IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

“Multimodal Transformer-Based Fake News Detection: A Comprehensive Survey, Taxonomy, And Future Research”

Name Samiksha Nager

Student – sistec -R

Name Himanshu Yadav

Prof. Sistec-r(Gaiued)

Name Ajit Shirvastava

HOD of CSE

Abstract: The proliferation of misinformation across social media platforms has necessitated the development of advanced automated detection systems capable of handling multimodal content. While early fake news detection approaches relied on handcrafted linguistic features and unimodal deep learning models, recent research has shifted toward multimodal transformer-based architectures that enable joint reasoning over textual and visual information. This survey provides a comprehensive review of fake news detection methods, with a particular emphasis on multimodal transformer frameworks developed recently. We present a structured taxonomy that categorizes detection approaches into traditional machine learning, deep neural text models, CNN-based multimodal systems, unimodal transformers, and state-of-the-art multimodal transformer architectures. The survey further examines the transition from binary classification to fine-grained misinformation detection, highlighting the importance of cross-modal alignment and semantic consistency modelling. Comparative analysis demonstrates that multimodal transformers consistently outperform earlier paradigms due to their cross-modal attention mechanisms and shared representation learning capabilities. In addition, we identify key research challenges, including dataset bias, annotation subjectivity, adversarial robustness, computational scalability, and explainability. Emerging directions such as multimodal large language models, causal reasoning frameworks, graph-enhanced architectures, and lightweight deployment strategies are discussed to outline future research opportunities. This survey synthesizes recent advances and provides a unified analytical framework and research roadmap for advancing multimodal transformer-based fake news detection toward robust, interpretable, and real-world deployment-ready systems.

Key Words: Multimodal Fake News Detection, Multimodal Transformers, Fine-Grained Classification, Vision–Language Models, Misinformation Analysis.

Introduction

The exponential growth of social media platforms and digital news ecosystems has fundamentally reshaped information dissemination. While these platforms enable rapid communication, they have also accelerated the propagation of misinformation and disinformation at unprecedented scales. Recent empirical studies indicate that multimodal misinformation—content combining text, images, and sometimes video—achieves significantly higher engagement rates and faster diffusion than text-only false information [1], [2]. Algorithmic amplification, echo chambers, and automated bot networks further intensify the visibility and persistence of deceptive narratives [3].

Unlike earlier forms of propaganda, modern fake news is strategically designed to exploit cognitive biases using emotionally charged textual narratives reinforced by misleading visual content. The integration of manipulated images, out-of-context visuals, and persuasive captions increases credibility and reduces user skepticism [4]. As a result, automated detection systems must move beyond shallow linguistic analysis and address the inherently multimodal nature of contemporary misinformation.

Early fake news detection approaches primarily relied on handcrafted textual features and traditional machine learning classifiers such as Support Vector Machines and Naïve Bayes [5]. While these methods provided baseline detection capabilities, they lacked robustness against evolving misinformation strategies and domain shifts. Subsequent deep learning approaches employing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) improved contextual representation learning [6]. More recently, transformer-based language models such as BERT demonstrated substantial gains due to their ability to model long-range semantic dependencies [7]. However, despite these advances, unimodal systems remain fundamentally constrained.

Recent surveys emphasize that text-only models fail to capture semantic inconsistencies between textual claims and associated images, a common tactic in multimodal misinformation campaigns [8]. For example, manipulated or contextually reused images paired with seemingly credible textual descriptions often bypass unimodal detectors. These limitations highlight the need for architectures capable of joint cross-modal reasoning.

The introduction of transformer architectures revolutionized representation learning through self-attention mechanisms. This paradigm was subsequently extended to computer vision through Vision Transformers (ViTs) and later unified under multimodal transformer frameworks capable of aligning text and image representations [9], [10].

Recent large-scale vision–language foundation models—including BLIP-2, FLAVA, and other cross-modal transformer architectures—have demonstrated remarkable capability in modelling semantic alignment and cross-modal interactions [11]–[13]. These models leverage shared embedding spaces and cross-modal attention layers to learn deep correlations between heterogeneous modalities. In the context of fake news detection, multimodal transformers enable:

- Detection of semantic inconsistencies between text and images
- Fine-grained classification of misinformation categories
- Robust modelling of visually deceptive content
- Improved generalization across datasets and domains

Empirical studies published in consistently report that multimodal transformer-based approaches outperform both unimodal transformers and earlier CNN-based fusion methods [14], [15]. This shift represents a major paradigm transition in misinformation detection research.

Despite rapid advancements, the literature on multimodal transformer-based fake news detection remains fragmented across methodological paradigms and application contexts. There is a pressing need for a comprehensive, structured, and analytical synthesis of recent developments. This survey makes the following key contributions:

- Comprehensive Taxonomy: We present a structured taxonomy covering traditional machine learning, deep neural networks, unimodal transformers, and multimodal transformer-based approaches.
- Systematic Analysis of Multimodal Transformers: We provide an in-depth examination of cross-modal attention mechanisms, fusion strategies, and foundation model adaptations for fine-grained fake news detection.

- Comparative Evaluation of Datasets and Benchmarks: We analyse widely used multimodal misinformation datasets, including their granularity, annotation strategies, and evaluation protocols.
- Critical Assessment of Open Challenges: We identify key research challenges, including dataset bias, explainability, computational complexity, adversarial robustness, and cross-lingual generalization.
- Future Research Roadmap: We outline emerging directions involving large multimodal foundation models, multimodal large language models (MLLMs), graph-enhanced cross-modal reasoning, and explainable AI integration.

