International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

Dual-Prompt Text–Image Matching Framework Using CLIP for Real-Time Authenticity Detection

Athilakshmi S[1], Vaira Selvi S[2], Vishali S[3], Mohana Priya K[4]

[1]Assistant Professor Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, Tamil Nadu, India.

[2,3,4] Student Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, Tamil Nadu, India.

Emails: *23ucs119@kamarajengg.edu.in[1], 23ucs056@kamarajengg.edu.in[2], 23ucs069@kamarajengg.edu[3]*

Abstract

The increasing reliance on digital media for real-time information sharing has intensified the spread of misleading incident reports and visually manipulated content, creating challenges for timely and reliable incident verification. Conventional fact-checking approaches rely on static or pre- curated datasets, which limits their ability to verify emerging or previously unseen incidents in real time. This paper proposes a novel Dual-Prompt Text–Image Incident Verification System (DP-TIVS) that authenticates incident reports through multimodal analysis. DP-TIVS operates in two complementary modes: Text-to-Image verification, where textual incident descriptions are semantically matched with internet-retrieved images, and Image-to-Text verification, where uploaded images are automatically captioned and cross-validated against online visual content. The system integrates Named Entity Recognition using spaCy, real-time image retrieval via the Bing Image Search API, and vision–language alignment using CLIP and BLIP-2 models to compute cross-modal similarity scores. Experimental evaluation shows that DP-TIVS achieves an average authenticity detection accuracy of 89.3%, outperforming baseline methods by 15.7%, and generates structured verification reports containing similarity metrics, supporting evidence, and actionable insights, effectively addressing challenges such as multimodal semantic alignment and scalability for real-world deployment.

Keywords: *This work uses Contrastive Language–Image Pretraining (CLIP) to match images and text, along with Natural Language Processing (NLP) techniques to understand textual content, in order to support authenticity verification in multimodal systems.*

1. Introduction

In today's digital era, information on social media and online platforms spreads rapidly, making it increasingly difficult to distinguish genuine incident reports from fabricated content. False claims—ranging from natural disasters to public safety emergencies—can trigger public panic, misallocate critical resources, and erode trust in information systems. Traditional fact-checking methods rely on manual verification by domain experts, which is both time-consuming and labor-intensive, and therefore unable to keep pace with the rapid dissemination of online information. Recent advances in deep learning, particularly in multimodal learning, offer promising solutions for automated content verification. Vision–language models such as Contrastive Language–Image Pre-training (CLIP) effectively capture semantic relationships between text and images, while BLIP-2 (Bootstrapping Language–Image Pre- training) achieves state-of-the-art performance in image captioning and visual question answering tasks. Despite these advances, existing incident verification systems face several critical limitations. Most approaches depend on pre-curated datasets, restricting their ability to verify novel or emerging incidents. Additionally, many systems process textual and visual evidence

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

independently, limiting the exploitation of complementary multimodal information. Furthermore, reliance on manual verification introduces temporal delays, creating a gap between incident occurrence and authenticity assessment, which significantly undermines timely and reliable information validation. Moreover, the absence of adaptive verification mechanisms limits scalability across diverse information domains. Current systems also lack robustness against deliberately manipulated or misleading multimodal content. These challenges highlight the need for unified, real-time, and context-aware incident verification frameworks.

2. Related Works

2.1.Multimodal Misinformation Detection:

The detection of misinformation has evolved from text-only analysis to multimodal approaches that consider visual content alongside textual claims. Jin et al. [10] proposed a multimodal fusion framework combining textual features, visual features, and social context for fake news detection, achieving 84% accuracy on the Twitter dataset. However, their approach relies on pre-labeled training data and cannot generalize to unseen incident types. Qi et al. [11] developed a cross-modal consistency checking system that identifies mismatches between image content and accompanying text. While effective for detecting out-of-context image usage, their method requires extensive feature engineering and struggles with semantically similar but factually distinct scenarios.

