation practices [36, 80], and benchmarks [94] that could be useful for describing misinformation strategies that are difficult to define solely in terms of factuality. At the same time, the exact relation between these concepts remains unclear: one-sided facts, selection bias, and framing bias describe similar mechanisms, but they do not define when biased but factual reporting becomes misinformation. Without a clearer distinction between these and other overlapping concepts, it will remain challenging to decide whether such content should be treated as misinformation, media bias, or both.

Taken together, the gaps identified in this section have different origins. Some misinformation strategies are under-represented because they lack operationalizable artifacts, making them technically harder to address. Others, such as strategic misinformation campaigns or the boundary to rumors, are absent because the detection literature has not yet engaged with corresponding theoretical concepts. In a third case, one-sided facts, the boundary between misinformation and adjacent phenomena such as bias remains undefined. Despite these different causes, the overall effect is the same: the field's current scope of video misinformation is narrower than that described by theoretical frameworks.

### 5.3 Algorithms

The proposed framework classifies algorithms in two ways: First, by how they claim that misinformation can be detected, and secondly, by how models analyze the content. Most inspected papers provide a detection paradigm, stating through which information or proxies they detect misinformation (Sec. 4.3.1). As shown in Sec. 4.5, 45 connections map manipulation strategies to detection paradigms, but these connections concentrate on two reasoning types.

Multimodal inconsistency and fabrication artifact detection check whether the content is consistent or shows manipulation traces. The manipulation traces connect logically to misinformation strategies that produce artifacts

(Sec. 5.2): decontextualization produces cross-modal mismatches, malicious editing and generative content leave signal-level traces, and dubbing introduces digital artifacts. However, manipulation trace based detection cannot detect strategies that lack artifacts, such as staged content, impersonation, one-sided facts, or emotional portrayal.

In contrast, external validation checks claims against external knowledge to verify their factuality. They are mostly used with wrong facts and correct facts wrong conclusion. However, they cannot address strategies where the deception does not rely on incorrect claims, such as emotional portrayal. Together, content integrity and external validation cover a subset of the misinformation strategies, and both are logically grounded in properties of the strategies they target.

The remaining paradigms, emotion/sentiment, style, intent, stance, source credibility, and social reaction, are grounded in statistical properties, rather than targeting specific misinformation strategies. Emotion/sentiment is supported by 19 papers, source credibility by empirical findings such as the verification gap reported by Hu et al. [38], and social reaction by engagement statistics reported by Qi et al. [66]. As shown in Sec. 4.5, statistical detection paradigms are used 53 times across the corpus, but with only 3 connections to misinformation strategies. Given that misinformation strategies differ in modality, artifact profile, and deception mechanism (Sec. 5.2), which strategies aggregate correlations can address is yet to be explored. Empirical evidence also shows that a deeper analysis is needed, as Papadopoulou et al. [64] report no correlation between sentiment and misinformation, and Zong et al. [109] show that comment features introduce spurious correlations between common keywords and labels. Social and credibility paradigms are further constrained: they require the content to spread and accumulate engagement before detection, and the signals they rely on are platform-dependent, as reported by Papadopoulou et al. [64]. Furthermore, how recommendation algorithms influence propagation is currently not discussed. 14 papers do not provide a definition, description or example of misinformation (Sec. 4.5). As such, they specify how they detect misinformation, but not what strategy they detect.

How much certain misinformation strategies are addressed by detection paradigms also varies. Staged content is defined by five papers, but is not connected to any detection paradigm. Correct facts wrong conclusion and one-sided facts each connect through a single paper, and emotional portrayal through two. Even for well-represented misinformation strategies, the connection density to detection paradigms varies: decontextualization connects in 65% of papers that define it, malicious editing in 48%, and generative content in only 30%. Therefore, we see that algorithms focus on detecting traces of certain misinformation strategies, while relying on more global heuristics for others.

#### 5.4 Datasets & Evaluation

As shown in Section 4.4, 42 out of 51 algorithms predict binary labels, collapsing the misinformation strategies into a single class. The datasets they train on follow this pattern: 19 out of 27 datasets provide binary veracity labels, and the most reused datasets, FakeSV [66] and FakeTT [11], provide binary annotations for in-the-wild data. Because many in-the-wild datasets collect misinformation without knowing which misinformation strategies they contain, the distribution of strategies is unknown. A model trained on in-the-wild data may learn to detect some misinformation strategies well, while others it may not detect at all, but binary evaluation cannot reveal any strategy-specific biases.

The per-strategy evaluation by Xu et al. [99] illustrates what binary metrics obscure: their model achieves near-perfect detection for generated content but performs at chance for linguistic manipulations such as entity substitution. Combined with the fact that they label generated content as misinforming, this pattern suggests the classifier learned to detect generation artifacts rather than misinformation, a distinction that binary evaluation alone could not reveal.

