applied sciences

MDPI

Article

Lightweight Multimodal Frameworks for Fake News Detection Using Cross-Modal Consistency Modeling and Parameter-Efficient Attention

Sarab AlMuhaideb *, Hajar Almeleehan, Jana Alamer, Sarah Aljuhani, Dimah Alotaibi and Athoog Alsuhaibany

Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; 444200728@student.ksu.edu.sa (H.A.); 444200767@student.ksu.edu.sa (J.A.); 444200927@student.ksu.edu.sa (S.A.); 444201161@student.ksu.edu.sa (D.A.); 444202989@student.ksu.edu.sa (A.A.)
* Correspondence: salmuhaideb@ksu.edu.sa

Abstract

The rapid proliferation of fake news poses significant societal and psychological risks, driving an urgent need for automated detection systems. While conventional unimodal methods fail to capture complex cross-modal relationships, existing multimodal frameworks are often computationally prohibitive for real-world deployment. To address this limitation, this paper proposes two lightweight, high-performance multimodal frameworks for fake news detection: Cross-Modal Consistency Modeling (CMCM) and Parameter-Efficient Cross-Modal Attention (PE-CMA). The CMCM architecture utilizes a similarity branch to explicitly model semantic alignment and misalignment between text–image pairs, effectively identifying mismatched content characteristic of fake news. Conversely, the PE-CMA architecture introduces a lightweight cross-modal attention mechanism combined with a learned token pooling technique. This compresses token representations into a highly informative subset prior to attention application, significantly lowering computational complexity while preserving cross-modal interactions. Both frameworks leverage MobileNetV3-Large and TinyBERT as highly efficient image and text encoders, respectively. Extensive evaluations on the Fakeddit and Weibo datasets demonstrate the superiority of the proposed methods. On Fakeddit, CMCM and PE-CMA achieve classification accuracies of 92.29% and 91.17%, respectively. On Weibo, PE-CMA and CMCM deliver accuracies of 91.81% and 91.06%, respectively. Ultimately, these architectures outperform several state-of-the-art baselines, offering an optimized balance between classification accuracy and computational efficiency.

Keywords: fake news detection; multimodal frameworks; cross-modal attention

Check for updates

Academic Editor: João M. F. Rodrigues

Received: 13 September 2026
Revised: 4 October 2026
Accepted: 26 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The exponential growth of digital communication platforms has revolutionized information dissemination, enabling news to propagate globally with unprecedented velocity. This rapid spread has significantly amplified the societal impact of misinformation, posing severe threats to public and social trust [1]. Empirical research indicates that fake news diffuses approximately six times faster than verified information and exhibits a 70% higher probability of being retweeted [2]. This profound disparity underscores the critical real-world impact of digital misinformation and highlights the urgent necessity for robust intervention mechanisms.

Manual fact-checking is inherently unscalable and cannot keep pace with the massive volume of real-time online content. Consequently, automated detection frameworks are essential to provide the computational speed and consistency required for real-time data monitoring, thereby alerting users and platforms prior to widespread viral dissemination. Modern misinformation is intrinsically heterogeneous, integrating diverse data modalities such as text, images, video, and audio. Effective automated systems must therefore transcend single-modality analysis to model these diverse data types within an integrated computational framework.

Early automated detection paradigms primarily relied on unimodal architectures [3–5]. Text-based approaches extract contextual and semantic features from written content to isolate linguistic patterns indicative of deception. While effective for semantic analysis, these models fail to evaluate visual components, such as manipulated images, which are often deployed to add emotional weight and reinforce misleading claims. Conversely, image-based unimodal systems [6,7] focus strictly on visual content to detect structural manipulations or anomalies. However, their efficacy is bounded because they neglect the accompanying textual context, which frequently contains critical semantic cues necessary for veracity verification.

To address these limitations, multimodal fake news detection frameworks combine textual and visual modalities to enable a holistic analysis of news content. Rather than evaluating features in isolation, multimodal systems exploit the cross-modal dependencies that deceptive content often relies upon to appear credible. By jointly processing text and images, these architectures can identify semantic inconsistencies and cross-modal misalignments that are undetectable when modalities are analyzed independently, thereby significantly enhancing classification accuracy.

Numerous studies [8–10] have validated the superiority of multimodal architectures, demonstrating that deep learning frameworks leveraging cross-modal fusion consistently outperform unimodal baselines. Despite these advancements, a critical research gap persists. The majority of state-of-the-art multimodal detection models [11,12] maximize classification accuracy by employing deep, highly parameterized architectures that are computationally prohibitive for real-world deployment. This limitation underscores an urgent need for lightweight, resource-efficient multimodal frameworks.

This work stems from a critical empirical observation in digital disinformation: fake news is fundamentally characterized by semantic dissonance or contextual mismatch between the textual claim and its accompanying visual asset [13,14], rather than the intrinsic complexity of either modality alone. Recognizing that cross-modal inconsistency carries more diagnostic weight than sheer scale, our proposed architectures deliberately shift the parameter budget away from oversized unimodal backbones. Instead, we focus architectural capacity entirely on capturing cross-modal alignment and interactions, ensuring that classification power emerges directly from modeling cross-modal consistency rather than expanding parameter count.

Lightweight architectures optimize computational efficiency by minimizing parameter counts and floating-point operations while maintaining high representation capacity. Examples include MobileNet [15] and GhostNet [16] for visual feature extraction, alongside TinyBERT [17] and MiniLM [18] for natural language processing. To preserve this efficiency during data integration, lightweight cross-modal fusion mechanisms are required to capture complex inter-modality interactions without introducing significant computational overhead. Additionally, incorporating an explicit similarity measurement branch is essential to model semantic alignment or misalignment between modalities, allowing the system to mathematically quantify cross-modal consistency. To address these challenges, this work focuses on designing lightweight multimodal architectures, specifically

the Cross-Modal Consistency Modeling (CMCM) and Parameter-Efficient Cross-Modal Attention (PE-CMA) frameworks, that integrate efficient attention fusion mechanisms with dedicated similarity branches to effectively capture cross-modal inconsistencies without compromising edge-device viability.

The primary contributions of this research are summarized as follows:

- We propose two novel, lightweight multimodal frameworks: Cross-Modal Consistency Modeling (CMCM) and Parameter-Efficient Cross-Modal Attention (PE-CMA), which intentionally shift computational budgets from oversized unimodal backbones to highly optimized cross-modal interactions using MobileNetV3-Large and TinyBERT as shared encoders.
- We design a dedicated parallel similarity measurement branch within the CMCM framework that leverages auxiliary alignment losses to mathematically quantify semantic dissonance, enforcing strict structural alignment and effectively isolating the cross-modal mismatches inherent in digital disinformation.
- We introduce a low-overhead attention mechanism in the PE-CMA framework that bypasses the quadratic complexity of standard cross-attention by integrating learned token pooling with low-rank query factorization, preserving fine-grained token-level mapping at a fraction of the computational cost.
- We provide a systematic architectural synthesis evaluating multiple projection-head alignment strategies and fusion topologies (including concatenation, Siamese, and Hadamard styles).

While certain isolated components have been individually studied, this work provides a unique architectural synthesis optimized for multimodal fake news detection. This includes evaluating distinct fusion topologies (Concatenation and Siamese Fusion) alongside projection-head alignment to map disparate modalities into a unified dimensional space. Comprehensive evaluations on the benchmark Fakeddit [19] and Weibo [20] datasets demonstrate state-of-the-art efficiency. Specifically, CMCM-Siamese achieves 92.29% accuracy on Fakeddit with only 19.29 M parameters, while PE-CMA attains 91.81% accuracy on Weibo with just 15.33 M parameters, as validated through detailed ablation, frozen-encoder probing, and comparative studies.

The remainder of this paper is structured as follows. Section 2 reviews the existing literature on multimodal fake news detection. Section 3 details our proposed architectures, and Section 4 outlines the experimental setup. We present our findings in Section 5 and analyze them in Section 6. Finally, Section 7 concludes the paper and suggests directions for future work.

2. Related Work

This section reviews the foundational literature in multimodal misinformation analysis. Early detection frameworks relied predominantly on unimodal paradigms, employing statistical, recurrent, and transformer backbones for textual verification [3–5,21–23] or forensic and convolutional feature extractors for visual manipulation detection [6,7]. However, as highlighted in recent taxonomies categorizing detection into small-model paradigms versus large language model-involved frameworks [24], single-channel representations cannot capture inter-modality discrepancies, making cross-modal fusion essential while balancing representation capacity with deployment feasibility. Accordingly, we focus on multi-source modeling across traditional and attention-based fusion strategies, examine cross-modal consistency paradigms via Siamese networks, and delineate the computational limitations addressed by our proposed architectures.

2.1. Multimodal Fake News Detection

To overcome the limitations of single-source analysis, traditional multimodal frameworks merge textual and visual features using feature-level concatenation [8,11,12,25–31] or decision-level ensembling [9,32–34].

In early feature-level approaches, networks combined convolutional or transformer text embeddings with VGG-19 visual vectors [35], using structures such as event adversarial game constraints to isolate event-independent factors across Twitter and Weibo datasets [25], or mapping aligned text–image matrices into a unified classification layer [26]. Generative architectures have also utilized feature concatenation within a shared latent layer of a multimodal variational autoencoder to map foundational cross-modal correlations during training [27]. When benchmarked on the Fakeddit dataset, concatenated dual-branch convolutional neural networks consistently outperformed standalone unimodal baselines [8]. Beyond simple feature-level concatenation, alternative aggregation strategies have been explored to manage modal dependencies; these include hybrid fusion approaches that unite early and late fusion paths through meta-learning classifiers [36], and architectures implementing both concatenation and maximum fusion logic alongside supplemental metadata features like image polarity and popularity [9]. Stacked ensemble paradigms integrating specialized language configurations, such as BERT [37] combined with ELECTRA [38] and NasNet [39] architectures, have succeeded in optimizing accuracy while restricting overall computational complexity across standard benchmarks [32]. Multi-view frameworks like bootstrapping multi-view representations have been introduced to deconstruct inputs by isolating textual narratives from discrete low-level visual patterns and higher-level image semantics [33].

To capture deeper semantic interdependencies and cross-modal correlations, contemporary architectures employ attention-based fusion mechanisms that dynamically weigh and integrate heterogeneous data streams. Transformer-based self-attention layers have been used to project image embeddings directly into textual token spaces, enabling joint multimodal interaction and surpassing traditional concatenation or late-fusion benchmarks [40]. Rather than relying on simple feature aggregation, advanced architectures implement co-attention layers to learn multi-domain dependencies across textual, spatial–visual, and frequency-domain representations [41], or integrate attention mechanics with bidirectional recurrent networks alongside image tampering algorithms to align multimodal consistency across text–image pairs [13]. Furthermore, cross-modal attention has been leveraged in combination with contrastive learning and optimal transport to minimize representation gaps and align feature distributions [14]. To resolve cross-source differences, multilevel frameworks deploy sequential cross-attention stages that map visual-semantic alignment before filtering redundant features with self-attention layers [42]. Attention mechanisms also allow models to dynamically adapt to data ambiguity; for instance, measuring text–image ambiguity via divergence metrics enables frameworks to selectively shift focus between distinct unimodal features and global cross-modal correlations based on content consistency [43]. Attention-driven fusion has also been embedded within multi-stage frameworks that mimic cognitive workflows, incorporating domain-specific terms, structural keyword generation, and summarization pipelines to simulate human thinking for fraud classification [44].

Recent multimodal fake news detection frameworks have predominantly prioritized predictive effectiveness by deploying high-capacity foundation models, generative pipelines, and complex ensembles. For instance, approaches aligning semantic and entity representations via multi-stage BERT and co-attention modules [45], or stacking vision-and-language transformers such as RoBERTa [46] and vision transformers (ViT) [47] through Variational Autoencoder bottlenecks and gated attention [48], achieve superior classification accuracy

at the expense of substantial computational and memory overhead. Similarly, incorporating auxiliary generative models, such as text-conditioned generative adversarial networks (GANs) [49] to synthesize images for visual consistency checks [50] or adversarial multimodal GAN architectures paired with legacy CNN-transformer ensembles [51], introduces severe latency bottlenecks and training instability. Even parameter-efficient adaptation and prompt-tuning strategies, including dual-stream shared prompts over ViT and XLM-RoBERTa [52] backbones [53] and interactive text-aware CLIP prompting [54], remain computationally tethered to hundred-million-parameter foundational encoders. Furthermore, deep multi-model ensembles stacking disparate architectures like ViLBERT [55], Transformer-XL [56], XLNet [57], and T5 [58] metaclassifiers [59] induce extreme parameter bloat that limits throughput.

2.2. Lightweight Detection Frameworks

To facilitate deployment on resource-constrained devices, contemporary lightweight frameworks minimize structural and computational complexity across deepfake detection [60,61], computer vision forensics [62], and multimodal alignment tasks [63–66]. In deepfake forensics, architectural footprints are constrained either by replacing heavy networks with concatenated handcrafted texture and keypoint descriptors evaluated via traditional classifiers [60], or by incorporating pyramid atrous aggregation and shuffle fusion mixers within streamlined convolutional neural networks (CNNs) to optimize parameter counts and FLOPs [61]. Similarly, unimodal frameworks achieve task-specific efficiency by substituting conventional convolutions with depthwise-separable layers to analyze localized motion residue windowing [62], or by leveraging asymmetric group convolution branches coupled with multi-scale coordinate attention to isolate complex plant disease patterns without severe computational overhead [67]. Beyond single-source inputs, heterogeneous multimodal systems balance high-dimensional streams through parameter-efficient encoding and streamlined cross-modal fusion. This includes pairing MobileNet backbones with partially frozen text layers for emotion recognition [63], replacing massive vision-language transformer layers with dictionary lookup transformations [64], using weight-sharing multi-branch vision transformers for remote sensing segmentation [65], and integrating specialized foundation models with low-rank adaptation layers for medical visual question answering [66]. Collectively, these methodologies demonstrate that partial network freezing, sub-network weight sharing, and specialized layer design can drastically reduce parameters with minimal loss in classification accuracy.

2.3. Applications of Siamese Networks

Siamese architectures serve as an effective paradigm for verification, matching, and tracking tasks by processing input pairs through identical weight-sharing subnetworks to map comparable latent embeddings [68]. Conventionally deployed for unimodal tasks like few-shot image classification and visual object tracking, these networks evaluate multi-level features using distance metrics or cross-correlation similarity spaces to locate targets across sequential frames [69–71]. To bridge cross-modal dependencies, contemporary frameworks adapt Siamese-inspired architectures to multimodal contexts to evaluate semantic consistency and alignment. In vision-language tracking, cross-modal bilinear fusion and dual co-attention are integrated into Siamese pipelines to capture higher-order relationships for spatial localization [72]. Siamese fusion strategies are uniquely suited for detecting semantic agreement and structural contradictions between heterogeneous streams; combining visual and textual features via concatenation, absolute differences, and Hadamard products enables models to accurately classify content alignment in complex multimodal

media [73]. Although Siamese networks have seen extensive utility across these similarity-based domains, their application remains unexplored within fake news detection.

