Interpreting Multimodal Fake News Detection Models: An Experimental Study of Performance Factors and Modality Contributions

Noha A. Saad Eldien[1], Wael H. Gomaa[2], Khaled T. Wassif[3], Hanaa Bayomi[4]
Faculty of Computers and Artificial Intelligence, Cairo University, Egypt[1,3,4]
Faculty of Computer Science, MSA University, Egypt[1]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt[2]
Faculty of Computers and Information Technology, Future University, Egypt[4]

***Abstract*—The widespread dissemination of multimodal misinformation requires models that can reason across textual and visual content while remaining interpretable. However, many existing multimodal fusion approaches implicitly assume uniform modality reliability, providing limited transparency into modality contributions. This study introduces TweFuse-W, a lightweight multimodal framework for fine-grained fake-news detection that reframes multimodal fusion as a modality reliability estimation problem, rather than merely merging modalities or explicitly modeling their interactions. TweFuse-W integrates BERTweet-based textual representations with Swin Transformer visual features using a sample-conditioned, learnable weighted-sum gate operating at the modality level, producing global reliability weights without cross-attention overhead. By explicitly parameterizing modality contributions during inference, the proposed approach provides intrinsic interpretability. Experiments on the six-class Fakeddit dataset show that TweFuse-W achieves a macro-F1 score of 0.838, outperforming simple concatenation (macro-F1 = 0.820). Analysis of the learned modality weights confirms meaningful interpretability, with textual representations dominating in Satire, Misleading, False Connection, and Imposter Content ($\alpha_T = 0.57$–0.62), while visual cues exert greater influence in Manipulated Content ($\alpha_V = 0.51$). Overall, these findings demonstrate that adaptive modality weighting enhances both predictive performance and model transparency, serving as a lightweight and interpretable *complementary fusion strategy* for multimodal fake-news detection.**

***Keywords*—Multimodal fake news detection; modality reliability modeling; adaptive fusion; interpretable fusion; lightweight multimodal models**

I. Introduction

The widespread circulation of misinformation on social media poses a significant threat to information credibility and public trust. As digital communication continues to expand, deceptive content increasingly influences political decision-making, public health, and societal stability worldwide [1]. Social media platforms such as Twitter and Facebook contribute to this issue by prioritizing engagement, which often results in misleading content spreading faster than verified information [2].

Misinformation today is rarely limited to text alone. False claims are often presented as multimodal posts in which images and captions work together to strengthen misleading messages. These posts may include manipulated images, misleading visual contexts, or mismatched image–text pairs that make false content appear more credible and harder to detect [3]. This growing reliance on multimodal presentation highlights the importance of reasoning across both text and images for effective fake-news detection.

Beyond a binary distinction between real and fake, misinformation arises from diverse underlying intentions that shape how content is generated and disseminated [4], [5]. The Fakeddit dataset [6] used in this study reflects this real-world complexity through a six-class taxonomy encompassing categories such as Satire, Misleading Content, False Connection, Imposter Content, and Manipulated Content . Accurately distinguishing among these fine-grained categories requires not only multimodal inputs but also an understanding of how textual and visual signals contribute differently across misinformation types.

Existing multimodal fake-news detection approaches largely fall into two broad fusion paradigms. The first relies on simple fusion strategies, such as concatenation or summation, which treat modalities independently and implicitly assume uniform modality importance across samples. The second paradigm employs attention-based or interaction-driven fusion mechanisms that explicitly model cross-modal dependencies, often at the expense of increased architectural complexity and computational overhead. While interaction-based methods can capture fine-grained feature alignments, both paradigms provide limited transparency into modality contributions and rarely analyze modality reliability, which is particularly critical in fine-grained misinformation classification settings.

In parallel, the choice of unimodal encoders plays a crucial role in multimodal misinformation detection. General-purpose language models such as BERT [7], RoBERTa [8], and DistilBERT [9] are not explicitly optimized for informal, noisy social-media text. In contrast, domain-specific language models such as BERTweet [10] are better suited to capturing platform-specific linguistic patterns, slang, and abbreviated expressions commonly found in misinformation posts. These observations reveal a key research gap: the need for an adaptive multimodal fusion framework that efficiently handles social-media language while providing interpretable insights into modality contributions.

Multimodal fusion in misinformation detection need not be limited to either static modality merging or explicit interaction modeling. An alternative perspective is to view fusion as a *modality reliability estimation problem*, where the relative contribution of each modality is adaptively regulated on a per-sample basis. This formulation emphasizes global, modality-level trust signals rather than localized feature interactions, enabling adaptive fusion behavior alongside transparent interpretation of multimodal reasoning.

Guided by this formulation, **TweFuse-W** is introduced as a lightweight multimodal fusion framework that employs a sample-conditioned, learnable weighted-sum gating mechanism operating at the modality level. Rather than relying on cross-attention or heuristic confidence measures, TweFuse-W produces explicit modality reliability weights as part of the forward inference process. This design maintains computational efficiency while providing intrinsic interpretability, allowing direct analysis of modality dominance across fine-grained misinformation categories.

The main contributions of this study are summarized as follows:

- An adaptive, sample-conditioned weighted-sum fusion mechanism is introduced to dynamically model modality importance, extending beyond fixed late-fusion strategies and manually tuned weighting schemes.
- A domain-aware multimodal architecture is presented that integrates BERTweet for textual encoding with a Swin Transformer for visual representation, tailored to the characteristics of social-media misinformation.
- A quantitative interpretability analysis is conducted based on class-level distributions of learned modality weights, providing insight into how textual and visual signals contribute across fine-grained misinformation categories.
- A controlled experimental evaluation is performed across multiple textual encoders under identical training conditions, enabling fair and systematic assessment of fusion strategies and their impact on detection performance.