The scope of this problem becomes apparent when examining the alignment between argued strategies and dataset annotations (Tab. 9). Of the 21 papers that trace a complete argumentation chain from a manipulation strategy through a detection paradigm to an implementation method, 15 train and evaluate exclusively on datasets that do

not distinguish between strategies. These models argue, for example, that they detect decontextualization through multimodal inconsistency, but evaluate on data that does not annotate whether a given sample is decontextualized or uses a different strategy. If a dataset over-represents one strategy, a model that detects only that strategy may outperform a more general model without the evaluation revealing the bias. The few papers that do use strategy-specific datasets introduce a different problem: they evaluate exclusively on their own data and do not additionally test on a generic in-the-wild benchmark [3, 53, 99]. Thus, it is impossible to assess whether the strategy-specific detection transfers to realistic conditions where multiple strategies co-occur and the distribution is uncontrolled.

From the dataset side, 5 of the 11 strategies have zero coverage with annotated labels (Tab. 8): staged content, one-sided facts, dubbing, impersonation, and emotional portrayal. These are precisely the strategies that Section 4.2 showed to have limited or no forensic artifacts, making them both harder to detect and harder to evaluate. However, without annotated data, models cannot be trained to detect specific strategies, and without strategy-level labels, existing models cannot be evaluated on whether they miss them. Notably, none of the self-reported limitations in the reviewed papers concern the granularity of dataset labels or the alignment between argued manipulation strategies and training annotations (Sec. 4.5.3). This suggests that the disconnect between what models claim to detect and what their evaluation can measure is not yet recognized as a problem within the field.

## 5.5 Limitations

First, the data for RQ1 and the framework were extracted and synthesized by one researcher. While this increased the consistency of the annotations and labels, it may introduce biases or errors. The extracted argumentation chains further depend on what papers explicitly state. As such, the framework does not reflect implicit reasoning or the authors' mental models. Finally, the framework does not capture misinformation strategies that are not addressed in the literature, as we derived it from how the computer science literature describes misinformation, not from in-the-wild examples.

## 5.6 Future directions

*5.6.1 Understanding video misinformation better.* To build robust detection models, the field should first focus on better defining and understanding video misinformation. By doing so, it can shift from conceptualizing detection paradigms on empirical evidence to building models on tested, proven definitions and characteristics of video misinformation. We identify two areas where a clear definition is needed. First, rumor is currently defined as unverifiable claims by theoretical frameworks, but video misinformation detection papers use rumor synonymously with misinformation (Sec. 5.2). Secondly, the boundary between one-sided facts and selection/framing bias is not clear (Sec. 5.2). Without clear boundaries, researchers and annotators can not clearly separate biased and misinforming information, or even decide to use a different approach, such as a scale between the two concepts. If adjacent fields like media bias are taken into account, the misinformation domain may benefit for describing, annotating, and benchmarking misinformation.

We also see two areas where the understanding of video misinformation strategies may be improved. First, theoretical frameworks define strategic misinformation campaigns, but detection methods do not address them. Because campaigns are targeted, distributed through multiple sources in a coordinated manner, or built to mimic how information permeates a discussion space, detection approaches that assume misinformation to be a single, unconnected event may be unable to detect it. Secondly, the proposed framework may overlook some misinformation strategies. Because we built the proposed framework inductively from the video misinformation detection literature, it captures only the strategies of misinformation that the literature currently recognizes. However, because the framework is content-oriented rather than algorithm-oriented (Sec. 5.1), new strategies of misinformation can be defined directly from examples found in the

wild. Thus, future work should analyze real-world examples to find misinformation strategies not defined in the current detection literature. To address the found misinformation strategies, new detection paradigms should also be defined.

*5.6.2 Addressing misinformation better.* To detect video misinformation better and more reliably in the future, we see three concrete points future work may address: First, finding new detection paradigms for misinformation strategies, secondly, including deepfake detection models, and third, utilizing the strengths of VLMs and LLMs better.

In our framework (Fig. 5), we find that misinformation strategies that produce few to no signal level artifacts, such as correct facts, wrong conclusion, staged content, or emotional portrayal have few to no detection paradigm defined for them (Sec. 5.2). Future detection algorithms for these misinformation strategies can therefore not rely on existing detection paradigms and should develop new ones, e.g., by analyzing semantic or contextual clues.

Furthermore, we see potential in deepfake detection algorithms because they give information on whether the content was generated or not, but so far, only Xu et al. [100] use a deepfake detector. However, they define generated content as misleading in their annotations, which mixes authenticity and veracity. Agarwal et al. [3] instead take into account that for generated content to be misleading, it must be presented to create the assumption that the person shown is really them. To bridge the gap between misinformation and deepfake detection, future work should take the presentation of the content into account, for example, whether the content is claimed not to be generated.

Finally, VLMs and LLMs are increasingly popular across various parts of the video misinformation detection pipeline (Sec. 4.3.2). So far, VLMs and LLMs are used for content analysis and context integration [13, 32, 35, 38, 52, 102], reasoning [35, 88, 102, 108], or extracting visual information [32, 52, 96, 100, 102, 108]. However, how well they adapt to their respective tasks and the risks involved in using them are currently unexplored. For example, hallucinations [54] or internal biases [89] may affect the detection of misinformation in unexpected ways and warrant further evaluation.

