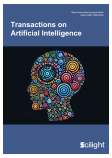
Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Review

Check for updates

A Comprehensive Survey of Multimodal Fake News Detection: Datasets, Methods, and Challenges

Guoyong Cai [1,2,*], Zhipeng Qiu [1], Guoxin Bi [1] and Qinghua Liu [1,2]

[1] School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[2] Guangxi Key Lab of Trusted Software, Guilin 541004, China
* Correspondence: ccgycai@guet.edu.cn

How To Cite: Cai, G.; Qiu, Z.; Bi, G.; et al. A Comprehensive Survey of Multimodal Fake News Detection: Datasets, Methods, and Challenges. *Transactions on Artificial Intelligence* **2026**, *2*(1), 131–160. https://doi.org/10.53941/tai.2026.100009

Received: 4 February 2026
Revised: 27 March 2026
Accepted: 3 April 2026
Published: 12 May 2026

Abstract: With the rapid proliferation of multimodal social media platforms, fake news has been increasingly disseminated through multiple modalities such as text, images, and videos, posing serious threats to social stability, public cognition, and these platforms' ecosystem. Existing unimodal fake news detection methods face much challenge in multimodal scenarios, as they can not capture fully cross-modal semantic correlations and inconsistencies. Multimodal fake news detection, which integrates heterogeneous information from text, visual, and audio modalities to explore inter-modal consistency and complementarity, has therefore become a major research focus in recent years. A comprehensive survey of recent advances in multimodal fake news detection is presented in the paper, which consists of systematically reviews of the fundamental concepts, detection tasks, and underlying technical principles in this field. Major benchmark datasets and commonly used evaluation metrics are also introduced, followed by a structured taxonomy of representative detection methods and a summary of their experimental results. Furthermore, the key challenges faced by current research are discussed and promising future research directions are outlined. Compared with existing surveys, this work presents a more comprehensive method categorization that emphasizes the evolution of detection techniques, offers clearer comparisons of datasets and experimental analyses, and provides more practical insights for researchers and practitioners in multimodal fake news detection.

Keywords: fake news detection; multimodal learning; feature extraction; deep learning

1. Introduction

With the rapid development of the Internet and social media technologies, news dissemination has become increasingly diversified, evolving from single-text formats toward multimodal fusion. Consequently, fake information is no longer conveyed through isolated textual descriptions but is instead reinforced by combinations of text with images, audio, and videos, substantially increasing its deceptive power [1]. Such multimodal fake news, characterized by attention-grabbing content and rapid dissemination, has posed significant threats to public cognition, social opinion stability, and even national security [2]. Studies have shown that during major public events, multimodal fake news spreads significantly faster than text-only misinformation and exhibits stronger misleading effects, highlighting the urgent need for effective technical approaches to accurately identify and mitigate its impact [3].

Traditional unimodal fake news detection methods suffer from inherent limitations. By relying solely on textual stylistic features or single-source propagation data, these approaches fail to capture semantic correlations and inconsistencies across different modalities. For instance, a news report claiming that a natural disaster has

occurred in a certain region may present logically coherent textual descriptions, while the accompanying image is actually taken from an unrelated event several years earlier. Unimodal methods, which lack the ability to verify cross-modal consistency, are prone to misclassification in such cases. As multimodal technologies become increasingly accessible, the fabrication of fake information has grown more sophisticated and concealed, further exposing the performance bottlenecks of unimodal detection methods in complex real-world scenarios. Against this backdrop, multimodal fake news detection has emerged as a promising research direction. By integrating heterogeneous information from text, images, audio, and videos, this paradigm aims to exploit inter-modal complementarity and consistency, thereby becoming a key pathway to improve detection accuracy.

In recent years, the field of multimodal fake news detection has witnessed substantial research progress. Technical paradigms have evolved from early approaches based on handcrafted features and traditional machine learning with simple feature fusion, to deep learning-based methods incorporating cross-modal attention mechanisms. Moreover, research has expanded beyond content-based features to include social context, external knowledge graphs, and propagation dynamics. In particular, the introduction of large language models (LLMs) has further enhanced the capability of models to reason about complex cross-modal inconsistencies [2]. Despite these advances, several critical challenges remain unresolved. The persistent "semantic gap" across modalities often limits fusion effectiveness; real-world data characteristics such as scarce fake samples and severe class imbalance hinder model generalization; the black-box nature of many deep models undermines interpretability and public trust; and model adaptability degrades significantly when confronted with emerging AI-generated multimodal fake content (AIGC). These issues indicate that multimodal fake news detection has yet to establish a systematic theoretical framework and standardized application paradigm.

From the perspective of existing survey studies [1–5], although several reviews have attempted to summarize progress in multimodal fake news detection, notable limitations persist. Some surveys focus on a single technical direction, such as deep learning models, while neglecting the complete research pipeline encompassing datasets, methodologies, and evaluation protocols [3,5]. Other reviews cover multiple dimensions but lack a clear logic of technological evolution in their method categorization, and fail to conduct systematic comparisons of experimental results across different datasets, making it difficult for researchers to identify core trends in the field [4]. Furthermore, recent advances, including collaborative reasoning with large language models, are insufficiently covered, and practical deployment challenges in real-world scenarios are rarely analyzed in depth [1,2]. As a result, researchers are often required to invest substantial effort in navigating fragmented literature, which hinders the rapid formation of a holistic understanding of the field.

To address these limitations, this paper presents a systematic and comprehensive survey of recent advances in multimodal fake news detection. The contributions of this work are threefold: First, we bridge the gap in existing surveys by constructing a unified framework that encompasses concept definitions, datasets and evaluation metrics, method categorization, experimental performance comparisons, and challenges with future directions, enabling researchers to quickly grasp the overall landscape of the field. Second, we propose a novel method taxonomy guided by both technological evolution and task scenarios, categorizing existing approaches into nine core directions and clarifying their respective strengths and applicable settings through comparative experimental analysis. Third, with a focus on real-world application requirements, we conduct an in-depth analysis of current bottlenecks related to dataset quality, model interpretability, and real-time detection efficiency, and outline future research directions that balance theoretical significance with practical feasibility. This survey aims to provide clear guidance for theoretical research on multimodal fake news detection, while also offering valuable insights for real-world deployment, ultimately contributing to the construction of a healthier and more trustworthy digital information ecosystem.

The remainder of this paper is organized as follows. Section 2 first defines fake news, its core characteristics, and detection tasks, with an emphasis on clarifying the concept, technical foundations, and application scenarios of multimodal fake news detection. Section 3 reviews major multimodal datasets, comparing their properties in terms of data sources, modality types, and application scenarios, and introduces commonly used evaluation metrics along with multimodal-specific assessment criteria to establish a unified evaluation standard. Section 4 presents the technical principles, innovations, and representative achievements of existing methods, and systematically compares their experimental performance on benchmark datasets through comprehensive tables. Section 5 summarizes key challenges such as dataset limitations and cross-domain generalization issues, and discusses future research directions including high-quality dataset construction and advanced multimodal fusion strategies. Finally, Section 6 concludes the paper by summarizing current achievements, outstanding challenges, and future prospects.

2. Background

Fake news detection is an important research direction at the intersection of information security and natural language processing, and its core conceptual framework has been continuously enriched with the advancement of related technologies. Clearly defining fake news, clarifying the scope of detection tasks, and understanding the unique characteristics of multimodal scenarios are fundamental prerequisites for conducting research in this field.

2.1. Fake News

Fake news refers to information that is intentionally generated by publishers or disseminators and has been verified to be inconsistent with objective facts. Such information is disseminated through various channels, including news reports, social media posts, and short videos, and is characterized by the combination of falsity and intentionality [3].

Falsity is reflected in fundamental deviations from real events, such as fabricating non-existent incidents, distorting key factual elements, or selectively presenting information out of context to mislead interpretation. Intentionality, on the other hand, relates to the underlying motivations of the disseminators, including generating traffic and economic benefits by inciting panic, influencing public attitudes through misinformation, or damaging the reputation of specific individuals, organizations, or social groups using false content.

2.2. Fake News Detection

Fake news detection aims to assess the authenticity of information through technical means, with the goal of rapidly identifying false content from massive information streams and curbing its further dissemination. As a comprehensive task, fake news detection not only requires analyzing the intrinsic content features of information but also involves modeling the dissemination context, user behaviors, and social interactions, making it considerably more complex than traditional text classification tasks [4].

Early detection approaches primarily relied on manual verification and rule-based matching, which depended heavily on expert knowledge or predefined features and suffered from low efficiency and limited coverage. With the advancement of machine learning techniques, detection methods gradually shifted toward automation by extracting linguistic and propagation-related features from text and employing classifiers for large-scale detection. In recent years, the adoption of deep learning has further improved detection performance. Neural network-based models are capable of automatically learning deep semantic representations and, when combined with social network analysis and external knowledge sources, form multidimensional detection frameworks that provide effective technical support for efficient fake news identification.

2.3. Multimodal Fake News Detection

Multimodal fake news detection refers to detection techniques that integrate multiple modalities, such as text, images, audio, and videos, with the aim of improving identification accuracy through cross-modal correlation analysis. Its emergence and development are driven by the diversification of fake news dissemination forms. Contemporary fake information no longer relies solely on textual descriptions but often enhances deceptiveness through multimodal combinations, such as pairing sensational text with irrelevant images to attract attention, or generating audio–visual content that contradicts textual narratives using deepfake technologies.

The core objective of multimodal detection lies in uncovering intrinsic relationships among different modalities. On the one hand, explicit inconsistencies can be identified through cross-modal consistency verification; on the other hand, feature representations can be enriched by exploiting the complementary information provided by different modalities. From a technical perspective, multimodal fake news detection involves addressing three key challenges: modal feature extraction, multimodal fusion strategies, and semantic alignment across modalities.

Currently, multimodal fake news detection has formed a deep learning–centered technical framework. For example, Transformer-based cross-modal attention models are capable of capturing fine-grained associations between text and images; graph neural networks can model dynamic interactions of multimodal information within propagation networks; and large language models enhance reasoning over complex cross-modal inconsistencies through knowledge augmentation. These approaches have demonstrated significant advantages in applications such as social media rumor detection and content moderation on news platforms, and have become essential technical solutions for combating multimodal fake information [5].

3. Datasets and Evaluation Metrics

3.1. Datasets

The construction and development of fake news detection datasets are closely aligned with the evolving requirements of detection techniques. As research progresses, dataset scale, coverage, and modal diversity have been continuously expanded. In the field of fake news detection, the richness and diversity of datasets play a critical role in model training and evaluation. Datasets with different sources, scales, and data types provide researchers with multidimensional perspectives for analysis and experimentation.

In this survey, we conduct a systematic review and analysis of widely used fake news detection datasets, as summarized in Table 1. The following subsection introduces several representative datasets in detail.

From the perspectives of data sources and language distribution, existing datasets can be broadly categorized into Chinese, English, and multilingual datasets. Chinese datasets are predominantly collected from the Weibo platform, such as Weibo-16, Weibo-20, and Weibo-21. These datasets cover both text-only and text–image multimodal settings and further incorporate propagation trees and user interaction information that reflect the characteristics of Chinese social media. As a result, they provide valuable support for studies on propagation patterns and social feedback in Chinese-language scenarios. In contrast, English datasets are more abundant and originate from a wider range of sources, including social media platforms such as Twitter, comprehensive platforms such as Reddit and Fakeddit, as well as professional fact-checking websites such as GossipCop and PolitiFact. Datasets derived from fact-checking websites tend to exhibit stronger domain specificity, particularly in vertical domains such as politics and entertainment. Although multilingual datasets are relatively limited in number, datasets such as PHEME and MM-COVID, which support both Chinese and English, provide essential data foundations for cross-lingual fake news detection research.

From the perspective of modality types, datasets can be divided into unimodal and multimodal categories. Unimodal datasets primarily focus on textual information, including Weibo-16, Twitter-15, and LIAR. These datasets were developed at an early stage and are available in large quantities, making them suitable for early research on textual feature extraction and propagation pattern analysis. However, due to their reliance on a single modality, they are insufficient for detecting complex multimodal fake content. Multimodal datasets, on the other hand, integrate additional modalities such as images, videos, and audio on top of textual data. For example, Weibo-21 extends earlier Weibo datasets by incorporating image modalities, FakeSV further includes video and audio modalities, and Fakeddit is characterized by its large-scale samples and rich multimodal annotations. These datasets better reflect the current dissemination patterns of multimodal fake news and have therefore become the primary data resources supporting recent research.

From the perspectives of annotation quality and modal consistency, the quality evolution of datasets exhibits a transition from “simple association” toward “deep alignment”. Early datasets, such as Twitter, Weibo-16, and Weibo, primarily relied on platform-generated tags or standard labeling processes; while these possess advantages in scale, there remains significant room for optimization in cross-modal deep semantic interaction and fine-grained alignment. In contrast, Weibo-21, GossipCop, FakeNewsNet, and MM-COVID have significantly enhanced the credibility and annotation quality of data samples through expert review, professional website verification, or authoritative institutional labeling. Notably, FakeSV achieved precise chronological synchronization and alignment across text, audio, and video modalities within a short-video context for the first time, effectively addressing the issue of semantic fragmentation between modalities found in earlier datasets.

From the perspective of fake news types, datasets are shifting from simple factual fabrication toward complex logical manipulation and fine-grained classification. Fakeddit provides high semantic complexity by constructing a three-level label system that covers various fine-grained deceptive types, such as satire, misleading content, and tampering. Similarly, LIAR overcomes the limitations of traditional binary classification through a six-level labeling scheme, placing greater emphasis on the degree of deviation in factual logic. Furthermore, PolitiFact, Snopes, and the multi-domain samples in Weibo-21 demonstrate high diversity in detecting factual tampering across politics, healthcare, and cross-domain scenarios, thanks to their rigorous chains of evidence or extensive domain coverage. This multi-dimensional quality assessment not only reveals the academic value of the datasets themselves but also provides a basis for judging their applicability and limitations in various complex scenarios.

Table 1. Summary of representative fake news detection datasets.