The remainder of this paper is organized as follows. Section II reviews related work on multimodal fake-news detection, fusion strategies, and domain-specific language modeling. Section III describes the TweFuse-W methodology, including the adaptive weighted-sum fusion mechanism and the multimodal encoding pipeline. Section IV outlines the experimental setup, dataset configuration, training protocol, and evaluation metrics. Section V presents the experimental results and the analysis of learned modality-weight patterns. Section VI discusses the implications of these findings. Finally, Section VII concludes the paper and outlines directions for future research.

II. Literature Review

The increasing impact of rapidly spreading fake news has motivated multimodal detection approaches that jointly analyze text and images. Early studies primarily adopted static fusion strategies, while later work shifted toward interaction-driven fusion using attention mechanisms to capture richer cross-modal dependencies.Recent work has also emphasized feature-level fusion strategies that aim to balance effectiveness and model complexity [11]. Table I provides a structured qualitative overview of representative multimodal fake-news detection approaches, organized according to the fusion taxonomy illustrated in Fig. 1. It summarizes the diversity of datasets, task definitions, and fusion strategies explored in prior work. The table presents a qualitative survey of existing multimodal fake-news detection approaches, emphasizing methodological trends rather than quantitative performance comparison

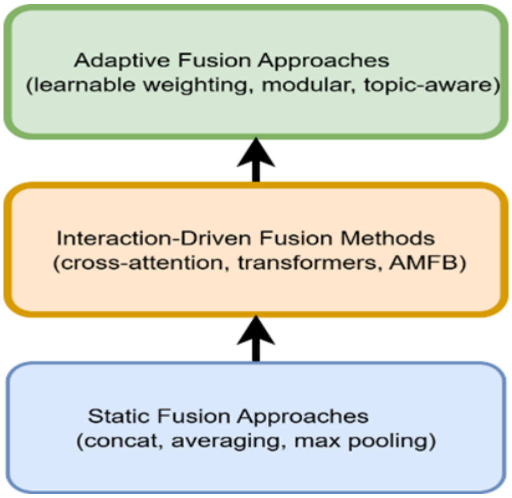
Adaptive Fusion Approaches
(learnable weighting, modular, topic-aware)
Interaction-Driven Fusion Methods
(cross-attention, transformers, AMFB)
Static Fusion Approaches
(concat, averaging, max pooling)

Fig. 1. Conceptual taxonomy of multimodal fusion strategies, ranging from static fusion approaches to interaction-driven and adaptive, learnable fusion methods.

A. Static Fusion Approaches

Early multimodal work processed each modality separately and merged them later using simple operations like concatenation or pooling. In study [12], visual CNNs were paired with textual encoders and evaluated with different pooling strategies, with max pooling performing best. Similarly, SpotFake [13] fused BERT and VGG-19 via simple concatenation, yielding moderate performance gains. However, such methods offer limited inter-modal interaction and cannot model modality importance. Later work like [14] used dot-product similarity to align text and image features, but still treated modalities as static and non-adaptive.

B. Interactive Fusion

Recent work has moved to interaction-driven fusion using attention and transformers to model deeper cross-modal relationships beyond simple concatenation. Transformer-based fusion now leads the field for its scalability and strong cross-modal modelling. Examples include DeBERTNext [15]FB [16], and ETMA [17], which combine transformer encoders with advanced attention mechanisms to achieve high accuracy. Recent surveys [18][19] also confirm transformers as the dominant approach in multimodal misinformation detection.

C. Adaptive Fusion

Parallel work has focused on adaptive and interpretable fusion, such as learnable intra/inter-modal weighting [18] and modular designs like M3DUSA [20]. Topic-aware transformers

[19] also improve fine-grained classification. However, surveys [21], [22] note that most multimodal models still offer limited transparency regarding modality contributions. Recent multimodal approaches for fine-grained fake news detection have explored diverse fusion strategies, including attention-based and modular architectures [4], [20], [23]. Efficiency-aware and domain-adaptive multimodal frameworks have recently gained attention, particularly in low-resource or dynamically evolving misinformation settings [24], [25].

D. Research Gaps and Problem Definition

Despite substantial progress from static fusion strategies to transformer-based multimodal models, several fundamental limitations remain. Simple fusion methods, such as concatenation or summation, implicitly assume uniform modality reliability and lack mechanisms to down-weight unreliable or misleading modalities on a per-sample basis. In contrast, attention-based architectures explicitly model cross-modal interactions but incur significant computational overhead and often provide limited transparency into modality-level contributions. Moreover, the majority of existing studies focus on binary real–fake classification, leaving fine-grained misinformation categories and systematic analysis of modality reliability largely underexplored.

This work addresses these gaps by formulating multimodal fake-news detection as a *modality reliability estimation problem* in a fine-grained classification setting. The task is to classify a multimodal social-media post, consisting of textual content and an associated image, into one of six misinformation categories, while explicitly accounting for the relative reliability of each modality.

Formally, a multimodal input sample is defined as

$$x = \{x^{(t)}, x^{(i)}\}, \tag{1}$$

where, $x^{(t)}$ denotes the textual input and $x^{(i)}$ denotes the corresponding visual input. The target label is defined as

$$y \in \{1, \dots, 6\}, \tag{2}$$

representing the six fine-grained misinformation classes. The objective is to learn a prediction function

$$f\Big(x^{(t)}, x^{(i)}\Big) \rightarrow y, \tag{3}$$

where, the contribution of each modality is regulated through *adaptive, sample-conditioned reliability weights*, denoted by $w_t(x)$ for text and $w_i(x)$ for images. These weights are designed to capture the relative trustworthiness of each modality for a given input instance, rather than merely facilitating feature aggregation or cross-modal interaction.