*5.6.3 Addressing potential shortcomings & risks.* One of the largest shortcomings stems from how current misinformation detection models are evaluated, which does not surface whether the models detect misinformation as they were designed (Sec. 5.4). Binary datasets that can contain any type of misinformation hide which artifacts and clues the model uses for detection (Sec. 5.3). Furthermore, the proportions of misinformation strategies in current in-the-wild datasets are unknown, which may lead to models scoring poorly because the artifacts and clues they are designed to exploit are less prevalent in the evaluation dataset. Notably, none of the papers' reported limitations concern the shortcomings of binary labels or the alignment between argued detection paradigms and training annotations (Sec. 4.5.3), suggesting that this disconnect is not yet recognized as a problem within the field. If the algorithms were evaluated on different misinformation strategies individually, the model's use of different artifacts and cues, and the approach's limitations, could be assessed in greater detail. We further see that some algorithms propose their own synthetic type-specific datasets, which they use to train and evaluate the proposed models (Sec. 4.4). As such, how well their detection method generalizes to real-world misinformation is unknown. Future work should therefore strive for more detailed evaluation, including various misinformation strategies in known proportions, so that the performance of the model's targeted misinformation strategies, as well as its generalization to other misinformation strategies, becomes clear.

While strategy-based evaluations of misinformation give information about where the model fails, why it fails remains opaque. We therefore advocate a shift towards more transparent architectures whose outputs are understandable to humans. So far, 42 out of 51 models predict binary labels, which, combined with the black-box nature of their algorithms, prevents humans from understanding, interpreting, or verifying the classifications. However, we already see models move towards more transparent model outputs: reasoning and grounding show where the misinformation stems from, thereby enabling manual checking of whether the decision is supported by the provided evidence. However, so far,

no paper has assessed whether the reasoning produced by detection models is useful for humans. If reasoning and grounding approaches were combined with evaluation methods and findings from misinformation intervention studies [33], humans could make their own decisions about whether the model output is true, thereby reducing the risk that misinformation detection models are misused to automatically censor what is uploaded to a platform.

Besides misuse of detection models for censorship, there may be other risks when deploying misinformation detection models for content moderation. For example, the hallucination and bias risks identified in Section 5.6.2 translate into concrete harms when models are deployed to filter content: false flags remove legitimate content, and systematic biases may disproportionately affect certain groups or topics. On the other hand, propagation/diffusion and social reaction detection paradigms require information from recipients who watch and interact with the misinformation (Sec. 4.3.1), which may negatively affect them and increase the likelihood of the misinformation resurfacing, as in the *Plandemic* case study (Sec. 1). Therefore, future work should be aware of the risks associated with automatic misinformation detection methods and strive towards solutions that are well-tested, transparent, and human-verifiable.

## 6 Conclusion

We presented a new framework for video misinformation detection that links how misinformation is conveyed to how it is detected. Our framework is built on a corpus of 51 papers, from which we extracted how misinformation deceives, how papers hypothesize it can be detected, and what features and cues they use to detect it, as well as the logical connections between the three. The resulting framework covers 11 misinformation strategies, 11 detection paradigms, and 18 implementation methods. Furthermore, we examined existing video misinformation detection datasets along dimensions of labeling granularity, annotation strategy, and strategy coverage.

Through this analysis, we found a core weakness in the current literature: papers motivate their architectures by referencing specific misinformation strategies, yet they evaluate on datasets that collapse all strategies onto a single binary label. As a result, it is impossible to verify whether a model actually detects misinformation as it was designed to, or whether it relies on shortcuts and other unwanted behavior.

Building on the gaps and shortcomings of the field identified in this review, future work should focus on four key areas of improvement: Understanding and defining misinformation better, building models that address misinformation more robustly and more transparently, evaluating detection models on misinformation strategies, and better assessing the risks of misinformation detection methods. Together, these four improvements should steer the field towards robust identification of misinformation, transparent outputs that humans can verify, and reduced risk of misuse.

## Acknowledgments

Special thanks to Isabella Habereeder, Christoph Mandl and Filip Kučera from the Media Bias Group for the discussions and feedback throughout the writing of this paper.

## References

- [1] S. Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM CSUR*, 57, (Nov. 2024), 1–29. doi:10.1145/3697349.
- [2] [n. d.] About Reuters Fact Check. Retrieved Feb. 18, 2026 from <https://www.reuters.com/fact-check/about/>.
- [3] Shruti Agarwal, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, and Anna Rohrbach. 2023. Watch those words: Video falsification detection using word-conditioned facial motion. In *WACV*, 4699–4708. doi:10.1109/WACV56688.2023.00469.
- [4] Esma Aïmeur, Sabine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13, 1, (Feb. 2023), 30. doi:10.1007/s13278-023-01028-5.