2.4. Discussion of Related Work

The literature highlights a definitive paradigm shift from unimodal architectures toward transformer-based multimodal systems to combat complex digital misinformation. While single-source models isolate distinct semantic patterns in text [3,22,23] or structural tampering in images [6,7], they fail to capture cross-modal dynamics in real-world contexts where unmanipulated images are weaponized through deceptive text alignment, such as the "False Connection" anomalies categorized in standard benchmarks like Fakeddit [19] and Weibo [10,14,74]. To bridge this gap, classical frameworks employ feature-level concatenation [8,10–12,30], mathematical variants [9,34], or decision-level ensembles [32,34], while advanced frameworks leverage attention mechanisms to explicitly align text tokens with visual features [14,42,44,74].

While these advanced frameworks successfully model text–image consistency, they inherently couple the alignment mechanisms directly into the deep feature fusion process. This formulation introduces severe computational bottlenecks, as aligning multimodal semantics relies on parameter-heavy cross-attention transformers [40,42], bidirectional recurrent networks [13], or complex optimal transport metrics [14] to iteratively evaluate full, uncompressed token sequences. Consequently, these traditional approaches suffer from quadratic computational complexity and severe parameter bloat, limiting their viability for resource-constrained deployments. In direct contrast to these heavily coupled formulations, our proposed architectures explicitly mitigate these bottlenecks through two distinct lightweight strategies. First, the CMCM architecture decouples consistency measurement from the fusion stage by implementing a parallel, lightweight similarity branch on shallow global embeddings, enforcing cross-modal consistency without the massive parameter overhead of traditional co-attention models. Second, to directly address the quadratic complexity of sequence alignment, the PE-CMA architecture introduces learned token pooling prior to cross-modal interaction. PE-CMA structurally compresses the sequence length of extracted features, drastically reducing GFLOPs and memory footprint without sacrificing the dynamic, fine-grained feature-weighing capabilities of traditional co-attention.

Concurrently, there is an evident architectural evolution toward substituting traditional CNNs with parameter-efficient transformers, such as utilizing BERT for text alongside ViT or parameter-optimized adaptations like LoRA to reduce inference overhead [11,42]. More recently, Siamese-style fusion configurations like SimCLIP have emerged to identify semantic contradictions by passing cross-modal inputs through operations such as concatenation, absolute differences, and Hadamard products [72,73]. However, a critical limitation persists: these Siamese approaches rely fundamentally on a shared latent embedding space predefined by joint models like CLIP [75]. When text and visual inputs are encoded via separate, unaligned network types, simple mathematical operations fail to expose underlying discordances. Consequently, a pronounced research gap persists: the existing literature overwhelmingly optimizes detection metrics by compounding structural complexity, frequently layering multi-stage foundation models, generative verification pipelines, and cumbersome ensembles, leaving a critical need for lightweight, low-latency architectures that achieve state-of-the-art accuracy through parameter-efficient cross-modal attention and contrastive alignment without relying on unwieldy foundation backbones.

To systematically address the limitations identified in prior work and establish a baseline for our framework and its subsequent conclusions, this study formalizes its objectives through the following three core research questions:

- RQ1: To what extent can a computationally streamlined, lightweight architecture effectively mitigate parameter overhead in multimodal fake news detection while maintaining competitive classification performance?
- RQ2: Does the explicit modeling of cross-modal discrepancies via a dedicated similarity evaluation branch enhance the detection of deceptive news, specifically in resolving contextually unrelated text–image pairs?
- RQ3: Can the synergistic integration of low-rank weight decomposition and learned token pooling minimize the computational complexity of cross-modal attention mechanisms without degrading their contextual representation capacity?

3. Proposed Multimodal Architectures

To address the computational and modeling limitations of existing multimodal fake news detection systems, this section introduces two lightweight architectures: the Cross-Modal Consistency Modeling (CMCM) framework and the Parameter-Efficient Cross-Modal Attention (PE-CMA) framework. Both models shift from conventional unimodal processing to an integrated multimodal pipeline, relying on optimized feature extractors. We first define the shared modality encoders, followed by a detailed mathematical formulation of the integration strategies and similarity measurements unique to CMCM and PE-CMA.

3.1. Modality Encoders

3.1.1. MobileNetV3-Large Visual Feature Extractor

To balance classification accuracy with edge-device computational efficiency, we employ MobileNetV3-Large [76], an architecture optimized via platform-aware Neural Architecture Search (NAS) [77] and NetAdapt [78]. The network is built upon inverted residual bottleneck blocks integrated with Squeeze-and-Excitation (SE) modules [79] and computationally efficient activation functions: hard-swish (h-swish(x), Equation (1)) and hard-sigmoid (h-sigmoid(x), Equation (2)) .

$$\text{h-swish}(x) = x \cdot \frac{\text{ReLU6}(x+3)}{6} \tag{1}$$

$$\text{h-sigmoid}(x) = \frac{\text{ReLU6}(x+3)}{6} \tag{2}$$

Processing a 224 × 224 × 3 RGB input, the network hierarchically extracts features through a sequence of early, mid, and deep depthwise convolutional bottlenecks. Instead of an exhaustive layer-by-layer breakdown, the progression can be summarized as a systematic reduction in spatial resolution coupled with channel expansion, shifting from standard ReLU to hard-swish activations in deeper layers to maximize representational capacity without inflating latency. Ultimately, a 1 × 1 convolution expands the final spatial features to 960 channels, followed by 7 × 7 global average pooling. This yields a compact semantic feature representation ready for cross-modal integration.

3.1.2. TinyBERT Textual Feature Extractor

To achieve high textual encoding efficiency, we utilize TinyBERT [17], a compact Transformer variant distilled from BERT that drastically reduces layer count and hidden dimensions while preserving performance. The network processes an input sequence of length L through an embedding layer—combining token, positional, and segment embeddings—yielding an $L \times 312$ feature matrix.

This matrix is iteratively refined across four identical Transformer encoder blocks, each comprising a multi-head self-attention (MHSA) mechanism and a feed-forward network

(FFN). The MHSA sublayer computes scaled dot-product attention across 12 parallel heads ($d_k = 26$) according to Equation (3):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V represent the query, key, and value matrices.

Following attention computation, the representations are normalized and passed into the FFN. Rather than detailing the standard linear expansions and compressions typical of Transformer blocks, we highlight the use of the Gaussian Error Linear Unit (GELU) activation function [80] within the FFN, approximated in Equation (4):

$$\text{GELU}(x) \approx 0.5x\left(1 + \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right]\right) \tag{4}$$

Standard residual connections and layer normalization are applied after both the MHSA and FFN sublayers to stabilize learning, ultimately producing contextualized token embeddings ready for cross-modal integration.

3.2. Cross-Modal Consistency Modeling (CMCM) Architecture

The design of the CMCM framework is motivated by the need for explicit, joint optimization of semantic fusion and structural alignment. While standard fusion strategies combine features blindly for a final classification head, they often fail to force the network to explicitly recognize whether an image–text pair belongs together. To resolve this, CMCM introduces a dedicated Similarity Measurement Branch running parallel to the fusion layer. By enforcing an auxiliary alignment loss (via Cosine Similarity or Symmetric Contrastive Loss), the network is explicitly penalized when processing mismatched multimodal content. This structural constraint guides the projection heads to map highly consistent pairs close together in the shared latent space, providing the subsequent classification layer with highly discriminative, consistency-aware feature representations without adding heavy transformer layers.

The proposed CMCM framework utilizes MobileNetV3 [76] and TinyBERT [17] as lightweight image and text encoders, balancing computational efficiency with accuracy. An overview of the proposed framework is provided in Figure 1. Raw text and image inputs are processed through modality-specific encoders to extract feature representations, which are subsequently mapped to a shared dimension via projection heads. During training, feature alignment is evaluated using a similarity function, while a fusion strategy combines the projected vectors to feed a final classification layer.

3.2.1. Modality Embeddings and Projection Heads

Prior to multimodal fusion, feature representations are extracted directly from the subnetwork backbones rather than their task-specific classification heads. Visual Embeddings are extracted from MobileNetV3 following global average pooling (with the final expansion layer and classification head omitted), yielding a compact 960-dimensional semantic image vector. Textual Embeddings are extracted from the final TinyBERT encoder layer by selecting the 312-dimensional [CLS] token embedding, which serves as the holistic context representation for the input text sequence.

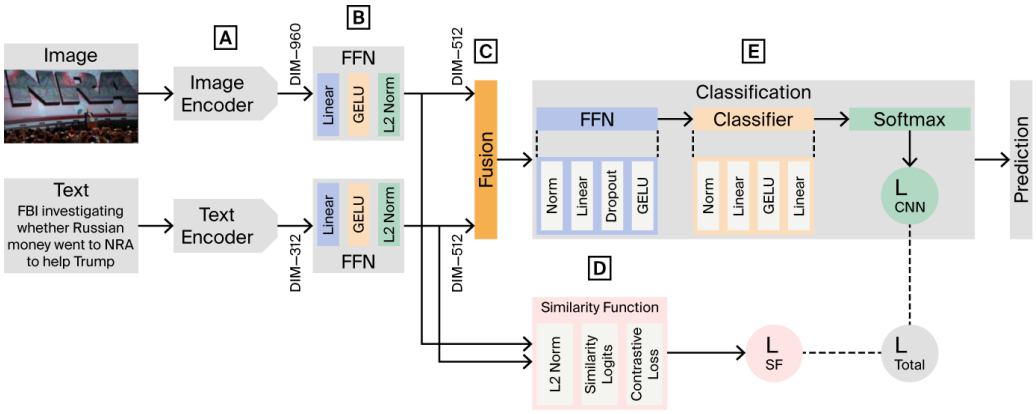


Figure 1. Overall architecture of the proposed Cross-Modal Consistency Modeling (CMCM) Architecture. (**A**) Image and text encoders. (**B**) Projection head. (**C**) Fusion module. (**D**) Similarity function. (**E**) Classification layer.

To align feature dimensions across modalities prior to fusion, independent shallow projection heads map the extracted embeddings into a shared 512-dimensional vector space [73,81]. Each projection head consists of a single linear layer, GELU non-linear activation (4), and L_2 normalization (5) [82]:

$$\hat{\mathbf{e}} = \frac{\mathbf{e}}{\sqrt{\sum_{i=1}^{d} e_i^2 + \epsilon}} \tag{5}$$

where $\mathbf{e}$ denotes the input vector and ϵ represents a small stability constant. This transforms the 312-dimensional TinyBERT embedding and 960-dimensional MobileNetV3 embedding into normalized, dimensionally compatible representations $E_{\text{text}} \in \mathbb{R}^{512}$ and $E_{\text{image}} \in \mathbb{R}^{512}$, respectively.

3.2.2. Fusion Strategies

To systematically evaluate the impact of multimodal integration on representation quality, we explore four feature-level fusion strategies [83]:

1. Concatenation Fusion: Preserves modality-specific characteristics by concatenating embeddings directly along the feature dimension [8–12], producing a 1024-dimensional joint vector E_{fused}:
$$E_{\text{fused}} = \text{Concat}(E_{\text{text}}, E_{\text{image}}) \tag{6}$$
2. Element-wise Maximum Fusion: Selects the maximum feature activation across aligned dimensions to highlight dominant cross-modal signals [9], yielding a 512-dimensional vector:
$$E_{\text{fused}}^{(i)} = \max\left(E_{\text{text}}^{(i)}, E_{\text{image}}^{(i)}\right) \tag{7}$$
3. Siamese-style Fusion: Following [73], combines joint representations, absolute differences, and element-wise products into a 2048-dimensional vector to capture semantic alignment and divergence:
$$E_{\text{fused}} = \left[\text{Concat}(E_{\text{text}}, E_{\text{image}}), |E_{\text{text}} - E_{\text{image}}|, E_{\text{text}} \odot E_{\text{image}}\right] \tag{8}$$
4. Hadamard Product with Element-wise Summation: Integrates unimodal features with cross-modal interactions by adding the Hadamard interaction term E_{interact} (9) to the original embeddings, yielding a 512-dimensional vector (10):
$$E_{\text{interact}} = E_{\text{text}} \odot E_{\text{image}} \tag{9}$$
$$E_{\text{fused}} = E_{\text{text}} + E_{\text{image}} + E_{\text{interact}} \tag{10}$$

3.2.3. Similarity Measurement Branch

To provide explicit cross-modal consistency cues for fake news detection, we incorporate a similarity measurement branch operating on the projected embeddings $\mathbf{e}_t$ and $\mathbf{e}_v$ [13,14]. Because cross-modal inconsistency frequently signals unverified or manipulated content, this branch calculates an alignment loss $\mathcal{L}_s$ based on matched versus mismatched image–text pairs. We evaluate two similarity-based loss functions:

1. Cosine Similarity Loss: Computes the normalized cosine similarity score $s \in [-1, 1]$ between projected text and image embeddings [13]:

$$s = \frac{\mathbf{e}_t \cdot \mathbf{e}_v}{\|\mathbf{e}_t\|\|\mathbf{e}_v\|} \tag{11}$$

The score s is converted into an alignment probability p_s via the sigmoid function:

$$p_s = \frac{1}{1 + e^{-s}} \tag{12}$$

Given a pair alignment label $y \in \{0, 1\}$, the similarity loss $\mathcal{L}_s$ is computed using binary cross-entropy:

$$\mathcal{L}_s = -[y \log(p_s) + (1 - y) \log(1 - p_s)] \tag{13}$$

2. Contrastive Loss: Computes scaled dot-product similarity distributions across a batch of N pairs using a temperature parameter τ [14]. The directional text-to-image ($p^{t \to i}$) and image-to-text ($p^{i \to t}$) probability matrices are defined as:

$$p_{ij}^{t \to i} = \frac{\exp\left(\frac{\mathbf{e}_t^i \cdot \mathbf{e}_v^j}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\mathbf{e}_t^i \cdot \mathbf{e}_v^k}{\tau}\right)}, \quad p_{ij}^{i \to t} = \frac{\exp\left(\frac{\mathbf{e}_v^i \cdot \mathbf{e}_t^j}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\mathbf{e}_v^i \cdot \mathbf{e}_t^k}{\tau}\right)} \tag{14}$$

The cross-entropy losses for both directions are evaluated against ground-truth pairing labels y_{ij}:

$$\mathcal{L}_{i \to t} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} \log p_{ij}^{i \to t}, \quad \mathcal{L}_{t \to i} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} \log p_{ij}^{t \to i} \tag{15}$$

The final contrastive similarity loss is given by the symmetric average:

$$\mathcal{L}_s = \frac{1}{2}(\mathcal{L}_{i \to t} + \mathcal{L}_{t \to i}) \tag{16}$$

3.2.4. Classification Layer and Total Objective

The fused feature representation E_{fused} is processed through a feed-forward network (FFN), a task-specific classifier, and a Softmax function to generate final class predictions.