III. Methodology

This study introduces **TweFuse-W**, an adaptive multimodal framework for fine-grained fake-news detection that prioritizes intrinsic interpretability. The framework integrates domain-specific textual representations extracted using *BERTweet* with visual features obtained from a *Swin Transformer*, enabling complementary multimodal reasoning. Rather than relying on static fusion rules or explicit cross-modal interaction modeling, TweFuse-W employs a learnable weighted-sum fusion mechanism that dynamically regulates the contribution of each modality on a per-sample basis. This formulation explicitly reflects the varying reliability of textual and visual information across different misinformation categories, while maintaining a lightweight fusion design. Fig. 2 illustrates the overall processing pipeline, including multimodal feature extraction, adaptive modality reliability weighting, and final classification. The remainder of this section details the dataset and preprocessing steps, unimodal feature extraction, adaptive fusion mechanism, and classification procedure.

A. Dataset Preparation

The *Fakeddit* dataset [12] is selected due to its aligned text–image pairs and fine-grained annotation schema, making it well suited for evaluating adaptive multimodal fusion strategies. The dataset contains approximately one million Reddit posts annotated under binary, three-class, and six-class settings. This study focuses on the highly imbalanced six-class taxonomy illustrated in Fig. 3.

Primary performance results are reported using the official Fakeddit train, validation, and test splits. In addition, a stratified 20% subset of the dataset is employed exclusively for controlled ablation and interpretability analyses. This subset enables repeated experimental runs under reduced computational cost while faithfully preserving the original class distribution.

Consistent with prior work, the full raw *Fakeddit* dataset is rarely used due to missing or unrecoverable image URLs [18], a limitation also observed during dataset inspection. Consequently, a clean, stratified 20% subset is adopted to preserve the original class distribution and content characteristics.

A Chi-square goodness-of-fit test confirms that the class-label proportions in the reduced subset do not differ significantly from those of the full dataset ($\chi^2 = 0.0004$, $p > 0.99$). Linguistic equivalence is assessed by comparing headline-length distributions using the Kolmogorov–Smirnov test and Cohen's d, yielding statistically indistinguishable results (KS = 0.0027, $p > 0.99$; $d = -0.0064$). Mean headline lengths are nearly identical (7.47 ± 5.59 vs. 7.51 ± 5.75 words). Visual consistency is also preserved, with comparable image availability rates (99.76% vs. 99.77%).

These analyses demonstrate that the reduced subset faithfully preserves the statistical, linguistic, and visual characteristics of the complete *Fakeddit* dataset. The exact subset indices and sampling scripts will be released publicly to ensure full reproducibility.

B. Data Preprocessing

The reduced subset is used for comparative experiments and in-depth analyses, including the investigation of learned modality-weight patterns, while the complete dataset is utilized exclusively to validate the proposed framework at scale.

A unified preprocessing pipeline is applied to both modalities to ensure consistency, reproducibility, and leakage-free

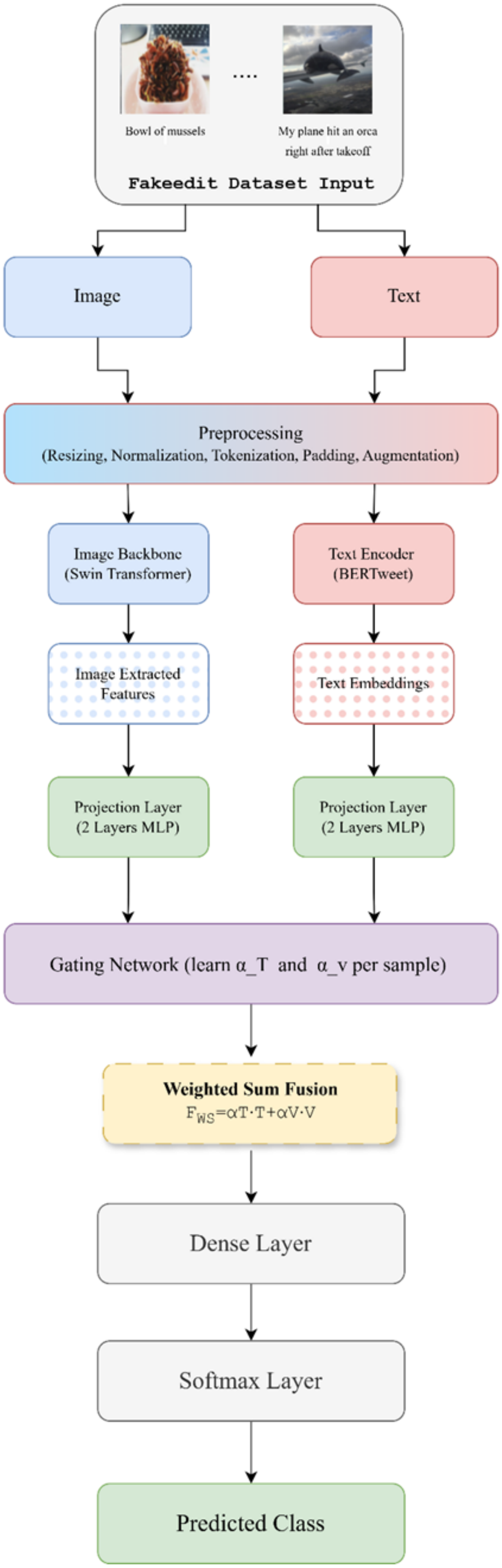
Bowl of mussels
....
My plane hit an orca right after takeoff
Fakeedit Dataset Input
Image
Text
Preprocessing
(Resizing, Normalization, Tokenization, Padding, Augmentation)
Image Backbone (Swin Transformer)
Text Encoder (BERTweet)
Image Extracted Features
Text Embeddings
Projection Layer (2 Layers MLP)
Projection Layer (2 Layers MLP)
Gating Network (learn α_T and α_v per sample)
Weighted Sum Fusion
F_{WS}=αT·T+αV·V
Dense Layer
Softmax Layer
Predicted Class

Fig. 2. Architecture of the *TweFuse-W* framework for multimodal fake news detection.