Dataset	Size	Source	Modality	Annotation Quality	Modal Consistency	Fake News Type	Description
Weibo-16 [6]	4664 (2313 fake news items)	Weibo	Text	Medium: Relies on official tags; lacks manual verification	Contains only text modality	Focuses on early social propagation features	Chinese dataset; includes propagation trees; suitable for studying dissemination patterns
Weibo-20 [7]	6362 (3161 fake news items)	Weibo	Text	High: Combined with user interaction and social exchange feedback verification.	Contains only text modality	Focuses on text style and propagation analysis	Contains user interaction data; supports social feedback analysis
Weibo-21 [8]	9128 (4488 fake news items)	Weibo	Text, Image	High: Expert review and emotional consistency check.	High: Deep semantic alignment of text and image	Covers complex fake news types across 9 domains	Multimodal dataset covering nine domains
Weibo [9]	9528 (4749 fake news items)	Weibo	Text, Image	Medium: Standard labeling process; contains some noise.	Medium: Basic image-text association; shallow semantic interaction	Primarily traditional image-text misleading types	Includes paired textual content and corresponding images
PHEME [10]	5802 (1972 fake news items)	Twitter	Text, Image	High: Based on journalist verification and debunking tags; supports cross-lingual validation.	Medium: Focuses on image-text factual association; covers multi-language semantic alignment	Centers on specific breaking events; fake logic is relatively concentrated	Supports English; multimodal dataset suitable for rumor detection
Twitter-15 [11]	1490 (370 fake news items)	Twitter	Text	Medium: Primarily relies on platform tags; focuses on propagation structure labeling.	Contains only single text modality	Focuses on early social media rumors; types are mostly text fabrication	English dataset; includes dissemination paths; suitable for propagation-based analysis
Twitter-16 [11]	818 (205 fake news items)	Twitter	Text	High: Combined with forwarding and comment data for multi-dimensional social cross-validation	Contains only single text modality	Focuses on social interaction behavior analysis; fake forms are relatively traditional	Includes retweet and comment data; supports social interaction analysis
Twitter [12]	about 17,000 tweets	MediaEval-15, MediaEval-16	Text, Image	Medium: Competition source data; contains some automated labeling noise	Medium: Possesses basic image-text correspondence; semantic association between modalities is relatively simple	Primarily covers misleading multimedia content; relatively little deepfake content	Multimodal dataset with multiple modalities; suitable for detecting manipulated media content on social platforms
GossipCop [13]	22,140	GossipCop	Text, Image	High: Professional website verification; focuses on authenticity judgment of celebrity news	Medium: Primarily based on textual claims; visual modality often serves as auxiliary display	Focuses on the entertainment field; fake logic is relatively concentrated on performance information	Focuses on the entertainment domain; emphasizes celebrity misinformation detection
PolitiFact [13]	3568	PolitiFact	Text, Image	Extremely High: Labeled by authoritative political fact-checking agencies; possesses a rigorous chain of evidence	Medium: Focuses on the correspondence between political claims and image facts	Covers complex factual tampering such as policy misleading and speech fabrication	Political domain; supports political misinformation research

Table 1. *Cont.*

Dataset	Size	Source	Modality	Annotation Quality	Modal Consistency	Fake News Type	Description
FakeNewsNet [14]	23,196 (5755 fake news items)	GossipCop, PolitiFact	Text, Image	High: Integrates multi-source fact-checking data; includes complete social context	Medium: Standard image-text association; supports multimodal feature fusion verification	Covers both political and entertainment domains; diverse forms of fake expressions	Multi-source dataset integrating social context and user engagement data
FakeSV [15]	3654 (1827 fake news items)	Douyin, Kuaishou	Text, Video, Audio	Extremely High: Includes manual double-audit and complete social feedback data	Extremely High: Precise chronological synchronization and alignment achieved across text, audio, and video modalities	Covers unique fake news types for short videos such as audio-visual inconsistency and deepfakes	Short-video dataset; multimodal with complete social interaction context
LIAR [16]	12,836	PolitiFact	Text	High: Professional fact-checkers labeled based on a six-level scale	Contains only text modality	High degree of refinement; includes logic such as “half-true, half-false”	Fine-grained labels; supports degree-based fake news classification
MM-COVID [17]	11,565 (3981 fake news items)	Multiple social platforms	Text, Image	High: Authoritative institutions labeled based on specific epidemic facts	High: Focuses on the consistency between epidemic-related images and text facts	Types are concentrated on medical rumors and misleading science popularization	COVID-19–focused dataset
Snopes [18]	4341	Snopes	Text	Extremely High: Sourced from veteran fact-checking websites; possesses a complete chain of evidence	Primarily validates the authenticity of textual statements	Covers multimodal fake claims across multiple fields such as politics, science, and culture	Cross-domain fact-checking; supports multi-domain fake news detection
Fakeddit [19]	More than 1 million samples	Reddit	Text, Image	High: Weak supervision combined with manual spot checks; massive labeling scale	Medium: Possesses image-text correspondence; however, some shallow associations exist due to the large scale	Three-level label system covering various types such as satire and tampering	Large-scale multimodal dataset; supports fine-grained classification and multiple fake news categories

From the perspectives of application scenarios and research objectives, datasets can also be categorized according to their focus on propagation analysis, domain-specific detection, or fine-grained classification. Datasets such as Weibo-16 and Twitter-16, which emphasize propagation analysis, are characterized by the inclusion of propagation trees, reposting records, and comment information, facilitating studies on diffusion paths and spread dynamics. Domain-specific datasets are constructed to meet the requirements of particular application scenarios. For instance, GossipCop focuses on the entertainment domain, PolitiFact targets political news, and MM-COVID is designed for pandemic-related misinformation. Samples in these datasets exhibit domain-specific content characteristics and deception patterns, enabling improved detection performance in vertical scenarios. In addition, datasets such as LIAR and Fakeddit support fine-grained classification by introducing detailed labels, such as degrees of falsity and types of misinformation, thereby overcoming the limitations of traditional binary classification and providing data support for more precise fake news governance.

From the perspective of research limitations, existing datasets exhibit significant boundaries of applicability while supporting diverse detection tasks. First, there are constraints in geographical and linguistic adaptation. Chinese datasets such as Weibo-16, Weibo-20, Weibo-21, and Weibo are deeply coupled with the propagation characteristics and contexts of domestic social media, making it difficult to directly migrate them to research on overseas social platforms. Although FakeSV fills the gap in domestic short-video detection, its content is entirely based on the Chinese environment, which limits its support for generalized research on cross-lingual or global fake news propagation. Second, there are limitations in modal integrity. While early datasets like Twitter-15, Twitter-16, LIAR, and Snopes possess advantages in propagation chains or fact-checking logic, the absence of key modalities such as vision and audio prevents them from being used to verify current mainstream cross-modal deep fusion algorithms. Furthermore, limitations exist in domain generalization and scenarios. Vertical domain datasets like GossipCop and PolitiFact have rigorous annotations, but their feature distributions are strongly biased toward specific fields. Consequently, models trained on such data show a significant decline in generalization performance when processing public health emergencies like MM-COVID or multi-domain breaking news like PHEME. Finally, there are constraints in data distribution and authenticity. Large-scale datasets such as Fakeddit and Twitter provide rich multimodal samples, but most manually constructed balanced datasets deviate severely from the reality of the extremely low proportion of fake news in real-world social media scenarios, which restricts model performance in handling real-world class imbalance issues. Additionally, as datasets like FakeNewsNet integrate multi-source data, they face technical challenges in maintaining data consistency and performing deduplication across different sources.

Overall, existing datasets have formed a relatively complete system in terms of scale, sources, and modal types, covering diverse needs ranging from single-text to full-modal fusion, and from general scenarios to vertical domains. However, deep deconstruction of annotation quality, modal consistency, and the diversity of fake news types reveals that while the evolution of datasets has significantly increased the complexity of detection tasks, severe challenges remain in adapting to real-world scenarios. The future construction of datasets needs not only to bridge the gaps in cross-lingual consistency and full-modal balance but also to establish standardized quality control systems and non-balanced sample libraries that more closely align with real-world distributions. This paradigm shift from being "data-driven" to "quality-and-scenario dual-driven" will be a key support for advancing multimodal fake news detection technology from laboratory evaluation toward large-scale, real-time social media governance.

3.2. Evaluation Metrics

Fake news detection is essentially a classification task, and its evaluation framework is primarily based on the confusion matrix, which quantifies four types of prediction outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Based on these quantities, multiple performance metrics are derived to assess detection effectiveness from different perspectives. The most commonly used metrics in existing studies include Accuracy, Precision, Recall, and the F1-score. Accuracy measures the proportion of correctly classified samples among all instances and is suitable for benchmark datasets with relatively balanced class distributions. Precision evaluates the proportion of samples predicted as fake news that are indeed fake, thereby reducing the risk of misclassifying real news as false. Recall focuses on the proportion of actual fake news samples that are correctly identified, helping to minimize the probability of missed detections. The F1-score, defined as the harmonic mean of Precision and Recall, balances the trade-off between the two and is more robust to class imbalance. As such, it has become a core metric for performance comparison across different detection methods and is widely adopted in evaluations on mainstream datasets such as FakeNewsNet and the Weibo series [20].

To accommodate the specific requirements of fake news detection tasks, evaluation metrics are often extended according to task characteristics. In multimodal detection scenarios, such as those involving the FakeSV

short-video dataset [15], modality consistency metrics are introduced to assess cross-modal coherence. These metrics evaluate semantic similarity between text and images, as well as temporal synchronization between audio and video, to determine whether contradictions exist across modalities, thereby providing a more comprehensive assessment of multimodal fusion models. For fine-grained classification datasets such as LIAR [16], Macro-F1 and Micro-F1 scores are commonly employed. Macro-F1 emphasizes balanced performance across all categories, while Micro-F1 reflects overall classification accuracy, enabling more precise evaluation of a model's ability to distinguish ambiguous classes such as partially true or partially false statements. In addition, for real-time detection scenarios, efficiency-related metrics, including inference time and throughput, have become important considerations for balancing detection performance with the response speed required for practical deployment.

4. Multimodal Fake News Detection Methods

Fake news detection methods have evolved alongside technological advances, exhibiting a clear progression from "unimodal to multimodal analysis, from data-driven approaches to knowledge-enhanced models, and from static feature analysis to dynamic modeling". Early methods primarily focused on the stylistic and semantic features of textual content. With the widespread adoption of social media, research has gradually incorporated multimodal information such as images and videos, as well as social context including user interactions and propagation networks. In recent years, the semantic understanding capabilities of large language models (LLMs), the structural modeling power of graph neural networks, and the logical reasoning abilities introduced by causal learning have further propelled technical upgrades. Meanwhile, the integration of external knowledge, such as knowledge graphs and factual evidence, has effectively addressed challenges related to the "semantic gap" and "missing factual information". In the following, we systematically review multimodal fake news detection methods from four core technical perspectives: content feature extraction-based approaches, factual verification-based approaches, propagation dynamics-based modeling approaches, and approaches focusing on detection performance and application-oriented optimization. For each category, we analyze the underlying technical principles, key innovations, and representative methods. It should be noted that as multimodal detection techniques co-evolve, certain intersections and overlaps exist between different technical paradigms. To maintain the systematicity of this survey, the classification logic of this paper is primarily based on the core driving features used by the models to identify fake information. For instance, if the core innovation of a model lies in the introduction of external structured knowledge for comparison, it is categorized under "fact-based verification"; if its core lies in leveraging the generative reasoning capabilities of large language models for comprehensive judgment, it is classified under "large language models". This classification aims to help readers grasp the field's trends through the main thread of technological evolution.

4.1. Content Feature Extraction-Based Methods

The core deceptiveness of multimodal fake news is often embedded within the content itself, including text, images, and audio. By directly extracting deceptive features from multimodal content, preliminary identification of fake information can be achieved. Methods in this category form the foundation of multimodal fake news detection. They do not rely on external knowledge sources or propagation-related data, but instead determine authenticity by analyzing correlations and inconsistencies across content modalities. As such, these approaches are particularly suitable for early-stage information dissemination scenarios where external data are limited or unavailable. In the following, we elaborate on content-level deceptive feature extraction from two perspectives: multimodal content fusion and entity-level verification, detailing the underlying extraction logic and technical implementations.

4.1.1. Multimodal Content Fusion-Based Methods

Multimodal content fusion-based methods aim to overcome the limitations of relying solely on textual information by integrating multiple sources of content, including text, images, audio, and videos, to achieve more comprehensive capture of deceptive features. Textual content features serve as the foundation of this category of methods and typically include linguistic style, sentiment, and topical information. Fusion strategies are selected according to the characteristics of different modalities. Early fusion integrates shallow features from text, visual, and audio modalities at the feature level, whereas late fusion combines the outputs of modality-specific models at the decision level. Attention-based fusion further focuses on identifying key inter-modal associations. In terms of enhancement strategies, contrastive learning improves the discriminability of multimodal representations by constructing positive and negative sample pairs, while adversarial learning enhances robustness to fake samples by generating adversarial perturbations. Together, these strategies contribute to improved detection accuracy and generalization performance.

(1) Fusion of Content Features.

Multimodal fusion typically begins with extracting core textual features, which are then jointly modeled with other modalities. Linguistic style features focus on capturing characteristic writing patterns of fake news, sentiment features emphasize emotional tendencies in both news content and user feedback, and topic features examine the consistency between news topics and content. These three types of features jointly form the core discriminative basis for content fusion. Cantini et al. [21] proposed a multimodal approach based on topic and linguistic style features to identify fake information on social media. By analyzing textual characteristics such as lexical complexity and sentence structure and combining them with visual features through early fusion, their method achieved performance improvements in fake news detection. Farhoudinia et al. [22] investigated the use of sentiment analysis for detecting COVID-19–related fake news on social media. By modeling sentiment features from both news content and user comments and leveraging sentiment consistency, their approach demonstrated strong generalization performance across different datasets. Qian et al. [23] proposed a multi-granularity fake news detection method that extracts textual features using RoBERTa and visual features using DenseNet, and captures inter-modal interactions through attention mechanisms. By fusing multimodal content features, their model effectively improves detection performance and shows notable advantages in fine-grained fake news identification scenarios. Building upon the aforementioned studies, a comparison of methods within the content fusion paradigm reveals significant mechanistic differences in terms of feature interaction depth and discriminative logic. The approach by Cantini et al. emphasizes early fusion at the feature level, achieving a preliminary integration of multimodal information through simple concatenation. While this mechanism is computationally efficient, it struggles to uncover deep cross-modal associations. In contrast, Qian et al. introduced a co-attention mechanism that utilizes multi-granularity feature extraction and dynamic weight allocation. This achieves a transition from "physical concatenation" to "semantic alignment" between modalities, thereby enabling more precise capture of subtle textual-visual contradictions. Furthermore, the core mechanism of the method by Farhoudinia et al. lies in utilizing the emotional consistency of user comments as an external verification logic, whereas the other two studies focus more on mining the endogenous features of the news content itself. This technological evolution, transitioning from shallow fusion to deep interaction and from global features to multi-granularity alignment, reflects the continuous optimization of content fusion methods in enhancing detection accuracy and processing complex semantic scenarios.