- [5] Naheed Akhtar, Mubbashar Saddique, Khurshid Asghar, Usama Ijaz Bajwa, Muhammad Hussain, and Zulfiqar Habib. 2022. Digital Video Tampering Detection and Localization: Review, Representations, Challenges and Algorithm. *Mathematics*, 10, 2, (Jan. 2022), 168. doi:[10.3390/math10020168](https://doi.org/10.3390/math10020168).
- [6] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimitar Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. In *COLING*.
- [7] Berat Biçer, Bahadır Durmaz, Serhat Aras, and Hamdi Dibeklioglu. 2025. DECEPTiON: Bridging gaps in in-the-wild deception research. *IEEE Transactions on Affective Computing*, 16, (July 2025), 3452–3464. doi:[10.1109/TAFFC.2025.3591205](https://doi.org/10.1109/TAFFC.2025.3591205).
- [8] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7, 1, (Mar. 2018), 71–86. doi:[10.1007/s13735-017-0143-x](https://doi.org/10.1007/s13735-017-0143-x).
- [9] Elena Broda and Jesper Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48, 2, (Apr. 2024), 139–166. doi:[10.1080/23808985.2024.2323736](https://doi.org/10.1080/23808985.2024.2323736).
- [10] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *ACM MM*. doi:[10.1145/3581783.3612426](https://doi.org/10.1145/3581783.3612426).
- [11] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. In *ACM MM*, 1351–1360. doi:[10.1145/3664647.3680663](https://doi.org/10.1145/3664647.3680663).
- [12] Donglin Cao, Xiong Tang, Yanghao Lin, and Dazhen Lin. 2025. Short video rumor detection based on causal graph. *Information Sciences*, 703, 121941, (June 2025). doi:[10.1016/j.ins.2025.121941](https://doi.org/10.1016/j.ins.2025.121941).
- [13] Liyuan Cao, Zihang Guo, and Huaiwen Zhang. 2025. Event consistency-aware robust fake news detection. In *ACM MM*. doi:[10.1145/3746027.3755417](https://doi.org/10.1145/3746027.3755417).
- [14] Angela Carrera-Rivera, William Ochoa, Felix Larrinaga, and Ganix Lasa. 2022. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9, (Jan. 2022), 101895. doi:[10.1016/j.mex.2022.101895](https://doi.org/10.1016/j.mex.2022.101895).
- [15] Lizhi Chen, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2025. Disconfounding fake news video explanation with causal inference. In *IJCAI*, 4842–4850. doi:[10.24963/ijcai.2025/539](https://doi.org/10.24963/ijcai.2025/539).
- [16] Lizhi Chen, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2025. Multimodal Fake News Video Explanation: Dataset, Analysis and Evaluation. (Apr. 2025).
- [17] Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. 2024. SafeWatch: An efficient safety-policy following video guardrail model with transparent explanations. In *ICLR*. doi:[10.48550/arXiv.2412.06878](https://doi.org/10.48550/arXiv.2412.06878).
- [18] Hyewon Choi and Youngjoong Ko. 2022. Effective fake news video detection using domain knowledge and multimodal data fusion on youtube. *Pattern Recognition Letters*, 154, (Feb. 2022), 44–52. doi:[10.1016/j.patrec.2022.01.007](https://doi.org/10.1016/j.patrec.2022.01.007).
- [19] Hyewon Choi and Youngjoong Ko. 2021. Using adversarial learning and biterm topic model for an effective fake news video detection system on heterogeneous topics and short texts. *IEEE access : practical innovations, open solutions*, 9, (Oct. 2021), 164846–164853. doi:[10.1109/ACCESS.2021.3122978](https://doi.org/10.1109/ACCESS.2021.3122978).
- [20] Carme Colomina, Héctor Sánchez Margalef, and Richard Youngs. 2021. The Impact of Disinformation on Democratic Processes and Human Rights in the World. European Parliament, Policy Department for External Relations. doi:[10.2861/59161](https://doi.org/10.2861/59161).
- [21] Luca D’Amiano, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2019. A PatchMatch-based dense-field algorithm for video copy–move detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, (Feb. 2019). doi:[10.1109/tcsvt.2018.2804768](https://doi.org/10.1109/tcsvt.2018.2804768).
- [22] [n. d.] Fake News Detection | European Data Protection Supervisor. Retrieved Mar. 1, 2026 from <https://www.edps.europa.eu/press-publications/publications/techsonar/fake-news-detection>.
- [23] Sheera Frenkel, Ben Decker, and Davey Alba. 2020. How the ‘Plandemic’ Movie and Its Falsehoods Spread Widely Online. *The New York Times*, (May 20, 2020). Retrieved Mar. 1, 2026 from <https://www.nytimes.com/2020/05/20/technology/plandemic-movie-youtube-facebook-coronavirus.html>.
- [24] Zhe Fu, Kanlun Wang, Wangjiaxuan Xin, Lina Zhou, Shi Chen, Yaorong Ge, Daniel Janies, and Dongsong Zhang. 2024. Detecting misinformation in multimedia content through cross-modal entity consistency: a dual learning approach. In *Pacific Asia Conference on Information Systems*.
- [25] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *CVPR*, 15180–15190. doi:[10.1109/CVPR52729.2023.01457](https://doi.org/10.1109/CVPR52729.2023.01457).
- [26] [n. d.] Google trends: fake news word history. Retrieved Sept. 8, 2025 from <https://trends.google.com/trends/explore?date=all&q=Fake%20News&hl=de>.
- [27] Suwani Gunasekara, Saumya Pareek, Ryan M. Kelly, and Jorge Goncalves. 2025. The Influence of Content Modality on Perceptions of Online Misinformation. In *CHI*, 1–10. doi:[10.1145/3706598.3713098](https://doi.org/10.1145/3706598.3713098).
- [28] Hanghui Guo et al. 2025. Consistent and invariant generalization learning for short-video misinformation detection. In *ACM MM*. doi:[10.1145/3746027.3755809](https://doi.org/10.1145/3746027.3755809).
- [29] Longqin Guo, Zeqian Chen, and Xiaoyang Liu. 2025. A fake news detection framework integrating multi-domain and multimodal features. *Neurocomputing*, 658, 131711, (Dec. 2025). doi:[10.1016/j.neucom.2025.131711](https://doi.org/10.1016/j.neucom.2025.131711).