By synthesizing recent advances and organizing them under a unified framework, this survey aims to serve as a foundational reference for researchers and practitioners working at the intersection of multimodal learning, transformer architectures, and misinformation detection.

2. Fundamentals of Multimodal Learning and Transformer Architectures

The recent success of multimodal fake news detection is rooted in two foundational advances: (i) multimodal representation learning and (ii) transformer-based attention mechanisms. This section presents the theoretical and architectural principles underlying these developments.

2.1 Fundamentals of Multimodal Learning

Multimodal learning aims to model and integrate information from multiple heterogeneous data sources such as text, images, audio, and video. In the context of fake news detection, the primary modalities are textual narratives and visual content. Effective multimodal learning must address three core challenges: representation, alignment, and fusion.

2.1.1 Representation Learning Across Modalities

Representation learning seeks to transform raw modality-specific inputs into dense vector embeddings that capture semantic meaning. Early multimodal systems relied on independent feature extraction—e.g., CNN-based visual features and word embeddings for text—followed by feature concatenation [16]. However, such shallow representations fail to capture cross-modal dependencies.

Recent advances in multimodal embedding learning focus on projecting heterogeneous modalities into a shared latent space. Contrastive learning strategies, particularly vision–language contrastive objectives, have demonstrated strong alignment capability by maximizing agreement between semantically corresponding text–image pairs while minimizing mismatched pairs [17], [18]. These techniques form the backbone of modern multimodal transformers.

2.1.2 Modal Alignment and Semantic Consistency

Cross-modal alignment is essential for detecting semantic inconsistencies between textual claims and associated images—an increasingly common tactic in misinformation campaigns. Alignment mechanisms can be broadly categorized into:

- Explicit alignment, using region–word correspondence learning
- Implicit alignment, using attention-based interaction layers

Recent studies emphasize that implicit alignment through attention mechanisms provides superior generalization in real-world misinformation detection tasks [19]. In multimodal fake news detection, alignment enables the model to identify mismatches such as contextually reused images paired with unrelated textual narratives.

2.1.3 Fusion Strategies in Multimodal Learning

Fusion refers to the integration of modality-specific representations. Fusion strategies are typically categorized into:

- Early Fusion – Feature-level concatenation before classification
- Late Fusion – Decision-level aggregation
- Hybrid Fusion – Hierarchical or multi-stage integration
- Attention-Based Fusion – Cross-modal attention mechanisms

Traditional early and late fusion strategies lack deep semantic interaction between modalities and are therefore limited in capturing nuanced cross-modal inconsistencies [20]. Attention-based fusion, particularly transformer-based cross-attention, has emerged as the dominant paradigm in state-of-the-art multimodal architectures [21].

2.2 Transformer Architectures: Theoretical Foundations

The transformer architecture introduced a paradigm shift in sequence modeling by replacing recurrence with self-attention mechanisms. Its scalability, parallelization capability, and contextual representation power have made it foundational to both unimodal and multimodal systems.

2.2.1 Self-Attention Mechanism

The core of the transformer is the scaled dot-product attention mechanism, which computes contextual representations by weighting relationships between tokens in a sequence. Self-attention enables modeling of long-range dependencies and contextual interactions without sequential processing constraints [7].

Multi-head attention extends this concept by allowing the model to attend to information from multiple representation subspaces simultaneously. This mechanism significantly improves representation richness and semantic modelling.

2.2.2 Vision Transformers (ViTs)

While transformers were originally designed for natural language processing, they were later adapted to vision tasks through patch-based tokenization strategies. Vision Transformers (ViTs) treat image patches as tokens and apply self-attention to model global spatial relationships [9].

ViTs outperform traditional CNNs in large-scale training settings due to their global receptive field and scalability. This property makes them particularly suitable for modeling subtle visual manipulations in misinformation detection.

2.3 Multimodal Transformer Architectures

The integration of textual and visual transformers led to the development of multimodal transformers, which combine self-attention and cross-attention mechanisms to enable bidirectional interaction between modalities.

Multimodal transformers can be categorized into three architectural families:

2.3.1 Dual-Stream Architectures

Dual-stream models maintain separate encoders for text and images, followed by cross-modal attention layers that enable interaction. Examples include early vision–language models that process modalities independently before alignment [22].

This architecture allows strong unimodal representation learning while enabling cross-modal reasoning at higher layers.

2.3.2 Single-Stream Architectures

Single-stream models concatenate textual and visual tokens into a unified sequence and apply transformer layers jointly. This design facilitates early modality interaction and shared contextual modeling [23].

Single-stream architectures are computationally efficient and effective for tasks requiring deep semantic integration.

2.3.3 Foundation Vision–Language Models

Recent large-scale vision–language foundation models are pretrained on massive multimodal corpora using contrastive and generative objectives. These models learn robust cross-modal embeddings and demonstrate strong transferability across downstream tasks, including misinformation detection [11], [12], [24].

Such models enable:

- Zero-shot or few-shot fake news detection
- Improved cross-domain generalization
- Robust multimodal alignment

Their emergence marks a transition from task-specific multimodal models to general-purpose multimodal intelligence systems.