The FFN applies layer normalization, linear projection, dropout regularization, and a GELU activation function (4), repeated N times to balance computational efficiency and feature extraction. The output is fed into a classifier head comprising layer normalization, a linear layer, GELU activation, and a final linear layer to project the features into C class logits (z). The raw logits are converted into class probabilities $\hat{y}_i$ via the Softmax function:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{17}$$

During training, the primary task is optimized using categorical cross-entropy loss:

$$\mathcal{L}_{\text{ce}}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{18}$$

To jointly leverage classification performance and cross-modal consistency, the model optimizes a weighted multi-task loss $\mathcal{L}_{\text{total}}$ combining classification loss with the alignment loss $\mathcal{L}_s$:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{ce}}(y, \hat{y}) + \beta \mathcal{L}_s, \quad \text{where } \alpha + \beta = 1 \tag{19}$$

During inference, final predictions are determined by $\arg\max_i(\hat{y}_i)$.

3.3. Parameter-Efficient Cross-Modal Attention Architecture (PE-CMA)

The Parameter-Efficient Cross-Modal Attention (PE-CMA) framework is engineered specifically to capture fine-grained token-level interactions without suffering from the quadratic computational complexity ($O(N^2)$) typical of standard cross-attention. PE-CMA resolves this computational bottleneck through a two-fold design strategy: first, it utilizes learned token pooling (via TokenLearner and attention-based mechanisms) to dynamically compress raw visual and textual tokens into a fixed set of $P = 8$ representative structural tokens. Second, it employs low-rank query projection to factorize query matrices while directly reusing keys and values from the opposing modality. This allows PE-CMA to achieve the expressive capability of deep bidirectional cross-attention at a fraction of the computational cost, as shown in Figure 2C.

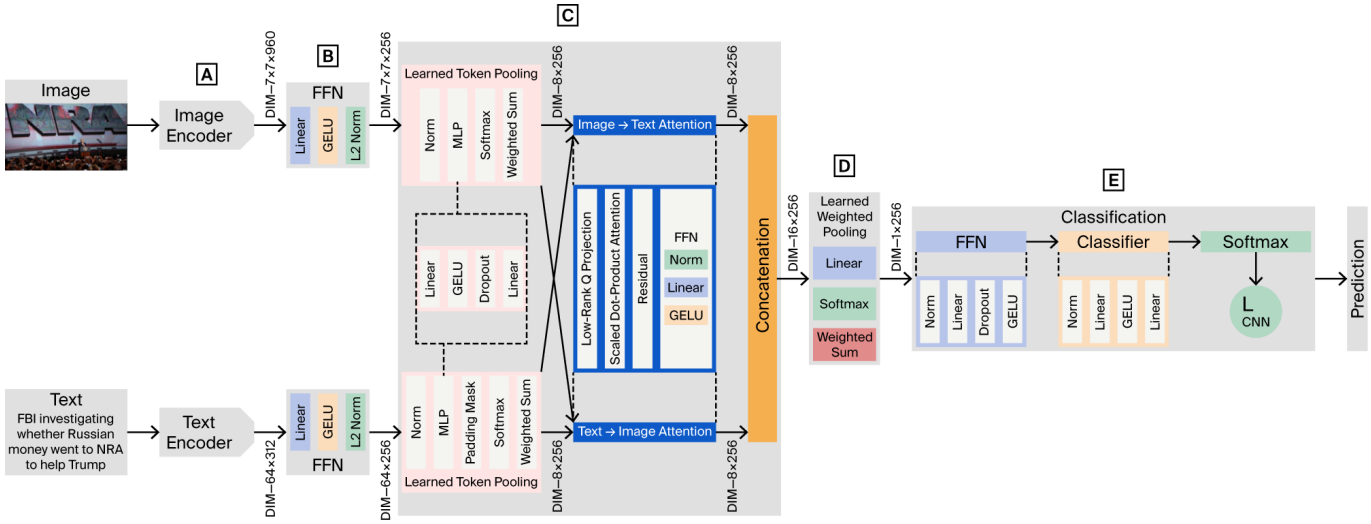


Figure 2. Overview of the proposed Parameter-Efficient Cross-Modal Attention (PE-CMA) architecture. (**A**) Image and text encoders. (**B**) Projection head. (**C**) Parameter-efficient Cross-modal Attention Block. (**D**) Adaptive weighted pooling layer. (**E**) Classification layer.

PE-CMA employs the same feature extraction backbones as CMCM (MobileNetV3 [76] and TinyBERT [17]), but retains token-level representations rather than pooled embeddings. For visual tokens, the global average pooling and classification layers from MobileNetV3 are omitted, yielding a $7 \times 7 \times 960$ feature map, treated as 49 spatial visual tokens of dimension 960. Conversely, the full contextualized sequence embeddings from TinyBERT are retained as textual tokens, bounded by sequence length L ($L = 64$ for Fakeddit; $L = 197$ for Weibo) with a hidden dimension of 312.

Using the same single linear layer and L_2-normalization design described in Section 3.2.1, the visual (49×960) and textual ($L \times 312$) token sequences are projected into a shared 256-dimensional latent space. Reducing embedding dimensions prior to cross-attention minimizes both parameter count and FLOPs, which scale linearly in low-rank attention mechanisms [84].

The Cross-Modal Attention (CMA) block establishes cross-modal interaction through learned token pooling and low-rank query projection [85], minimizing computational overhead by compressing token sequence lengths prior to attention calculations.

3.3.1. Learned Token Pooling

To mitigate the quadratic cost of attending over raw token sequences (49 visual, 64 or 197 textual), each modality is independently compressed into $P = 8$ representative tokens:

1. Visual Token Pooling: Utilizing TokenLearner (specifically TokenLearnerModuleV11) [86], spatial feature maps $X \in \mathbb{R}^{H\times W\times d}$ are modulated by $P = 8$ learned spatial attention maps $\alpha_i(X)$, broadcast across channels via $\gamma(\cdot)$, and spatially aggregated via global average pooling $\rho(\cdot)$:

$$z_i = \rho(X \odot \gamma(\alpha_i(X))), \quad Z_{\text{vis}} = [z_i]_{i=1}^{P} \in \mathbb{R}^{P\times d} \tag{20}$$

2. Textual Token Pooling: Following attention-based pooling principles [87], textual tokens $X \in \mathbb{R}^{L\times d}$ are normalized across the sequence length via a lightweight scoring network $f_\theta(\cdot)$ to generate $P = 8$ compressed tokens $Z_{\text{text}} \in \mathbb{R}^{P\times d}$:

$$A = \text{softmax}(f_\theta(X)), \quad Z_{\text{text}} = AX \tag{21}$$

3.3.2. Low-Rank Query Projection and Bidirectional Attention

To minimize parameter complexity, only query projections W_Q are learned, while key (K) and value (V) representations are directly reused from the opposite modality's pooled tokens [85]. Query matrices Q_i for each modality $i \in \{t, v\}$ are factorized using a low-rank bottleneck of rank r [88]:

$$Q_i = Z_i W_1^{(i)} W_2^{(i)} + b_{Q_i} \tag{22}$$

where the pooled input sequence $Z_i \in \mathbb{R}^{P\times d}$, and the projection matrices are defined as $W_1^{(i)} \in \mathbb{R}^{d\times r}$ and $W_2^{(i)} \in \mathbb{R}^{r\times d_q}$.

Using text-to-image interaction as an example, scaled dot-product attention weights $\alpha_{t\to v}$ and attended outputs A_t are evaluated as:

$$\alpha_{t\to v} = \text{softmax}\left(\frac{Q_t K_v^\top}{\sqrt{d_k}}\right), \quad A_t = \alpha_{t\to v} V_v \tag{23}$$

The output is processed through a residual connection, layer normalization, and a GELU-activated linear projection:

$$E_t = \text{Norm}(A_t + Z_t), \quad O_t = \text{GELU}(\text{Linear}(E_t)) \tag{24}$$

Symmetric operations yield updated visual representations O_i. The final cross-modal representation $F \in \mathbb{R}^{2P\times d}$ concatenates both refined sequence vectors along the token dimension:

$$F = \text{Concat}(O_t, O_i) \tag{25}$$

4. Materials and Methods

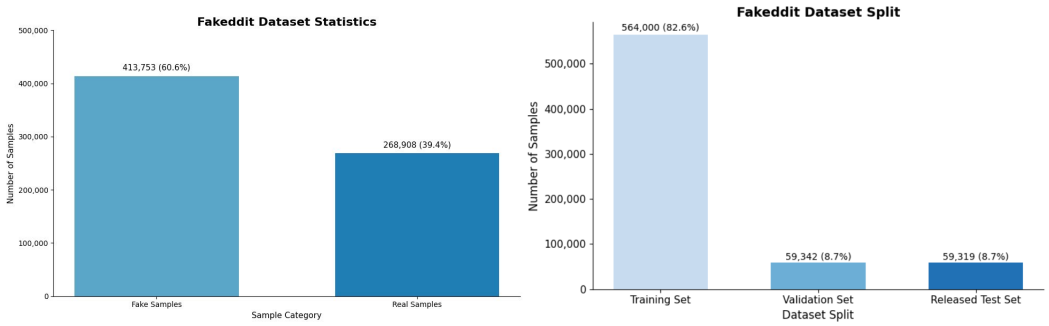
We first introduce the datasets and preprocessing protocols, detailing the preparation of the Fakeddit and Weibo benchmarks. We then discuss the encoder selection and preliminary experiments that justified the choice of lightweight backbones. Finally, we provide the experimental setup and implementation details, specifying the procedures for model selection and hyperparameter tuning alongside the performance measures used for model evaluation.

4.1. Datasets and Preprocessing

4.1.1. Fakeddit Dataset

To train and evaluate the proposed architectures, we utilize the Fakeddit dataset [19], a large-scale multimodal repository sourced from Reddit using the pushshift.io API [89]. Each submission in this dataset contains an image, its corresponding text caption (omitting accompanying articles), and various metadata.

The dataset comprises 1,063,106 total samples collected from 22 thematic subreddits. For this study, we isolate the 682,661 multimodal samples (text–image pairs), which are distributed into 413,753 fake samples and 268,908 real samples (Figure 3a). The official dataset split is preserved, allocating approximately 82.6% for training, 8.7% for validation, and 8.7% for testing (Figure 3b).



(**a**) (**b**)

Figure 3. Fakeddit dataset class distribution and split: (**a**) Distribution of real vs. fake multimodal samples in the Fakeddit dataset. (**b**) Fakeddit Dataset Split.

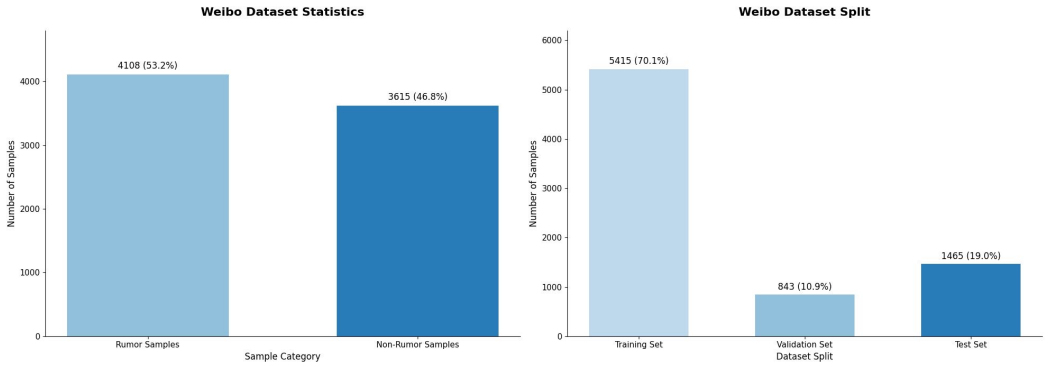
While Fakeddit offers fine-grained labeling hierarchies (2-way, 3-way, and 6-way classification) to categorize specific misinformation types such as satire, misleading content, and manipulated images, our study focuses exclusively on binary (2-way) classification. Under this scheme, a sample is labeled as true only if both the image and corresponding text are accurate and semantically aligned; all other instances (e.g., false connections, manipulated content) are aggregated into the fake class. Although the dataset includes comments and metadata, our experimental pipeline strictly utilizes only the image and text modalities.

4.1.2. Weibo Dataset

To evaluate the proposed architectures on a non-English corpus, we utilize the Weibo dataset [20], sourced from the Chinese microblogging platform Sina Weibo [90]. Ground truth annotations are established via Sina Weibo's official rumor debunking system: false posts are labeled as rumors by a review committee, while real posts are verified by the Xinhua News Agency [91].

Following the data filtering and splitting protocols established in prior multimodal studies [25,92], we utilize a refined subset of 7723 samples, comprising 4108 rumors (fake) and 3615 non-rumors (real) (Figure 4a). We utilized the predefined training, validation, and testing splits provided in the publicly available Event Adversarial Neural Networks (EANN) repository [25]. To explicitly prevent data leakage, the original EANN study partitioned the dataset at a 7:1:2 ratio strictly at the event level using a single-pass clustering method. This event-based separation guarantees that no common events are shared across the training (5415 samples), validation (843 samples), and testing (1465 samples) sets (Figure 4b).

While the original dataset includes complex social context features (e.g., reposts, user metadata), our experimental pipeline strictly isolates the text and image modalities to focus on visual–linguistic consistency.



(**a**) (**b**)

Figure 4. Weibo dataset class distribution and split: (**a**) Distribution of real vs. fake multimodal samples in the Weibo dataset. (**b**) Weibo Dataset Split.

4.1.3. Data Preprocessing Pipeline

Identical preprocessing pipelines were applied to both datasets to ensure compatibility with the selected encoders.

Visual inputs were converted to standard RGB format, resized to 224×224 pixels, transformed into tensors [93], and normalized utilizing ImageNet dataset statistics. To enhance model generalization and mitigate positional or lighting biases, dynamic (on-the-fly) data augmentations, specifically random cropping and color jittering, were applied exclusively during the training phase [94].

Textual data was tokenized using the WordPiece algorithm [95], which handles out-of-vocabulary words via subword mapping and eliminates the need for traditional stemming or lemmatization [96]. Sequences were bookended with standard '[CLS]' and '[SEP]' tokens, and uniform tensor dimensions were enforced via padding (using '[PAD]') or truncation. The maximum sequence length was strictly bounded according to the respective dataset characteristics: 64 tokens for Fakeddit and 197 tokens for Weibo. Finally, sequences and their corresponding attention masks were converted into numerical identifiers to ensure the encoder exclusively processed valid tokens.