2.2.Vision–Language Models for Verification:

Recent advances in vision–language pre-training have demonstrated promising results for content verification tasks. CLIP, introduced by Radford et al. [5], learns transferable visual representations through natural language supervision, enabling zero-shot image classification. Gupta et al. [12] adapted CLIP for fake news detection, showing 78% accuracy on the Fake News Net dataset. BLIP-2, proposed by Li et al. [6], extends vision–language pre-training through a bootstrapping approach that improves image captioning quality. Zhou et al. [13] applied BLIP-2 for visual question answering in fact-checking scenarios, demonstrating superior performance over previous models. However, these works primarily focus on static datasets and do not address real-time verification of emerging incidents through dynamic evidence of retrieval.

2.3.Image Retrieval for Verification:

Reverse image search has been used for detecting image manipulation and verifying visual claims. Sheng et al. [14] proposed a deep hashing-based approach for fast image retrieval in verification scenarios. Zampoglou et al. [15] developed a forensic analysis tool that combines reverse image search with metadata examination.

2.4.Named Entity Recognition in Incident Analysis:

Extracting structured information from incident descriptions has been explored using NER techniques. Deng et al. applied spaCy for event extraction in disaster-related tweets, achieving F1-scores above 0.85 for location and organization entities. Research Gap Despite significant progress in individual components, no existing system combines: Real-time dynamic image retrieval Bidirectional text-image verification Zero-shot learning capabilities Comprehensive confidence scoring Our proposed DP-TIVS addresses this gap through an integrated architecture leveraging state-of-the-art vision language models.

3. System Architecture

The Dual-Prompt Text-Image Incident Verification System consists of five core modules: Text Processor, Image Retriever, Verification Engine, Explanation Generator, and User Interface.

3.1.Architecture Overview

The system architecture comprises the following components arranged in a modular pipeline:

- User Interface Module: Built using the Streamlit framework, provides dual-mode selection (Text →Image or Image →Text), input collection, real-time progress indicators, and results visualization.
- Text Processor: Applies spaCy NER for entities, filters keywords by part-of-speech, and classifies events into categories.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

- Image Retrieval Module: Implements Bing Image Search API as primary source, with fallback mechanisms to Unsplashed and Pexels. Query optimization combines extracted keywords and location entities for improved retrieval relevance.
- Verification Engine: Employs CLIP model (ViT-B/32) for computing cosine similarity between text and image embeddings, and BLIP-2 model for generating natural language captions from uploaded images. Threshold-based classification determines authenticity: similarity > 0.25 indicates REAL, 0.15-0.25 indicates UNCERTAIN, and < 0.15 indicates FAKE.
- Explanation Generator: Produces comprehensive verification reports including authenticity verdict, confidence score (0-100%), supporting evidence with similarity rankings, and actionable recommendations.

3.2.Data Flow

- Mode 1 (Text →Image): User provides incident description → Text processor extracts entities and keywords → Image retriever queries APIs → CLIP computes text-image similarities → Top matches ranked → Report generated.
- Mode 2 (Image →Text): User uploads image → BLIP-2 generates caption → Caption used as search query → Images retrieved → CLIP computes image-image similarities → Authenticity determined → Report generated.