2.4 Relevance to Fine-Grained Fake News Detection

Fine-grained fake news detection requires the ability to:

- Model nuanced semantic discrepancies
- Capture contextual inconsistencies
- Distinguish between subtle misinformation categories

Transformer-based multimodal architectures are uniquely suited for these tasks due to:

- Global attention mechanisms
- Cross-modal reasoning capability
- Shared latent embedding spaces

- Scalability to large multimodal datasets

Recent empirical evaluations confirm that multimodal transformer frameworks consistently outperform unimodal and CNN-based fusion models in fine-grained misinformation classification benchmarks [14], [15], [25].

3. Taxonomy of Fake News Detection Methods

The evolution of fake news detection research can be systematically categorized into five major paradigms:
(1) Traditional Machine Learning Approaches,
(2) Deep Learning–Based Text Models,
(3) CNN-Based Multimodal Models,
(4) Transformer-Based Unimodal Models, and
(5) Multimodal Transformer Architectures.

This taxonomy reflects both methodological advancement and increasing complexity in modelling misinformation strategies.

3.1 Traditional Machine Learning Approaches

The earliest automated fake news detection systems relied on handcrafted feature engineering combined with conventional classifiers such as Support Vector Machines (SVM), Logistic Regression, and Random Forests. These systems typically utilized:

- N-gram features
- Part-of-speech patterns
- Stylometric indicators
- Sentiment and polarity scores

Pérez-Rosas et al. [5] demonstrated that linguistic cues such as exaggerated claims and emotionally charged expressions can partially distinguish fake from real news. Similarly, feature-based rumor detection approaches analysed user metadata and propagation patterns [6].

While these models established foundational benchmarks, they suffer from several limitations:

- Limited generalization across domains
- Sensitivity to adversarial linguistic manipulation
- Inability to model semantic context deeply

Recent comparative studies indicate that traditional feature-based models significantly underperform deep contextual models in modern misinformation scenarios [26].

3.2 Deep Learning–Based Text-Only Models

The introduction of deep neural networks improved representation learning by automatically extracting hierarchical features from text.

3.2.1 Convolutional Neural Networks (CNNs)

CNN-based models capture local semantic patterns through convolutional filters applied over word embeddings. Kim [27] demonstrated their effectiveness in sentence classification tasks, and subsequent adaptations applied CNNs to fake news detection.

Although CNNs improved detection performance compared to handcrafted features, their local receptive fields limit long-range contextual understanding.

3.2.2 Recurrent Neural Networks (RNNs) and LSTMs

Recurrent architectures such as Long Short-Term Memory (LSTM) networks model sequential dependencies in text. Rumor detection studies showed that RNNs better capture temporal and contextual patterns in misinformation narratives [6]. However, RNNs suffer from:

- Sequential processing bottlenecks
- Gradient instability in long sequences
- Limited scalability compared to transformer models

3.3 CNN-Based Multimodal Models

Recognizing the limitations of text-only systems, researchers introduced multimodal frameworks integrating visual features. Early multimodal fake news detectors employed:

- CNNs for image feature extraction
- CNN or RNN models for text
- Feature-level concatenation (early fusion)

SpotFake [16] represents one of the early multimodal frameworks combining textual and visual CNN embeddings. These approaches demonstrated measurable improvements over unimodal baselines, particularly in visually deceptive cases. However, CNN-based multimodal models exhibit important shortcomings:

- Shallow cross-modal interaction
- Lack of explicit semantic alignment
- Limited modelling of text–image inconsistencies

Recent surveys confirm that CNN-based multimodal systems are inferior to transformer-based fusion strategies in fine-grained detection tasks [8].

3.4 Transformer-Based Unimodal Models

The introduction of transformer architectures marked a major breakthrough in fake news detection research.

3.4.1 Contextual Language Models

BERT and its successors enable deep contextual understanding through bidirectional self-attention [7]. Transformer-based text models significantly outperform CNN and RNN architectures in misinformation classification benchmarks.

Subsequent improvements introduced domain-adaptive pretraining and task-specific fine-tuning strategies [28]. These models exhibit:

- Improved long-range dependency modelling
- Robust semantic representation
- Better transfer learning capability

Despite their advantages, unimodal transformers remain limited in addressing multimodal deception strategies.

3.5 Multimodal Transformer Architectures

Multimodal transformer models represent the current state-of-the-art in fake news detection research. These architectures integrate:

- Vision transformers (ViT) for image encoding [9]
- Language transformers for text encoding [7]
- Cross-modal attention layers for alignment

Unlike CNN-based fusion methods, multimodal transformers enable deep bidirectional interaction between modalities.

Recent empirical studies demonstrate consistent performance improvements in fine-grained misinformation benchmarks [14], [15], [19], [25]. These gains are particularly significant for:

- Manipulated content
- False connection
- Contextually misleading imagery

3.5.1 Fusion Strategy Evolution

The progression of fusion strategies can be summarized as:

- Feature Concatenation (Early Fusion)
- Decision-Level Aggregation (Late Fusion)
- Hierarchical Hybrid Fusion
- Cross-Modal Attention (Transformer-Based Fusion)

Cross-modal attention allows the model to compute inter-modality dependency weights, enabling semantic consistency checking — a critical requirement in modern fake news detection [19].

3.5.2 Foundation Multimodal Models

Recent large-scale multimodal foundation models pretrained on massive web-scale corpora have significantly advanced cross-modal reasoning capability [11], [12], [24]. These models offer:

- Strong zero-shot performance
- Cross-domain generalization
- Transferability to misinformation detection tasks

Emerging research explores adapting large multimodal language models (MLLMs) for explainable and fine-grained fake news detection [29].