(2) Enhancement Strategies.

Enhancement strategies optimize the quality of multimodal representations through techniques such as contrastive learning and adversarial learning. Contrastive learning improves feature discriminability by constructing positive and negative sample pairs, enabling models to identify subtle differences between real and fake news. Adversarial learning, on the other hand, enhances robustness to fake samples by generating perturbations that reduce misclassification caused by modal noise. Chen et al. [24] proposed a contrastive learning-based multimodal fake news detection method that improves the discriminative power of multimodal features through positive–negative sample construction. By aligning textual and visual representations using contrastive objectives, their approach achieved notable performance gains. Shen et al. [25] introduced a multimodal fake news detection framework combining contrastive learning and optimal transport theory, referred to as MCOT. Their method aligns textual and visual embeddings via contrastive learning while leveraging optimal transport to optimize distribution alignment across modalities, resulting in more effective multimodal fusion. Xing et al. [26] proposed a multimodal fake news detection method based on bidirectional semantic enhancement and adversarial networks. This research introduced a dual adversarial learning framework during the training stage, which optimizes feature representations by using adversarial samples to perturb the model, significantly enhancing the generalization ability and stability of multimodal detection. Throughout the aforementioned enhancement strategies, although all aim to optimize feature quality, their underlying mechanisms exhibit significant differences. Both the methods of Chen et al. and Shen et al. center on contrastive learning, yet the former focuses on achieving basic alignment of cross-modal features through the construction of simple positive and negative sample pairs. In contrast, the MCOT framework proposed by the latter further incorporates optimal transport theory to optimize the mapping relationship of different multimodal feature spaces from the perspective of global distribution alignment, thereby addressing the limitations of traditional contrastive learning in processing complex semantic distributions. Distinct from these, the adversarial learning mechanism adopted by Xing et al. does not enhance discriminability through sample alignment, but rather utilizes a dual adversarial learning framework to introduce perturbative samples during training, forcing the model to learn more robust representations. This technological evolution, transitioning from pure feature alignment to distribution optimization and adversarial robustness,

reflects the diverse entry points of enhancement strategies in addressing cross-modal semantic gaps and noise in fraudulent samples.

4.1.2. Entity-Based Methods

Entity-based methods take core entities in news content as the central focus for detection and assess news authenticity by verifying entity correctness, the rationality of inter-entity relationships, and the consistency of entity attributes. Entities typically include key elements such as persons, events, locations, temporal expressions, and numerical values. These methods first extract entities from news content using named entity recognition techniques, and then evaluate the degree of alignment between extracted entities and real-world information through entity linking and relational reasoning. As a result, entity-based approaches are particularly effective for detecting fake news involving entity fabrication and relationship manipulation. Gleenski et al. [27] investigated entity consistency in the dissemination of fake news by analyzing the alignment between entities mentioned in news articles and real-world information. By leveraging entity linking and consistency checking techniques, their approach effectively identified entity fabrication and relationship manipulation in fake news. Li et al. [28] proposed an entity-oriented multimodal alignment and fusion network for fake news detection. Their method encapsulates textual and visual entities using capsule networks and employs an improved dynamic routing algorithm to achieve cross-modal entity alignment, enabling accurate authenticity judgments. This approach is particularly suitable for fake news that requires entity-centric comparison and verification. Qi et al. [29] introduced a knowledge graph–enhanced fake news detection method that evaluates news authenticity by verifying relationships between entities mentioned in news content and those in a knowledge graph. By linking news entities to structured knowledge graph information and performing relational reasoning, their method effectively detects fake news involving entity fabrication and relationship manipulation. Looking across the aforementioned entity-based methods, the core mechanistic differences lie in the focus of entity verification and the complexity of the alignment logic. The study by Glenski et al. focuses on the external alignment of entities, emphasizing the verification of mentions against real-world data to identify fabrication. In contrast, Li et al. introduced a structural alignment mechanism, utilizing capsule networks and dynamic routing to achieve deep semantic matching between textual and visual entities. Furthermore, Qi et al. realized a transition toward a knowledge-enhanced reasoning mechanism by linking news entities to structured knowledge graphs to verify inter-entity relationships. This technological evolution reflects the continuous deepening of entity-based methods in capturing fine-grained factual contradictions.

4.2. Fact-Based Verification Methods

Relying solely on content feature extraction is often insufficient for addressing fake news that is partially true or characterized by factual ambiguity. For example, a news article may present logically coherent text and semantically consistent text-image pairs, yet lack authoritative factual support. Fact-based verification methods address this limitation by incorporating external authoritative knowledge or retrieving concrete evidence, thereby establishing a verification loop that connects news content with real-world facts. By mitigating issues related to factual insufficiency in multimodal content, these approaches significantly enhance the credibility of detection results and have become essential techniques in domains that require professional knowledge support, such as politics, healthcare, and finance. Fact-based verification methods primarily include external knowledge enhancement and evidence matching, which are highly consistent in their objective: constructing a “content-to-fact” verification loop. The slight difference between the two lies in the fact that “external knowledge” focuses on supplementing background semantics, while “evidence matching” focuses on retrieving directly corresponding authoritative evidence. In the following, we introduce fact-based verification methods from these two perspectives: external knowledge enhancement and evidence matching, and discuss their respective technical pathways.

4.2.1. External Knowledge-Based Methods

External knowledge-based methods enhance fake news detection by incorporating structured or unstructured knowledge beyond the news content itself, thereby supplementing missing background information and addressing issues such as semantic ambiguity and unknown facts. External knowledge sources mainly include knowledge graphs and external knowledge bases. Knowledge graphs store structured entity–relation information, while external knowledge bases comprise unstructured authoritative documents and domain-specific databases. These two types of knowledge are integrated with news representations through techniques such as knowledge embedding and semantic association, thereby improving detection performance in professional domains.

(1) Knowledge Graphs.

Knowledge graph-based methods verify the authenticity of news entities and the rationality of their relationships by leveraging structured entity–relation networks. Typically, entities are first extracted from news content and then matched against entity attributes and relations in a knowledge graph to identify fake characteristics such as non-existent entities and contradictory relationships, making these methods particularly suitable for entity-dense fake news scenarios. Zhang et al. [30] proposed a knowledge-aware attention network that identifies entity mentions in news articles and aligns them with entities in a knowledge graph. By incorporating entities and their contextual information as external knowledge and employing a two-level attention mechanism to evaluate knowledge importance, their method effectively fuses semantic representations of news content with knowledge-level representations, leading to improved detection performance. Xie et al. [31] proposed a fake news detection method based on a knowledge graph-enhanced heterogeneous graph neural network. By constructing a heterogeneous graph that integrates news entities, knowledge entities, and semantic relationships, this approach effectively compensates for the lack of external factual knowledge in content-only models, significantly improving both the accuracy and interpretability of fake news detection. Hao et al. [32] introduced a fine-grained knowledge graph-enhanced multimodal fake news detection framework (FKGFND). This framework deeply integrates textual, visual, and knowledge graph representations by performing fine-grained modeling of images to extract embedded text and person details while associating them with background knowledge. Simultaneously, it constructs a ternary knowledge graph to structurally link textual entities, image-based persons, and background knowledge, effectively identifying cross-modal semantic contradictions and entity falsification, thereby enhancing detection precision and interpretability. Looking across the aforementioned knowledge graph-based methods, the core mechanistic differences lie primarily in the dimensions of knowledge fusion and the granularity of modeling. The mechanism of Zhang et al. focuses on supplementation at the semantic level, embedding entity contexts as auxiliary features into text representations through attention mechanisms, which represents a "feature-augmented" fusion. In contrast, Xie et al. shifted toward spatial structural modeling by constructing heterogeneous graphs containing news and knowledge entities, utilizing graph neural networks to capture complex non-linear semantic relationships, thereby significantly enhancing the model's logical reasoning capabilities. Furthermore, Hao et al. achieved a further breakthrough in modeling granularity. Their mechanism not only attends to textual entities but also performs fine-grained deconstruction of images, establishing ternary associations between visual elements such as persons and embedded text and structured knowledge. This enables the identification of deeper semantic contradictions between images and text. This technological evolution, progressing from simple feature embedding to heterogeneous graph modeling and further to cross-modal fine-grained association, reflects the continuous deepening of knowledge graph methods in processing multi-source heterogeneous information and complex fact-checking.

(2) External Knowledge Bases.

External knowledge base-based methods verify news facts by leveraging unstructured authoritative documents and databases, such as official reports, domain guidelines, and fact-checking website archives. These methods associate news content with external knowledge through semantic retrieval and text matching techniques, enabling the identification of fake news involving factual errors and data manipulation, and are particularly suitable for scenarios that require professional knowledge support. Hu et al. [33] proposed a fact-checking approach based on external knowledge bases by linking entities in news content to resources such as Wikidata to verify authenticity. By comparing entities mentioned in news articles with those in knowledge bases, their method identifies fake characteristics such as factual errors and data manipulation, thereby improving detection performance to some extent. Fu et al. [34] introduced a multimodal fake news detection method that combines external knowledge with user interaction features. By incorporating structured information from external knowledge bases as background knowledge, their approach enhances the model's understanding of news content and improves detection capability. Han et al. [35] proposed a multimodal fake news detection model based on external knowledge bases and contrast-driven feature enhancement, termed EC-Fake. This model employs a semantic information decoupling module to separate foreground objects and background context in news images, applies contrast-driven feature enhancement for multi-level feature fusion, and strengthens key representations to improve fusion effectiveness and detection accuracy. Furthermore, the model leverages large language models to extract external knowledge from news content and integrates it with multimodal features, thereby effectively enhancing both interpretability and accuracy. Integrating the aforementioned methods based on external knowledge bases, significant mechanistic differences are observed in both knowledge acquisition techniques and cross-modal collaborative logic. The approach by Hu et al. emphasizes a singular "entity alignment" mechanism, identifying factual errors by matching news entities with unstructured knowledge bases, which represents a direct fact-checking pathway. In contrast, Fu et al. introduced a "multi-source feedback" mechanism, combining

background common sense from external knowledge bases with dynamic user interaction features to supplement static knowledge with social context. Distinguishing itself from these, Han et al.'s EC-Fake model achieves a transition from "external enhancement" to "endogenous decoupling and knowledge-driven" mechanisms; by decoupling image subjects and utilizing Large Language Models (LLMs) to extract targeted knowledge, it achieves higher-dimensional semantic alignment and logical traceability. This technological evolution, progressing from simple entity comparison to multi-source feature fusion and further to generative knowledge enhancement, reflects the continuous deepening of external knowledge base methods in enhancing detection robustness and deep semantic understanding.

4.2.2. Evidence-Based Methods

Evidence-based methods adopt fact checking as the core logic and perform fake news detection through a three-stage pipeline: identifying verifiable claims, retrieving authoritative evidence, and assessing consistency between evidence and news content. Specifically, these approaches first locate claims within news articles that require verification, then obtain authoritative evidence through search engines, fact-checking platform APIs, or external repositories, and finally evaluate contradictions between news content and retrieved evidence via semantic matching and logical reasoning. Such methods are particularly suitable for fake news that requires explicit factual validation. Yu et al. [36] proposed a dual evidence-aware fake news detection framework, termed DEP-FEND, which enhances detection accuracy by integrating evidence-aware vectors derived from the news environment with those obtained from external knowledge bases. Cheung et al. [37] introduced an automated fact-checking approach that combines instruction-following language models with external evidence retrieval. Their method retrieves evidence relevant to input claims using search engines and leverages this external information to augment the knowledge of pre-trained language models, enabling more accurate prediction of claim veracity. Zheng et al. [38] proposed the RAV framework for real-time fact-checking of zero-shot multimodal social media claims. This method employs hybrid evidence retrieval to obtain textual and visual evidence from authoritative knowledge bases and achieves joint optimization of retrieval and verification through end-to-end training, precisely uncovering fake information in image-text claims. Regarding multi-source evidence processing. Wu et al. [39] developed a tightly coupled graph reasoning network for tabular unstructured evidence, achieving precise verification of complex factual claims by modeling fine-grained associations between table entries. Furthermore. Wu et al. [40] constructed a unified evidence-enhanced reasoning system that improves the integrity of the evidence chain through multi-round retrieval and information completion mechanisms, effectively addressing the failure of verification caused by sparse original evidence. To tackle the challenge of heterogeneous graph modeling in evidence reasoning. Wu et al. [41] proposed the Heuristic Heterogeneous Graph Reasoning Network (HHGRN). This method innovatively combines heuristic reasoning with heterogeneous graph neural networks to construct a heterogeneous graph that integrates news claims to be verified, multi-source authoritative evidence, entity relationships, and other types of nodes and associations. By using heuristic rules to guide information propagation and feature aggregation among graph nodes, its effectiveness was validated on authoritative fact-checking datasets such as FEVEROUS, providing a new technical pathway for fake news detection in complex evidence scenarios. In summary, evidence-based methods exhibit significant mechanistic differences in terms of retrieval depth and the complexity of reasoning structures. The DEP-FEND framework by Yu et al. emphasizes a dual-dimensional "feature enhancement" mechanism, improving discriminative accuracy by fusing evidence vectors from news contexts and knowledge bases, which essentially represents a static superposition of feature spaces. In contrast, the RAV framework by Zheng et al. implements an "end-to-end interaction" mechanism, addressing the challenge of real-time cross-modal evidence matching through the joint optimization of multimodal hybrid retrieval and verification. In the series of studies by Wu et al., the logic of evidence checking has evolved from simple vector fusion to a "structured reasoning" mechanism: ranging from tightly coupled graph modeling for tabular evidence and an enhancement system for completing evidence chains through multi-round retrieval, to the heuristic-guided heterogeneous graph reasoning employed in HHGRN. This mechanistic shift reflects that such methods not only focus on the "presence or absence" of evidence but also excavate the "logical consistency" between pieces of evidence through heuristic rules and graph neural structures, thereby demonstrating stronger robustness compared to traditional retrieval methods when dealing with complex factual claims.