To evaluate the cosine similarity loss, binary semantic alignment targets were derived from the native annotations of each dataset. For the Fakeddit dataset, samples labeled as "false connection", where the image explicitly does not support the text, were mapped to the mismatched class, while all other remaining samples in the dataset were mapped to the matched class. Because the Weibo dataset lacks fine-grained alignment annotations, we utilized its global labels: the "non-rumor" class was mapped as a matched pair, while the "rumor" label served as a heuristic proxy for a mismatched pair. This aligns with observations that Weibo rumors frequently rely on out-of-context images [25] and that cross-modal inconsistency indicates fabricated news [97]. However, we acknowledge that this acts as a noisy proxy, as some fake news samples inherently exhibit strong semantic alignment through custom-edited images.

For the contrastive loss, negative pairs are created artificially during training via in-batch negative sampling, following the formulation by [14]. Specifically, matched pairs correspond to the image and text belonging to the same original dataset sample. During training, each text embedding is compared with all image embeddings within the same

mini-batch. The image associated with the same sample is treated as the matched pair, while images from the other samples in the mini-batch are treated as mismatched pairs.

4.2. Encoder Selection and Preliminary Experiments

To identify the most efficient modality extractors, preliminary evaluations were conducted on subsets of the Fakeddit dataset [19] using an NVIDIA T4 GPU and optimizing for Cross-Entropy loss. As detailed in Tables 1 and 2, model performance and efficiency are evaluated across four key metrics: total trainable parameters in millions (Params (M)), floating-point operations in billions (FLOPs (G)), classification accuracy on the validation split, and total wall-clock training duration in minutes (Train Time (min)).

Image Encoder Selection: Three architectures were evaluated on a 10% Fakeddit subset (approximately 37,600 non-corrupted images, utilizing an 80:20 internal training-to-validation split): MobileViT-V2 [98] from Hugging Face [99], alongside GhostNetV2 [16] and MobileNetV3-Large [76] from the TIMM library [100]. Models were trained for 5 epochs utilizing a batch size of 32 and a learning rate of 1×10^{-4} (Table 1).

Table 1. Comparative evaluation of image encoder backbones across validation accuracy and computational cost on Fakeddit.

Model	Params (M)	FLOPs (G)	Accuracy (%)	Train Time (min)
GhostNetV2	4.88	0.362	75.69	29
MobileNetV3-Large	5.48	0.220	74.76	26
MobileViT-V2	4.39	3.700	75.47	40

While GhostNetV2 marginally outperformed in accuracy (<1% difference), MobileNetV3-Large demonstrated a superior efficiency trade-off, requiring significantly fewer FLOPs and the shortest training time. MobileViT-V2 was excluded due to its prohibitive computational overhead. Consequently, MobileNetV3-Large was selected as the visual backbone.

Text Encoder Selection: Five Transformer-based models were evaluated on 470,754 Fakeddit text samples (excluding sequences $\leq$ 2 words), utilizing an 80:10:10 internal training–validation–test split. The evaluation included ALBERT-base [101], TinyBERT [17], DistilBERT [102], and MiniLM [18,99], as well as DeBERTa-v3 [103] configured with Low-Rank Adaptation (LoRA) [104]. Sequences were capped at 128 tokens, and models were trained for 3 epochs with a batch size of 16 and a learning rate of 2×10^{-5} (Table 2).

Table 2. Comparative evaluation of text encoder backbones across validation accuracy and computational cost on Fakeddit.

Model	Params (M)	FLOPs (G)	Accuracy (%)	Train Time (min)
ALBERT-base	12.0	21.8	85.52	186
TinyBERT	14.5	1.2	86.39	35
MiniLM	22.0	– [1]	87.02	43
DistilBERT	66.0	22.0	87.13	92
DeBERTa-v3 + LoRA	184.0	50.0	85.62	161

[1]—indicates missing or unreported FLOPs in the original MiniLM literature.

TinyBERT achieved the most optimal balance of efficiency and classification performance. Although MiniLM and DistilBERT provided marginal accuracy improvements, they incurred substantially higher parameter counts and longer training times. ALBERT and DeBERTa-v3 yielded lower accuracies while requiring extensive computational time. Therefore, TinyBERT was selected to fulfill the lightweight architectural requirements for the text modality.

4.3. Experimental Setup and Implementation Details

Experiments were implemented in Python 3.13.16 using Google Colab with an NVIDIA A100 GPU runtime. Local preprocessing and preliminary tasks utilized a Windows 11 laptop (Intel Core i7-1360P, 16 GB RAM).

4.3.1. Model Selection and Hyperparameter Tuning

Following architecture construction, models undergo end-to-end training, with hyperparameters optimized via Optuna [105] using the AdamW optimizer [106] on validation sets. Table 3 lists fixed configurations.

Table 3. Static training configurations and architectural parameters across datasets.

Hyperparameter	Value
Projection Head Dimension	512 (CMCM), tuned in PE-CMA
Projection Head Layer	1 Layer
Projection Head & Classification Activation	GELU
TinyBERT Token Length	64 (Fakeddit), 197 (Weibo)
Optimizer Type	AdamW
Loss Function	Cross-Entropy Loss

The search spaces for hyperparameter tuning are summarized in Table 4 for the CMCM architecture and Table 5 for the PE-CMA architecture.

Table 4. Hyperparameter search space defined for Optuna optimization of the CMCM architecture.

Hyperparameter	Range of Values
Batch Size	8, 16, 32, 64, 128
Classification FFN Blocks	1, 2
Dropout Rate	0.0, 0.1, 0.2, 0.4
Learning Rate	$10^{-2}, 10^{-3}, 10^{-4}$
Fusion Strategies	Concatenation, Element-Wise Max, Siamese Style, Hadamard Product with Element-Wise Sum
Loss Weights (α, β)	(0.8, 0.2), (0.7, 0.3)
Alignment Loss Type	Contrastive Loss, Cosine Similarity Loss
Contrastive Loss Temperature	0.1, 0.07

Table 5. Hyperparameter search space defined for Optuna optimization of the PE-CMA architecture.

Hyperparameter	Range of Values
Batch Size	8, 16, 32, 64, 128
Classification FFN Blocks	1, 2
Dropout Rate	0.0, 0.1, 0.2, 0.4
Learning Rate	$10^{-2}, 10^{-3}, 10^{-4}$
Shared Projection Dimension	256, 512
Low-rank Decomposition Rank	16, 32, 45

4.3.2. Evaluation Methodology

To evaluate classification performance and computational efficiency, the following metrics are utilized. For all binary metrics, the positive class is defined as 'Fake' for the Fakeddit dataset and 'Rumor' for the Weibo dataset:

- Accuracy (Acc): Measures the overall proportion of correct predictions made by the classifier [107]:
$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{26}$$
where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.
- Precision (P): Quantifies the proportion of correctly predicted positive outcomes among all predicted positives [107]:
$$P = \frac{TP}{TP + FP} \tag{27}$$
- Recall (R): Measures the proportion of actual positive outcomes correctly identified by the model [107]:
$$R = \frac{TP}{TP + FN} \tag{28}$$
- F1-Score ($F1$): Computes the harmonic mean of precision and recall, serving as a robust measure for imbalanced datasets [107]:
$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{29}$$
- Parameters and FLOPs: Model complexity is evaluated via total trainable weights (Parameters) [108] and floating-point operations per second (FLOPs), which sum the arithmetic operations required for a single forward pass across all layers [109].
- Inference Time: Measures the average wall-clock time required by a trained model to process a single sample during testing [110].

Statistical significance ($\alpha = 0.05$) across fusion strategies, ablation settings, and architectures is determined using IBM SPSS Statistics 32 via the Wilcoxon signed-rank test [111] and the Friedman test [112], with the Holm–Bonferroni correction applied to all post hoc p-values. To isolate the contribution of individual architectural components and design choices, dedicated ablation experiments are conducted, with detailed configurations and findings presented alongside the evaluation in Section 5 .

It is important to distinguish the evaluation methodology used for the preliminary experiments from that of the final architectures. While all preliminary evaluations averaged results across ten independent training runs using random initial seeds to ensure structural stability, the final CMCM and PE-CMA models were each based on a single trained instance. To assess performance variability and conduct rigorous statistical comparisons between these final models without the computational burden of repeated end-to-end training runs, we employed a paired bootstrapping technique. In all statistical comparisons, we paired performance samples using a shared, fixed pool of 30 bootstrap resamples across the compared models; the same bootstrap instances generated the corresponding paired results. Consequently, this bootstrap procedure specifically measures the robustness of the predictions against sample variance, rather than variance arising from random weight initialization or optimization stochasticity.

5. Results

We first present the outcomes of the hyperparameter tuning and configuration selection phase to establish optimal baseline settings. We then analyze the optimized baseline performance and evaluation, followed by ablation studies and component analysis that isolate the performance impact of individual architectural components. Next, we assess feature transferability through a frozen pre-trained encoder evaluation. We then conduct a CPU-based inference and quantization analysis to validate the architectures' lightweight

design and viability for edge deployment, before finally providing a comparative evaluation with existing studies to contextualize our performance against current benchmarks.

5.1. Hyperparameter Optimization and Architectural Configuration Selection

Hyperparameters were optimized over 50 trials using Optuna [105] to maximize validation accuracy on the Fakeddit dataset. The optimization process incorporated early stopping via a MedianPruner, configured with 5 startup trials, 3 warmup steps, and an evaluation interval of 1 step. For the Weibo dataset experiments, the batch size was set to 64 and the maximum token length was restricted to 197 [33].

5.1.1. CMCM Tuning Outcomes and Fusion Strategy Selection

Optuna tuning across four fusion strategies yielded the optimal configurations and validation accuracies presented in Table 6, showing the Optimal Batch Size, Number of Feed-Forward Neural Network (FFN) Blocks, Projection layer Dropout (Proj. Dropout), Classification layer Dropout (Cls. Dropout), Learning Rate (LR), α, β, Alignment Loss Type (Align. Loss), Temperature (Temp.), and the Corresponding Best Validation Accuracy (Val. Acc). Concatenation achieved the highest validation accuracy at 88.67%, closely followed by Hadamard product with element-wise summation (88.60%), Siamese-style (88.24%), and element-wise maximum (88.04%).

Table 6. Optimal hyperparameter configurations and corresponding validation accuracies across different CMCM feature fusion strategies.

Fusion Strategy	Batch Size	FFN Blocks	Proj. Drop.	Cls. Drop.	LR	α	β	Align. Loss	Temp.	Val. Acc (%)
Concatenation	128	1	0.2	0.0	10^{-4}	0.8	0.2	Contrastive	0.1	88.67
Element-wise Max.	32	2	0.2	0.2	10^{-4}	0.8	0.2	Contrastive	0.1	88.04
Siamese-style	128	1	0.1	0.2	10^{-4}	0.7	0.3	Contrastive	0.07	88.24
Hadamard + Sum	32	1	0.2	0.2	10^{-4}	0.8	0.2	Contrastive	0.07	88.60

A global Friedman test across 10 random seeds indicated a statistically significant difference among the fusion strategies ($p < 0.001$). Mean ranks identified concatenation and Siamese-style as the top performers (both 3.30), followed by Hadamard product with sum (2.40) and element-wise maximum (1.00).

Pairwise Wilcoxon signed-rank tests with Holm–Bonferroni correction ($\alpha = 0.05$) confirmed that concatenation and Siamese-style significantly outperform element-wise maximum, while differences involving the Hadamard product and between concatenation and Siamese-style ($p = 0.96$) are not statistically significant (Table 7). Nonetheless, concatenation and Siamese fusion were selected as the primary fusion strategies due to their consistently superior performance.

Table 7. Pairwise Wilcoxon signed-rank test comparisons of feature fusion strategies with Holm–Bonferroni correction ($\alpha = 0.05$).

Fusion Strategy	Concatenation	Element-Wise Max.	Siamese-Style
Element-Wise Max.	0.03 ↑	–	–
Siamese-Style	0.96	0.03 ←	–
Hadamard Product + Sum	0.11	0.03 ←	0.12

Cells marked with dashes (–) indicate redundant pairwise comparisons. Arrows indicate the direction of superior performance for statistically significant differences ($p < 0.05$).

5.1.2. PE-CMA Tuning Outcomes and Attention Configuration Selection

Hyperparameter optimization via Optuna for the PE-CMA framework on the Fakeddit dataset yielded an optimal configuration consisting of a batch size of 128, a single feed-forward neural network block, a learning rate of 10^{-4}, and a shared dimension of 256 with a rank of 45. Additionally, the dropout rates for the projection and classification layers were established at 0.0 and 0.2, respectively. As detailed in Table 8, this parameterization achieved a validation accuracy of 88.01%.

Table 8. Optimal hyperparameter configuration and validation performance for the PE-CMA framework on the Fakeddit dataset determined via Optuna optimization.

Batch Size	FFN Blocks	Proj. Drop.	Cls. Drop.	Learning Rate	Shared Dim.	Rank	Val. Acc (%)
128	1	0.0	0.2	10^{-4}	256	45	88.01

Evaluating multiple PE-CMA configurations varying pooled tokens and shared dimensions revealed minimal performance differences on Fakeddit (Table 9). Wilcoxon signed-rank tests confirmed no statistically significant differences ($p > 0.05$) compared to the 8-token, 256-dimension baseline (Table 10). Consequently, the 8-256 configuration was selected for parameter efficiency and lower computational cost.

Table 9. Performance comparison and computational complexity of PE-CMA configurations across varying numbers of pooled tokens and shared projection dimensions on the Fakeddit dataset.

Tokens	Dim.	Acc (%)	P (%)	R (%)	F1 (%)	Train Time	Params (M)	GFLOPs	ms/Sample
8	256	91.14	89.30	88.46	88.88	3 h 18 min	18.25	1.41	1.82
8	512	91.26	89.87	88.27	89.06	3 h 34 min	19.19	1.43	1.96
32	256	91.14	89.32	88.19	88.75	5 h 21 min	18.53	1.42	1.86
49	512	91.22	89.22	88.68	88.95	3 h 04 min	19.12	1.46	1.91

Table 10. Wilcoxon signed-rank test pairwise comparisons of PE-CMA attention configurations against the baseline (8 pooled tokens, 256 dimensions) on the Fakeddit dataset.

Comparison	Z	*p*-Value
32-256 vs. 8-256	0.000	1.000
49-512 vs. 8-256	−0.420	0.674
8-512 vs. 8-256	−0.700	0.484

Notation *A-B* indicates *A* pooled tokens and a shared projection dimension of *B*. Baseline configuration is 8-256.