TABLE I. SUMMARY OF REPRESENTATIVE MULTIMODAL FAKE NEWS DETECTION APPROACHES

Study & Year	Dataset	Task Type	Model Used	Fusion Strategy	Notes
[12] (2020)	Fakeddit	Binary, 6-class	Basic CNN + BERT / InferSent	Concatenation, mean & max pooling	Simple pooling; weak inter-modal interaction
[26] (2023)	Twitter, Weibo	Binary	ResNet (image), BERT (text)	Contrastive + attention	Binary classification task only
[21] (2024)	Fakeddit, MFND	3-class	ResNet50, BERT	Graph-based fusion	High memory consumption
[27] (2024)	Twitter, Weibo	Binary	ERNIE-M, AlexNet	Multi-head attention	Complex multilayer fusion; limited to binary task
[18] (2024)	Reduced Fakeddit & Gossip datasets	Binary	BERT, ResNet50	Early, joint, and late fusion	Dataset reduction without verifying preserved class distribution
[19] (2025)	PolitiFact, GossipCop	Binary	BERTweet, ViT	Topic-aware multimodal cross-attention fusion	No explicit handling of data imbalance

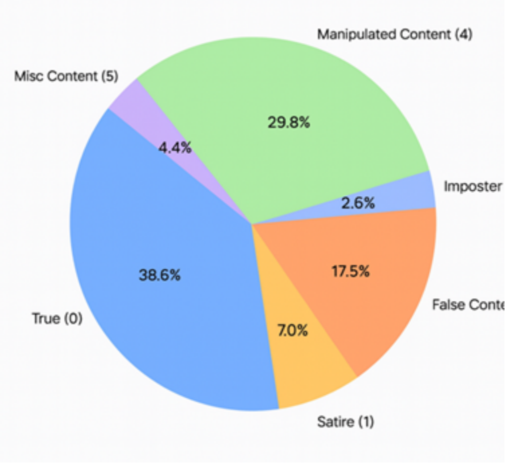

Fig. 3. Class distribution across the Fakeddit dataset categories, highlighting the dominance of True and Manipulated Content instances.

evaluation. All preprocessing and augmentation steps are performed independently within each data split.

For the textual modality, headlines are tokenized using the corresponding encoder tokenizer (BERT, DistilBERT, or BERTweet) with matched normalization and token-handling settings. Sequences are padded or truncated to a maximum length of 80 tokens based on corpus statistics. Lightweight textual augmentation is applied during training to improve generalization, with semantic consistency enforced by retaining only augmented samples with cosine similarity greater than or equal to 0.90 relative to the original text.

For the visual modality, images are resized to 224×224 pixels and normalized using ImageNet statistics. Corrupted or unreadable samples are removed. Controlled visual augmentations, including horizontal flipping, random cropping, and mild color jittering, are applied during training.

C. Feature Extraction

In all experiments, multimodal representations are extracted using frozen pretrained encoders to ensure training stability, fair comparison across model variants, and full reproducibility. Freezing the encoders also isolates the contribution of the fusion mechanism by eliminating confounding effects introduced by encoder fine-tuning.

For the textual modality, each headline is processed using the language model under evaluation (BERT, DistilBERT, or BERTweet), with all encoder parameters kept fixed. The final `[CLS]` token representation is used as the sentence-level embedding. To standardize dimensionality across textual encoders, each embedding is passed through a lightweight projection network that maps it into a unified 768-dimensional latent space.

For the visual modality, images are encoded using a frozen Swin Transformer, and the pooled backbone output is used as the visual representation. Since the native dimensionality and scale of visual features differ from those of textual embeddings, a matching projection layer is applied to map the visual features into the same 768-dimensional latent space. This alignment prevents artifacts arising from mismatched feature distributions and ensures compatibility during multimodal fusion.

Projection heads for both modalities are implemented as two-layer feed-forward blocks (`Linear` → `ReLU` → `Linear`), mapping the native encoder outputs to the shared latent space. Specifically, the Swin Transformer visual features (1024-dimensional) are projected to 768 dimensions, while the BERTweet embeddings (768-dimensional) are passed through an equivalent projection head to standardize representational capacity. Projection weights are not shared across modalities, allowing each branch to adapt independently while maintaining dimensional comparability. This design preserves modality-specific representation independence while ensuring alignment for subsequent fusion operations.

D. Fusion

A central objective of this study is to examine how adaptive modality weighting influences multimodal reasoning and interpretability in fine-grained fake-news detection. As a reference baseline, standard feature concatenation is included, as it represents a widely adopted early-fusion strategy in multimodal modeling. Given textual features T and visual features V, concatenation is defined as:

$$F_{\text{concat}} = [T; V] \tag{4}$$

As shown in Eq. (4), simple concatenation forms a fixed joint representation by stacking text and image embeddings into a single high-dimensional vector. Although straightforward to implement, this formulation assigns equal importance to both modalities for all samples and provides no mechanism to adapt modality contributions or interpret fusion behavior when one modality is more informative or reliable than the other.