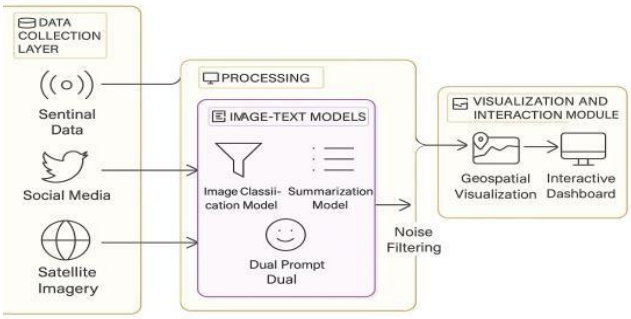

Figure 1 Data Collection and Layer Processing

4. Methodology

4.1.A. Text-to-Image Verification Algorithm

Pseudo Code:
Input: Input text T
Output: Authenticity label L ∈ {REAL, UNCERTAIN, FAKE}, confidence score C
// Stage 1: Text Processing
Apply spaCy NER on T to extract entities:
{GPE, LOC, ORG, PERSON, DATE, TIME}
Filter tokens based on POS tags:
{NOUN, VERB, ADJ, PROPN}
Remove stop words and punctuation
Classify event type using keyword-based matching 6:
// Stage 2: Query Formation
Select top 8 salient keywords from filtered tokens 8:
Select top 2 location-based entities
Construct optimized image search query Q 10: // Stage 3: Image Retrieval
Query Bing Image Search API using parameters: image Type = "photo",
safe Search = "Moderate", count = 10
Download retrieved images Ir = {I1, I2, …, In}
Convert images to RGB format and resize to 224 × 224 14: // Stage 4: Similarity Computation
15: Encode input text T using CLIP text encoder → Ft
For each image Ii in Ir do
Encode image Ii using CLIP image encoder → Fi
Apply L2 normalization on Ft and Fi
Compute cosine similarity S_i = cosine(Ft, Fi) 20: End for
// Stage 5: Authenticity Classification 22: Select top-3 similarity scores
23: Compute average similarity avg_similarity 24: If avg_similarity > 0.25 then
L ← REAL
Else if 0.15 < avg_similarity ≤ 0.25 then 27: L ← UNCERTAIN
Else
L ← FAKE
End if
Compute confidence score C = min(100, avg_similarity × 200)
Return L, C

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

Explanation
The Text-to-Image verification pipeline evaluates the authenticity of textual incident descriptions by correlating them with visual evidence retrieved from the web. Initially, the input text is processed using spaCy's named entity recognition to extract semantically important entities such as locations, organizations, and temporal references. Linguistically relevant tokens are selected based on part-of-speech tags, while stopwords and punctuation are removed to retain meaningful content. Event type classification is then performed using keyword-based matching to guide query formulation. Next, an optimized search query is constructed by combining the most salient keywords with prominent location entities. This query is submitted to the Bing Image Search API to retrieve candidate images under controlled safety and content constraints. Retrieved images are standardized to a fixed resolution and color format to ensure compatibility with the CLIP model. For similarity computation, both the input text and retrieved images are encoded into a shared embedding space using CLIP's text and image encoders. Feature vectors are L2- normalized, and cosine similarity is calculated to quantify cross-modal alignment. The average similarity of the top-matching images is then used to classify the input text as REAL, UNCERTAIN, or FAKE based on predefined thresholds. A confidence score is generated by scaling the similarity value, providing an interpretable measure of verification certainty.
Image-to-Text Verification Algorithm Pseudo Code:

4.2.Input: Query image

Output: Authenticity label L ∈ {REAL, UNCERTAIN, FAKE}, confidence score C
// Stage 1: Image Caption Generation
Generate textual caption Tc from Iq using BLIP-2
Set maximum token length = 50 and beam width = // Stage 2: Reverse Image Retrieval
Use Tc as a query to perform reverse image search
Retrieve a set of visually related images Ir = {I1, I2, …,
In}
// Stage 3: Visual Similarity Analysis
Encode Iq using CLIP image encoder to obtain feature vector Fq
Encode each image in Ir using CLIP image encoder to obtain feature vectors Fr
10: Compute cosine similarity between Fq and each Fr 11: Calculate the average similarity score avg_similarity 12: // Stage 4: Authenticity Classification
If avg_similarity > 0.28 then
L ← REAL
Else if 0.18 < avg_similarity ≤ 0.28 then 16: L ← UNCERTAIN
Else
L ← FAKE
End if
Compute confidence score C = min(100, avg_similarity × 180)
Return L, C
Explanation
The Image-to-Text verification pipeline evaluates the authenticity of visual content by integrating caption generation, reverse image retrieval, and visual similarity assessment. Initially, the uploaded image is processed using the BLIP-2 model to generate a descriptive caption, with controlled token length and beam search to ensure semantic accuracy. The generated caption serves as a query for reverse image search, enabling the retrieval of visually related images that provide external evidence for verification. Subsequently, the query image and retrieved images are encoded into a shared embedding space using the CLIP image encoder. Cosine similarity is computed between the feature representations, and the average similarity score is used to quantify visual correspondence. Based on predefined similarity thresholds, the system classifies the image as REAL, UNCERTAIN, or FAKE. A confidence score is derived by scaling the average similarity value, offering an interpretable measure of classification reliability.