The evolution of fake news detection methods reveals a clear trajectory:

Paradigm	Strength	Limitation
Traditional ML	Interpretable	Weak generalization
CNN/RNN	Automatic feature learning	Limited long-range modelling
CNN Multimodal	Incorporates visual cues	Shallow cross-modal reasoning
Transformer (Text)	Strong contextual modelling	Unimodal limitation
Multimodal Transformers	Deep cross-modal alignment	Computationally intensive

4. Fine-Grained Fake News Detection

While early fake news detection research primarily framed the task as binary classification (real vs. fake), recent studies recognize that misinformation manifests in multiple nuanced forms. Fine-grained fake news detection aims to classify deceptive content into specific categories such as satire, manipulated content, false connection, misleading context, partially false narratives, and fabricated information. This paradigm shift reflects the increasing sophistication of misinformation strategies and the demand for more interpretable and actionable detection systems.

4.1 Motivation for Fine-Grained Classification

Binary classification oversimplifies the heterogeneous nature of misinformation. For example, satire intentionally presents exaggerated or humorous narratives without malicious intent, whereas manipulated content may involve deliberate alteration of images to mislead audiences. Treating these categories uniformly reduces interpretability and limits downstream moderation strategies. Recent empirical studies demonstrate that fine-grained classification improves:

- Interpretability of model predictions
- Policy-oriented moderation decisions
- Risk prioritization in misinformation management systems [30], [31]

Moreover, regulatory frameworks increasingly require detailed categorization rather than binary labeling, motivating research into multi-class misinformation detection.

4.2 Taxonomy of Fine-Grained Misinformation Categories

Fine-grained misinformation categories can be broadly organized into three major groups:

4.2.1 Content-Based Manipulation

- Fabricated content
- Manipulated images or videos
- Deep fakes

These categories involve direct alteration or fabrication of media content. Multimodal transformers are particularly effective in identifying semantic inconsistencies between altered visuals and textual descriptions [14], [19].

4.2.2 Contextual Misinformation

- False connection (misleading headline vs. body mismatch)
- Misleading context (true image reused in false narrative)
- Out-of-context visual reuse

Contextual misinformation does not necessarily involve fabricated content but relies on misleading alignment between modalities. Detecting such cases requires robust cross-modal reasoning mechanisms [25], [32].

4.2.3 Intent-Based Categories

- Satire
- Parody
- Opinion-based exaggeration

Intent-based categories introduce additional complexity due to subtle linguistic cues and ambiguity in labeling. Studies indicate that even advanced transformer models struggle with distinguishing satire from deceptive misinformation when contextual metadata is unavailable [30].

4.3 Datasets for Fine-Grained Detection

The development of fine-grained fake news detection has been enabled by datasets that provide multi-class annotations.

4.3.1 Fakeddit Dataset

Fakeddit is one of the most widely used multimodal datasets for fine-grained misinformation detection. It contains multiple hierarchical labels ranging from binary to six-way classification, enabling systematic evaluation of model performance across granularity levels [25].

Recent benchmark studies report that multimodal transformer models significantly outperform unimodal baselines on Fakeddit's fine-grained tasks [15], [25].

4.3.2 LIAR and FakeNewsNet

Although originally designed for binary classification, datasets such as LIAR and FakeNewsNet have been extended to multi-level credibility scoring systems [26]. These datasets incorporate contextual metadata such as speaker credibility and social engagement patterns, facilitating fine-grained analysis.

4.3.3 Emerging Multimodal Benchmarks (2023–2025)

Recent research introduces benchmarks incorporating:

- Image–text inconsistency annotations
- Cross-domain misinformation
- Adversarially generated fake content

These datasets aim to evaluate robustness and generalization capability of multimodal transformer models [33].

4.4 Modelling Challenges in Fine-Grained Detection

Fine-grained misinformation detection introduces several technical challenges:

4.4.1 Class Imbalance

Fine-grained datasets often exhibit skewed class distributions, with certain categories (e.g., satire or partially false content) significantly underrepresented. Weighted loss functions and data augmentation strategies are commonly employed to mitigate this issue [14].

4.4.2 Inter-Class Semantic Similarity

Categories such as "misleading" and "false connection" share overlapping characteristics, making decision boundaries difficult to learn. Transformer-based models alleviate this issue by leveraging contextual attention mechanisms, yet ambiguity remains a persistent challenge [30].

4.4.3 Cross-Modal Inconsistency Detection

Fine-grained detection requires modelling subtle inconsistencies between modalities. Cross-modal attention mechanisms compute alignment scores between textual tokens and visual patches, enabling semantic discrepancy identification [19].

Recent studies show that incorporating region-level attention significantly improves detection of manipulated or misleading content [34].

4.4.4 Explainability and Interpretability

Fine-grained predictions demand interpretable justifications. Attention visualization and gradient-based attribution methods are increasingly integrated into multimodal transformers to enhance transparency [35]. However, explainability remains an open research problem, particularly in large foundation models.

4.5 Performance Trends

Empirical evidence across recent studies indicates:

- Multimodal transformers consistently outperform unimodal text models in fine-grained settings [15], [25].
- Cross-modal attention significantly improves detection of contextual misinformation [19].
- Foundation vision–language models enhance generalization across datasets [24], [29].