To enable sample-conditioned fusion, TweFuse-W employs an adaptive weighted-sum fusion gate that learns instance-specific modality contributions. The fused representation is computed as:

$$F_{\text{ws}} = \alpha_t T + \alpha_v V, \qquad \alpha_v = 1 - \alpha_t \tag{5}$$

as defined in Eq. (5), where α_t and α_v denote the relative importance assigned to the textual and visual modalities, respectively. These modality weights are predicted by a lightweight gating network that takes the concatenated feature vector $[T; V]$ as input. The gating function produces two logits that are normalized using a softmax operation, as expressed in Eq. (6):

$$[\alpha_t, \alpha_v] = \text{Softmax}\left(g([T; V])\right) \tag{6}$$

where, $g(\cdot)$ denotes a two-layer feed-forward network with the structure `Linear`–`ReLU`–`Linear`. The bias terms of the output layer are initialized to zero, yielding an approximately uniform modality weighting at the start of training. During optimization, the gating network learns instance-specific reliability shifts based on the content of each multimodal sample.

In implementation, the gating network receives the concatenated 1536-dimensional feature vector $[T; V]$ and is defined as `Linear`(1536→128) → `ReLU` → `Linear`(128→2). This design enables dynamic modulation of modality contributions while maintaining a lightweight architecture that intentionally avoids deep interaction or cross-attention modules. As a result, observed performance differences can be attributed primarily to adaptive modality weighting rather than increased model capacity.

Beyond performance gains, the explicit modality weights (α_t, α_v) provide a transparent and interpretable signal. These weights enable analysis of modality-reliance behavior at both the sample and category levels, offering insights into how different misinformation classes depend on textual or visual evidence.

To support interpretability analysis, the modality weights (α_t, α_v) are recorded during test-time inference after the fusion step and before classification. For each test sample, the corresponding sample index, ground-truth label, predicted label, and modality weights are stored. These values are subsequently aggregated to compute sample-level distributions and class-wise statistics, facilitating systematic analysis of modality-reliance patterns across fine-grained misinformation categories.

E. Classification Layer

After obtaining the fused feature vector F from the fusion module, the model predicts the news category using a linear softmax classifier. The probability distribution over the six Fakeddit classes is computed according to Eq. (7):

$$\hat{y} = \text{Softmax}\left(W_c F + b_c\right) \tag{7}$$

As shown in Eq. (7), W_c and b_c denote the learnable parameters of the classification layer. This formulation provides a direct mapping from the fused multimodal representation to class logits and ensures a simple and consistent prediction interface across all fusion variants.

Model training employs the standard cross-entropy loss, which is applied uniformly across all experimental configurations. This consistent optimization setup ensures that any observed performance differences can be attributed to the fusion strategy rather than to variations in the classifier design or loss function. As a result, fair comparison is maintained between baseline fusion methods and the proposed adaptive weighted-sum fusion approach.

All reported evaluation metrics are directly reproducible from the corresponding training logs and evaluation scripts.

F. Evaluation and Experiments

This section describes the experimental setup, evaluation protocol, and performance analysis of TweFuse-W on the six-class Fakeddit benchmark.

G. Evaluation Criteria

Evaluation emphasizes macro-averaged metrics to account for the severe class imbalance inherent in the six-class Fakeddit taxonomy (see Fig. 3). Macro-F1 is adopted as the primary performance indicator, complemented by macro-precision and macro-recall. In addition, class-wise F1 scores are reported to analyze performance variations across individual misinformation categories.

Beyond predictive performance, evaluation explicitly examines interpretability by analyzing the modality weights produced by the adaptive fusion mechanism defined in Eq. (6). During inference, the learned textual and visual weights are recorded for each test sample and aggregated at the class level. This aggregation enables quantitative assessment of how modality importance varies across misinformation types, providing a fine-grained view of fusion behavior that remains largely underexplored in prior multimodal fake-news detection studies.

H. Experimental Setup, Model Complexity, and Reproducibility

All experiments are conducted in the Kaggle cloud environment using NVIDIA GPUs with CUDA support. The implementation is based on PyTorch 2.0, with HuggingFace Transformers used for textual encoders and the `timm` library employed for the Swin Transformer visual backbone.

Following the dataset validation procedure described in Section III, primary performance evaluation is conducted using the official Fakeddit train, validation, and test splits. A stratified 20% subset is employed exclusively for controlled ablation studies and interpretability analyses. Although the dataset provides predefined 80/10/10 train–validation–test partitions, the subset is sampled independently within each split to preserve original boundaries and class distributions while reducing computational cost and preventing data leakage.

All experiments are executed with a fixed random seed (42) to ensure deterministic behavior. To enable fair comparison, all baseline models and TweFuse-W share an identical training configuration. Optimization is performed using AdamW with a learning rate of 3×10^{-5} and a linear warm-up followed by decay. Models are trained for up to 20 epochs with a batch size of 32, using early stopping based on validation loss with a patience of four epochs.

To ensure that observed performance differences arise from fusion behavior rather than architectural capacity, model complexity is explicitly controlled. As summarized in Table II, all pretrained encoder backbones are frozen during training, and trainable parameters are limited to lightweight projection heads, the adaptive fusion gate defined in Eq. (6), and the linear classifier defined in Eq. (7). The total number of trainable parameters remains below 2.8 million, confirming that TweFuse-W maintains a lightweight design and enabling fair comparison with concatenation-based baselines.

This standardized training protocol ensures that observed performance differences can be attributed to the fusion strategy rather than to variations in optimization, initialization, or hyperparameter tuning.

For reproducibility, the complete implementation executed on Kaggle—including preprocessing, training, and evaluation scripts—is publicly available at:

I. Comparative Baselines

To contextualize the proposed adaptive fusion strategy, TweFuse-W is evaluated against multimodal baselines that reflect common practices in image–text misinformation detection. All models share the same preprocessing pipeline, frozen encoder backbones, projection layers, and training configuration, ensuring that performance differences can be attributed specifically to the fusion mechanism rather than architectural or optimization discrepancies.