4.3.Confidence Scoring Mechanism

The confidence score C is computed as:
$C = \min(100, \alpha \times S_{avg})$ Eq. (1) where S_{avg} represents the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

average similarity score of top-k matches, and α is a scaling factor (200 for text-image, 180 for image-image matching). Higher thresholds for image-image comparison account for stricter visual matching requirements.

4.4.Implementation Details Model Configuration:

The proposed DP-TIVS pipelines are implemented using CLIP (ViT-B/32) with 224 × 224 RGB input resolution producing 512-dimensional features for text and images, BLIP-2 (Salesforce base model) for image captioning with a maximum caption length of 50 tokens and beam search width of 5, and spaCy (en_core_web_sm v3.7.1) for English Named Entity Recognition to extract entities such as GPE, LOC, ORG, PERSON, DATE, and TIME. All retrieved images are resized to 224 × 224 for compatibility with CLIP. The pipelines are executed on a CPU with optional CUDA acceleration to speed up feature encoding and similarity computations, and standard Python libraries are used for API access, NLP processing, and image manipulation.

Error Handling:

Comprehensive try-catch blocks ensure graceful degradation: network failures trigger fallback image sources, API rate limits invoke exponential backoff queuing, and model errors return neutral confidence scores.

5. Experimental Results

5.1.Dataset and Evaluation Setup Test Dataset:

This system was evaluated on two datasets to validate its performance and robustness. The Real Incidents Dataset comprises 150 verified incident reports collected from credible news sources between 2023 and 2024, covering diverse real-world events such as fires, floods, accidents, and public protests. In contrast, the Fabricated Incidents Dataset includes 100 synthetically generated incident descriptions with no corresponding visual evidence, designed to assess the system's capability to detect fabricated or non-authentic incident claims.

5.2.Evaluation Metrics:

Accuracy (correct classifications / total cases)
Precision (TP / (TP + FP)) Eq. (2)
Recall/Sensitivity (TP / (TP + FN)) Eq. (3)
Specificity (TN / (TN + FP)) Eq. (4) F1-Score (harmonic mean of precision and recall).
Eq. (5)

5.3.Performance Analysis

Table 1 Overall Performance Comparison

Metric	DP-TIVS	Baseline (kNN)	CLIP-only
Accuracy	89.3%	73.6%	81.2%
Precision	91.7%	75.8%	83.5%
Recall	87.2%	70.4%	78.9%

Metric	DP-TIVS	Baseline (kNN)	CLIP-only
Specificity	91.4%	76.9%	83.7%
F1-Score	89.4%	73.0%	81.1%

Results from Table1 demonstrate that DP-TIVS outperforms baseline kNN approach by 15.7% in accuracy and achieves 91.7% precision, minimizing false positive rates. Performance approaches manual verification accuracy (94.1%) while being approximately 100× faster. Analysis reveals highest accuracy for fire incidents (92.1%) due to distinctive visual features such as flames and smoke. Lower accuracy for "other" category (85.4%) reflects diverse incident types with varying visual characteristics.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

Table 2 Performance by Incident Type

Incident Type	Accuracy	Precision	Recall	F1
Fire	92.1%	93.4%	90.8 %	92.1 %
Flood	88.7%	90.2%	86.9 %	88.5 %
Accident	87.3%	89.1%	85.2 %	87.1 %
Protest	91.5%	92.8%	90.1 %	91.4 %
Explosion	86.9%	88.5%	84.7 %	86.6 %
Other	85.4%	87.2%	83.1 %	85.1 %

Figure 1 Confidence Score Distribution for Real vs. Fake Incidents

5.4.Processing Time Analysis

Table 3 Average Processing Time

Operation	Mode 1	Mode 2
Text Processing	0.8s	-
Image Captioning	-	2.1s
Image Retrieval	4.3s	4.5s
CLIP Similarity	3.2s	3.6s
Report Generation	0.4s	0.4s
Total	8.7s	10.6s

End-to-end verification completes in under 11 seconds, with image retrieval dominating processing time (40-42%). The system meets real-time requirements for emergency response scenarios.