4.3. Propagation Dynamics-Based Methods

The propagation patterns of multimodal fake news on social media differ significantly from those of real news. For example, fake news often spreads through mechanisms such as rapid reposting by bot accounts and the aggregation of highly polarized emotional comments, whereas the dissemination of real news relies more on

organic user interactions. Propagation dynamics-based methods aim to capture abnormal diffusion patterns of fake information by modeling the relationships among users, news items, and propagation networks. These approaches are well suited for large-scale and dynamically evolving social media environments, and can complement content-based and fact-based methods to further improve detection accuracy. In the following, we introduce propagation dynamics analysis from two perspectives: social context extraction and graph neural network-based modeling.

4.3.1. Social Context-Based Methods

Social context-based methods focus on the dissemination environment of fake news on social media and the characteristics of user interactions, and perform detection by exploiting relational information among users, news, and propagation networks. User interaction features constitute the core of this category of methods and include sentiment polarity of comments, temporal patterns of reposting, and propagation paths induced by follower relationships. By jointly modeling user profiles and network structures, these methods capture distinctive social dissemination patterns of fake news. In practice, fake news often exhibits characteristics such as rapid reposting by bot accounts and the clustering of extreme emotional comments, which differ markedly from the natural diffusion of real news. Dellys et al. [42] proposed a multimodal fake news detection framework based on a fusion attention mechanism that integrates textual, visual, and social context features. Their approach incorporates social contextual information such as user profiles, posting status, social reputation, and geographic location, and combines these with textual and image features using a ViLBERT-based attention mechanism, resulting in improved detection performance. Raza et al. [43] introduced a Transformer-based fake news detection framework that jointly models news content and social context to learn informative representations for veracity prediction. By leveraging user profiles, likes, comments, and reposting behaviors, and integrating these with textual content features, their method employs zero-shot learning to estimate user credibility, significantly improving both detection accuracy and timeliness. Mehta et al. [44] roposed a graph neural network-based social context representation enhancement method. By constructing an interaction graph that connects news sources, articles, and users, and embedding social interaction information into the graph, their approach designs reasoning operators to uncover latent associations such as document similarity and user engagement patterns, enabling accurate modeling of fake news propagation logic. This method demonstrates superior performance across multiple tasks. Looking across the aforementioned social context-based methods, the core mechanistic differences are reflected in the modeling dimensions of social features and their interaction logic. The approach by Dellys et al. employs a "multidimensional feature concatenation" mechanism, aligning static contexts such as user profiles and geographic locations with content features through attention mechanisms, focusing on utilizing social backgrounds to enhance content representation. In contrast, Raza et al. introduced a "user credibility assessment" mechanism based on logical reasoning, utilizing the Transformer architecture to capture dynamic user interaction behaviors, such as likes and retweets, to achieve quantitative discrimination of user reputation through zero-shot learning. Furthermore, Mehta et al. realized a transition toward a "structured relationship mining" mechanism; by constructing complex relationship graphs and designing specialized reasoning operators, they transformed originally isolated social information into a global representation endowed with propagation logic. This evolution, moving from background feature enhancement to dynamic reputation assessment and further to graph structure relational reasoning, reflects the continuous deepening of these methods in characterizing the complex interaction patterns of social media.

4.3.2. Graph Neural Network-Based Methods

Graph neural network (GNN)-based methods abstract elements in fake news detection scenarios into graph structures and leverage models such as graph convolutional networks (GCN), graph attention networks (GAT), and temporal graph attention networks (TGAT) to capture both local and global relational features among nodes. These approaches address the limitations of traditional methods in modeling structured relationships. Graph structures may include news–user interaction graphs, entity–relation knowledge graphs, and propagation trees composed of nodes and diffusion paths. By aggregating node features through graph convolution and focusing on critical associations via attention mechanisms, GNN-based methods are particularly suitable for fake news detection scenarios characterized by complex propagation networks and dense relational structures. Su et al. [45] proposed a hypergraph neural network-based fake news detection method named Hy-DeFake, which constructs hypergraphs to capture complex high-order relationships in online social networks. Compared with conventional GNNs, hypergraph neural networks are more effective in representing and modeling complex social interactions, leading to improved detection performance. Malik et al. [46] introduced an ensemble GNN model, referred to as EGNN, for fake news detection. This model integrates user engagement patterns and textual features, and employs

multiple GNN variants—including standard GNNs, graph attention networks (GAT), and bidirectional graph convolutional networks (BiGCN)—to capture diverse user interaction behaviors. Lakzaei et al. [47] proposed a graph neural network-based semi-supervised fake news detection method, termed LOSS-GAT. Their approach adopts a two-stage label propagation strategy, where a GNN is first used as an initial classifier to categorize news into real and fake classes, followed by graph structure enhancement through structural augmentation techniques and the introduction of stochasticity in local neighborhood aggregation. This method achieves significant performance improvements across multiple datasets, particularly in scenarios with limited labeled data. Looking across the aforementioned graph neural network (GNN)-based methods, the core mechanistic differences are primarily reflected in the modeling dimensions of graph structures and their information propagation strategies. The approach by Su et al. realized a leap from ordinary graphs to a "high-order association" mechanism; by constructing hypergraphs, they broke the limitations of binary relationships, enabling the capture of complex non-linear interactions among multiple nodes in social networks. In contrast, Malik et al. focused on a "multi-paradigm ensemble" mechanism, aggregating user engagement patterns and textual features from different dimensions by combining various GNN variants such as GAT and BiGCN, aiming to broaden the model's coverage of diverse interaction behaviors. Distinct from these, Lakzaei et al. introduced a "semi-supervised label propagation" mechanism, the core of which lies in utilizing graph structure augmentation techniques and the randomness of neighborhood aggregation to achieve predictions through structured derivation in scenarios where labeled data are scarce. This evolution, progressing from topological structure upgrading to model architecture ensemble and further to learning paradigm optimization, reflects the continuous deepening of graph neural network methods in processing complex social topologies and low-resource detection tasks.

4.4. Detection Performance and Application-Oriented Optimization Methods

As multimodal fake news detection moves toward real-world deployment, simply improving detection accuracy is no longer sufficient. Real-time detection requires a careful balance between accuracy and efficiency; detection in emerging domains is often constrained by limited labeled data; and gaining public trust necessitates addressing the black-box nature of many models. Methods focusing on detection performance and application-oriented optimization tackle these practical challenges by incorporating techniques such as model lightweighting, causal reasoning, and large language model-based enhancement. These approaches improve the practicality, robustness, and generalization ability of detection systems and serve as key enablers for transitioning multimodal fake news detection from laboratory research to real-world applications. In the following, we review scenario-oriented optimization strategies from three perspectives: efficiency optimization, reliability enhancement, and generalization improvement.

4.4.1. Knowledge Distillation-Based Methods

Knowledge distillation-based methods aim to address the imbalance between detection accuracy and computational efficiency by transferring knowledge from complex teacher models to lightweight student models. This paradigm improves inference speed and deployment flexibility while maintaining comparable detection performance. Typically, teacher models adopt sophisticated architectures, whereas student models are lightweight variants. Knowledge transfer is achieved through techniques such as temperature scaling, feature distillation, and attention transfer, making these methods particularly suitable for real-time detection scenarios. Lan et al. [48] proposed an improved knowledge distillation approach that enhances the efficiency and accuracy of knowledge transfer from large teacher models to lightweight student models through label correction and data selection. Their method mitigates inaccurate supervision by correcting erroneous teacher predictions and selecting appropriate training samples for distillation. Mu et al. [49] introduced a self-supervised distillation learning method, termed SSDL, for multimodal fake information detection. By adopting self-supervised representation learning strategies, SSDL optimizes multiple objectives, including task-agnostic representation consistency assessment and task-specific multimodal knowledge estimation. Chen et al. [50] proposed a text-centric fake news detection model based on external knowledge distillation, referred to as LEKD. By incorporating external knowledge and leveraging the natural language understanding and generation capabilities of large language models (LLMs), LEKD enhances the detection performance of small language models (SLMs). The framework consists of five components: an LLM-based knowledge enrichment module, a pretrained SLM knowledge encoder, a graph-based semantic-aware feature alignment module, a knowledge cross-distillation module, and an MLP-based detector. Experimental results demonstrate that this method achieves new state-of-the-art performance on real-world datasets such as Weibo and GossipCop. Looking across the aforementioned knowledge distillation-based methods, the core mechanistic differences are primarily reflected in the sources of knowledge transfer and their optimization

pathways. The approach by Lan et al. emphasizes a "supervisory signal optimization" mechanism; by introducing label revision and data selection during the distillation process, it aims to resolve noise interference present in the teacher model's original output, thereby enhancing the fitting precision of the student model. In contrast, Mu et al. realized a transition toward a "self-supervised multi-task" mechanism, utilizing task-agnostic representation consistency assessment to strengthen the model's underlying perception of multimodal features and reducing reliance on task-specific annotations. Distinct from these, Chen et al. introduced a "generative enhancement from large models" mechanism, injecting external knowledge extracted by Large Language Models (LLMs) into small models as higher-dimensional guidance signals; this addresses the insufficiency of traditional lightweight models in fact-checking depth through cross-modal semantic alignment. This evolution, progressing from supervisory signal refinement to self-supervised feature learning and further to cross-scale model knowledge alignment, reflects the continuous exploration of knowledge distillation methods in balancing detection efficiency with the depth of logical reasoning.

4.4.2. Causal Learning-Based Methods

Causal learning-based methods go beyond traditional correlation-based modeling by explicitly uncovering causal relationships in fake news dissemination and detection. By distinguishing causal factors from spurious correlations, these approaches enhance model robustness. Typically, causal graphs are constructed to model causal relationships among disseminators, content, and audiences, intervention-based experiments are employed to identify key causal factors, and counterfactual reasoning is used to validate causal effects of features. Such methods are particularly suitable for complex propagation scenarios. Chen et al. [51] proposed a multimodal fake news detection approach based on causal intervention and counterfactual reasoning. By constructing causal graphs to analyze relationships among news content, propagation paths, and user comments, their method identifies causal features of fake news. Counterfactual reasoning is further employed to verify causal effects, leading to improved robustness and detection accuracy. Gong et al. [52] introduced a causal debiasing-based fake news detection method that mitigates the influence of environment-biased samples using structural causal models (SCMs) and reweighting strategies. By addressing distribution shifts of features and labels across different news domains in fake news propagation graphs, their approach improves generalization performance on unseen domains. Cheng et al. [53] proposed a causal understanding-based method for analyzing fake news dissemination on social media. By applying causal inference theory to identify and alleviate selection bias, their work provides deeper insights into causal relationships between user attributes and fake news propagation. This approach demonstrates strong theoretical and empirical performance and contributes to protecting society from the harmful effects of misinformation. Looking across the aforementioned causal learning-based methods, the core mechanistic differences are reflected in the objectives of causal modeling and the strategies for addressing bias. The approach by Chen et al. emphasizes a "feature effect verification" mechanism; by employing counterfactual reasoning, it directly quantifies the causal contribution of content and propagation features to the judgment results, aiming to eliminate irrelevant correlational interference. In contrast, Gong et al. realized a transition toward an "environmental debiasing" mechanism, utilizing structural causal models to identify and weaken domain-specific confounding factors, thereby addressing the issue of distribution shift when models are applied across different domains. Distinct from these, Cheng et al. focused on a "dissemination selection bias" mechanism, mitigating observational biases caused by user attributes through causal inference to deeply characterize the causal logic between user behavior and information diffusion. This evolution, progressing from internal feature verification to external environmental debiasing and further to dissemination bias correction, reflects the multidimensional exploration of causal learning methods in enhancing the robustness and explanatory depth of detection models.

4.4.3. Large Language Model-Based Methods

Large Language Model (LLM)-based methods leverage the powerful semantic understanding, contextual modeling, and generative reasoning capabilities of LLMs to revolutionize the fake news detection paradigm. The core advantages of these methods lie in zero-shot and few-shot learning, LLM-assisted evidence acquisition, and deep logical reasoning enhancement. Through paradigms such as "pre-training and fine-tuning", "prompt engineering", "retrieval-augmented generation (RAG)", and "reverse reasoning", LLMs effectively address the bottlenecks of traditional models in areas such as semantic ambiguity, data scarcity, cross-modal contradictory reasoning, and insufficient interpretability. Consequently, this approach stands as one of the most promising research directions in the current field. Although some LLM-based methods involve evidence retrieval, the essential difference lies in their utilization of the natural language reasoning capabilities of LLMs rather than

simple feature comparison; therefore, this paper categorizes them under the LLM domain to emphasize this paradigm shift.

(1) Zero-Shot and Few-Shot Learning.

Zero-shot and few-shot learning rely on the general semantic understanding capabilities of LLMs to perform fake news detection in scenarios with limited or no labeled data. Zero-shot learning requires no labeled samples and directly assesses news veracity using the knowledge acquired during LLM pretraining. Few-shot learning, in contrast, adapts LLMs to specific scenarios using a small number of labeled examples through prompt tuning or model fine-tuning, making it particularly suitable for new domains or low-resource language settings. Hu et al. [54] investigated the role of large language models in fake news detection, with a particular focus on zero-shot and few-shot learning scenarios. Their study employed the GPT-3.5-turbo model and evaluated its performance under different prompting strategies, including zero-shot prompting, zero-shot chain-of-thought prompting, and few-shot prompting. Experimental results demonstrated that few-shot prompting achieved strong performance on both Chinese and English datasets. Liu et al. [55] proposed a few-shot fake news detection framework based on large language models, termed DAFND, which enhances LLM performance from both internal and external perspectives. The framework consists of a detection module, an investigation module, a judgment module, and a decision module that collaboratively improve fake news detection performance in low-resource settings. Chalehchaleh et al. [56] introduced an LLM-based data augmentation framework to address data scarcity in multilingual fake news detection. Their approach utilizes the Llama3 model to generate synthetic news samples through zero-shot and few-shot prompting strategies, and experimental results indicate that LLM-based data augmentation can effectively improve detection performance. Looking across the aforementioned zero-shot and few-shot learning methods, the core mechanistic differences lie primarily in the depth of knowledge utilization and the application logic of Large Language Models (LLMs). The research by Hu et al. emphasizes a "prompt-triggering" mechanism, activating the model's inherent pre-trained knowledge by designing various types of prompts, which essentially serves as a form of knowledge probing without a structured framework. In contrast, the DAFND framework proposed by Liu et al. realizes a transition toward a "multi-module collaborative reasoning" mechanism; by simulating a human-like decision-making chain of investigation and judgment, it upgrades the LLM from a simple classifier to an intelligent system with investigative capabilities, significantly enhancing logical rigor in low-resource environments. Distinct from these, Chalehchaleh et al. adopted a different approach by utilizing a "generative data augmentation" mechanism, where the LLM serves as an auxiliary tool to provide synthetic samples for downstream detection models, aiming to address the scarcity of underlying data in multilingual scenarios. This evolution, progressing from direct prompt-based judgment to structured collaborative reasoning and further to auxiliary sample expansion, reflects the continuous deepening of zero or few-shot detection methods in excavating LLM potential and adapting to diversified task scenarios.