J. Multimodal Concatenation Baselines

Three multimodal baselines are constructed using standard concatenation-based fusion, as defined in Eq. (4). Each baseline combines a fixed Swin Transformer visual encoder with one textual encoder—BERT, DistilBERT, or BERTweet. Textual `[CLS]` embeddings and pooled visual features are projected into a shared latent space and concatenated prior to classification. This design provides a controlled reference that applies uniform, sample-invariant modality weighting and lacks adaptive fusion behavior, while maintaining comparable model complexity to TweFuse-W (Table II).

K. Adaptive Weighted-Sum Fusion (TweFuse-W)

TweFuse-W introduces adaptivity through the lightweight gating mechanism described in Eq. (6), which operates on the joint multimodal representation. For each input sample, the gate predicts normalized weights for the textual and visual modalities, which are then applied to compute the weighted-sum fusion defined in Eq. (5). This formulation enables dynamic, sample-conditioned modulation of modality importance without introducing deep cross-modal interaction modules.

By comparing TweFuse-W against concatenation-based baselines under comparable parameter budgets (Table II), the experimental analysis isolates the contribution of adaptive modality weighting. In addition, the learned modality weights are logged and analyzed at the class level, enabling systematic investigation of modality-reliance patterns across fine-grained misinformation categories and directly supporting the interpretability objectives of the proposed framework.

IV. Results

All primary performance results reported in this section are obtained from evaluation on the official Fakeddit test split. TweFuse-W is evaluated on the six-class Fakeddit benchmark following the official evaluation protocol. As reported in Table VI, TweFuse-W achieves a Macro-F1 score of 0.838, outperforming the standard concatenation baseline 0.820. This result demonstrates that adaptive, sample-conditioned modality weighting provides a consistent performance advantage over uniform fusion strategies under standard benchmark conditions.

To gain deeper insight into class-level behavior, robustness across text encoders, and fusion interpretability, additional experiments are conducted on a stratified 20% subset of the Fakeddit dataset. Under this controlled setting, TweFuse-W achieves a Macro-F1 of 0.854, with particularly strong performance on categories characterized by clearer linguistic or visual cues, such as *Manipulated Content* (F1 = 0.963) and *Misleading* (F1 = 0.885), as detailed in Table V. Competitive performance is also maintained on more ambiguous categories, including *False Connection* (F1 = 0.746) and *Imposter Content* (F1 = 0.783). These subset-based results are used exclusively for methodological validation and interpretability analysis rather than for primary benchmark comparison.

As summarized in Table III, TweFuse-W consistently outperforms concatenation-based baselines across all evaluated textual encoders. The largest improvement is observed for the BERTweet–Swin configuration, where Macro-F1 improves from 0.833 to 0.854, highlighting the benefit of combining domain-specific text representations with adaptive multimodal fusion. Consistent performance trends across encoders are further illustrated in Fig. 4.

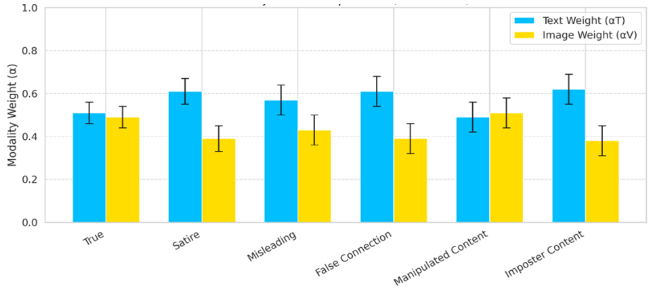

Fig. 4. Learned modality weights (α_T and α_V) across the six Fakeddit classes.

A modest decrease in absolute performance is observed when transitioning from the stratified subset to the full official dataset. This behavior is expected due to the increased diversity, noise, and difficulty of the complete benchmark. Importantly, TweFuse-W continues to outperform the concatenation baseline under this stricter evaluation setting, confirming that the observed gains generalize beyond subset-based experiments.

Detailed class-wise precision, recall, and F1-scores are reported in Table V. To analyze fusion behavior, the learned modality weights (α_T, α_V) are recorded for each test sample and aggregated per class. The resulting statistics, presented in Table IV, reveal interpretable modality-reliance patterns.

TABLE II. TRAINABLE COMPONENTS AND PARAMETER COUNTS OF THE TWEFUSE-W FRAMEWORK

Module	Description	Input → Output	Params
Text projection head	Two-layer MLP (`Linear–ReLU–Linear`) on BERTweet	768 → 768 → 768	1,181,184
Image projection head	Two-layer MLP (`Linear–ReLU–Linear`) on Swin	1024 → 768 → 768	1,377,792
Gating network	Fusion gate (`Linear–ReLU–Linear–Softmax`)	1536 → 128 → 2	196,994
Classification head	Linear classifier	768 → 6	4,614
Total trainable parameters	–	–	**2,760,584**

Text-dominant categories such as *Satire*, *Misleading*, *False Connection*, and *Imposter Content* rely more heavily on textual cues, whereas *True* and *Manipulated Content* exhibit more balanced text–image contributions. The low variance of modality weights across samples indicates stable and consistent fusion behavior.