- [30] Zhiwei Guo, Yang Li, Zhenguo Yang, Xiaoping Li, Lap-Kei Lee, and Qing Li. 2024. Cross-modal attention network for detecting multimodal misinformation from multiple platforms. *IEEE Transactions on Computational Social Systems*, 11, (Mar. 2024), 4920–4933. doi:10.1109/TCSS.2024.373661.
- [31] Michael Gusenbauer. 2024. Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Research Synthesis Methods*, 15, 5, (June 2024), 802–817. doi:10.1002/jrsm.1729.
- [32] Linfeng Han, Xiaoming Zhang, Tianbo Wang, Yun Liu, and Zhiqiang Dong. 2026. Enhancing large language model for fake news video detection via cross-modal retrieval. *Information Processing & Management*, (Mar. 2026). doi:10.1016/j.ipm.2025.104471.
- [33] Katrin Hartwig, Frederic Doell, and Christian Reuter. 2024. The landscape of user-centered misinformation interventions - a systematic literature review. *ACM Computing Surveys*, 56, 11, (July 2024), 1–36. doi:10.1145/3674724.
- [34] Angie Drobnic Holan. [n. d.] The Principles of the Truth-O-Meter: How we fact-check. Retrieved Feb. 17, 2026 from <https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/>.
- [35] Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *WWW*, 4684–4698. doi:10.1145/3696410.3714559.
- [36] Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: a Case Study on Media Bias Detection. In *NAACL Findings*, 1370–1386. doi:10.18653/v1/2025.findings-naacl.75.
- [37] Rui Hou, Veronica Perez-Rosas, Stacy Loeb, and Rada Mihalcea. 2020. Towards Automatic Detection of Misinformation in Online Medical Videos. In *ICMI*, 235–243. doi:10.1145/3340555.3353763.
- [38] Lingtong Hu, Zituo Wang, Jiayi Zhu, Yifan Hu, and Xianbing Wang. 2025. MAGE-fend: Multimodal adaptive fusion with guidance from LLM expertise for fake news detection on short video platforms. *Knowledge-Based Systems*, 329, (Nov. 2025), 114298. doi:10.1016/j.knosys.2025.114298.
- [39] X. Huang, T. Ma, H. Tang, and H. Rong. 2025. Knowledge-enhanced dynamic scene graph attention network for fake news video detection. *IEEE Transactions on Multimedia*, PP, 99.0, 1–14. doi:10.1109/TMM.2025.3623491.
- [40] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Computer Vision - Eccv 2018, Pt Xi*, 106–124. doi:10.1007/978-3-030-01252-6\_7.
- [41] Cherilyn Ireton and Julie Posetti, (Eds.) 2018. *Journalism, "Fake News" & Disinformation: Handbook for Journalism Education and Training*. United Nations Educational, Scientific and Cultural Organization.
- [42] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. 2021. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23, 5, (May 2021), 1301–1326. doi:10.1177/1461444820959296.
- [43] Junho Kim, Yongjun Shin, Gyeongho Jung, and Hyunchul Ahn. 2025. A hybrid detection method for YouTube fake news using related video data. *Engineering Applications of Artificial Intelligence*, 156, A, (Sept. 2025), 111130. doi:10.1016/j.engappai.2025.111130.
- [44] Rodney Kinney et al. 2025. The Semantic Scholar Open Data Platform. (Apr. 2025).
- [45] Xiangzheng Kong, Zhi Zeng, Chenxi Zhu, Zihan Ma, and Minnan Luo. 2025. Harmony in chaos: a progressive noise-resilient network for robust fake news video detection. In *ICME*, 1–6. doi:10.1109/ICME59968.2025.11208997.
- [46] Litty Koshy and S. Prayla Shyry. 2025. Detection of tampered real time videos using deep neural networks. *Neural Computing and Applications*, 37, 11, 7691–7703. doi:10.1007/s00521-024-09988-1.
- [47] Rina Kumari, Vipin Gupta, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2023. Emotion aided multi-task framework for video embedded misinformation detection. *Multimedia Tools and Applications*, 83, (Oct. 2023), 37161–37185. doi:10.1007/s11042-023-17208-6.
- [48] Mingxin Li, Yuchen Zhang, Haowei Xu, Xianghua Li, Chao Gao, and Zhen Wang. 2025. Learning complex heterogeneous multimodal fake news via social latent network inference. In *AAAI*, 433–441. doi:10.1609/aaai.v39i1.32022.
- [49] Ruofan Li, Wei Zhang, and Yong Liu. 2025. DAFSVFND: Dual attention fusion network for fake news detection on short video platforms. In *ICDAR*, 629–646. doi:10.1007/978-3-032-04624-6\_37.
- [50] 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, (Apr. 2022), 6092. doi:10.1038/s41598-022-10117-y.
- [51] Yili Li, Jian Lang, Rongpei Hong, Qing Chen, Zhangtao Cheng, Jia Chen, Ting Zhong, and Fan Zhou. 2025. REAL: Retrieval-augmented prototype alignment for improved fake news video detection. In *ICME*, 1–6. doi:10.1109/ICME59968.2025.11209008.
- [52] Fang Liu, Yili Li, Jian Lang, Rongpei Hong, and Fan Zhou. 2026. Enhancing fake news video detection with self-driven question-answer from LLMs. *Information Processing & Management*, (Mar. 2026). doi:10.1016/j.ipm.2025.104432.
- [53] Fuxiao Liu, Yaser Yacoub, and Abhinav Shrivastava. 2023. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. In *EACL*, 178–188. doi:10.18653/V1/2023.EACL-MAIN.14.
- [54] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A Survey on Hallucination in Large Vision-Language Models. (May 2024).
- [55] Ying Lu and Naiwei Yao. 2025. A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network. *Scientific Reports*, 15, 1, (July 2025). doi:10.1038/s41598-025-05702-w.
- [56] 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 *ICWSM*, 651–662. doi:10.1609/icwsml.v16i1.19323.