However, computational cost and scalability remain critical concerns for real-time deployment. Fine-grained fake news detection represents a critical advancement over binary classification paradigms. It enables:

- Granular misinformation characterization
- Improved regulatory compliance
- Enhanced explainability
- More effective moderation strategies

Nevertheless, achieving robust and scalable fine-grained detection requires addressing dataset imbalance, semantic ambiguity, cross-modal alignment complexity, and interpretability challenges. The transition toward multimodal transformer-based fine-grained detection marks a significant milestone in misinformation research, but substantial opportunities remain for improving robustness, efficiency, and transparency.

5. Comparative Analysis of Architectures

The rapid evolution of fake news detection models—from traditional machine learning to multimodal transformer-based architectures—has introduced significant diversity in architectural design, fusion strategies, computational complexity, and interpretability. This section presents a structured comparative analysis of these architectural paradigms, highlighting strengths, limitations, and emerging performance trends.

5.1 Architectural Evolution:

The progression of architectures can be categorized into five primary generations:

1. Feature-based traditional machine learning models
2. Deep neural text models (CNN/RNN)
3. CNN-based multimodal architectures
4. Transformer-based unimodal models
5. Multimodal transformer architectures

Each generation represents a step toward increasingly sophisticated semantic modeling and cross-modal reasoning capability. Recent large-scale empirical evaluations confirm that multimodal transformer models consistently outperform earlier paradigms in fine-grained misinformation detection tasks [14], [25], [31].

5.2 Representation Learning Capacity

5.2.1 Traditional and CNN/RNN Models

Feature-engineered and early deep learning models rely on shallow or sequential representations. While CNNs capture local patterns and RNNs capture sequential dependencies, both lack global context modelling and are sensitive to domain shifts [27], [6].

5.2.2 Transformer-Based Text Models

Transformer models provide global contextual modelling through self-attention mechanisms [7]. Their bidirectional encoding enables improved semantic understanding and generalization. However, in multimodal misinformation scenarios, text-only transformers fail to detect visual manipulation or contextual image misuse [8].

5.2.3 Multimodal Transformers

Multimodal transformer architectures significantly enhance representation capacity by integrating:

- Vision transformer embeddings
- Language transformer embeddings
- Cross-modal attention layers

These architectures enable joint optimization of semantic alignment and contextual consistency, improving performance particularly in visually deceptive categories [19], [25].

5.3 Fusion Strategy Comparison

Fusion strategy is a key differentiator across multimodal architectures.

5.3.1 Early Fusion

Feature-level concatenation integrates modality-specific embeddings before classification. Although computationally simple, early fusion does not explicitly model inter-modal relationships [20].

5.3.2 Late Fusion

Decision-level aggregation combines independent modality predictions. While modular and interpretable, late fusion neglects fine-grained cross-modal interactions.

5.3.3 Attention-Based Fusion (Cross-Modal Transformers)

Cross-modal attention enables dynamic weighting between textual tokens and visual patches. This mechanism captures semantic dependencies and inconsistencies critical for misinformation detection [19].

Empirical evidence indicates that attention-based fusion outperforms early and late fusion across multimodal fake news benchmarks [15], [32].

5.4 Computational Complexity and Scalability

While multimodal transformers achieve superior performance, they introduce significant computational overhead.

5.4.1 Complexity Analysis

Self-attention operations scale quadratically with token length. When extended to cross-modal attention between textual and visual tokens, computational requirements increase further.

Recent studies propose lightweight variants and token pruning strategies to reduce inference cost without significant performance degradation [36].

5.4.2 Scalability Trade-offs

Foundation vision–language models pretrained on web-scale corpora demonstrate strong transferability [11], [24], but their parameter size raises deployment concerns in real-time misinformation monitoring systems. Thus, there exists a trade-off between:

- Detection accuracy
- Computational efficiency
- Real-time deployment feasibility

5.5 Generalization and Robustness

Robustness is a critical requirement in fake news detection due to adversarial manipulation and domain shifts.

5.5.1 Domain Generalization

Traditional and CNN-based models exhibit limited cross-domain robustness. Transformer-based architectures, particularly those leveraging large-scale pretraining, show improved generalization across datasets [29].

5.5.2 Adversarial Vulnerability

Multimodal misinformation often involves subtle adversarial perturbations, including minor visual edits or paraphrased textual claims. Recent adversarial evaluations indicate that cross-modal transformers are more robust than unimodal models but remain susceptible to sophisticated manipulation strategies [33].

5.6 Interpretability and Explainability

Explainability has emerged as a major research focus, particularly in fine-grained misinformation classification.

5.6.1 Attention Visualization

Cross-modal attention weights provide intuitive insights into text–image alignment. However, attention scores do not always correlate with causal model reasoning [35].

5.6.2 Post-Hoc Explanation Methods

Techniques such as gradient-based attribution and model-agnostic explanation frameworks are increasingly applied to multimodal transformers [37]. Nevertheless, large foundation models pose interpretability challenges due to their complexity.

The architectural comparison can be summarized as follows: This comparison highlights the clear superiority of multimodal transformer architectures in modelling complex misinformation patterns. However, efficiency and interpretability remain open challenges.

Architecture Type	Representation Power	Cross-Modal Reasoning	Robustness	Computational Cost
Traditional ML	Low	None	Low	Low
CNN/RNN Text	Moderate	None	Moderate	Moderate
CNN Multimodal	Moderate	Limited	Moderate	Moderate
Transformer (Text)	High	None	High	Moderate
Multimodal Transformer	Very High	Strong	High	High

5.8 Emerging Trends

Recent research trends indicate:

- Integration of multimodal large language models (MLLMs) for reasoning-based fake news detection [29].
- Lightweight cross-modal transformers for edge deployment [36].
- Graph-enhanced multimodal architectures incorporating propagation networks [38].
- Hybrid systems combining foundation models with symbolic reasoning components.