5.2. Optimized Baseline Performance and Evaluation

Utilizing the optimal configurations identified during the Optuna optimization phase, the CMCM (Concatenation and Siamese variants) and PE-CMA frameworks were evaluated by training on the full training subsets of the Fakeddit and Weibo datasets. These evaluations were conducted on an NVIDIA A100 GPU with early stopping monitoring validation accuracy, set to a patience of 10 epochs. To accommodate language differences, language-specific TinyBERT variants (English for Fakeddit and Chinese for Weibo) were deployed, introducing minor parameter variations across the respective datasets. Table 11 details the resulting performance metrics and computational costs, with optimum scores indicated in bold.

Table 11. Quantitative performance, parameter efficiency, and computational overhead of the optimized CMCM and PE-CMA architectures on the Fakeddit and Weibo benchmarks.

Dataset	Arch.	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs	Train Time	ms/Sample
Fakeddit	CMCM-Concat.	92.03	**90.31**	89.50	89.90	18.77	1.39	5 h 15 min	1.80
	CMCM-Siamese	**92.29**	89.04	**91.84**	**90.42**	19.29	1.39	5 h 35 min	1.83
	PE-CMA	91.17	88.14	89.79	88.96	18.25	1.41	5 h 5 min	1.82
Weibo	CMCM-Concat.	90.17	88.83	**92.59**	90.67	15.84	4.077	6 min 31 s	2.93
	CMCM-Siamese	91.06	91.06	91.67	91.36	16.36	4.08	11 min 11 s	3.04
	PE-CMA	**91.81**	**93.44**	90.48	**91.94**	15.33	4.11	10 min 54 s	2.84

5.2.1. Performance Evaluation on the Fakeddit Benchmark

On Fakeddit, CMCM with Siamese fusion achieved the highest overall accuracy (92.29%), recall (91.84%), and F1-score (90.42%), while CMCM with Concatenation achieved top precision (90.31%). Wilcoxon signed-rank tests confirmed statistically significant performance differences among all three models ($p < 0.001$). CMCM's superiority on Fakeddit stems from its similarity branch; because Fakeddit features frequent text–image mismatches, explicitly modeling cross-modal consistency is highly effective. Siamese fusion captures agreement and disagreement via concatenation, absolute difference, and Hadamard product operations, aligning well with the dataset's characteristics.

Conversely, PE-CMA delivered superior computational efficiency—requiring fewer parameters (18.25 M) and the shortest training time (5 h 5 min)—due to learned token pooling reducing sequence lengths from 49/64 to 8 tokens per modality, coupled with a 256-dimensional shared projection space.

The confusion matrices (Figure 5) demonstrate that specialized architectures effectively mitigate class bias and ensure true generalization across evaluation benchmarks. While the CMCM-Concatenation model exhibits a slight classification skew with a higher propensity for false positives, the CMCM-Siamese and PE-CMA configurations achieve superior classification equilibrium and compact error profiles by enforcing cross-modal consistency and low-rank cross-attention.

5.2.2. Performance Evaluation on the Weibo Benchmark

On Weibo, PE-CMA outperformed both CMCM variants, attaining the highest accuracy (91.81%), precision (93.44%), and F1-score (91.94%) with minimal parameter overhead and the lowest inference latency (2.84 ms/sample). Wilcoxon signed-rank tests (Table 12) showed a significant difference between CMCM-Siamese and CMCM-Concat ($p = 0.001$), whereas differences between PE-CMA and CMCM variants were not statistically significant ($p > 0.05$).

Table 12. Pairwise Wilcoxon signed-rank test comparisons of classification performance among PE-CMA and CMCM variants on the Weibo dataset.

Comparison	*p*-Value
CMCM-Concat. vs. CMCM-Siamese	<0.01
PE-CMA vs. CMCM-Siamese	0.36
PE-CMA vs. CMCM-Concat.	0.28

Unlike Fakeddit, Weibo lacks a primary focus on unrelated text–image pairs, reducing the utility of CMCM's similarity branch. PE-CMA excels here by leveraging attention-based interaction modeling to learn richer cross-modal relationships, aligning with the existing literature emphasizing advanced fusion for Weibo [10,33,41,43].

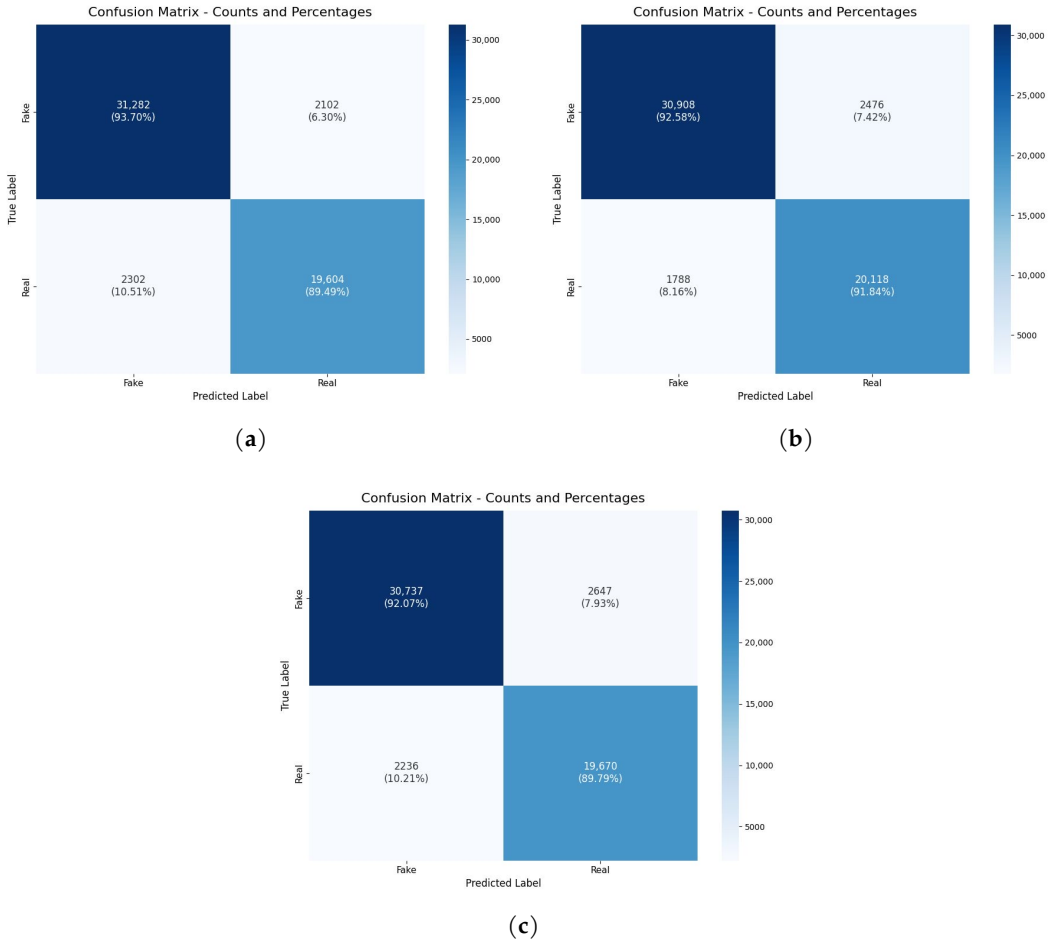


Figure 5. Confusion matrices of the proposed approaches on the Fakeddit dataset: (**a**) CMCM-Concatenation Fusion, (**b**) CMCM-Siamese Fusion, and (**c**) PE-CMA.

The confusion matrices illustrated in Figure 6 evaluate the class-wise prediction performance of the proposed frameworks on the Weibo dataset. CMCM-Concatenation (Figure 6a) achieves strong performance on genuine news classification with a 92.59% true negative rate, though it suffers from a higher false-negative rate (12.41% of fake news instances misclassified as real). CMCM-Siamese (Figure 6b) effectively mitigates this issue by enforcing explicit alignment, improving fake news detection accuracy to 90.41% and establishing a highly balanced diagnostic error profile between both classes. PE-CMA (Figure 6c) provides the highest sensitivity toward deceptive content, successfully isolating 93.23% of fake news instances while maintaining a low false-positive rate.

5.3. Ablation Studies and Component Analysis

Ablation studies were conducted to isolate component contributions for the best-performing architecture on each dataset: CMCM-Siamese on Fakeddit and PE-CMA on Weibo (Table 11).

For CMCM-Siamese, evaluations assessed four variations: (1) Textual Branch Only (excluding MobileNetV3 and fusion), (2) Visual Branch Only (excluding TinyBERT and fusion), (3) Excluding the Similarity Measurement Branch (omitting consistency scoring and alignment loss), and (4) PCA Projection Head Replacement (substituting trainable projections with non-trainable principal component analysis (PCA)).

For PE-CMA, evaluations assessed five structural variants: (1) Without Token Pooling (applying attention directly to all 49 image and 197 text tokens instead of 8), (2) Without Cross-modal Attention (omitting direct attention interactions prior to fusion), (3) Average-

Based Final Pooling (replacing learned pooling with standard average pooling), (4) Without Cross-modal Attention Block (relying exclusively on projected unimodal representations), and (5) Full-Rank Attention (replacing the proposed low-rank projections with standard full-rank weight matrices).

To ensure a fair comparison, all ablation variants were trained using identical hyperparameters and early stopping criteria (monitoring validation accuracy with a patience of 10 epochs) to the complete proposed models.

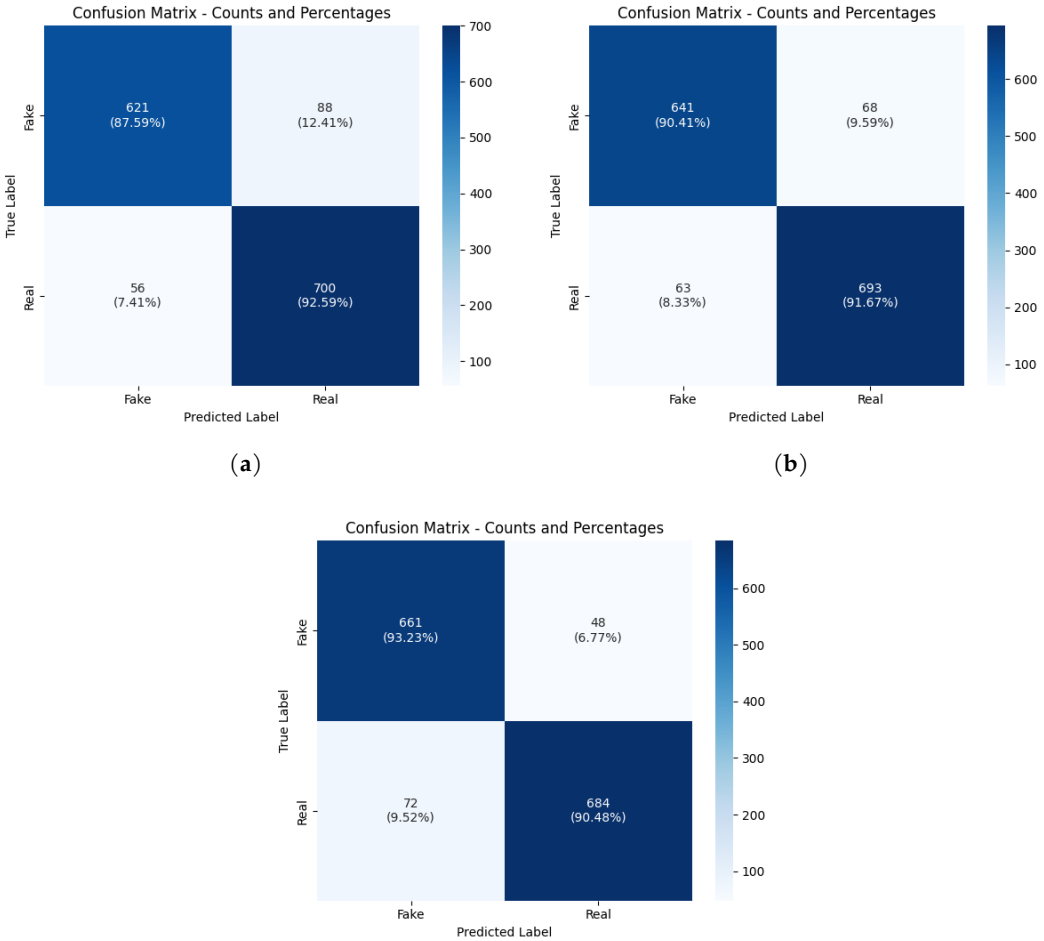


Figure 6. Confusion matrices of the proposed fusion approaches on the Weibo dataset: (**a**) CMCM-Concatenation Fusion, (**b**) CMCM-Siamese Fusion, and (**c**) PE-CMA.

5.3.1. CMCM-Siamese Fusion Ablations (Fakeddit Dataset)

Unimodal evaluations on Fakeddit demonstrated that using only the textual branch achieved 87.36% accuracy in 20 min (15.04M parameters), whereas the visual branch alone yielded 84.09% accuracy in 10 h 24 min (3.99M parameters) (Table 13), where values in bold indicate the optimum scores. While text contributed more strongly, the full multimodal CMCM-Siamese model achieved superior accuracy (92.29%), confirmed as statistically significant via Wilcoxon signed-rank tests ($p < 0.001$).

Excluding the similarity measurement branch reduced accuracy to 91.22% ($p < 0.001$), confirming that explicitly modeling cross-modal consistency helps identify misleading content (e.g., unrelated text–image pairs). Replacing the learnable projection head with non-trainable PCA decreased accuracy to 91.49% ($p < 0.001$), demonstrating that task-specific learned projections are essential.

Table 13. Ablation analysis evaluating the contributions of unimodal modalities, similarity measurement, and projection head design against the full CMCM-Siamese model on the Fakeddit dataset.

Ablation Experiment	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs	Train Time	ms/Sample
Textual Branch Only	87.36	83.48	84.87	84.17	15.04	1.163	20 min	0.105
Visual Branch Only	84.09	79.51	80.63	80.06	3.99	0.225	10 h 24 min	6.953
Excluding Similarity Branch	91.22	87.86	90.32	89.08	19.29	1.387	5 h 29 min	1.702
PCA Projection Head	91.49	**89.54**	88.90	89.22	18.23	1.387	5 h 40 min	1.810
Full Multimodal Model (for comparison)								
CMCM - Siamese Fusion	**92.29**	89.04	**91.84**	**90.42**	19.29	1.387	5 h 35 min	1.832

5.3.2. PE-CMA Ablations (Weibo Dataset)

Ablations on Weibo assessed the components of PE-CMA (Table 14), where values in bold indicate the optimum scores. Removing initial token pooling reduced accuracy to 89.96% ($p = 0.194$). Removing cross-modal attention dropped accuracy to 89.01% ($p < 0.001$), highlighting its role in modeling inter-modality interactions. Replacing final learned pooling with average pooling decreased accuracy to 91.06% ($p < 0.001$), showing the effectiveness of learnable weighting. Removing the complete cross-modal attention block reduced accuracy to 89.76% ($p < 0.001$). Finally, substituting the proposed low-rank projection with a standard full-rank cross-modal attention mechanism yielded a lower accuracy of 90.85% ($p = 0.003$); critically, the full-rank baseline also increased inference latency (4.521 ms/sample vs. 2.837 ms/sample) and training time. Overall, these findings validate the design of PE-CMA in capturing rich cross-modal relationships efficiently.