- [57] [n. d.] Misinformation and Disinformation, UNHCR. Retrieved Dec. 13, 2024 from <https://www.unhcr.org/innovation/wp-content/uploads/2022/02/Factsheet-4.pdf>.
- [58] Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. 2021. “Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. *American Behavioral Scientist*, 65, 2, (Feb. 2021), 180–212. doi:10.1177/0002764219878224.
- [59] Erica Mourão, João Felipe Pimentel, Leonardo Murta, Marcos Kalinowski, Emilia Mendes, and Claes Wohlin. 2020. On the performance of hybrid search strategies for systematic literature reviews in software engineering. *Information and Software Technology*, 123, (July 2020), 106294. doi:10.1016/j.infsof.2020.106294.
- [60] Scott Neuman. 2020. Seen ‘Plandemic’? We Take A Close Look At The Viral Conspiracy Video’s Claims. *NPR*, (May 8, 2020). Retrieved Mar. 1, 2026 from <https://www.npr.org/2020/05/08/852451652/seen-plantemic-we-take-a-close-look-at-the-viral-conspiracy-video-s-claims>.
- [61] [n. d.] Oxford Word of the Year 2016 | Oxford Languages. Retrieved Sept. 8, 2025 from <https://languages.oup.com/word-of-the-year/2016/>.
- [62] Matthew J Page et al. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, (Mar. 2021), n160. doi:10.1136/bmj.n160.
- [63] Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading Metadata Detection on YouTube. In *ECIR 2019*, 140–147. doi:10.1007/978-3-030-15719-7\_18.
- [64] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online Information Review*, 43, 1, (Feb. 2019), 72–88. doi:10.1108/OIR-03-2018-0101.
- [65] Tim Polzehl, Vera Schmitt, Nils Feldhus, Joachim Meyer, and Sebastian Möller. 2023. Fighting disinformation: Overview of recent AI-based collaborative human-computer interaction for intelligent decision support systems. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 267–278. doi:10.5220/0011788900003417.
- [66] 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 *AAAI*, 14444–14452. doi:10.1609/AAAI.V37I12.26689.
- [67] 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 *ACL Findings*, 11947–11959. doi:10.18653/v1/2023.findings-acl.756.
- [68] Yohanna Rodriguez-Ortega, Dora M. Ballesteros, and Diego Renza. 2021. Copy-Move Forgery Detection (CMFD) Using Deep Learning for Image and Video Forensics. *Journal of Imaging*, 7, 3, (Mar. 2021), 59. doi:10.3390/jimaging7030059.
- [69] Ana Romero-Vicente and Ira Pragnya Senapati. 2025. Platforms’ Policies on Climate Change Misinformation. EU DisinfoLab, (July 2025). Retrieved Mar. 1, 2026 from [https://eu.boell.org/sites/default/files/2025-07/final\\_20250722-platforms-policies-on-climate-change-misinformation\\_1.pdf](https://eu.boell.org/sites/default/files/2025-07/final_20250722-platforms-policies-on-climate-change-misinformation_1.pdf).
- [70] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*.
- [71] Saminder Dhesei, Laura Fontes, Pedro Machado, Farhad Fassihi Tash, and David Ada Adama. 2023. Mitigating adversarial attacks in deepfake detection: An exploration of perturbation and AI techniques.
- [72] Gautam Kishore Shahi, Amit Kumar Jaiswal, and Thomas Mandl. 2024. FakeClaim: a multiple platform-driven dataset for identification of fake news on 2023 israel-hamas war. In *Advances in Information Retrieval, Ecir 2024, Pt V*, 66–74. doi:10.1007/978-3-031-56069-9\_5.
- [73] Shaikh Akib Shahriyar and Matthew K. Wright. 2022. Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*, 13–18. doi:10.1145/3494109.3527194.
- [74] 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)*, 899–908. doi:10.1109/BigData52589.2021.9671928.
- [75] Lanyu Shang, Yang Zhang, Yawen Deng, and Dong Wang. 2025. MultiTec: a data-driven multimodal short video detection framework for healthcare misinformation on TikTok. *IEEE Transactions on Big Data*, 11, (Jan. 2025), 2471–2488. doi:10.1109/TBDATA.2025.3533919.
- [76] K. Sitara and Babu M. Mehtre. 2016. Digital video tampering detection: An overview of passive techniques. *Digital Investigation*. doi:10.1016/j.dii.2016.06.003.
- [77] Snopes. [n. d.] Fact Check Ratings. Retrieved Feb. 18, 2026 from <https://www.snopes.com/fact-check-ratings/>.
- [78] Timo Spinde. 2025. *Automated Detection of Media Bias: From the Conceptualization of Media Bias to Its Computational Classification*. Springer Fachmedien. doi:10.1007/978-3-658-47798-1.
- [79] Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2024. The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. (Jan. 2024).
- [80] Timo Spinde, Christina Kreuter, Wolfgang Gaismaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2024. Do You Think It’s Biased? How to Ask for the Perception of Media Bias. In *JCDL*, 61–69. doi:10.1109/JCDL52503.2021.00018.
- [81] [n. d.] Spotting disinformation: Six tactics used to fool us. Retrieved Jan. 21, 2026 from <https://www.europarl.europa.eu/topics/en/article/20250227STO27081/spotting-disinformation-six-tactics-used-to-fool-us>.
- [82] Marianna Spring. 2020. Coronavirus: ‘Plandemic’ virus conspiracy video spreads across social media. *BBC*, (May 8, 2020). Retrieved Mar. 1, 2026 from <https://www.bbc.com/news/technology-52588682>.
- [83] 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, (Nov. 2021), 301–319. doi:10.1093/jcmc/zmb010.
- [84] Yoo Yeon Sung, Jordan Boyd-Graber, and Naemul Hassan. 2023. Not all fake news is written: a dataset and analysis of misleading video headlines. In *EMNLP*. doi:10.18653/v1/2023.emnlp-main.1010.