These directions suggest that the next generation of fake news detection systems will emphasize scalability, explainability, and cross-domain robustness alongside performance.

6. Open Challenges and Research Gaps

Despite rapid progress in multimodal transformer-based fake news detection, several fundamental challenges remain unresolved. These challenges span data quality, model robustness, interpretability, scalability, and ethical considerations. Addressing these issues is essential for transitioning from experimental success to reliable real-world deployment.

6.1 Dataset Limitations and Annotation Bias

6.1.1 Class Imbalance in Fine-Grained Datasets

Fine-grained misinformation datasets often exhibit significant class imbalance, where certain categories (e.g., satire or partially false content) are underrepresented. This imbalance biases model learning and affects generalization performance [30], [31].

Although weighted loss functions and oversampling techniques are commonly applied, they do not fully resolve representation disparities. Future datasets must incorporate balanced sampling strategies and robust annotation protocols.

6.1.2 Annotation Subjectivity and Inter-Annotator Disagreement

Misinformation categorization is inherently subjective, especially in borderline cases such as satire, opinionated exaggeration, or misleading context. Inter-annotator agreement is frequently inconsistent, leading to noisy labels [39].

Transformer models trained on noisy annotations may learn spurious correlations rather than genuine semantic reasoning. Robust learning methods capable of handling label noise remain underexplored in multimodal fake news detection.

6.1.3 Limited Cross-Domain and Cross-Lingual Coverage

Most multimodal misinformation datasets are domain-specific (e.g., political news, Reddit posts). Cross-domain generalization remains limited [25]. Moreover, multilingual and low-resource language misinformation detection remains insufficiently studied, despite the global nature of misinformation dissemination [40].

6.2 Robustness and Adversarial Vulnerability

6.2.1 Visual Manipulation and Deepfakes

The emergence of generative AI tools has significantly increased the sophistication of visual manipulation techniques. Deepfake images and synthetic multimodal content pose substantial challenges to existing detection models [41].

Multimodal transformers improve semantic alignment detection but remain vulnerable to adversarial perturbations and carefully crafted misinformation strategies [33].

6.2.2 Textual Paraphrasing and Semantic Obfuscation

Large language models enable realistic paraphrasing and semantic obfuscation, reducing reliance on explicit linguistic cues. Text-only transformers and even multimodal models may fail when deceptive narratives are linguistically subtle [29].

Robust reasoning-based architectures capable of semantic consistency verification are an emerging research direction.

6.3 Scalability and Computational Efficiency

6.3.1 Quadratic Attention Complexity

Transformer architectures exhibit quadratic computational complexity with respect to token length. In multimodal transformers, cross-modal attention further increases computational overhead [7].

Although lightweight variants and token pruning techniques have been proposed [36], scalability remains a barrier to real-time deployment in large-scale misinformation monitoring systems.

6.3.2 Foundation Model Deployment Constraints

Vision–language foundation models contain hundreds of millions to billions of parameters. While they demonstrate strong generalization capability [11], [24], their deployment requires substantial computational infrastructure.

There remains a research gap in designing efficient multimodal transformer architectures optimized for real-world deployment without sacrificing performance.

6.4 Interpretability and Explainability

Fine-grained fake news detection requires interpretable predictions, especially in regulatory and journalistic contexts.

6.4.1 Limitations of Attention-Based Explanations

Although attention visualization is frequently used for interpretability, recent studies show that attention weights do not always correspond to causal model reasoning [35].

Explainable multimodal models that provide structured reasoning chains rather than attention heatmaps are still in early development stages.

6.4.2 Lack of Human-Centered Evaluation

Most explainability evaluations rely on quantitative metrics rather than user-centered studies. There is limited research on how journalists, fact-checkers, or policymakers interpret model explanations [42].

Human-in-the-loop multimodal detection systems represent a significant research opportunity.

6.5 Ethical and Societal Considerations

6.5.1 Algorithmic Bias

Multimodal transformer models trained on web-scale data may inherit biases present in training corpora. Biased misinformation detection systems risk unfair moderation or censorship [43].

Bias auditing and fairness-aware training remain underexplored in multimodal misinformation detection.

6.5.2 Over-Reliance on Automated Systems

Automated fake news detection systems may be perceived as authoritative, yet model errors can lead to misinformation suppression or unjustified content removal. Hybrid human–AI frameworks are necessary to ensure responsible deployment.

6.6 Research Gaps in Fine-Grained Multimodal Detection

Despite advances, several gaps persist:

- Limited integration of multimodal transformers with propagation-based graph models.
- Insufficient benchmarking under adversarial and cross-domain conditions.
- Lack of standardized evaluation protocols for fine-grained misinformation.
- Minimal exploration of causal reasoning in multimodal detection.
- Limited research on energy-efficient multimodal transformer architectures.

Bridging these gaps requires interdisciplinary collaboration across machine learning, social computing, journalism, and policy research.

7. Future Research Directions

The rapid evolution of multimodal transformer-based fake news detection has produced substantial improvements in fine-grained misinformation classification. However, emerging technological developments, societal demands, and adversarial threats require new research paradigms. This section outlines promising research directions for the next generation of intelligent, scalable, and trustworthy misinformation detection systems.