Table 14. Ablation study evaluating the impact of token pooling, cross-modal attention, and learned pooling mechanisms on the classification performance and efficiency of the PE-CMA framework using attention-based fusion on the Weibo dataset.

Ablation Experiment	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs	Train Time	ms/Sample
Without Token Pooling	89.96	89.65	90.48	90.06	15.29	4.135	8 min 44 s	2.902
Without Cross-modal Attention	89.01	86.24	**93.65**	89.79	15.15	4.108	8 min 13 s	2.801
Average-based Final Pooling	91.06	93.34	89.02	91.13	15.33	4.110	6 min 21 s	2.819
Without Attention Block	89.76	92.92	86.77	89.74	15.11	4.103	6 min 16 s	2.870
Full-Rank Attention	90.85	91.25	91.01	91.13	15.68	4.112	12 min 27 s	4.521
Full Multimodal Model (for comparison)								
PE-CMA	**91.81**	**93.44**	90.48	**91.94**	15.33	4.110	10 min 54 s	2.837

5.4. Evaluation of Frozen Pre-Trained Encoders and Feature Transferability

To determine whether pre-trained feature extractors can eliminate the need for domain-specific encoder fine-tuning, a frozen-encoder evaluation (linear probing) was conducted across both architectures (Table 15, where bold values denote optimal performance per framework). In this setting, the backbone encoders remained strictly frozen while only the task-specific projection heads and classification layers were optimized, utilizing early stopping monitored via validation accuracy, with a patience of 3 epochs. The evaluated configurations included joint vision–language models—MobileCLIP [113], CLIP [75], and SigLIP2 [114]—as well as decoupled modality representations pairing MiniLM [18] with DeiT-Tiny [115] and DINOv2 [116]. Frozen CMCM-Siamese experiments were conducted on Fakeddit, whereas PE-CMA experiments utilized Weibo with language-appropriate alternatives (e.g., Chinese-CLIP). It is important to note that the fine-tuned baseline results reported in Table 15 differ from the fully optimized main results presented

earlier in Table 11 (e.g., reporting 89.01% rather than 91.81% accuracy for PE-CMA). This variance arises directly from the distinct experimental settings required for this specific comparative analysis. To ensure a fair and consistent evaluation against the frozen-encoder configurations, both of the fine-tuned baseline models (CMCM-Siamese and PE-CMA) in Table 15 were restricted to only 3 training epochs. In contrast, the optimal results reported in Table 11 reflect the full training setup, where models were trained to complete convergence utilizing early stopping with a patience of 10 epochs.

Table 15. Comparative evaluation of frozen pre-trained encoder backbones against the fine-tuned baseline for CMCM-Siamese on Fakeddit and PE-CMA on Weibo.

Arch.	Image Encoder	Text Encoder	Encoder Setting	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs	Train Time	ms/Sample
CMCM	MobileCLIP	MobileCLIP	Frozen	91.49	87.58	91.49	89.50	55.64	3.711	2 h 5 min	1.913
	CLIP	CLIP	Frozen	**92.78**	**90.29**	**91.64**	**90.96**	153.12	6.790	3 h 13 min	1.854
	SigLIP2	SigLIP2	Frozen	88.47	84.72	86.50	85.60	377.29	22.35	3 h 21 min	1.804
	DeiT-Tiny	MiniLM	Frozen	89.32	86.06	87.18	86.61	29.89	7.470	6 h 19 min	1.735
	DINOv2	MiniLM	Frozen	89.86	86.77	87.80	87.28	46.48	8.220	3 h 37 min	1.755
	MobileNetV3	TinyBERT	Fine-Tuned	92.31	89.58	91.21	90.38	19.29	1.387	2 h 36 min	1.798
PE-CMA	Chinese-CLIP	Chinese-CLIP	Frozen	87.30	89.47	84.55	87.42	189.27	33.70	3 min 36 s	3.177
	SigLIP2	SigLIP2	Frozen	78.23	77.91	80.69	79.27	376.19	22.17	9 min 25 s	1.269
	DeiT-Tiny	MiniLM	Frozen	84.64	83.74	87.17	85.42	123.97	5.310	5 min 34 s	2.869
	DINOv2	MiniLM	Frozen	85.53	88.53	82.67	85.50	140.51	8.910	13 min 1 s	3.025
	MobileNetV3	TinyBERT	Fine-Tuned	**89.01**	**89.61**	**89.02**	**89.32**	15.33	4.110	2 min 50 s	4.852

Within CMCM-Siamese, the frozen CLIP backbone achieved the highest overall detection metrics (92.78% accuracy, 90.96% F1-score), yielding a statistically significant improvement over the fine-tuned MobileNetV3–TinyBERT baseline ($p < 0.001$). However, this modest gain incurred an extreme computational penalty: frozen CLIP required 153.12 M parameters and 6.79 GFLOPs, representing an eight-fold increase in parameter footprint and a five-fold increase in FLOPs relative to the baseline (19.29M parameters, 1.39 GFLOPs). All other frozen configurations (MobileCLIP, SigLIP2, DeiT-Tiny–MiniLM, and DINOv2–MiniLM) significantly underperformed the fine-tuned baseline ($p < 0.001$).

On the Weibo dataset with PE-CMA, the fine-tuned MobileNetV3–TinyBERT pair surpassed all frozen counterparts—including Chinese-CLIP, SigLIP2, DeiT-Tiny–MiniLM, and DINOv2–MiniLM—achieving 89.01% accuracy and an 89.32% F1-score with the lowest parameter footprint (15.33 M) ($p < 0.001$).

This probing evaluation indicates that while massive foundation models like CLIP provide strong frozen feature representations, they impose prohibitive parameter and memory requirements that hinder edge or real-time deployment. In contrast, end-to-end fine-tuning of compact backbones (MobileNetV3 and TinyBERT) establishes a superior Pareto efficiency between classification accuracy and inference overhead, confirming that robust multimodal fake news detection does not strictly necessitate unwieldy foundation models.

5.5. CPU Evaluation and Edge-Deployment Viability

To substantiate the lightweight design of the proposed architectures and assess their practical viability for deployment on resource-constrained or edge devices, we conducted inference evaluations on a standard CPU environment. Following established methodologies for efficient edge deployment [117], we applied post-training dynamic quantization to convert the model weights from 32-bit floating-point (FP32) to 8-bit integer (INT8) precision. Table 16 presents the model size, CPU latency per sample, and corresponding classification performance for both the PE-CMA and CMCM-Siamese frameworks. Note that the baseline FP32 metrics reported in this CPU environment exhibit slight deviations from

the GPU-accelerated results in Table 11. Because all experimental settings apart from the hardware remained exactly identical, this variance is an expected artifact of hardware-level numerical differences. Specifically, differing floating-point accumulation strategies between the standard CPU and the NVIDIA A100 GPU introduce minor deviations in the output logits due to the non-associativity of floating-point arithmetic, which can occasionally alter predictions for borderline samples [118,119].

As demonstrated, applying INT8 quantization to the PE-CMA architecture dramatically reduces the model storage footprint by over 75% (from 175.15 MB down to 42.78 MB) while fully preserving classification performance (achieving 91.40% accuracy and 91.68% F1-score). Similarly, the CMCM-Siamese framework achieved a 25.59% reduction in model size (from 73.83 MB to 54.93 MB) while maintaining highly comparable predictive performance. We note a slight increase in CPU latency for the INT8 models; this is an expected artifact of dynamic quantization, which computes activation scaling factors on-the-fly during inference. Regardless, the absolute inference latencies (between 55 ms and 65 ms per sample) confirm that both architectures are highly capable of real-time processing. These results provide concrete empirical evidence that the proposed frameworks successfully balance predictive robustness with the stringent memory and computational constraints required for edge deployment.

Table 16. Evaluation of model size, CPU latency, and classification performance under standard FP32 and Post-Training Dynamic INT8 Quantization.

Model	Precision	Model Size (MB)	CPU Latency (ms)	Accuracy (%)	F1-Score (%)
PE-CMA	FP32 (Original)	175.15	55.43	91.40	91.64
	INT8 (Quantized)	42.78	57.09	91.40	91.68
CMCM-Siamese	FP32 (Original)	73.83	55.00	92.90	91.18
	INT8 (Quantized)	54.93	65.02	92.40	90.36

5.6. Benchmarking Against Existing Studies

Performance comparisons against state-of-the-art benchmarks on the Fakeddit and Weibo datasets are divided into two tables to ensure fair evaluation. Table 17 summarizes the baseline models we successfully reproduced under unified hardware, software, preprocessing, and data-splitting conditions. Conversely, Table 18 presents results compiled directly from the existing literature for models lacking publicly available source code.

Crucially, we must highlight that the results in Table 18 are based on indirect comparisons from the literature and cannot serve as definitive evidence of the proposed models' performance advantages. In both tables, values in bold indicate the optimum scores among the respective groups.

On Fakeddit (Figure 7), CMCM achieves competitive performance against reproduced benchmarks (Table 17), achieving 92.29% accuracy and a 90.42% F1-score with a fraction of the computational overhead of larger baselines. While indirectly cited models in Table 18 like BMR [33], GAMED [31], and Uppada et al. [9] report higher F1 scores, they rely on heavy backbones exceeding 150M–250M parameters. CMCM requires only 19.29M parameters and 1.39 GFLOPs, offering a superior balance of accuracy and efficiency. CMCM also demonstrates favorable training and inference scaling against complex reproduced models like MMBT [40] and Giron et al. [29].

On Weibo (Figure 8), PE-CMA validates the efficacy of low-rank weight decomposition and learned token pooling. Achieving 91.81% accuracy and a 91.94% F1-score with only 15.33 M parameters, PE-CMA matches or exceeds heavier reproduced attention frameworks

like MCAN [41] at a substantially lower computational cost, while outperforming classical concatenation models like MVAE [27] and CAFE [43].

Table 17. Direct performance comparison, parameter footprint, and computational complexity of the proposed frameworks against reproduced state-of-the-art models evaluated under unified experimental conditions.

Dataset	Model	Year	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs	Train Time	ms/Sample
Fakeddit	MVAE [27]	2019	86.24	82.17	83.84	82.75	148.80	19.64	6 h 21 min	0.04
	MMBT [40]	2020	90.58	90.09	90.27	90.18	169.20	14.73	18 h 12 min	32.80
	Giron et al. [29]	2025	88.63	88.62	88.63	88.62	151.00	14.55	10 h 38 min	2.01
	CMCM (Proposed)	2026	**92.29**	**89.04**	**91.84**	**90.42**	**19.29**	**1.39**	**5 h 35 min**	1.83
Weibo	MVAE [27]	2019	79.79	79.79	100.00	88.76	148.80	19.64	6 min 12 s	0.15
	MCAN [41]	2021	92.11	92.01	92.68	92.34	244.26	31.05	18 min 23 s	5.53
	CAFE [43]	2022	88.05	88.48	88.36	88.42	22.24	3.67	6 min 26 s	0.20
	BMR [33]	2023	**94.95**	**95.23**	**94.97**	**95.10**	229.98	77.28	6 h 9 min	61.17
	PE-CMA (Proposed)	2026	91.81	93.44	90.48	91.94	**15.33**	**4.11**	**10 min 54 s**	2.84

Table 18. Indirect performance comparison of models cited directly from the literature. Note: These results are based on indirect comparisons from the literature due to closed-source implementations and cannot serve as definitive evidence of the proposed models' performance advantages. Parameter counts for certain models were estimated from their component architectures ("–" indicates unreported data).

Dataset	Model	Year	Acc (%)	P (%)	R (%)	F1 (%)	Params (M)	GFLOPs
Fakeddit	EANN [2] [25]	2018	87.70	90.43	88.11	89.26	147.00 [1]	19.63 [3]
	BMR [2] [33]	2023	91.65	94.34	92.88	93.61	229.98	77.28
	Uppada et al. [9]	2023	91.94	93.76	92.83	93.29	156.00 [1]	–
	ELD-FN [28]	2024	88.83	93.54	90.29	91.89	210.00 [1]	–
	GAMED [31]	2025	93.93	93.55	93.71	93.63	252.00 [1]	–
Weibo	EANN [25]	2018	82.70	84.70	81.20	82.90	147.00 [1]	19.63 [3]
	SpotFake [26]	2019	89.23	–	–	–	123.00 [1]	30.42 [3]
	MFND-CMM [10]	2025	90.30	91.40	92.30	91.80	232.25 [1]	–

[1] Estimated from reported backbone components. [2] Results sourced from GAMED [31]. [3] Results sourced from BMR [33].

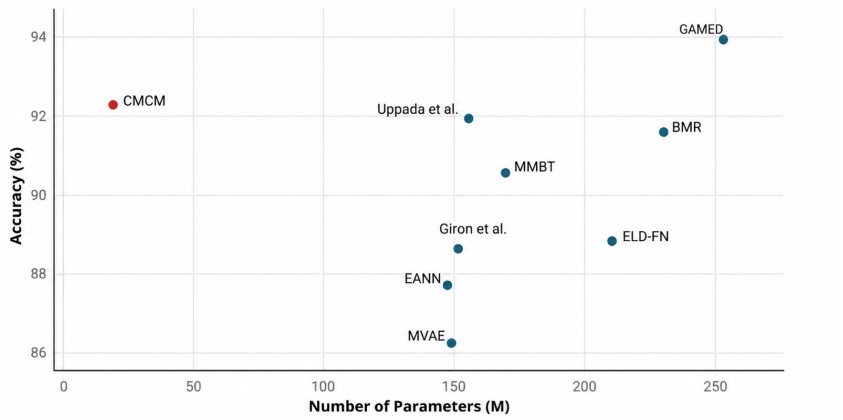


Figure 7. Trade-off between classification accuracy and model parameter footprint (in millions) across multimodal models on the Fakeddit dataset. The proposed CMCM (red) establishes high parameter efficiency relative to baseline methods (blue): MVAE [27], EANN [25], Giron et al. [29], ELD-FN [28], MMBT [40], BMR [33], Uppada et al. [9], and GAMED [31].