- [85] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. 2017. Defining “Fake News”: A typology of scholarly definitions. *Digital journalism*, 6, 2, (Aug. 2017), 137–153. doi:10.1080/21670811.2017.1360143.
- [86] [n. d.] The Washington Post’s guide to manipulated video. Retrieved May 15, 2025 from <https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/>.
- [87] Martijn Visser, Nees Jan van Eck, and Ludo Waltman. 2021. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2, 1, (Apr. 2021), 20–41. doi:10.1162/qss\_a\_00112.
- [88] Junxi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025. FakeSV-VLM: Taming VLM for Detecting Fake Short-Video News via Progressive Mixture-Of-Experts Adapter. In *EMNLP*. doi:10.18653/v1/2025.findings-emnlp.257.
- [89] Sibowang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2026. VLBiasBench: A Comprehensive Benchmark for Evaluating Bias in Large Vision-Language Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (Apr. 2026), 1–14. doi:10.1109/TPAMI.2026.3683747.
- [90] Wenhao Wang, Mingxin Li, Jiao Qiao, Haotong Du, Xianghua Li, Chao Gao, and Zhen Wang. 2025. MFAE: Multimodal Feature Adaptive Enhancement for Fake News Video Detection. In *CIKM*. doi:10.1145/3746252.3761344.
- [91] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *SIGKDD*, 849–857. doi:10.1145/3219819.3219903.
- [92] Zituo Wang, Lingtong Hu, Jiayi Zhu, Donggyu Kim, and Xiaojing Bo. 2025. Learning to live with COVID-19: Informative fictions of TikTok misinformation and multimodal video analysis. *Social Science Computer Review*. doi:10.1177/08944393251366232.
- [93] Claire Wardle and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. Council of Europe Strasbourg.
- [94] Martin Wessel, Tomáš Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2023. Introducing MBIB – the first Media Bias Identification Benchmark Task and Dataset Collection. In *SIGIR*, 2765–2774. doi:10.1145/3539618.3591882.
- [95] Jevin D. West and Carl T. Bergstrom. 2021. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118, 15, (Apr. 2021), e1912444117. doi:10.1073/pnas.1912444117.
- [96] Z. Xia, S. Zhao, J. Han, and J. Chen. 2025. Knowledge-augmented contrastive learning and multi-modal fusion for fake news detection service in social network. In *IEEE International Conference on Web Services (ICWS)*, 984–995. doi:10.1109/ICWS67624.2025.00130.
- [97] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating misinformation in the era of generative AI models. In *ACM MM*, 9291–9298. doi:10.1145/3581783.3612704.
- [98] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metzke, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *EMNLP*, 6787–6800. doi:10.18653/v1/2021.emnlp-main.544.
- [99] Qingzheng Xu, Huiqiang Chen, Heming Du, Hu Zhang, Szymon Lukasik, Tianqing Zhu, and Xin Yu. 2024. M3A: A multimodal misinformation dataset for media authenticity analysis. In *Computer Vision and Image Understanding*. doi:10.1016/j.cviu.2024.104205.
- [100] Qingzheng Xu, Heming Du, Szymon Lukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025. MDAM3: a misinformation detection and analysis framework for multitype multimodal media. In *WWW*, 5285–5296. doi:10.1145/3696410.3714498.