(2) LLM-Assisted Evidence Acquisition.

LLM-assisted evidence acquisition exploits the retrieval and generation capabilities of LLMs to obtain authoritative evidence for fact checking in fake news detection. LLMs can automatically invoke search engines to retrieve external evidence or generate structured evidence representations, and subsequently assess the consistency between news content and evidence. This paradigm overcomes the inefficiency of traditional, manual evidence collection processes and is particularly suitable for real-time fake news detection scenarios that require timely factual support. Li et al. [57] proposed the FactAgent-Pro framework, an LLM agent system for fake news detection. Through an iterative process of claim decomposition, multi-source retrieval, and evidence aggregation, this framework automatically invokes search engines to obtain authoritative evidence and fuses internal LLM knowledge with external retrieval results for cross-verification, significantly enhancing the accuracy and reliability of complex news claim verification. Li et al. [58] investigated the use of retrieval-augmented large language models (LLMs) for COVID-19 fact-checking; this method employs OpenAI's GPT-4 as the backbone model and incorporates evidence from external knowledge bases through a retrieval-augmented generation (RAG) system to improve the precision of fact-checking. Khaliq et al. [59] introduced the RAGAR framework, which relies on RAG technology and leverages LLMs to autonomously generate targeted queries to precisely complete missing information required for news verification; simultaneously, it moves beyond a single predefined corpus by integrating multi-source authoritative evidence from the Internet and structured knowledge bases, significantly improving the robustness and comprehensiveness of evidence retrieval and effectively addressing the data source limitations of traditional retrieval methods. Looking across the aforementioned methods for LLM-assisted evidence acquisition, significant mechanistic differences are observed in the interactivity of retrieval strategies and the breadth of evidence aggregation. The FactAgent-Pro framework by Li et al. emphasizes a "multi-round iterative agent" mechanism. By simulating a human fact-checker's process of claim decomposition and

autonomous searching, it achieves a cycle of verification between internal and external knowledge, enabling it to handle logically complex and lengthy assertions. In contrast, the research by Li et al. follows the typical "Retrieval-Augmented Generation (RAG)" standard paradigm, focusing on using domain-specific authoritative knowledge bases as semantic constraints to complete the LLM's knowledge boundaries through unidirectional information retrieval. Furthermore, the RAGAR framework by Khaliq et al. advances an "active querying and multi-source heterogeneous integration" mechanism. Its core no longer relies on a single corpus; instead, it performs dynamic sampling between the Internet and structured knowledge bases via targeted queries autonomously generated by the LLM, thereby outperforming traditional passive retrieval in terms of evidence robustness and comprehensiveness. This evolution, progressing from agent-based iterative verification to domain-specific RAG constraints and further to broad-spectrum multi-source active retrieval, reflects continuous innovation in addressing the bottlenecks of timeliness in evidence acquisition and data source coverage.

(3) LLM-Empowered Reasoning Enhancement.

With the deepening of research, the role of LLMs in fake news detection is evolving from simple feature extractors toward intelligent hubs possessing deep logical reasoning and interpretability modeling capabilities. Deng et al. [60] proposed the D2et network, a denoising-enhanced multimodal detection framework. This framework utilizes LLMs to mine "local authenticity" and "global insight" labels, generating high-order semantic guidance signals to filter redundant noise in multimodal data and accurately capture deceptive signals in short-video scenarios. Zhang et al. [61] introduced the Negative Reasoning framework, a detection paradigm designed to exploit model hallucinations. By actively inducing and capturing logical inconsistencies through controlled negative reasoning chains, this method transforms the unique hallucination characteristics of LLMs into a tool for falsification, effectively revealing the inherent absurdity of fake news content. Liu et al. [62] proposed the TELLER framework, a trustworthy fake news detection system. This framework implements an explainable, generalizable, and controllable detection process through LLMs, utilizing the common-sense reasoning power of large models to perform deep decomposition of news claims, significantly enhancing the logical transparency of judgment results while improving detection accuracy. Zheng et al. [63] proposed a Rationale-Augmented detection method, utilizing Large Vision-Language Models (LVLMs) to achieve a leap from prediction to analysis. This method guides the model to generate detailed judgment rationales and analytical processes, combining multimodal feature fusion with deep semantic explanation to effectively address the difficulty traditional models face in providing concrete evidence. Regarding the aforementioned LLM-empowered reasoning enhancement methods, they exhibit significant mechanistic differences in logical mining depth and detection credibility. The D2et network focuses on a "semantic denoising" mechanism, optimizing the purity of multimodal fusion through global guidance signals generated by the LLM. The Negative Reasoning framework shifts toward a "reverse logical verification" mechanism, providing a new dimension for detecting hidden forged content by inducing internal model contradictions. The TELLER framework embodies a "systematic controllable reasoning" mechanism, achieving standardization and traceability of the detection process through logical decomposition. Finally, the Rationale-Augmented method establishes a "deep semantic analysis" mechanism, driving the evolution of models from single-category prediction to detailed verdict analysis. These paradigm shifts enable LLMs to serve as an irreplaceable logical hub in handling complex, variable, and highly adversarial fake news tasks.

Across the aforementioned LLM-based detection paradigms, the technical routes exhibit significant differentiated characteristics in terms of fusion logic and the depth of knowledge utilization, primarily manifesting as three types of technical solutions: "Endogenous Knowledge Triggering", "Exogenous Factual Anchoring", and "Deep Logical Inversion". First, the "Endogenous Knowledge Triggering" logic based on prompt engineering (e.g., DAFND) centers on activating the general semantic common sense and cross-modal logical reasoning capabilities accumulated during the LLM's pre-training phase through meticulously designed prompts. Its fusion logic is characterized as "content self-consistency verification", which utilizes the model's innate linguistic intuition to judge the sensationalism, logical contradictions, or visual forgery traces of news. Since this route requires no external data interaction, it offers an extremely fast response speed and is best suited for handling breaking fake news with obvious stylistic anomalies or logical loopholes, possessing significant advantages particularly in the early stages of dissemination when immediate external evidence is lacking. Second, the "Exogenous Factual Anchoring" logic based on Retrieval-Augmented Generation (RAG) (e.g., RAGAR) positions the LLM as a "verification agent" that obtains real-time evidence via retrieval plugins from search engines or authoritative databases. Its fusion logic is characterized as "evidence consistency comparison", which calibrates the model's judgment by injecting external facts into the context, effectively mitigating the "hallucination" problem of generative models. This route is more applicable to verification scenarios involving highly professional and time-sensitive information, such as political or economic facts or medical data, providing interpretable support with

cited sources to meet the rigorous requirements of serious news checking for evidence chain integrity. Finally, the "Deep Logical Inversion" logic based on reasoning enhancement (e.g., TELLER) transcends simple feature comparison. Its core lies in constructing explicit mechanisms for logical decomposition and contradiction discovery. Its fusion logic is characterized as "discriminative semantic deconstruction", achieving a leap from category prediction to deep rational analysis by actively inducing logical conflicts or generating rational analysis reports. This technical solution is most suitable for dealing with highly disguised fake information, such as AIGC-generated content, providing high transparency and controllable decision-making evidence for complex detection tasks. The refinement of these technical solutions not only clarifies the application boundaries of LLMs in multimodal environments but also provides differentiated technical choices for fake news across different domains and dissemination stages.

4.5. Summary of Fake News Detection Methods

As a core means of addressing complex social media environments, the research paradigm of multimodal fake news detection has achieved a transition from single-dimensional feature stacking toward deep semantic collaborative reasoning. Through the review of various detection methods mentioned above, it is evident that the effectiveness of multimodal detection depends not only on the extraction quality of single-modal features but more significantly on the model's capability to capture cross-modal relational logic and its depth of adaptation to specific application scenarios. In order to further clarify the developmental trajectory of various technologies and their practical value in real-world governance, this section will provide a systematic summary focusing on two dimensions: the logical hierarchy of technological evolution and the mapping relationship between methods and application scenarios.

4.5.1. The Evolution of Multimodal Fusion Technology

Multimodal fake news detection techniques have gradually evolved from early paradigms based on independent unimodal modeling and simple feature concatenation toward approaches emphasizing deep cross-modal interactions and multi-dimensional information collaboration. Current research primarily advances along four core directions: optimization of feature extraction, modeling of cross-modal correlations, enhancement with external knowledge, and multi-path integration for performance improvement. Performance variations of representative methods on three widely used benchmark datasets—Weibo, Twitter, and PHEME—reflect not only model adaptability across different languages and application scenarios, but also highlight the critical role of multimodal fusion strategies in detection effectiveness. All methods follow the initial experimental settings at the time of their respective dataset releases, thereby ensuring the maximum reference value for comparison. Below, we provide a concise overview of representative multimodal fake news detection models.

- att-RNN [9]: An attention-based recurrent neural network that integrates textual, visual, and social context features. Textual information is modeled using LSTM, visual features are extracted via the pretrained VGG19 model, and social context features are incorporated to enhance inter-modal correlation learning.
- EANN [64]: An event adversarial neural network framework designed to learn transferable representations across different events. It consists of a multimodal feature extractor, an event discriminator, and a fake news detector, enabling effective adaptation to diverse event-specific detection scenarios.
- MVAE [65]: A multimodal variational autoencoder that exploits variational autoencoders (VAEs) to learn latent representations of news text and images. By fusing these modalities into a unified news representation, MVAE facilitates effective fake news identification.
- SpotFake [66]: A lightweight detection model that employs BERT to learn textual features and extracts visual features using a VGG model. Fake news detection is performed through direct concatenation of textual and visual features, resulting in a simple and deployable architecture.
- CAFE [67]: A model that extracts textual features using BERT and visual features using ResNet-34, and introduces a cross-modal uncertainty weight α to adaptively fuse textual features, visual features, and text–image joint representations, thereby optimizing modality weight allocation.
- MCAN [68]: A multimodal co-attention network that extracts features from text, spatial visual domains, and frequency domains, and performs deep fusion through stacked shared attention layers to strengthen fine-grained cross-modal associations.
- HMCAN [69]: A hierarchical multimodal contextual attention network that jointly models multimodal contextual information and hierarchical textual semantics within a unified deep architecture, making it suitable for complex semantic scenarios.

- LIMR [70]: A detection model focusing on both intra-modal and inter-modal relationship mining. By strengthening internal associations within text and image modalities and enhancing cross-modal interactions, LIMR improves fake news identification accuracy and demonstrates stable performance across multiple datasets.
- HCCIN [71]: A human cognition–inspired consistency inference network that simulates human judgment processes. It extracts textual semantics, visual features, and modal temporal information, and dynamically adjusts modality weights through a multi-dimensional consistency verification module, making it particularly effective in handling subtle text–image inconsistencies.
- MMFN [72]: A multimodal fake news detection model based on multi-granularity information fusion. It extracts textual and visual features at different granularities and integrates multi-source information through a multi-stage fusion strategy, balancing global semantics and local details.
- FSUR [73]: A frequency-spectrum-based multimodal rumor detection model that argues frequency-domain representations are more effective for multimodal representation and fusion. By extracting frequency-spectrum features from text and images, FSUR enhances fusion effectiveness and improves detection performance.
- AAR [74]: A model that leverages multimodal large language models to generate adversarial arguments for detection. By producing contradiction-oriented arguments related to news content, AAR supplements evidential inconsistencies and further improves fake news detection accuracy.

Table 2 summarizes the core experimental results of the aforementioned mainstream multimodal fake news detection models on three commonly used datasets, providing an intuitive demonstration of the performance of different technical technical routes on representative benchmarks. It should be noted that all results are directly taken from the public reports of the original publications; although each study is based on the same dataset names, there may be differences in specific training and testing set split ratios, data cleaning and filtering rules, and evaluation implementation details. Therefore, this table is intended to show the performance evolution trends of different technical routes and does not possess strict horizontal comparability.

Table 2. Summary of fake news detection methods on common datasets.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo [9]	att-RNN [9]	0.788	0.862	0.686	0.764	0.738	0.890	0.807
	EANN [64]	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE [65]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake [66]	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	CAFE [67]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCAN [68]	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	HMCAN [69]	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	LIMR [70]	0.900	0.882	0.823	0.847	-	-	-
	HCCIN [71]	0.915	0.921	0.973	0.946	0.892	0.941	0.916
	MMFN [72]	0.923	0.921	0.926	0.924	0.924	0.920	0.922
	FSUR [73]	0.901	0.922	0.892	0.906	0.879	0.913	0.895
	AAR [74]	0.931	0.902	0.927	0.914	-	-	-
Twitter [12]	att-RNN	0.682	0.780	0.615	0.689	0.603	0.770	0.676
	EANN	0.715	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	MCAN	0.809	0.889	0.765	0.822	0.732	0.871	0.795
	HMCAN	0.897	0.971	0.801	0.878	0.853	0.979	0.912
	LIMR	0.831	0.836	0.832	0.830	-	-	-
	HCCIN	0.918	0.922	0.909	0.915	0.874	0.982	0.925
	MMFN	0.935	0.960	0.856	0.905	0.924	0.980	0.951
	FSUR	0.952	0.983	0.938	0.960	0.901	0.984	0.940
	AAR	0.924	0.910	0.925	0.919	-	-	-
Pheme [10]	att-RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	MVAE	0.852	0.806	0.719	0.760	0.817	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	CAFE	0.861	0.812	0.645	0.719	0.875	0.943	0.908
	MCAN	0.867	0.819	0.821	0.820	0.887	0.885	0.886
	HMCAN	0.881	0.830	0.838	0.834	0.910	0.905	0.907
	HCCIN	0.904	0.846	0.861	0.853	0.916	0.923	0.919
	AAR	0.928	-	-	0.923	-	-	-

From the performance of mainstream multimodal fake news detection models on the three representative datasets—Weibo, Twitter, and PHEME—as reported in Table 2, an overall trend can be observed that reflects the evolution of model architectures and fusion strategies in multimodal fake news detection. Early models such as

att-RNN and EANN primarily relied on simple feature concatenation or basic modality interactions, which limited their ability to capture deep cross-modal correlations and resulted in relatively modest overall performance. For example, on the Weibo dataset, att-RNN achieved an Accuracy of only 0.788 and an F1-score of 0.764 for the Fake News category. Similarly, EANN obtained a Fake News F1-score of merely 0.617 on the Twitter dataset, indicating its difficulty in handling complex multimodal scenarios.