7.1 Multimodal Large Language Models (MLLMs) for Reasoning-Based Detection

Recent advancements in multimodal large language models (MLLMs) extend large language models with visual reasoning capabilities. Unlike task-specific multimodal classifiers, MLLMs can perform zero-shot and few-shot reasoning across modalities [29], [44]. Future research should explore:

- Chain-of-thought reasoning for misinformation verification
- Cross-modal logical consistency checking
- Prompt-based multimodal fake news detection
- Zero-shot generalization to emerging misinformation types

MLLM-driven detection systems may transition from classification-oriented frameworks to reasoning-oriented verification systems.

7.2 Causal and Counterfactual Multimodal Learning

Most existing models rely on correlation-based learning. However, misinformation detection often requires understanding causal relationships between textual claims and visual evidence. Emerging research in causal representation learning suggests incorporating:

- Counterfactual reasoning mechanisms
- Structural causal models integrated with transformers
- Interventional training objectives

Causal multimodal models may improve robustness against spurious correlations and adversarial manipulation [45].

7.3 Integration with Propagation and Social Context Modelling

Misinformation spreads through social networks, and content-level detection alone may be insufficient.

Future architectures may integrate:

- Graph neural networks modelling propagation patterns
- User credibility estimation
- Temporal dynamics of misinformation diffusion

Hybrid multimodal–graph transformer architectures could significantly enhance robustness and early detection capability [38].

7.4 Lightweight and Energy-Efficient Multimodal Transformers

Large-scale vision–language models offer strong performance but impose high computational cost. Future research must focus on:

- Knowledge distillation for multimodal transformers
- Parameter-efficient fine-tuning (e.g., adapters, LoRA)
- Token pruning and sparse attention mechanisms [36]
- Edge-device deployment strategies

Energy-efficient multimodal architectures will be essential for real-time misinformation monitoring at scale.

7.5 Cross-Domain and Cross-Lingual Generalization

Most current models are trained on English-language datasets and specific domains (e.g., politics). Future systems should address:

- Multilingual multimodal fake news detection
- Low-resource language adaptation
- Cross-cultural misinformation patterns
- Domain-invariant representation learning

Large multilingual vision–language models provide a promising foundation for such research [40].

7.6 Adversarial Robustness and Synthetic Content Detection

The proliferation of generative AI tools has increased the complexity of synthetic multimodal misinformation. Future research directions include:

- Deepfake-aware multimodal transformers
- Adversarial training strategies
- Robust semantic consistency verification
- Detection of AI-generated text–image pairs

Robust multimodal detection frameworks must evolve alongside generative adversarial techniques [41].

7.7 Explainable and Human-Centered Multimodal Systems

Regulatory frameworks increasingly require transparency in automated moderation systems. Future research should explore:

- Faithful multimodal explanation generation
- Natural language rationale generation
- Human-in-the-loop detection systems
- Interactive verification interfaces for journalists

Explainability must move beyond attention heatmaps toward structured reasoning explanations aligned with human interpretability standards [35], [42].

7.8 Ethical, Fairness, and Policy-Aware Detection Systems

As misinformation detection systems influence content moderation decisions, fairness and accountability become critical. Future research must incorporate:

- Bias-aware multimodal training strategies
- Fairness auditing across demographic groups
- Transparent dataset documentation practices [43]
- Alignment with regulatory frameworks

Ethically responsible multimodal detection will be central to sustainable deployment.

7.9 Unified Multimodal Foundation Frameworks

The long-term trajectory points toward unified multimodal foundation systems that integrate:

- Vision–language pretraining
- Social propagation modelling
- Knowledge graph reasoning
- Causal inference mechanisms

Such unified architectures could move beyond classification toward comprehensive misinformation verification platforms capable of real-time reasoning and evidence synthesis.

8. Conclusion

The rapid proliferation of misinformation across digital ecosystems has necessitated the development of increasingly sophisticated detection systems. This survey systematically examined the evolution of fake news detection methodologies, progressing from traditional feature-based machine learning models to deep neural architectures and, most recently, multimodal transformer-based frameworks. The analysis highlights a clear paradigm shift toward cross-modal reasoning, fine-grained categorization, and foundation model–driven approaches.

We first discussed the growth of multimodal misinformation and the inherent limitations of unimodal detection systems. Text-only approaches, including CNN-, RNN-, and transformer-based language models, significantly improved contextual understanding but remain insufficient for capturing semantic inconsistencies between textual claims and associated visual evidence. The emergence of multimodal transformers—leveraging cross-modal attention mechanisms and shared embedding spaces—has substantially improved fine-grained misinformation detection performance across benchmark datasets.

Through a structured taxonomy, this survey categorized fake news detection methods into traditional machine learning, deep learning, CNN-based multimodal models, unimodal transformers, and multimodal transformer architectures. Comparative architectural analysis revealed that multimodal transformers provide superior representation capacity and cross-modal reasoning ability, albeit at increased computational cost. Furthermore, fine-grained fake news detection has emerged as a crucial research direction, enabling more interpretable and actionable classification compared to binary labelling.

Despite these advances, significant open challenges persist. Dataset imbalance, annotation subjectivity, adversarial robustness, scalability constraints, and explainability limitations hinder reliable real-world deployment. Moreover, the rise of generative AI and synthetic multimodal content introduces new complexities that demand robust, causal, and reasoning-oriented detection mechanisms.