- [101] Xiong Xu, Shuai Tang, Mingcheng Zhu, Peisong He, Sirui Li, and Yun Cao. 2023. A novel model compression method based on joint distillation for deepfake video detection. In *Journal of King Saud University-Computer and Information Sciences*. doi:10.1016/j.jksuci.2023.101792.
- [102] Bingjian Yang, Danni Xu, Kaipeng Niu, Wenxuan Liu, Zheng Wang, and Mohan Kankanhalli. 2025. A New Dataset and Benchmark for Grounding Multimodal Misinformation. In *ACM Multimedia*. doi:10.1145/3746027.3758191.
- [103] Ming Yin, Wei Chen, Dan Zhu, and Jijiao Jiang. 2025. Enhancing video rumor detection through multimodal deep feature fusion with time-synccomments. *Information Processing & Management*, 62, 103935, (Jan. 2025), 1, (Jan. 2025). doi:10.1016/j.ipm.2024.103935.
- [104] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C. Kot. 2024. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. In *IEEE Transactions on Dependable and Secure Computing*, 4327–4342. doi:10.1109/TDSC.2024.3352049.
- [105] Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *ACM MM*, 6492–6500. doi:10.1145/3664647.3681673.
- [106] Zhi Zeng, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. Understand, refine and summarize: Multi-view knowledge progressive enhancement learning for fake news video detection. In *ACM MM*. doi:10.1145/3746027.3754551.
- [107] Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025. IMOL: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *ACL*, 30921–30933. doi:10.18653/v1/2025.acl-long.1494.
- [108] Yuchen Zhang, Mingxin Li, Chao Gao, and Xianghua Li. 2025. Confidence breeds success: Improving fake news video detection via LVLM-assisted inference. In *ICME*, 1–6. doi:10.1109/ICME59968.2025.11209223.
- [109] Linlin Zong, Wenmin Lin, Jiahui Zhou, Xinyue Liu, Xianchao Zhang, Bo Xu, and Shimin Wu. 2025. Text-guided fine-grained counterfactual inference for short video fake news detection. In *AAAI*, 1237–1245. doi:10.1609/aaai.v39i1.32112.
- [110] Linlin Zong, Shilin Sui, Wenjun Liang, Wanyu Song, Linlin Tian, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2025. CH-SV: a benchmark for multi-type chinese harmful short video detection. In *ACM MM*. doi:10.1145/3746027.3758279.

# CiteAssist

## CITATION SHEET

Generated with [citeassist.uni-goettingen.de](https://citeassist.uni-goettingen.de)

### BibTeX Entry

```
@misc{braun2026,  
  author={Braun, Florian and Hoang, Truc and Demartini, Gianluca and Echizen, Isao and  
    Spinde, Timo},  
  title={Video Misinformation Detection: A Systematic Review of Manipulation Tactics,  
    Datasets and Algorithms},  
  howpublished={preprint},  
  year={2026},  
  month={05},  
  url={https://media-bias-research.org/wp-content/uploads/2026/05/braun2026.pdf},  
  journal={ACM Computing Surveys (under review)},  
  date={2026},  
  publisher={Association for Computing Machinery (ACM)}}  
}
```

### Online Access

**Official Publication** <https://media-bias-research.org/wp-content/uploads/2026/05/braun2026.pdf>