Mid-stage models, including MVAE, SpotFake, and CAFE, improved modality-specific feature extraction and fusion strategies, leading to more stable performance across multiple datasets. SpotFake achieved an Accuracy of 0.869 on the Weibo dataset, with F1-scores exceeding 0.86 for both Fake News and Real News categories. CAFE further improved performance on the Twitter dataset, reaching an Accuracy of 0.806, while maintaining F1-scores above 0.80 for both classes, demonstrating the effectiveness of enhanced feature extraction and fusion mechanisms.

Advanced models employing deep collaborative fusion mechanisms consistently achieved superior performance, with each model realizing breakthroughs through distinct fusion strategies. MCAN deeply integrates textual features with spatial and frequency-domain visual representations via stacked shared attention layers. HMCAN focuses on joint modeling of hierarchical semantics and multimodal contextual information, strengthening hierarchical cross-modal interactions. HCCIN simulates human cognitive reasoning by introducing a multi-dimensional consistency verification module that dynamically adjusts modality weights. LIMR emphasizes both intra-modal relationship mining and cross-modal interaction modeling to enhance semantic alignment. MMFN adopts multi-granularity feature extraction and integrates global and local information through multi-stage fusion. FSUR exploits the advantages of frequency-spectrum features to improve multimodal representation and fusion effectiveness. AAR, leveraging multimodal large language models, generates adversarial arguments to supplement cross-modal contradictory evidence. Among these methods, HCCIN achieved an Accuracy of 0.915 and a Fake News F1-score of 0.946 on the Weibo dataset, and an Accuracy of 0.918 with a Fake News F1-score of 0.915 on the Twitter dataset. By incorporating multimodal co-attention, hierarchical semantic modeling, and cognition-inspired consistency reasoning, HCCIN effectively captures fine-grained text–image correlations and reduces misclassification caused by modality inconsistency. MMFN reached an Accuracy of 0.935 on the Twitter dataset, with F1-scores exceeding 0.90 for both classes. FSUR further improved performance on Twitter, achieving an Accuracy of 0.952 and a Fake News F1-score of 0.960, indicating that frequency-spectrum features offer advantages in multimodal representation and fusion and highlighting their potential in complex multimodal scenarios. AAR, benefiting from adversarial arguments generated by multimodal large language models, achieved Accuracies of 0.931 and 0.924 on the Weibo and Twitter datasets, respectively, demonstrating strong representativeness among comparable methods.

From a dataset perspective, models generally perform better on the Chinese Weibo dataset than on the English Twitter dataset. For instance, att-RNN achieved an Accuracy of 0.788 on Weibo, which is significantly higher than its Accuracy of 0.682 on Twitter. This discrepancy may be attributed to the more complete modal composition of the Weibo dataset and the degree of model adaptation to Chinese semantic characteristics. As a bilingual multimodal dataset, PHEME places greater demands on cross-lingual semantic alignment. Although HCCIN achieved a lower Fake News F1-score on PHEME (0.853) compared to Weibo (0.946), it still maintained relatively strong performance, reflecting its potential for cross-lingual generalization.

Overall, the observed performance differences among these models not only highlight the impact of architectural design and fusion strategies on multimodal information utilization, but also reveal the influence of dataset language characteristics and modal composition on detection effectiveness. These insights provide clear directions for future model optimization and dataset construction.

4.5.2. Technical Characteristics and Application Scenarios

The technical characteristics of multimodal fake news detection methods determine their applicability boundaries across different social media scenarios. For short-video platforms such as Douyin or Kuaishou, methods incorporating temporal consistency verification or frequency-domain feature analysis exhibit higher adaptability due to the strong audio-visual synchronization and complex temporal correlations of the information. The detection logic in such scenarios focuses on identifying subtle manipulation traces between audio and video, as simple image-text comparisons struggle to capture forgery signals within video streams. In contrast, information on image-text social media like Weibo or Twitter is characterized by fragmentation and a high dependency on social interaction feedback. Consequently, multimodal co-attention mechanisms or social context modeling schemes are more effective at capturing fine-grained contradictions between text and static images, while leveraging social signals such as retweets and comments to enhance judgment results.

In serious news domains with high professional barriers and rigorous factual requirements, such as politics or healthcare, the detection logic shifts from semantic comparison toward factual verification. Methods based on factual verification or knowledge graph enhancement possess significant advantages in these scenarios; their technical core lies in verifying the accuracy of claims through external authoritative knowledge bases and providing explainable support with cited sources. For early-stage breaking events or highly disguised AIGC (AI-Generated Content) scenarios, where immediate external evidence is lacking and visual flaws are minimal, research focus has shifted toward zero-shot reasoning or logical enhancement technologies powered by Large Language Models (LLMs). By performing deep deconstruction of content self-consistency through the model's endogenous common-sense reasoning, these approaches can tackle highly disguised fake information and provide decision-making evidence with logical transparency. This technical segmentation based on scenario characteristics clarifies the application boundaries of various detection methods and provides a theoretical basis for the differentiated governance of multimodal fake news.

5. Challenges and Future Directions

5.1. Challenges

Although multimodal fake news detection integrates multiple sources of information such as text and images and overcomes the limitations of unimodal approaches, it still faces multifaceted challenges arising from modality characteristics, data quality, and model design. The major challenges can be summarized as follows.

(1) Dataset limitations

From the perspective of research limitations, existing multimodal datasets exhibit significant shortfalls across multiple dimensions. First, there is a deficiency in multimodal integrity, as most datasets remain confined to the dual-modal combination of text and images. According to statistics from the representative datasets in Table 1, only the FakeSV dataset covers video and audio modalities, accounting for less than 10% of the total, which makes it difficult to support in-depth research on complex multimodal fake content such as modern short videos. Second, data privacy risks are prominent, as acquiring sensitive data such as user behavior and propagation paths can easily infringe upon privacy. Third, there is an imbalance in cross-linguistic and cross-cultural coverage, with existing mainstream datasets being highly concentrated in English. Statistical analysis of the representative datasets in Table 1 shows that English datasets account for over 80%, while data for low-resource languages is extremely scarce, and semantic differences across different cultural backgrounds are generally overlooked. Fourth, there is a severe imbalance in category distribution. Manually constructed experimental datasets typically lean toward a balanced 1:1 distribution of positive and negative samples. However, this is seriously decoupled from the actual social media environment, where the actual proportion of fake news is usually much lower than that of real news. This extreme discrepancy in category distribution leads models to easily develop algorithmic biases during training, resulting in a significant drop in accuracy when identifying fake content during practical deployment. To mitigate these multi-dimensional shortfalls of multimodal datasets, existing research is advancing in the directions of modal supplementation, data augmentation, and cross-linguistic adaptation. The FakeSV dataset constructed by Qi et al. [15] supplements text, video, and audio modalities in Chinese short-video scenarios and includes complete social context information, partially addressing the issue of insufficient multimodal completeness. The MCOT framework proposed by Shen et al. [25] employs contrastive learning to construct positive and negative text–image pairs and combines optimal transport theory to align modality feature distributions, effectively alleviating model bias caused by class imbalance. The MM-COVID dataset developed by Li et al. [17] supports bilingual Chinese–English text–image data and provides foundational resources for cross-lingual fake news detection, pointing the way toward future multilingual dataset construction despite limited low-resource samples.

(2) Cross-domain challenges.

Most existing multimodal fake news detection models are trained for specific domains and struggle to adapt to domain-specific variations in deceptive patterns. For example, political fake news emphasizes factual contradictions and source credibility, whereas entertainment fake news focuses more on sensational headlines, image manipulation, and propagation popularity. When applied across domains, models often fail to rapidly learn new deception patterns, resulting in significant accuracy drops and a lack of generalizable cross-domain feature adaptation mechanisms. To address these challenges, researchers have explored domain adaptation and feature optimization strategies. Nan et al. [8] proposed the MDFEND framework, which incorporates domain-adaptive feature extraction modules capable of handling multimodal text–image data across domains such as politics and entertainment, effectively reducing domain shift–induced errors. Xie et al. [31] further proposed a knowledge graph-enhanced heterogeneous graph neural network model. By integrating knowledge graphs into fake news

detection, this approach leverages external knowledge to mitigate feature distribution discrepancies across different domains, thereby enhancing cross-domain generalization capabilities. Gong et al. [52] further proposed a causal debiasing-based cross-domain detection method that separates domain-independent causal features from domain-specific confounding factors using structural causal models and reweighting strategies, partially mitigating performance degradation in cross-domain scenarios.

(3) Lack of model interpretability.

Most multimodal fusion models exhibit black-box behavior, with only a few studies incorporating interpretability mechanisms. When a model classifies news as fake, it often cannot clearly explain whether the decision is based on text–image semantic inconsistency, video manipulation artifacts, or abnormal propagation patterns, undermining user trust. More critically, interpretability outputs themselves may be vulnerable to adversarial manipulation, further weakening system credibility and hindering real-world deployment. To enhance interpretability, recent studies have leveraged knowledge integration, feature visualization, and logical traceability. Hao et al. [32] proposed a multimodal fake news detection framework that leverages fine-grained knowledge graphs to enhance the interpretability of modal fusion, providing a fundamental basis for the model to output its judgment rationale. The EC-Fake model introduced by Han et al. [35] separates image foreground objects from background context through semantic decoupling, accurately localizes manipulated regions, and employs large language models to extract external knowledge and generate natural language explanations. The FactAgent framework proposed by Li et al. [58] adopts a three-stage process—claim decomposition, authoritative evidence retrieval, and contradiction comparison—to produce traceable explanations with cited sources. Wu et al. [75] proposed the MFIR framework, which constructs an explicit reasoning module by deeply mining inconsistent features between text and images. This framework not only identifies semantic conflicts between modalities but also intuitively presents the basis for judgment, thereby providing logical interpretability support for fake news identification.

(4) Insufficient depth of multimodal fusion.

Many existing fusion strategies remain at a shallow level, relying on feature concatenation or simple weighting schemes without exploiting deep semantic correlations across modalities. Alternatively, modality-specific feature extractors are designed independently, resulting in limited inter-modal interaction and poor capture of subtle cross-modal inconsistencies. Moreover, current fusion methods struggle with temporal asynchrony in complex multimodal combinations, further constraining fusion quality. To address these limitations, researchers have explored deep interaction mechanisms and temporal-aware designs. The MCOT framework by Shen et al. [25] aligns text and image embedding spaces via contrastive learning and optimizes feature distribution alignment using optimal transport, overcoming shallow fusion constraints. The MCAN framework proposed by Wu et al. [68] extracts features from textual, spatial visual, and frequency domains and constructs dynamic cross-modal associations through stacked shared attention layers. The HCCIN framework by Wu et al. [71] simulates human cognitive reasoning by jointly modeling textual semantics, visual features, and temporal modality information, dynamically adjusting modality weights through multi-dimensional consistency verification to capture subtle cross-modal contradictions.

(5) Limited fine-grained detection.

Most detection tasks are restricted to binary classification, failing to capture finer-grained deception levels. Mature binary classifiers often perform poorly on fine-grained tasks, which require new models capable of distinguishing increasingly ambiguous class boundaries. Fine-grained detection is inherently challenging, as class distinctions become blurred and demand more precise multi-class decision functions and feature representations. To address this limitation, researchers have developed fine-grained label schemes and enriched annotations. The LIAR dataset by Wang et al. [16] introduces five-level veracity labels, ranging from true to completely false, providing a foundation for fine-grained detection. The Fakeddit dataset by Nakamura et al. [19] includes over one million samples with hierarchical labels capturing both deception type and degree across text–image modalities. Chalehchaleh et al. [56] further proposed an Llama3-based fine-grained data augmentation framework that generates synthetic samples with graded veracity scores via few-shot prompting, improving detection performance.

(6) Insufficient early detection capability.

In the early stages of fake news dissemination, only source-level multimodal content is available, with little or no social interaction data. Existing models often over-rely on propagation features, resulting in low early-stage detection accuracy. Moreover, deep semantic contradictions and publisher credibility cues are difficult to extract at early stages, causing missed intervention opportunities and exacerbating social harm, particularly in critical scenarios such as pandemics and elections. To improve the current situation where early detection of fake news over-relies on propagation data and suffers from low accuracy, researchers have optimized detection efficiency

from perspectives such as propagation modeling, real-time evidence acquisition, and few-shot and incremental learning. The Hy-DeFake framework by Su et al. [45] constructs hypergraph neural networks using news and users as nodes and early reposts and likes as hyperedges, achieving an early detection accuracy of 82% within one hour of propagation. Shao et al. [76] proposed an active learning framework that reduces labeling requirements and adapts dynamically to social media evolution, improving early detection timeliness and generalization. Suryawanshi et al. [77] proposed the FakeIDCA method, which is based on incremental learning and concept drift adaptation. This approach enables the rapid identification of emerging fake news patterns without the need for extensive historical data, effectively enhancing the recognition capability for unknown fake samples during the early stages of dissemination and providing support for early intervention.

(7) Limited model adaptability and generalization.

Models often struggle to adapt to emerging deception techniques, particularly AIGC-generated multimodal fake content, leading to sharp performance drops due to unseen feature distributions. Cross-platform generalization is also weak, as differences in content style and propagation mechanisms hinder rapid adaptation. To enhance robustness, researchers have explored dynamic adaptation and multilingual optimization strategies. Gong et al. [52] proposed a causal debiasing approach that separates core deceptive features from domain-specific confounders using structural causal models. Hu et al. [54] leveraged the zero-shot and few-shot learning capabilities of GPT-3.5-turbo to rapidly adapt to AIGC-generated fake content without retraining. Chalehchaleh et al. [56] proposed an LLM-based data augmentation framework that uses English prompts to guide adaptation to low-resource languages, achieving F1-score improvements of 10–13% in multilingual settings.