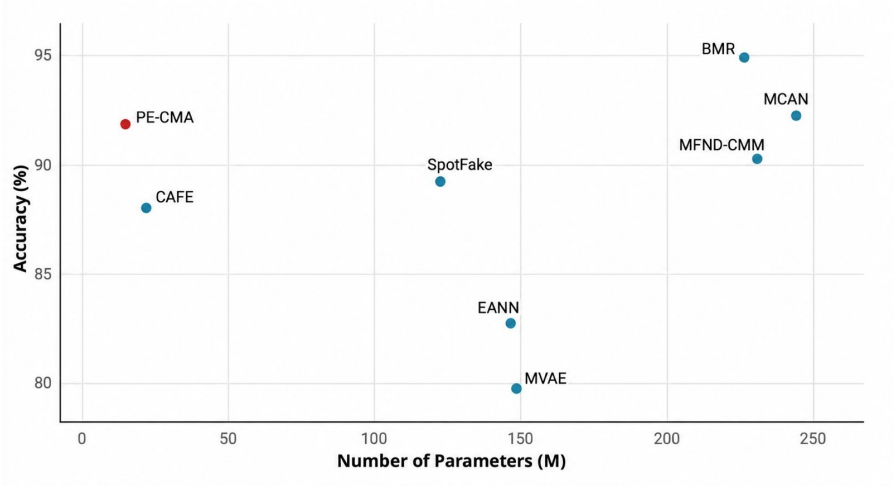


Figure 8. Trade-off between classification accuracy and model parameter footprint (in millions) across multimodal models. The proposed PE-CMA (red) achieves superior parameter efficiency relative to baseline methods (blue): CAFE [43], SpotFake [26], EANN [25], MVAE [27], MFND-CMM [10], BMR [33], and MCAN [41].

While baseline architectures like MVAE [27] or CAFE [43] report fast inference times, their benchmarks frequently rely on pre-extracted offline features rather than end-to-end processing. In contrast, CMCM and PE-CMA handle complete end-to-end forward passes through encoders, projection layers, and fusion mechanisms. These results confirm RQ1 and RQ3, demonstrating that lightweight multimodal designs achieve robust fake news detection with drastically reduced computational requirements.

6. Discussion

This work introduced two lightweight multimodal architectures—CMCM (Concatenation and Siamese variants) and PE-CMA—designed to balance classification performance and computational efficiency in fake news detection.

To answer our first research question regarding whether lightweight multimodal architectures can achieve competitive performance with reduced computational requirements, our evaluations demonstrate that heavy foundational models (e.g., BMR, MMBT, and MCAN) are not strictly necessary. While certain heavyweight baseline models report marginally higher absolute accuracy and F1-scores on the Fakeddit and Weibo datasets, they achieve these gains at the cost of massive parameter bloat, often exceeding 150M to 250M parameters. Consequently, the primary strength of the proposed CMCM and PE-CMA frameworks is not establishing a new absolute ceiling for raw detection metrics, but rather delivering an efficient accuracy–complexity trade-off. This is additionally supported by evaluations of frozen pre-trained encoders, which indicated that while large foundation models like CLIP can yield high accuracy on specific benchmarks, they demand significantly greater computational complexity, reinforcing the value of fine-tuned, lightweight encoder pairings like MobileNetV3 and TinyBERT.

To answer our second research question regarding whether explicitly modeling cross-modal inconsistency through a similarity branch improves fake news detection, we evaluated the CMCM-Siamese architecture on Fakeddit—a dataset that frequently features semantically unrelated or mismatched text–image pairs where CMCM-Siamese achieved the top accuracy of 92.29%. The ablation study demonstrated that removing the similarity branch caused a statistically significant drop in accuracy from 92.29% to 91.22% ($p < 0.001$), confirming that explicitly modeling cross-modal inconsistency is vital for detecting deceptive alignment. Conversely, on Weibo where misleading content manifests through subtler

semantic manipulations rather than outright visual-textual mismatch, the similarity branch provides less utility, leading to the selection of PE-CMA.

To answer our third research question regarding whether combining learned token pooling with lightweight cross-modal attention can reduce computational complexity while preserving performance, the PE-CMA architecture emerged as the superior model on the Weibo dataset, attaining 91.81% accuracy with only 15.33M parameters. Crucially, our ablation studies confirmed that removing the initial token pooling mechanism did not result in a statistically significant degradation of predictive accuracy. Rather than serving as a primary driver of classification performance, this finding validates the structural purpose of token pooling. While the absolute end-to-end computational cost of the model remains heavily dominated by the foundational processing requirements of the dense unimodal backbones, optimizing the cross-modal interaction phase is vital. By reducing sequence lengths from raw token counts down to 8 tokens per modality prior to cross-modal interaction, PE-CMA drastically reduces the localized computational burden within the attention block by mitigating the quadratic complexity of standard attention. This effectively prevents the fusion stage from becoming a severe computational bottleneck, rather than attempting to reduce the baseline cost of the feature extractors. Consequently, PE-CMA successfully preserves the contextual representation capacity necessary to outperform classical concatenation baselines, proving that attention-based multimodal interactions can be made computationally light through low-rank structural designs without sacrificing performance.

While these lightweight frameworks significantly lower the computational footprint compared to heavy unimodal backbones, it is crucial to explicitly bound their applicability and objectively acknowledge their architectural shortcomings. The CMCM framework is highly applicable to datasets characterized by blatant visual-textual mismatches. As our quantitative analysis demonstrates, Fakeddit exhibits a severe 18.93% mismatch rate (120,527 out of 636,836 samples, based on explicit annotations). This high baseline mismatch makes CMCM's strict consistency penalty highly effective on such data. However, its primary shortcoming is its dependence on the feature representation capacity of MobileNetV3-Large and TinyBERT. If these compact encoders fail to capture highly abstract semantic cues, the downstream similarity branch processes sub-optimal embeddings, while the strict auxiliary alignment losses risk over-regularizing the joint latent space in cases where manipulations are highly subtle.

Concurrently, the PE-CMA framework excels in environments requiring nuanced, attention-based semantic alignment. Our zero-shot Chinese-CLIP analysis, using a strict similarity threshold derived from the 5th percentile of the non-rumor training split, reveals that Weibo exhibits a notably lower mismatch rate of 10.64% (822 out of 7723 samples). This confirms that Weibo rumors frequently utilize misappropriated images that nonetheless maintain high surface-level semantic alignment, a nuance that PE-CMA is ideally suited to capture. Yet, its major shortcoming lies in the aggressive nature of the learned token pooling and low-rank query factorization. While this structural compression successfully mitigates quadratic complexity, it inherently risks obscuring micro-level or highly localized semantic anomalies (such as subtle image tampering) that might be discarded during the pooling phase. Consequently, utilizing these frameworks effectively requires calibrating the alignment loss weights, rank dimensions, and pooling factors based on the specific manipulative characteristics of the target data.

7. Conclusions

This research addressed the limitations of unimodal and computationally heavy multimodal fake news detection models by developing lightweight, efficient architectures supported by rigorous experimental evaluation.

This work demonstrates that lightweight multimodal designs utilizing efficient backbones like MobileNetV3-Large and TinyBERT achieve robust fake news detection performance while drastically reducing parameter size and GFLOPs compared to heavier state-of-the-art benchmarks. Explicitly modeling text–image discrepancies via a similarity measurement branch yields statistically significant performance gains, proving essential for datasets characterized by cross-modal inconsistencies. Additionally, integrating learned token pooling with lightweight cross-modal attention successfully lowers the computational complexity of attention mechanisms while preserving strong classification accuracy.

Despite these successes, the study's scope is subject to certain limitations. First, to avoid over-generalization, we emphasize that these findings are strictly bounded by the structural characteristics of the Fakeddit and Weibo datasets; the observed accuracy–complexity trade-offs may vary when applied to other social media platforms featuring different misinformation modalities, such as highly sophisticated deepfakes or purely AI-generated text. Second, as objectively analyzed, these models exhibit specific architectural shortcomings: CMCM may over-regularize subtle manipulations, while PE-CMA's aggressive token pooling risks discarding localized micro-anomalies.

The study faced logistical constraints, including linguistic variations necessitating different TinyBERT versions across datasets. While several key baselines were successfully reproduced for direct evaluation under unified conditions, the lack of publicly available source code for certain closed-source models restricted those specific comparisons to reported literature metrics. Future work will extend these frameworks by incorporating rich contextual metadata, user interaction patterns, and native video modalities to capture even more comprehensive signals of online misinformation.

Author Contributions: Conceptualization, H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; methodology, S.A. (Sarab AlMuhaideb), H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; software, H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; validation, H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; formal analysis, H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; writing—original draft preparation, S.A. (Sarab AlMuhaideb); writing—review and editing, S.A. (Sarab AlMuhaideb); visualization, H.A., J.A., S.A. (Sarah Aljuhani), D.A. and A.A.; supervision, S.A. (Sarab AlMuhaideb). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ongoing Research Funding program (ORF-2026-2189), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Ethical review and approval were waived for this study by the Permanent Committee for Scientific Research Ethics at King Saud University (Reference No: KSU-HE-25-1355, approved on 19/11/2025G, corresponding to 28/05/1447H) because the research utilizes open-source public datasets and Python code for deep learning techniques to detect fake news, which does not involve direct intervention with human participants or animal subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Fakeddit dataset used in this study is publicly available at https://github.com/entitize/Fakeddit (accessed on 28 September 2026). The Weibo dataset used in this study is available through the Event Adversarial Neural Networks (EANN) repository at https://github.com/yaqingwang/EANN-KDD18 (accessed on 28 September 2026).

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Munusamy, S.; Syasyila, K.; Shaari, A.H.; Pitchan, M.A.; Kamaluddin, M.R.; Jatnika, R. Psychological factors contributing to the creation and dissemination of fake news among social media users: A systematic review. *BMC Psychol.* **2024**, *12*, 673. [CrossRef] [PubMed]
2. Denniss, E.; Lindberg, R. Social media and the spread of misinformation: Infectious and a threat to public health. *Health Promot. Int.* **2025**, *40*, daaf023. [CrossRef] [PubMed]
3. Bhattacharjee, R.; Sufian, M.A.; Talha, M.; Ahmed, M.; Tasnim, F.; Sara, S.; Islam, K.M. Title-based fake news detection using LSTM. In Proceedings of the International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 20–22 December 2024; pp. 173–178. [CrossRef]
4. Sharma, P.; Sahu, R. Fake news detection using deep learning based approach. In Proceedings of the International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 10–11 August 2023; pp. 651–656. [CrossRef]
5. Salem, D.R.; Abdullah, A.A.; AlHabshy, A.A.; ElDahshan, K.A. Detecting Fake News Images Using a Hybrid CNN-LSTM Architecture. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 185–193 . [CrossRef]
6. Alqurashi, L.S.; AlMuraytib, S.; Qarout, R.; Zamzami, N. Fake Image Detection in Fake News Using Convolutional Neural Network (CNN). In Proceedings of the International Conference on Innovation in Artificial Intelligence and Internet of Things (AIIT), Jeddah, Saudi Arabia, 7–8 May 2025; pp. 1–8. [CrossRef]
7. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Learning rich features for image manipulation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1053–1061. [CrossRef]
8. Segura-Bedmar, I.; Alonso-Bartolome, S. Multimodal Fake News Detection. *Information* **2022**, *13*, 284. [CrossRef]
9. Uppada, S.K.; Patel, P.; Bhatnagar, S. An image and text-based multimodal model for detecting fake news in OSNs. *J. Intell. Inf. Syst.* **2023**, *61*, 367–393. [CrossRef] [PubMed]
10. Guo, Y.; Li, B.; Zhen, K.; Liu, J.; Li, G.; Wang, Q.; Liu, Y.J. Consistency-heterogeneity balanced fake news detection via cross-modal matching. *IEEE Trans. Artif. Intell.* **2025**, *6*, 1787–1796. [CrossRef]
11. Mura, D.A.; Usai, M.; Loddo, A.; Sanguinetti, M.; Zedda, L.; Di Ruberto, C.; Atzori, M. Is it fake or not? A comprehensive approach for multimodal fake news detection. *Online Soc. Networks Media* **2025**, *47*, 100314. [CrossRef]
12. Wang, H.; Yang, Y.; Han, S.; He, Z.; Yang, W. A Multimodal Fake News Detection Model Based on Cross-Image Semantic Fusion. In Proceedings of the 20th International Conference on Mobility, Sensing and Networking (MSN), Harbin, China, 20–22 December 2024; pp. 738–745. [CrossRef]
13. Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* **2021**, *58*, 102610. [CrossRef] [PubMed]
14. Shen, X.; Huang, M.; Hu, Z.; Cai, S.; Zhou, T. Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. *Front. Comput. Sci.* **2024**, *6*, 1473457. [CrossRef]
15. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv **2017**, arXiv:1704.04861. [CrossRef]
16. Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982. [CrossRef]
17. Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In Proceedings of the Findings of the Association for Computational Linguistics (EMNLP), Online, 16–20 November 2020; pp. 4163–4174. [CrossRef]
18. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5776–5788.
19. Nakamura, K.; Levy, S.; Wang, W.Y. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; pp. 6149–6157. Available online: https://aclanthology.org/2020.lrec-1.755/ (accessed on 27 September 2026).
20. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In Proceedings of the 25th ACM International Conference on Multimedia (MM), New York, NY, USA, 23–27 October 2017; pp. 795–816. [CrossRef]
21. Thota, A.; Tilak, P.; Ahluwalia, S.; Lohia, N. Fake News Detection: A Deep Learning Approach. *SMU Data Sci. Rev.* **2018**, *1*, 1–20 . Available online: https://scholar.smu.edu/datasciencereview/vol1/iss3/10 (accessed on 27 September 2026). [CrossRef]
22. Subhash, P.M.; Gupta, D.; Palaniswamy, S.; Venugopalan, M. Fake news detection using deep learning and transformer-based model. In Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; pp. 1–6. [CrossRef]