TABLE III. MACRO-F1 COMPARISON BETWEEN CONCATENATION AND ADAPTIVE WEIGHTED-SUM FUSION

Model	Concat	Weighted Sum
BERT + Swin	0.824	0.840
DistilBERT + Swin	0.810	0.832
BERTweet + Swin	0.833	0.854

TABLE IV. LEARNED TEXT (α_T) AND CISION (α_V) WEIGHTS USING TWEFUSE-W

Class	α_T (mean ± std)	α_V (mean ± std)
True	0.51 ± 0.05	0.49 ± 0.05
Satire	0.61 ± 0.06	0.39 ± 0.06
Misleading	0.57 ± 0.07	0.43 ± 0.07
False Connection	0.61 ± 0.07	0.39 ± 0.07
Manipulated Content	0.49 ± 0.07	0.51 ± 0.07
Imposter Content	0.62 ± 0.07	0.38 ± 0.07

TABLE V. CLASS-WISE PRECISION, RECALL, AND F1-SCORES OF TWEFUSE-W

Class	Precision	Recall	F1-score
True	0.861	0.854	0.858
Satire	0.868	0.852	0.860
Misleading	0.894	0.877	0.885
False Connection	0.770	0.723	0.746
Manipulated Content	0.965	0.961	0.963
Imposter Content	0.801	0.767	0.783

TABLE VI. PERFORMANCE ON THE OFFICIAL FAKEDDIT TEST SPLIT (SIX-CLASS TASK)

Model	Macro-F1
Concat (BERTweet + Swin)	0.820
TweFuse-W	0.838

V. DISCUSSION

The experimental results demonstrate the complementary importance of domain-aware language modeling and adaptive multimodal fusion for fine-grained fake-news detection. Configurations using BERTweet consistently achieve the highest Macro-F1 scores (Table III), confirming that domain-specific pretraining is critical for handling the informal and context-dependent language prevalent in social media content.

Beyond textual modeling, the results highlight the essential role of visual information. Visually grounded categories such as *Manipulated Content* are inadequately addressed by uniform fusion strategies that assign equal importance to all modalities. The adaptive weighted-sum fusion directly addresses this limitation by learning instance-level modality importance.

The confusion matrix in Fig. 5 further supports the interpretability analysis by revealing class-specific error patterns. Visually grounded categories, particularly *Manipulated Content*, exhibit strong diagonal dominance, indicating reliable predictions when visual cues are informative. In contrast, most misclassifications occur among semantically related, text-dominant categories, a trend that aligns with the modality-weight distributions reported in Table IV. This consistency suggests that the adaptive fusion mechanism assigns modality importance in a manner that reflects the underlying characteristics of each misinformation category.

Ablation results in Table VII show that removing adaptive fusion leads to a substantial performance drop, while replacing BERTweet with DistilBERT further degrades performance. These findings validate the individual contributions of both domain-aware text encoding and adaptive fusion. These findings align with recent observations that emphasize the growing challenge posed by increasingly sophisticated and generative forms of misinformation [28].

VI. CONCLUSION AND FUTURE WORK

This study introduces TweFuse-W, an adaptive multimodal framework for fine-grained fake-news detection that reconsiders how modality contributions should be modeled and interpreted. By complementing existing fusion paradigms, TweFuse-W shows that lightweight, sample-conditioned modality weighting can achieve strong and consistent performance across diverse misinformation categories. More importantly, the framework reframes multimodal fusion from an interpretability-centered perspective, enabling systematic analysis of modality reliability and providing insight into how textual and visual evidence contribute to fine-grained misinformation decisions.

The experimental results indicate that adaptive weighted-sum fusion provides a practical and effective alternative to uniform concatenation and more complex fusion strategies. Across multiple encoder configurations and evaluation settings, TweFuse-W consistently improves Macro-F1 performance while preserving a compact and interpretable design. These findings highlight a favorable efficiency–interpretability balance in multimodal fake-news detection.

Beyond performance, this work offers a reliability-aware perspective on multimodal fusion. Analysis of the learned modality weights shows that modality importance varies systematically across misinformation categories, indicating that modality reliability is input-dependent rather than fixed. This

TABLE VII. Ablation Study of TweFuse-W on the Fakeddit Dataset

Configuration	Description	Macro-F1 (%)	Δ (%)
Full Model	BERTweet + Swin (adaptive weighted sum)	0.854	—
w/o Weighted Sum	Concatenation fusion	0.833	−2.46
w/ DistilBERT	Replace BERTweet encoder	0.832	−2.57
Text Only	BERTweet encoder only	0.775	−9.26
Image Only	Swin Transformer only	0.700	−18.03

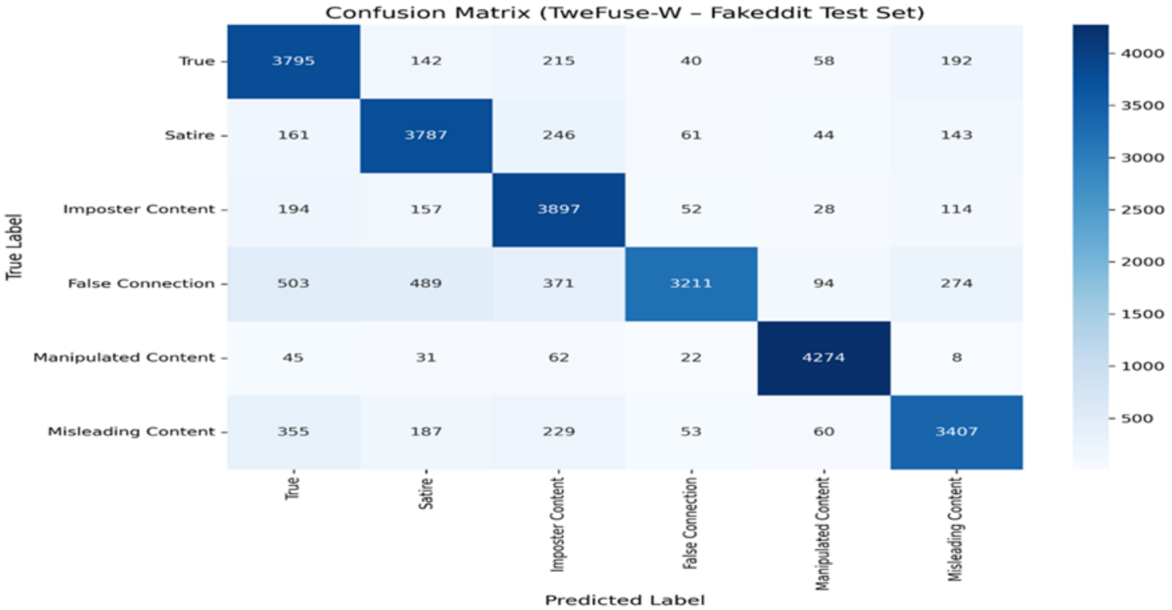

Fig. 5. Confusion matrix of TweFuse-W on the six-class fakeddit dataset.

highlights the value of modeling fusion as an adaptive reliability estimation process rather than a static feature combination.