(8) Real-time detection constraints.

Multimodal models often involve complex architectures with modality-specific feature extractors, leading to high computational costs and long inference latency. For example, multi-stream Transformer models may incur inference delays of several hundred milliseconds, exceeding the second-level response requirements of social media platforms. While lightweight solutions exist, they often come at the cost of significant accuracy degradation. To balance real-time performance and accuracy, researchers have explored model compression and knowledge distillation. Lan et al. [48] proposed an improved distillation method that corrects teacher model errors through label refinement and data selection, achieving a threefold inference speedup with only a 2.3% accuracy drop. The SSDL framework by Mu et al. [49] reduces model parameters by 60% through self-supervised representation learning and shortens inference latency to under 50 ms. The LEKD model by Chen et al. [50] distills external knowledge extracted by large language models into small language models and simplifies feature extraction modules (e.g., replacing VGG19 with MobileNet), achieving 89.2% accuracy on the Weibo dataset while being four times faster than conventional multimodal models.

5.2. Future Directions

To address the eight core challenges identified in Section 5.1 and propel multimodal fake news detection technology toward a more efficient and reliable new stage, future research must achieve systematic breakthroughs across dimensions such as data quality, model mechanisms, and application scenarios. This section proposes nine research directions that possess an explicit targeted mapping logic with the aforementioned core challenges as detailed in Table 3, aiming to overcome existing bottlenecks through continuous technological evolution.

(1) Constructing high-quality multimodal datasets.

To address the core bottleneck of dataset limitations, research should advance the development of full-modal datasets by incorporating dynamic modalities such as audio and video. By integrating multi-dimensional information including text, images, audio, propagation networks, and user interactions, a content-and-behavior dual-driven dataset architecture can be formed. During the construction process, standardized best practices and quality control criteria must be followed. First, a standardized annotation pipeline should be established, employing expert review and double-blind labeling mechanisms while assessing inter-annotator agreement to ensure the objectivity and accuracy of labels from the source. Second, multi-dimensional semantic consistency verification must be strengthened; during the data cleaning phase, a combination of automated tools and manual review should be used to rigorously verify the semantic alignment of cross-modal information and eliminate noisy samples with modal fragmentation. Simultaneously, privacy protection and compliance mechanisms should be established, utilizing technologies such as federated learning and differential privacy to ensure the legality of data utilization [78]. Finally, category distribution and scenario simulation should be optimized. Data augmentation can mitigate imbalance issues, and the construction of long-tail distribution sample libraries that align with real-world social media patterns is essential. By establishing a comprehensive quality assessment system across

dimensions such as modal integrity, annotation accuracy, and scenario coverage, researchers can ensure that datasets possess practical reference value for real-world governance.

(2) Advancing multilingual and cross-domain detection.

To address the issues of cross-domain feature shift and imbalanced multilingual coverage, researchers can utilize multilingual pre-trained models for multilingual detection. By optimizing the semantic representation of low-resource languages, the model's cross-lingual adaptation capability can be enhanced through transfer learning [79]. For cross-domain detection, domain-adaptive techniques should be developed to enable models to rapidly extract domain-invariant multimodal features and reduce performance degradation caused by domain shifts. Moreover, cross-lingual and cross-domain benchmark datasets with unified evaluation protocols should be constructed to promote fairness and generalization.

(3) Enhancing model reasoning capability and interpretability.

To address the issue of low trust caused by the lack of model interpretability, researchers should prioritize enhancing reasoning capabilities by integrating causal inference and external knowledge graphs (e.g., factual databases and source credibility repositories). This enables models not only to identify falsehoods but also to support detection conclusions with robust logic and factual evidence [80]. Interpretability should be further enhanced by incorporating attention visualization and feature attribution techniques to generate intuitive detection reports, such as highlighting text–image inconsistencies, emphasizing manipulated video regions, and tracing anomalous propagation nodes. User studies should be conducted to optimize the presentation of explanations (e.g., natural language reports and visual analytics), improving user comprehension and trust. Meanwhile, robust and attack-resistant interpretability mechanisms should be developed to prevent explanation manipulation.

(4) Deepening multimodal fusion and semantic alignment.

To overcome insufficient multimodal fusion depth and the challenges of audio-visual asynchrony, researchers should explore deep fusion architectures. By utilizing cross-modal pre-trained models (e.g., CLIP) to construct a unified semantic space, the fine-grained correlations between modalities can be captured through contrastive learning and cross-modal self-attention mechanisms [81]. Temporal-aware dynamic fusion methods should be developed to address temporal asynchrony among text, video, and audio by incorporating temporal attention mechanisms and ConvLSTM, dynamically adjusting modality weights and interaction patterns. Multimodal consistency verification should also be strengthened by designing detection modules based on frequency-domain and physiological features to accurately identify subtle manipulation traces in deepfake content.

(5) Promoting fine-grained fake news detection.

To address the deficiency in fine-grained detection that leads to the inability to distinguish complex types of falsehoods, a fine-grained classification system should be constructed to categorize both the types and degrees of falsity. Furthermore, regression models should be introduced to perform quantitative scoring of the degree of falsehood [82]. Fine-grained feature extraction techniques should be developed, with specialized detection modules tailored to different deception categories. In addition, fine-grained annotated datasets should be constructed to provide richer supervision for model training, supporting precise governance and differentiated intervention strategies.

(6) Optimizing early-stage fake news detection.

To address the issues of insufficient early detection capability and high dependency on propagation data, researchers can integrate multi-dimensional early features. By combining news content features with static attributes, the reliance on propagation data can be effectively reduced. Few-shot learning and incremental learning techniques should be employed to rapidly learn deceptive patterns at early dissemination stages, improving early detection accuracy. Dynamic propagation prediction models can be constructed by incorporating historical diffusion data and trending event signals to proactively identify high-risk fake news and provide a timely intervention window [83].

(7) Enhancing model adaptability and generalization.

To address the issue of insufficient model adaptability, dynamic adaptation mechanisms should be developed. By employing incremental learning and continual learning technologies, models can update their feature extraction capabilities in real-time to counter emerging multimodal fake content generated by AIGC [77]. Meta-learning techniques can be leveraged to allow rapid adaptation to new scenarios and platforms with minimal labeled data, reducing retraining costs. In addition, a universal multimodal feature repository should be constructed to extract domain- and platform-invariant deceptive patterns, enhancing cross-scenario generalization and reducing dependence on specific datasets.

(8) Improving model lightweighting and real-time detection capability.

To address the high-latency bottlenecks in the large-scale real-time governance of social media, future research must focus on resolving the trade-off between detection accuracy and inference efficiency. First, model architectures should be simplified by developing universal multimodal fusion frameworks that replace redundant modal-specific modules with a single unified encoder, thereby reducing computational complexity at the structural source. Second, model compression technologies should be deeply advanced by combining methods such as knowledge distillation, parameter quantization, and pruning to enhance processing speed while strictly controlling accuracy loss [84]. Furthermore, it is essential to overcome the deployment cost pressures in practical applications by exploring hardware-aware design and edge computing deployment. By leveraging specialized chip optimization and edge node inference, "second-level detection" responses can be achieved. This engineering trajectory, progressing from algorithmic lightweighting to hardware-collaborative optimization, will effectively alleviate the computational resource burden during large-scale real-time monitoring and propel multimodal detection technology from laboratory environments toward authentic, industrial-grade application scenarios.

(9) Integrating large language models.

To address the bottlenecks of models in complex semantic understanding and general knowledge, the strong reasoning capabilities of Large Language Models (LLMs) such as GPT-4 and Qwen can be leveraged to assist in multimodal fake news detection. LLMs can be used to generate logical analysis reports of multimodal content, assessing semantic consistency and factual plausibility. Through prompt engineering, LLMs can identify emerging multimodal fake content in zero-shot and few-shot scenarios [85]. Furthermore, collaborative frameworks combining LLMs and multimodal fusion models should be developed, where LLMs provide factual knowledge and reasoning support, while multimodal models capture visual and audio manipulation cues, achieving complementary strengths.

Table 3. Mapping between future research directions and core challenges.

	Future Research Direction	Core Challenges Addressed	Mapping Logic & Strategic Objectives
(1)	Construction of High-Quality Datasets	Dataset Limitations	Establishing standardized annotation and consistency verification systems to resolve modal missingness and sample distribution imbalance.
(2)	Multilingual and Cross-domain Detection	Cross-domain Issues	Mitigating English data bias and enhancing the transferability of models across different cultural contexts and vertical domains.
(3)	Enhanced Reasoning and Interpretability	Lack of Model Interpretability	Introducing causal reasoning to break the "black-box" nature of models, providing traceable evidentiary support and logical explanations.
(4)	Deepened Multimodal Semantic Alignment	Insufficient Multimodal Fusion Depth	Resolving audio-visual asynchrony and deep semantic fragmentation while excavating high-order correlations between modalities.
(5)	Advancement of Fine-grained Detection	Deficiency in Fine-grained Detection	Breaking the limitations of binary classification to achieve precise identification of complex deceptive forms such as "half-true, half-fake."
(6)	Optimization of Early Detection Capability	Insufficient Early Detection Capability	Reducing dependence on propagation features and capturing weak signals during the incubation period to achieve timely intervention.
(7)	Enhanced Adaptability and Generalization	Insufficient Model Adaptability and Generalization	Countering emerging forgery content generated by AIGC and enhancing model robustness within dynamic social environments.
(8)	Lightweighting and Real-time Detection	Real-time Detection Issues	Optimizing architectural redundancy and resolving high-latency bottlenecks to meet the demands of large-scale real-time monitoring on social media.
(9)	Integration with Large Language Models	Cross-domain Issues, Insufficient Generalization, Lack of Interpretability, etc.	Leveraging the common-sense reasoning and general capabilities of LLMs to bridge the semantic understanding gaps of small-scale models.

6. Conclusions

As a core technology for addressing complex misinformation, multimodal fake news detection has evolved from early approaches based on simple multimodal feature concatenation to advanced paradigms that achieve deep collaborative fusion through graph neural networks and Transformer architectures. By effectively integrating textual semantics, visual features, audio temporal signals, and social propagation information, recent methods have substantially improved detection accuracy for multimodal fake content. Significant breakthroughs have been achieved in areas such as cross-modal pretraining, deepfake detection, and large language model–assisted

reasoning, with several models reporting F1-scores exceeding 95% on public fake news benchmarks, laying a solid foundation for practical deployment.

This survey provides a systematic and comprehensive overview of the multimodal fake news detection landscape. Starting from a clear definition of core concepts, datasets, and evaluation metrics, we establish a structured research framework. Following the trajectory of technological evolution, existing detection methods are categorized into four major directions: content feature-based methods, fact-based verification methods, propagation dynamics-based modeling, and application-oriented optimization methods. Through comparative analysis of representative approaches on typical benchmark datasets, we clarify the strengths and applicable boundaries of each category. Furthermore, we conduct an in-depth examination of current challenges related to data quality, model design, fusion strategies, and real-world deployment, and outline forward-looking research directions accordingly.

Looking ahead, breakthroughs are expected through the construction of high-quality full-modality datasets, deeper multimodal fusion and semantic alignment, enhanced model interpretability and generalization, and the integration of large language models to strengthen reasoning capabilities. These advances have the potential to move multimodal fake news detection from passive identification toward proactive intervention in misinformation propagation. Continued innovation in this field will not only provide essential technical support for social media information governance, but also play a crucial role in maintaining public trust, safeguarding political stability, and protecting public health. In critical scenarios such as pandemic response, election oversight, and public safety, accurate identification of multimodal fake information can effectively reduce misleading decision-making and prevent societal harm. With the deep integration of cross-modal learning and information security technologies, multimodal fake news detection is poised to become a key breakthrough area in artificial intelligence, offering vital support for building a healthy and trustworthy digital information ecosystem.

Author Contributions

G.C.: literature research, writing, revision, and supervision. Z.Q.: literature research, writing, and revision. G.B.: literature research and revision. Q.L.: literature research. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China, grant number 62366010, the Guangxi Natural Science Foundation, grant number 2024GXNS-FAA010374, and the Innovation Project of GUET Graduate Education, grant number 2025YCXS048.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were created or analyzed in this study.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Gemini to polish the language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Li, X.; Qiao, J.; Yin, S.; et al. A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*, 2658–2675.

2. Alghamdi, J.; Luo, S.; Lin, Y. A comprehensive survey on machine learning approaches for fake news detection. *Multimed. Tools Appl.* **2024**, *83*, 51009–51067.
3. Nasser, M.; Arshad, N.I.; Ali, A.; et al. A Systematic Review of Multimodal Fake News Detection on Social Media Using Deep Learning Models. *Results Eng.* **2025**, *26*, 104752.
4. Yang, S.; Li, X.; Du, Y. A Survey of Fake News Detection Methods Based on Online Social Networks. *J. Xihua Univ.* **2025**, *44*, 37–46. (in Chinese).
5. Tufchi, S.; Yadav, A.; Ahmed, T. A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities. *Int. J. Multimed. Inf. Retr.* **2023**, *12*, 28.
6. Ma, J.; Gao, W.; Mitra, P.; et al. Detecting Rumors from Microblogs with Recurrent Neural Networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI), New York, NY, USA, 9–15 July 2016; pp. 3818–3824.
7. Zhang, X.Y.; Cao, J.; Li, X.R.; et al. Mining Dual Emotion for Fake News Detection. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 3465–3476.
8. Nan, Q.; Cao, J.; Zhu, Y.C.; et al. MDFEND: Multi-Domain Fake News Detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM), Gold Coast, QLD, Australia, 1–5 November 2021; pp. 3343–3347.
9. Jin, Z.; Cao, J.; Guo, H.; et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 795–816.
10. Zubiaga, A.; Liakata, M.; Procter, R. Exploiting Context for Rumor Detection in Social Media. In Proceedings of the International Conference on Social Informatics (SocInfo), Oxford, UK, 13–15 September 2017; pp. 109–123.
11. Ma, J.; Gao, W.; Wong, K.F. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 708–717.
12. Boididou, C.; Papadopoulos, S.; Zampoglou, M.; et al. Detection and Visualization of Misleading Content on Twitter. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 71–86.
13. Wu, Y.; Tang, Y.; Fan, C.; et al. MDVT: A Multimodal Fake News Detection Framework Based on Vision Transformer. In Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing, Sanya China, 27–29 December 2023.
14. Shu, K.; Mahudeswaran, D.; Wang, S.H.; et al. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188.
15. Qi, P.; Bu, Y.Y.; Cao, J.; et al. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI), Washington, DC, USA, 7–14 February 2023; pp. 14444–14452.
16. Wang, W.Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 422–426.
17. Li, Y.; Jiang, B.; Shu, K.; et al. Toward a Multilingual and Multimodal Data Repository for COVID-19 Disinformation. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Online, 10–13 December 2020; pp. 4325–4330.
18. Popat, K.; Mukherjee, S.; Strötgen, J.; et al. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW Companion), Perth, WA, Australia, 3–7 April 2017; pp. 1003–1012.
19. Nakamura, K.; Levy, S.; Wang, W.Y. Fakeddit: A New Multimodal Benchmark Dataset for Fine-Grained Fake News Detection. In Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; pp. 6149–6157.
20. Liu, H.; Chen, S.; Cao, S.; et al. Fake News Detection Based on Multimodal Learning. *Comput. Sci. Explor.* **2023**, *17*, 2015–2029.
21. Cantini, R.; Cosentino, C.; Kilanioti, I.; et al. Unmasking Deception: A Topic-Oriented Multimodal Approach to Uncover False Information on Social Media. *Mach. Learn.* **2025**, *114*, 13.
22. Farhoudinia, B.; Ozturkcan, S.; Kasap, N. Emotions Unveiled: Detecting COVID-19 Fake News on Social Media. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 932.
23. Qian, L.; Xu, R.; Zhou, Z. MRDCA: A Multimodal Approach for Fine-Grained Fake News Detection through Integration of RoBERTa and DenseNet Based upon Fusion Mechanism of Co-Attention. *Ann. Oper. Res.* **2025**, *348*, 257–278.
24. Chen, W.; Cai, F.; Guo, Y.; et al. Contrastive Learning of Cross-Modal Information Enhancement for Multimodal Fake News Detection. *Complex Intell. Syst.* **2025**, *11*, 303.