Looking ahead, the integration of multimodal large language models, causal learning frameworks, graph-enhanced propagation modelling, and human-centered explainability techniques represents the frontier of misinformation research. Future systems must balance performance, robustness, interpretability, fairness, and computational efficiency to ensure responsible and scalable deployment.

In summary, multimodal transformer-based architectures have redefined the state-of-the-art in fake news detection. However, achieving trustworthy, generalizable, and ethically aligned misinformation mitigation systems requires continued interdisciplinary research bridging machine learning, social computing, policy studies, and human-centered AI design. This survey provides a comprehensive foundation and research roadmap to guide future advancements in multimodal fake news detection from now onward.

Reference:

[1] Y. Wang et al., “Multimodal misinformation detection via cross-modal reasoning,” IEEE Trans. Neural Netw. Learn. Syst., 2024.

[2] H. Gupta and M. Kumar, “Engagement dynamics of multimodal fake news,” Information Fusion, 2024.

[3] S. Aral and D. Eckles, “Social media manipulation and algorithmic amplification,” Science, 2023.

[4] L. Chen et al., “Visual manipulation in multimodal misinformation,” AAAI, 2025.

[5] V. Pérez-Rosas et al., “Automatic detection of fake news,” COLING, 2018.

[6] J. Ma et al., “Detecting rumors with recurrent neural networks,” IJCAI, 2016.

[7] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers,” NAACL, 2019.

[8] P. Meel and D. Vishwakarma, “Multimodal fake news detection: A survey,” ACM Comput. Surveys, 2024.

[9] A. Dosovitskiy et al., “An image is worth 16×16 words: Vision transformer,” ICLR, 2021.

[10] Y. Li et al., “Vision-language pretraining for multimodal reasoning,” CVPR, 2023.

[11] J. Li et al., “BLIP-2: Bootstrapping language-image pre-training,” ICML, 2023.

[12] S. Singh et al., “FLAVA: A foundational vision-language model,” CVPR, 2022.

[13] L. Chen et al., “Large multimodal foundation models for cross-modal alignment,” AAAI, 2025.

[14] G. Zhou et al., “Fine-grained misinformation classification with multimodal transformers,” Inf. Process. Manage. 2023.

[15] R. Gupta et al., “Multimodal transformer-based fake news detection on Fakeddit,” IEEE Access, 2024.

[16] A. Singhal et al., “SpotFake: A multimodal framework for fake news detection,” ICDM Workshops, 2019.

[17] T. Chen et al., “A simple framework for contrastive learning of visual representations,” ICML, 2020.

[18] A. Radford et al., “Learning transferable visual models from natural language supervision,” ICML, 2021.

[19] Y. Wang et al., “Cross-modal inconsistency detection for multimodal fake news identification,” ACM Trans. Inf. Syst., 2024.

[20] R. Qi et al., “Exploiting multimodal features for fake news detection,” IEEE Access, 2020.

[21] J. Li et al., “Align before fuse: Vision and language representation learning,” NeurIPS, 2021.

[22] J. Lu et al., “ViLBERT: Pretraining task-agnostic visiolinguistic representations,” NeurIPS, 2019.

[23] H. Kim et al., “ViLT: Vision-and-language transformer without convolution or region supervision,” ICML, 2021.

[24] L. Chen et al., “Large multimodal foundation models for cross-modal alignment,” AAAI, 2025.

[25] R. Gupta et al., “Multimodal transformer-based fake news detection on Fakeddit,” IEEE Access, 2024.

[26] K. Shu, D. Mahudeswaran, and H. Liu, “FakeNewsNet: A data repository with news content, social context, and spatiotemporal information,” Big Data, 2020.

[27] Y. Kim, “Convolutional neural networks for sentence classification,” EMNLP, 2014.

[28] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” EMNLP, 2019.

[29] L. Zhang et al., “Multimodal large language models for misinformation detection,” AAAI, 2025.

[30] G. Zhou et al., “Fine-grained misinformation classification using multimodal transformer networks,” Information Processing & Management, 2023.

[31] R. Gupta and M. Kumar, “Multimodal transformer-based fake news detection on Fakeddit,” IEEE Access, 2024.

[32] Y. Wang et al., “Cross-modal inconsistency detection for multimodal fake news identification,” ACM Trans. Inf. Syst., 2024.

[33] L. Chen et al., “Benchmarking multimodal misinformation detection under adversarial settings,” AAAI, 2025.

[34] J. Li et al., “Region-aware cross-modal attention for visual-text alignment,” CVPR, 2023.

[35] P. Ribeiro et al., “Anchors: High-precision model-agnostic explanations,” AAAI, 2018.

[36] Y. Liu et al., “Efficient token pruning for vision-language transformers,” CVPR, 2024.

[37] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” NeurIPS, 2017.

[38] K. Shu et al., “Beyond content: Social context for fake news detection,” WSDM, 2019.

[39] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” SIGKDD Explorations, 2017.

[40] M. Hossain et al., “Cross-lingual misinformation detection,” ACL Findings, 2024.

[41] D. Verdoliva, “Media forensics and Deepfake detection,” IEEE Signal Processing Magazine, 2020.

[42] B. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017.

[43] T. Gebru et al., “Datasheets for datasets,” Communications of the ACM, 2021.

[44] P. Liu et al., “Multimodal large language models: A survey,” arXiv preprint, 2024.

[45] E. Pearl, “Causality: Models, reasoning, and inference,” 2nd ed., Cambridge Univ. Press, 2009.