23. Kamble, R.A.; Pareek, V.; Jain, T. Fake news detection using machine learning and deep learning. In Proceedings of the International Conference on Smart Power Control Renewable Energy (ICSPCRE), Rourkela, India, 19–21 July 2024; pp. 1–6. [CrossRef]
24. Hu, T.; Lei, Y.; Wang, Y.; Ma, R.; Yin, X.; Liu, M.; Zhao, R. Beyond Words and Pictures: A Survey of Multimodal Fake News Detection. *Expert Syst.* **2026**, *43*, e70260. [CrossRef]
25. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. 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 (KDD), London, UK, 19–23 August 2018; pp. 849–857. [CrossRef]
26. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. SpotFake: A Multi-modal Framework for Fake News Detection. In Proceedings of the 5th International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 39–47. [CrossRef]
27. Khattar, D.; Singh, J.; Gupta, M.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the World Wide Web Conference (WWW), San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921. [CrossRef]
28. Luqman, M.; Faheem, M.; Ramay, W.Y.; Saeed, M.K.; Ahmad, M.B. Utilizing Ensemble Learning for Detecting Multi-Modal Fake News. *IEEE Access* **2024**, *12*, 15037–15049. [CrossRef]
29. Girón, A.; Huertas-Tato, J.; Camacho, D. Multimodal Visio-Lingual Content Analysis to Detect Fake Content on Reddit. In *Proceedings of the Intelligent Data Engineering and Automated Learning (IDEAL)*; Springer: Cham, Switzerland, 2025; pp. 143–154. [CrossRef]
30. Mohawesh, R.; Al-Bahadili, H.; Qaralleh, A.; Al-Dhaqm, A.; Al-Zoubi, A.A. Truth Be Told: A Multimodal Ensemble Approach for Enhanced Fake News Detection in Textual and Visual Media. *J. Big Data* **2025**, *12*, 197. [CrossRef]
31. Shen, L.; Long, Y.; Cai, X.; Razzak, I.; Chen, G.; Liu, K.; Jameel, S. GAMED: Knowledge Adaptive Multi-Experts Decoupling for Multimodal Fake News Detection. In Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM), Hannover, Germany, 10–14 March 2025; pp. 586–595. [CrossRef]
32. Singh, P.; Srivastava, R.; Rana, K.P.S.; Kumar, V. SEMI-FND: Stacked ensemble based multimodal inferencing framework for faster fake news detection. *Expert Syst. Appl.* **2023**, *215*, 119302. [CrossRef]
33. Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; Ge, S. Bootstrapping multi-view representations for fake news detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 5384–5392. [CrossRef]
34. Lin, S.Y.; Chen, Y.C.; Chang, Y.H.; Lo, S.H.; Chao, K.M. Text–image multimodal fusion model for enhanced fake news detection. *Sci. Prog.* **2024**, *107*, 1–31. [CrossRef] [PubMed]
35. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015. Available online: https://arxiv.org/abs/1409.1556 (accessed on 27 September 2026).
36. Hamed, S.K.; Aziz, M.J.A.; Yaakub, M.R. Improving data fusion for fake news detection: A hybrid fusion approach for unimodal and multimodal data. *IEEE Access* **2024**, *12*, 112412–112425. [CrossRef]
37. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
38. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of the 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020. Available online: https://arxiv.org/abs/2003.10555 (accessed on 27 September 2026).
39. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710. [CrossRef]
40. Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; Testuggine, D. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv* **2020**, arXiv.1909.02950. [CrossRef]
41. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. 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. [CrossRef]
42. Yadav, A.; Gaba, S.; Khan, H.; Budhiraja, I.; Singh, A.; Singh, K.K. ETMA: Efficient transformer-based multilevel attention framework for multimodal fake news detection. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 5015–5027. [CrossRef]
43. Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; Shang, L. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the ACM Web Conference 2022, New York, NY, USA, 25–29 April 2022; pp. 2897–2905. [CrossRef]

44. Tian, K.; Rao, G.; Wang, X.; Yu, M.; Zhang, J.; Zhang, L. CMFNThinker: A Novel Cross-source Multi-modal Fake News Detection Model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025), Piscataway, NJ, USA, 6–11 April 2025; p. 10889602. [CrossRef]
45. Shang, W.; Song, K.; Ji, J.; Yi, T.; Cai, J.; Li, X. Semantic Space Aligned Multimodal Fake News Detection. *Inf. Fusion* **2026**, *125*, 103469. [CrossRef]
46. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
47. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
48. Chakarverti, M.; Goswami, A.; Kaliyar, R.; Yadav, A. Improving fake news detection concatenating multimodal features with transformers-based deep learning models. *Online Soc. Netw. Media* **2026**, *53*, 100348. [CrossRef]
49. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
50. Gupta, A.; Singh, M. Multimodal fake news detection using semantic-visual alignment and attention fusion. *SN Comput. Sci.* **2026**, *7*, 305. [CrossRef]
51. Ji, Z. Fake news detection method based on multimodal fusion and generative adversarial network. *Syst. Soft Comput.* **2026**, *8*, 200477. [CrossRef]
52. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020; pp. 8440–8451. [CrossRef]
53. Liu, S.; Zeng, H.; Guo, J.; Zhang, C.; Lee, I.K.; Zhang, Y.; Li, X. Dual-stream multimodal shared prompt model for fake news detection. *Inf. Process. Manag.* **2026**, *63*, 104759. [CrossRef]
54. Liu, S.; Wang, Y.; Guo, J.; Zeng, H.; Lee, I.K.; Zhang, Y.; Li, X. Text-aware CLIP: A generalized method for deepfake and fake news detection. *IEEE Trans. Multimed.* **2026**, *28*, 8413–8426. [CrossRef]
55. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Red Hook, NY, USA, 8–14 December 2019; Volume 32, pp. 5754–5764. Available online: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html (accessed on 27 September 2026).
56. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28 July–2 August 2019; pp. 2978–2988. [CrossRef]
57. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. XLNet: Generalized autoregressive pretraining for language understanding. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Red Hook, NY, USA, 8–14 December 2019; Volume 32, pp. 5753–5763. Available online: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (accessed on 27 September 2026).
58. Adewumi, T.; Sabry, S.; Abid, N.; Liwicki, F.; Liwicki, M. T5 for hate speech, augmented data, and ensemble. *Sci* **2023**, *5*, 37. [CrossRef]
59. Abdullah, M.; Zan, H.; Javed, A.; Sohail, M.; Mamyrbayev, O.; Turysbek, Z.; Eshkiki, H.; Caraffini, F. A multimodal ensemble-based framework for detecting fake news using visual and textual features. *Mathematics* **2026**, *14*, 360. [CrossRef]
60. Yasir, S.M.; Kim, H. Lightweight Deepfake Detection Based on Multi-Feature Fusion. *Appl. Sci.* **2025**, *15*, 1954. [CrossRef]
61. Guo, H.; Wang, L.; Li, B.; Guo, Z. TinyDF: Tiny and effective model for deepfake detection. In *Proceedings of the Advanced Intelligent Computing Technology and Applications (ICIC)*; Huang, D.S.; Chen, W.; Pan, Y.; Chen, H., Eds.; Springer: Singapore, 2025; pp. 247–256. [CrossRef]
62. Sandhya.; Kashyap, A. A Light Weight Depthwise Separable Layer Optimized CNN Architecture for Object-Based Forgery Detection in Surveillance Videos. *Comput. J.* **2024**, *67*, 2270–2285. [CrossRef]
63. Liu, P.; Wang, X. A Lightweight Multi-modal Emotion Recognition Network Based on Multi-task Learning. In Proceedings of the International Conference on Neuromorphic Computing (ICNC), Wuhan, China, 15–17 October 2021; pp. 368–372. [CrossRef]
64. Lou, Q.; Hsu, Y.C.; Uzkent, B.; Hua, T.; Shen, Y.; Jin, H. Lite-MDETR: A Lightweight Multi-Modal Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12196–12205. [CrossRef]
65. Wang, T.; Chen, G.; Zhang, X.; Liu, C.; Wang, J.; Tan, X.; Zhou, W.; He, C. LMFNet: Lightweight Multimodal Fusion Network for high-resolution remote sensing image segmentation. *Pattern Recognit.* **2025**, *164*, 111579. [CrossRef]

66. Alsinglawi, B.; McCarthy, C.; Webb, S.; Fluke, C.; Saidy, N.T. A Lightweight Large Vision-Language Model for Multimodal Medical Images. In *Proceedings of the International Conference Advanced Information Networking and Applications (AINA)*; Springer: Cham, Switzerland, 2026; pp. 219–230. [CrossRef]
67. Xiao, Y.; Wu, J.; Ma, C. GA-GhostNet: A Lightweight CNN Model for Identifying Pests and Diseases Using a Gated Multi-Scale Coordinate Attention Mechanism. *Eng. Lett.* **2024**, *32*, 1281–1290.
68. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese Neural Networks for One-Shot Image Recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2. Available online: https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf (accessed on 27 September 2026).
69. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; pp. 850–865. [CrossRef]
70. Jiang, W.; Zhang, S.; He, K. Siamese Transformer Networks for Few-Shot Image Classification. *Appl. Soft Comput.* **2026**, *194*, 114966. [CrossRef]
71. Han, S.; Gao, L.; Wu, Y.; Wei, T.; Wang, M.; Cheng, X. SiamSMN: Siamese Cross-Modality Fusion Network for Object Tracking. *Information* **2024**, *15*, 418. [CrossRef]
72. Ye, P.; Xiao, G.; Liu, J. Multimodal Features Alignment for Vision–Language Object Tracking. *Remote Sens.* **2024**, *16*, 1168. [CrossRef]
73. Huertas-Tato, J.; Koutlis, C.; Papadopoulos, S.; Camacho, D.; Kompatsiaris, I. A CLIP-Based Siamese Approach for Meme Classification. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–8. Learning, Lille, France, 2015 [CrossRef]
74. Lu, Y.; Yao, N. A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network. *Sci. Rep.* **2025**, *15*, 20544. [CrossRef] [PubMed]
75. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763. Available online: https://proceedings.mlr.press/v139/radford21a.html (accessed on 27 September 2026).
76. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324. [CrossRef]
77. Zoph, B.; Le, Q.V. Neural Architecture Search with Reinforcement Learning. *arxiv* **2017**, arXiv:1611.01578. [CrossRef]
78. Yang, T.J.; Howard, A.; Chen, B.; Zhang, X.; Go, A.; Sandler, M.; Sze, V.; Adam, H. NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–304. [CrossRef]
79. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
80. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2023**, arXiv:1606.08415. [CrossRef]
81. Gupta, K.; Ajanthan, T.; van den Hengel, A.; Gould, S. Understanding and Improving the Role of Projection Head in Self-Supervised Learning. *arXiv* **2022**, arXiv:2212.11491. [CrossRef]
82. Wang, F.; Xiang, X.; Cheng, J.; Yuille, A.L. NormFace: L_2 Hypersphere Embedding for Face Verification. In Proceedings of the ACM International Conference on Multimedia (MM), Mountain View, CA, USA, 23–27 October 2017; pp. 1041–1049. [CrossRef]
83. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
84. Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; Carreira, J. Perceiver: General Perception with Iterative Attention. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 4651–4664. Available online: https://proceedings.mlr.press/v139/jaegle21a.html (accessed on 27 September 2026).
85. Shou, Y.; Liu, H.; Cao, X.; Meng, D.; Dong, B. A Low-Rank Matching Attention Based Cross-Modal Feature Fusion Method for Conversational Emotion Recognition. *IEEE Trans. Affect. Comput.* **2024**, *16*, 1177–1189. [CrossRef]
86. Ryoo, M.S.; Piergiovanni, A.; Arnab, A.; Dehghani, M.; Angelova, A. TokenLearner: What Can 8 Learned Tokens Do for Images and Videos? *arXiv* **2022**, arXiv.2106.11297. [CrossRef]
87. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set Transformer: A Framework for Attention-Based Permutation-Invariant Neural Networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 3744–3753. Available online: https://proceedings.mlr.press/v97/lee19d.html (accessed on 27 September 2026).
88. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.P. Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2247–2256. [CrossRef]

89. Pushshift. Pushshift API. 2025. Available online: https://pushshift.io/ (accessed on 17 May 2026).
90. Ma, F. *Information Communication*; Morgan and Claypool Publishers: San Rafael, CA, USA, 2015. [CrossRef]
91. Xinhua News Agency. Official News Agency of China. 2025 Available online: http://www.xinhuanet.com/ (accessed on 30 November 2025).
92. Lu, W.; Tong, Y.; Ye, Z. DAMMFND: Domain-Aware Multimodal Multi-View Fake News Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 32036–32044. [CrossRef]
93. Stevens, E.; Antiga, L.; Viehmann, T. *Deep Learning with PyTorch*; Manning Publications: Shelter Island, NY, USA, 2020; ISBN 978-1-61729-526-3
94. Shorten, C.; Khoshgoftaar, T. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
95. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144. [CrossRef]
96. Chaerul Haviana, S.F.; Mulyono, S.; Badie'Ah. The Effects of Stopwords, Stemming, and Lemmatization on Pre-trained Language Models for Text Classification: A Technical Study. In Proceedings of the International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Jember, Indonesia, 20–21 September 2023; pp. 521–527. [CrossRef]
97. Zhou, X.; Wu, J.; Zafarani, R. SAFE: Similarity-aware multi-modal fake news detection. In Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2020), Singapore, 11–14 May 2020; pp. 354–367. [CrossRef]
98. Mehta, S.; Rastegari, M. Separable Self-Attention for Mobile Vision Transformers. *arXiv* **2022**, arXiv:2206.02680. [CrossRef]
99. Face, H. Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX. 2024. Available online: https://huggingface.co/docs/transformers (accessed on 27 September 2026).
100. Wightman, R. PyTorch Image Models (timm). 2024. Available online: https://github.com/huggingface/pytorch-image-models (accessed on 27 September 2026).
101. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv* **2020**, arXiv:1909.11942. [CrossRef]
102. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2020**, arXiv:1910.01108. [CrossRef]
103. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv* **2021**, arXiv:2111.09543. [CrossRef]
104. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. [CrossRef]
105. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [CrossRef]
106. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101. [CrossRef]
107. Vujović, Z. Classification Model Evaluation Metrics. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 599–606. [CrossRef]
108. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 27 September 2026).
109. Paoletti, M.E.; Haut, J.M.; Tao, X.; Plaza, J.; Plaza, A. FLOP-Reduction Through Memory Allocations Within CNN for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5938–5952. [CrossRef]
110. Reddi, V.J.; Cheng, C.; Kanter, D.; Mattson, P.; Schmuelling, G.; Wu, C.J.; Anderson, B.; Breughe, M.; Charlebois, M.; Chou, W.; et al. MLPerf Inference Benchmark. In Proceedings of the Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 30 May–3 June 2020; pp. 446–459. [CrossRef]
111. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
112. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
113. Vasu, P.K.A.; Pouransari, H.; Faghri, F.; Vemulapalli, R.; Tuzel, O. MobileCLIP: Fast Image-Text Models Through Multi-Modal Reinforced Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15963–15974. [CrossRef]
114. Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.F.; Abdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv* **2025**, arXiv:2502.14786. [CrossRef]
115. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation Through Attention. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; Volume 139, pp. 10347–10357. [CrossRef]

116. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features Without Supervision. *arXiv* **2024**, arXiv.2304.07193. [CrossRef]
117. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2704–2713. [CrossRef]
118. NVIDIA Corporation. *NVIDIA A100 Tensor Core GPU Architecture*; Technical report; NVIDIA Corporation: Santa Clara, CA, USA, 2020. Available online: https://www.megware.com/fileadmin/user_upload/LandingPage%20NVIDIA/nvidia-ampere-architecture-whitepaper.pdf (accessed on 4 October 2026).
119. PyTorch Contributors. Reproducibility: PyTorch 2.14 Documentation. 2026. Available online: https://docs.pytorch.org/docs/2.14/notes/randomness.html (accessed on 4 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.