25. Shen, X.; Huang, M.; Hu, Z.; et al. Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. *Front. Comput. Sci.* **2024**, *6*, 1473457.
26. Xing, Y.; Zhai, C.; Che, Z.; et al. A Multimodal Fake News Detection Model Based on Bidirectional Semantic Enhancement and Adversarial Network Under Web3.0. *Electronics* **2025**, *14*, 3652.
27. Glenski, M.; Weninger, T.; Volkova, S. Propagation from Deceptive News Sources: Who Shares, How Much, How Evenly, and How Quickly? *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 1071–1082.
28. Li, P.; Sun, X.; Yu, H.; et al. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Trans. Multimed.* **2022**, *24*, 3455–3468.
29. Qi, P.; Zhao, Y.; Shen, Y.; et al. Two Heads Are Better than One: Improving Fake News Video Detection by Correlating with Neighbors. In Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 11947–11959.
30. Zhang, Y.; Su, X.; Wu, J.; et al. EmoKnow: Emotion- and Knowledge-Oriented Model for COVID-19 Fake News Detection. In Proceedings of the 19th International Conference on Advanced Data Mining and Applications (ADMA), Cham, Switzerland, 2023; pp. 352–367.
31. Xie, B.; Ma, X.; Wu, J.; et al. Knowledge Graph Enhanced Heterogeneous Graph Neural Network for Fake News Detection. *IEEE Trans. Consum. Electron.* **2023**, *70*, 2826–2837.
32. Hao, R.; Luo, H.; Li, Y. Multi-Modal Fake News Detection Enhanced by Fine-Grained Knowledge Graph. *IEICE Trans. Inf. Syst.* **2024**, *E107-D*, 1234–1245.
33. Hu, L.; Yang, T.; Zhang, L.; et al. Compare to the Knowledge: Graph Neural Fake News Detection with External Knowledge. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Online, 1–6 August 2021; pp. 754–763.
34. Fu, L.; Liu, S. Multimodal Fake News Detection Incorporating External Knowledge and User Interaction Feature. *Adv. Multimed.* **2023**, *2023*, 8836476.
35. Han, M.; Li, J.; Chen, Y.; et al. EC-Fake: A Fake News Detection Model Based on External Knowledge and Contrast-Driven Feature Augmentation. *Neurocomputing* **2025**, *132*, 131214.
36. Yu, W.; Ge, J.; Chen, Z.; et al. Research on Fake News Detection Based on Dual Evidence Perception. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108271.
37. Cheung, T.H.; Lam, K.M. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. In Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 31 October–3 November 2023; pp. 846–853.
38. Zheng, L.; Li, C.; Zhang, X.; et al. Evidence Retrieval Is Almost All You Need for Fact Verification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 9274–9281.
39. Wu, L.; Wang, K.; Nie, K.; et al. TFGIN: Tight-Fitting Graph Inference Network for Table-Based Fact Verification. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–26.
40. Wu, L.; Wang, L.; Zhao, Y. Unified Evidence Enhancement Inference Framework for Fake News Detection. In Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju, South Korea, 3–9 August 2024; pp. 6541–6549.
41. Wu, L.; Yu, D.; Liu, P.; et al. Heuristic Heterogeneous Graph Reasoning Networks for Fact Verification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 14959–14973.
42. Dellys, H.N.; Mokeddem, H.; Sliman, L. On the Integration of Social Context for Enhanced Fake News Detection Using Multimodal Fusion Attention Mechanism. *AI* **2025**, *6*, 78.
43. Raza, S.; Ding, C. Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach. *Int. J. Data Sci. Anal.* **2022**, *13*, 335–362.
44. Mehta, N.; Pacheco, M.L.; Goldwasser, D. Tackling Fake News Detection by Continually Improving Social Context Representations Using Graph Neural Networks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022; pp. 1363–1380.
45. Su, X.; Yang, J.; Wu, J.; et al. Hy-DeFake: Hypergraph Neural Networks for Detecting Fake News in Online Social Networks. *Neural Netw.* **2025**, *187*, 107302.
46. Malik, A.; Behera, D.K.; Hota, J.; et al. Ensemble Graph Neural Networks for Fake News Detection Using User Engagement and Text Features. *Results Eng.* **2024**, *24*, 103081.
47. Lakzaei, B.; Chehreghani, M.H.; Bagheri, A. LOSS-GAT: Label Propagation and One-Class Semi-Supervised Graph Attention Network for Fake News Detection. *Appl. Soft Comput.* **2025**, *174*, 112965.
48. Lan, W.; Cheung, Y.; Xu, Q.; et al. Improve Knowledge Distillation via Label Revision and Data Selection. *IEEE Trans. Cogn. Dev. Syst.* **2025**, *17*, 1377–1388.

49. Mu, M.; Das Bhattacharjee, S.; Yuan, J. Self-Supervised Distilled Learning for Multi-Modal Misinformation Identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; pp. 2819–2828.
50. Chen, X.; Huang, X.; Gao, Q.; et al. Enhancing Text-Centric Fake News Detection via External Knowledge Distillation from LLMs. *Neural Netw.* **2025**, *187*, 107377.
51. Chen, Z.; Hu, L.; Li, W.; et al. Causal Intervention and Counterfactual Reasoning for Multi-Modal Fake News Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, Toronto, ON, Canada, 9–14 July 2023; pp. 627–638.
52. Gong, S.; Sinnott, R.; Qi, J.; et al. Unseen Fake News Detection through Causal Debiasing. In Proceedings of the ACM on Web Conference 2025 (WWW); Sydney, NSW, Australia, 28 April–2 May 2025; pp. 981–985.
53. Cheng, L.; Guo, R.; Shu, K.; et al. Causal Understanding of Fake News Dissemination on Social Media. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Singapore, 14–18 August 2021; ACM: New York, NY, USA, 2021; pp. 148–157.
54. Hu, B.; Sheng, Q.; Cao, J.; et al. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; AAAI Press: Palo Alto, CA, USA, 2024; pp. 22105–22113.
55. Liu, Y.; Zhu, J.; Liu, X.; et al. Detect, Investigate, Judge and Determine: A Knowledge-Guided Framework for Few-Shot Fake News Detection. In Proceedings of the 2025 IEEE International Conference on Data Mining (ICDM), Washington DC, DC, USA, 12–15 November 2025; pp. 477–486.
56. Chalehchaleh, R.; Farahbakhsh, R.; Crespi, N. Addressing Data Scarcity in Multilingual Fake News Detection: An LLM-Based Dataset Augmentation Approach. *Soc. Netw. Anal. Min.* **2025**, *15*, 92.
57. Li, D.; Li, F.; Song, B.B.; et al. IMRRF: Integrating Multi-Source Retrieval and Redundancy Filtering for LLM-Based Fake News Detection. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Miami, FL, USA, 10–15 June 2025; pp. 9127–9142.
58. Li, H.; Huang, J.; Ji, M.; et al. Use of Retrieval-Augmented Large Language Model for COVID-19 Fact-Checking: Development and Usability Study. *J. Med. Internet Res.* **2025**, *27*, e66098.
59. Khaliq, M.A.; Chang, P.Y.C.; Ma, M.; et al. RAGAR, Your Falsehood Radar: RAG-Augmented Reasoning for Political Fact-Checking Using Multimodal Large Language Models. In Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER), Miami, FL, USA, 16–21 June 2024; pp. 280–296.
60. Deng, X.; Yu, P.; Qian, S.; et al. Denoising-Enhanced Multimodal Detection Network for Fake News Video Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2026**. https://doi.org/10.1109/TCSVT.2026.3672169.
61. Zhang, C.; Feng, Z.; Zhang, Z.; et al. Is LLMs Hallucination Usable? LLM-Based Negative Reasoning for Fake News Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 1031–1039.
62. Liu, H.; Wang, W.; Li, H.; et al. TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 15556–15583.
63. Zheng, X.; Zeng, Z.; Wang, H.; et al. From Predictions to Analyses: Rationale-Augmented Fake News Detection with Large Vision-Language Models. In Proceedings of the ACM on Web Conference 2025, Sydney, NSW, Australia, 28 April–2 May 2025; pp. 5364–5375.
64. Wang, Y.; Ma, F.; Jin, Z.; et al. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.
65. Khattar, D.; Goud, J.S.; Gupta, M.; et al. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
66. Singhal, S.; Shah, R.R.; Chakraborty, T.; et al. SpotFake: A Multi-Modal Framework for Fake News Detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data, Singapore, 11–13 September 2019; pp. 39–47.
67. Chen, Y.; Li, D.; Zhang, P.; et al. Cross-Modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 2897–2905.
68. Wu, Y.; Zhan, P.; Zhang, Y.; et al. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 2560–2569.
69. Qian, S.; Wang, J.; Hu, J.; et al. Hierarchical Multi-Modal Contextual Attention Network for Fake News Detection. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021; pp. 153–162.

70. Singhal, S.; Pandey, T.; Mrig, S.; et al. Leveraging Intra- and Inter-Modality Relationship for Multimodal Fake News Detection. In Companion Proceedings of the Web Conference 2022, Virtual Event, 25–29 April 2022; pp. 726–734.
71. Wu, L.; Liu, P.; Zhao, Y.; et al. Human Cognition-Based Consistency Inference Networks for Multi-Modal Fake News Detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 211–225.
72. Zhou, Y.; Yang, Y.; Ying, Q.; et al. Multi-Modal Fake News Detection on Social Media via Multi-Grained Information Fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, Thessaloniki, Greece, 12–15 June 2023; pp. 343–352.
73. Lao, A.; Zhang, Q.; Shi, C.; et al. Frequency Spectrum Is More Effective for Multimodal Representation and Fusion: A Multimodal Spectrum Rumor Detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; pp. 18426–18434.
74. Zheng, X.; Luo, M.; Wang, X. Unveiling Fake News with Adversarial Arguments Generated by Multimodal Large Language Models. In Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025; pp. 7862–7869.
75. Wu, L.; Long, Y.; Gao, C.; et al. MFIR: Multimodal Fusion and Inconsistency Reasoning for Explainable Fake News Detection. *Inf. Fusion* **2023**, *100*, 101944.
76. Shao, Z.; Cai, G.; Liu, Q.; et al. An Active Learning Framework for Continuous Rapid Rumor Detection in Evolving Social Media. In Proceedings of the 2024 International Joint Conference on Neural Networks, Yokohama, Japan, 30 June–5 July 2024; pp. 1–8.
77. Suryawanshi, S.; Goswami, A.; Patil, P. FakeIDCA: Fake News Detection with Incremental Deep Learning Based Concept Drift Adaption. *Multimed. Tools Appl.* **2024**, *83*, 28579–28594.
78. Jayakody, N.; Mohammad, A.; Halgamuge, M.N. Fake News Detection Using a Decentralized Deep Learning Model and Federated Learning. In Proceedings of the IECON 2022—48th Annual Conference of the IEEE Industrial Electronics Society, Brussels, Belgium, 17–20 October 2022; pp. 1–6.
79. Ye, Y.; Feng, X.; Yuan, Z.; et al. CC-TUNING: A Cross-Lingual Connection Mechanism for Improving Joint Multilingual Supervised Fine-Tuning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 10–15 August 2025; pp. 19036–19051.
80. Han, L.; Zhang, X.; Zhou, Z.; et al. A Multifaceted Reasoning Network for Explainable Fake News Detection. *Inf. Process. Manag.* **2024**, *61*, 103822.
81. Yan, F.; Zhang, M.; Wei, B.; et al. SARD: Fake News Detection Based on CLIP Contrastive Learning and Multimodal Semantic Alignment. *J. King Saud Univ. Comput. Inf. Sci.* **2024**, *36*, 102160.
82. Jin, Y.; Wang, X.; Yang, R.; et al. Towards Fine-Grained Reasoning for Fake News Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 22 February–1 March 2022; pp. 5746–5754.
83. Wang, Z.; Pan, S.; Yang, Z. DBPR: Dynamic Bidirectional Propagation Relationship Graph Convolution Network for Fake News Detection on Social Media. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Honolulu, HI, USA, 1–4 October 2023; pp. 254–260.
84. Wei, Z.; Pan, H.; Qiao, L.; et al. Cross-Modal Knowledge Distillation in Multi-Modal Fake News Detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 4733–4737.
85. Habib, M.A.; Wadud, M.A.H.; Mridha, M.F.; et al. LLM-Powered Multimodal Reasoning for Fake News Detection. *Comput. Mater. Contin.* **2026**, *87*, 77.