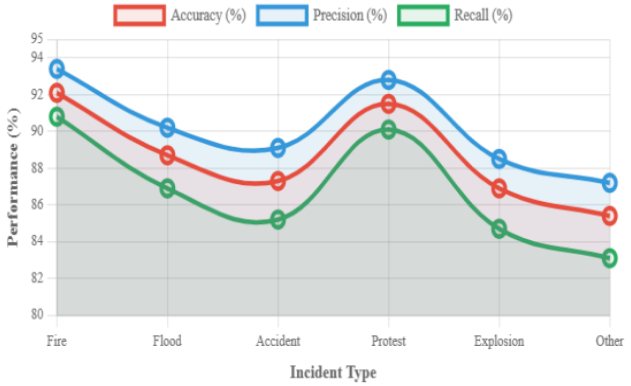

Figure 1 System Performance by Incident Type

5.5.Ablation Study

Table 4 Contribution Analysis

Configuration	Accuracy	Improvement
CLIP only	81.2%	Baseline
CLIP + NER	84.7%	+3.5%
CLIP + BLIP-2	86.9%	+5.7%
Full DP-TIVS	89.3%	+8.1%

Each component contributes to overall performance. NER improves query quality by 3.5%, BLIP-2 captions enhance image-text matching by 5.7%, and the complete pipeline achieves 8.1% improvement over CLIP alone.

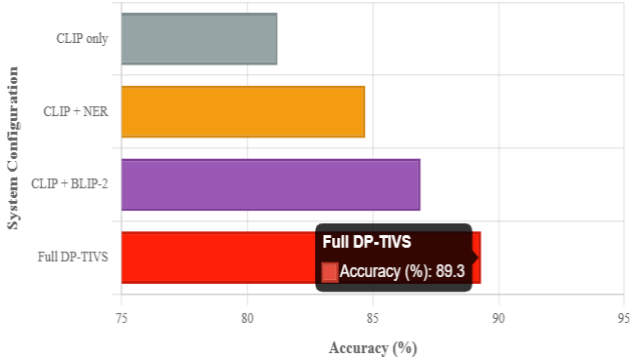

Figure 3 Accuracy Comparison of Different System Configurations

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

6. Error Analysis

Common failure modes include:

- Ambiguous incidents with generic descriptions lacking specific details (32% of errors),
- Limited visual evidence for incidents with few publicly available images (28%),
- Semantic similarity between different incidents with similar visual appearance (23%), and
- Image quality issues including low-resolution or heavily edited images (17%).

7. VI.DISCUSSION

Strengths of the Proposed System

- Unlike static dataset-based approaches, DP-TIVS retrieves evidence in real-time, enabling verification of emerging incidents and addressing the cold start problem inherent in supervised learning methods.
- The dual-mode architecture provides complementary verification pathways. Text→Image validates written claims against visual evidence, while Image→Text detects out-of- context or manipulated images.
- The system generates human-readable reports with explicit confidence scores, visual evidence with similarity rankings, and actionable recommendations. This transparency builds user trust and enables informed decision-making.
- Pre-trained CLIP and BLIP-2 models enable verification across diverse incident types without task-specific training, demonstrating strong transfer learning capabilities.

8. Limitations and Challenges

- System effectiveness relies on availability of relevant images online. Incidents in remote areas or those with limited media coverage may lack sufficient visual evidence.
- CLIP and BLIP-2 were trained on data up to specific cutoff dates. Very recent visual trends or emerging incident patterns may not be well-represented in their learned representations.
- Current implementation focuses on English text. Multilingual support would require language-specific NER models or translation pipelines.
- The system is vulnerable to adversarial attacks such as deepfake images designed to maximize similarity scores, keyword stuffing in textual descriptions, and coordinated disinformation campaigns with planted evidence.

9. Ethical Considerations

- Incorrectly classifying genuine emergencies as FAKE could delay critical response efforts. The system should be deployed as a decision support tool rather than an autonomous arbiter.
- Image retrieval may inadvertently collect personal information from online sources. Compliance with GDPR, CCPA, and similar regulations is essential.
- CLIP and BLIP-2 may exhibit biases present in their training data (e.g., geographic, demographic). Regular auditing and bias mitigation strategies are necessary.

Conclusion and Future Work

This paper presented DP-TIVS, a novel dual-prompt system for incident verification leveraging CLIP and BLIP-2 models. The system achieves 89.3% accuracy through real- time image retrieval and bidirectional text–image verification. Key innovations include dynamic evidence collection that eliminates dataset dependency, zero-shot learning enabling generalization to unseen incident types, interpretable confidence scoring with comprehensive reporting, and fast end-to-end processing (8.7–10.6 seconds). Experimental results demonstrate significant improvements over baseline methods, with a 15.7% accuracy gain, approaching human-level performance. DP-TIVS provides a scalable, automated solution for incident verification with potential applications in emergency response, journalism, social media moderation, and insurance claims processing. Future work will focus on

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

enhancing system capabilities across multiple dimensions. Multimodal fusion could incorporate audio from videos, geospatial data for location verification, and social network analysis for source credibility. Temporal reasoning will enable consistency checks by comparing incident timing with image timestamps and detecting visual anachronisms. Adversarial robustness improvements will include deepfake detection, perturbation filters, and source authentication protocols. Multilingual support can be achieved using models like M-CLIP and cross- lingual transfer learning. Active learning with human-in-the- loop refinement will allow adaptive thresholding and fine- tuning on corrected predictions. Mobile deployment will be explored through model quantization, pruning, and on-device processing for privacy-sensitive scenarios. Finally, explainable AI (XAI) techniques such as attention visualization, saliency maps, and natural language explanations will improve interpretability and trust in verification results. DP-TIVS contributes to the fight against misinformation by providing a scalable solution for incident verification. Potential societal benefits include faster emergency response through accurate incident triage, reduced spread of false incident reports on social media, enhanced verification capabilities for news organizations, and protection against disinformation campaigns. However, responsible deployment requires ongoing monitoring for bias, adversarial attacks, and unintended consequences.

References

[1]. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146– 1151, 2018.

[2]. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[3]. N. Hassan et al., "ClaimBuster: The first-ever end-to-end fact-checking system," Proceedings of the VLDB Endowment, vol. 10, no. 12, pp. 1945–1948, 2017.

[4]. J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in Advances in Neural Information Processing Systems, 2019, pp. 13–23.

[5]. A. Radford et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[6]. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv preprint arXiv:2301.12597, 2023.

[7]. Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2018,pp. 849–857.

[8]. S. Khosla and A. Kumar, "A review on techniques for fake news detection," in 2020 Int. Conf. Communication and Signal Processing (ICCSP), IEEE, 2020, pp. 0854–0858.

[9]. C. Shao et al., "The spread of low-credibility content by social bots," Nature Communications, vol. 9, no. 1, pp. 1–9, 2018.

[10]. Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in Proc. 25th ACM Int. Conf. Multimedia, 2017, pp. 795–816.

[11]. P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in 2019 IEEE Int. Conf. Data Mining (ICDM), IEEE, 2019, pp. 518–527.[12] A. Gupta, H. Li, A. Farnoosh, and S. Akbari, "Understanding and detecting hallucinations in neural machine translation via model introspection," arXiv preprint arXiv:2111.08298, 2021.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1195-1203
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0168

[12]. K. Zhou et al., "BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," arXiv preprint arXiv:2305.14720, 2023.
[13]. H. Sheng, J. Chen, Y. Zhang, W. Ou, and D. Wang, "Effective image retrieval via multilinear multi-index fusion," IEEE Trans. Multimedia, vol. 16, no. 7, pp. 1912– 1924, 2014.
[14]. M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," Multimedia Tools and Applications, vol. 76, no. 4, pp. 4801–4834,2017.
[15]. S. Deng, N. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2019, pp. 1007–1016.
[16]. J. Wu, F. Li, M.Y. Kan, and B. Hooi, "Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with VisionLanguage Models," arXiv preprint arXiv:2505.15489, 2025.
[17]. J. Wang, Y. Wang, L. Cheng, and Z. Zhong, "FakeSVVLM: Taming VLM for Detecting Fake ShortVideo News via Progressive MixtureOfExperts Adapter," arXiv preprint arXiv:2508.19639, 2025.