While the framework emphasizes simplicity and transparency, future work may integrate adaptive weighting with richer interaction mechanisms and extend evaluation to larger-scale and multi-seed settings. Overall, this study advances multimodal fake-news detection by showing that adaptive, interpretable modality weighting is both effective and conceptually meaningful, offering a complementary perspective to existing fusion approaches.

References

[1] V. Fionda, "Logic-based analysis of fake news diffusion on social media," *Social Network Analysis and Mining*, vol. 15, no. 1, p. 59, 2025.

[2] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection models: highlighting the factors affecting model performance and the prominent techniques used," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023.

[3] C. V. Mohan and N. Chinnasamy, "An automated multimodal hybrid system for web content fact-checking based on bert language model and convolutional neural network," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 16, 2024.

[4] L. Qian, R. Xu, and Z. Zhou, "Mrdca: a multimodal approach for fine-grained fake news detection through integration of roberta and densenet based upon fusion mechanism of co-attention," *Annals of Operations Research*, vol. 348, no. 1, pp. 257–278, 2025.

[5] S. M. Dwivedi and S. B. Wankhade, "Deep learning based semantic model for multimodal fake news detection," *International Journal of Intelligent Engineering & Systems*, vol. 17, no. 1, 2024.

[6] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854*, 2019.

[7] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: what we know about how bert works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, and D. Chen, "Roberta: a robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[10] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: a pre-trained language model for english tweets," in *Proc. EMNLP System Demonstrations*, 2020, pp. 9–14.

[11] F. Kou, B. Wang, H. Li, C. Zhu, L. Shi, J. Zhang, and L. Qi, "Potential features fusion network for multimodal fake news detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[12] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection," in *Proc. 12th Language Resources and Evaluation Conference*, 2020, pp. 6149–6157.

[13] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh, "Spotfake: a multi-modal framework for fake news detection," in *Proc. IEEE 5th International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 39–47.

[14] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing & Management*, vol. 58, no. 5, p. 102610, 2021.

[15] K. Saha and Z. Kobti, "Debertnext: a multimodal fake news detection framework," in *Proc. International Conference on Computational Science*, 2023, pp. 348–356.

[16] R. Kumari and A. Ekbal, “Amfb: attention-based multimodal factorized bilinear pooling for multimodal fake news detection,” *Expert Systems with Applications*, vol. 184, p. 115412, 2021.

[17] A. Yadav, S. Gaba, H. Khan, I. Budhiraja, A. Singh, and K. K. Singh, “Etma: efficient transformer-based multilevel attention framework for multimodal fake news detection,” *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5015–5027, 2023.

[18] S. Y. Lin, Y. C. Chen, Y. H. Chang, S. H. Lo, and K. M. Chao, “Text–image multimodal fusion model for enhanced fake news detection,” *Science Progress*, vol. 107, no. 4, p. 00368504241292685, 2024.

[19] R. Cantini, C. Cosentino, I. Kilanioti, F. Marozzo, and D. Talia, “Unmasking deception: a topic-oriented multimodal approach to uncover false information on social media,” *Machine Learning*, vol. 114, no. 1, p. 13, 2025.

[20] L. Martirano, C. Comito, M. Guarascio, F. S. Pisani, and P. Zicari, “M3dusa: a modular multi-modal deep fusion architecture for fake news detection on social media,” *Social Network Analysis and Mining*, vol. 15, no. 1, p. 53, 2025.

[21] M. Nasser, N. I. Arshad, A. Ali, H. Alhussian, F. Saeed, A. Da’u, and I. Nafea, “A systematic review of multimodal fake news detection on social media using deep learning models,” *Results in Engineering*, vol. 26, p. 104752, 2025.

[22] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, “A review of fake news detection models: Highlighting the factors affecting model performance and the prominent techniques used,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023.

[23] R. Mohawesh, I. Obaidat, A. A. AlQarni, A. A. Aljubailan, M. A. Al-Shannaq, H. B. Salameh, A. Al-Yousef, A. A. Saifan, S. M. Alkhushayni, and S. Maqsood, “Truth be told: a multimodal ensemble approach for enhanced fake news detection in textual and visual media,” *Journal of Big Data*, vol. 12, no. 1, p. 197, 2025.

[24] I. Ni’mah, R. Wijayanti, A. Santosa, A. Jarin, T. Sampurno, M. T. Uliniansyah, M. Fang, V. Menkovski, and M. Pechenizkiy, “A simple contrastive embedding framework for low-resource fake news detection,” *Neural Computing and Applications*, vol. 37, no. 26, pp. 21 407–21 433, 2025. [Online]. Available: https://doi.org/10.1007/s00521-025-11467-0

[25] X. Wang, J. Meng, D. Zhao, X. Meng, and H. Sun, “Fake news detection based on multi-modal domain adaptation,” *Neural Computing and Applications*, vol. 37, no. 7, pp. 5781–5793, 2025. [Online]. Available: https://doi.org/10.1007/s00521-024-10896-7

[26] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, “Cross-modal contrastive learning for multimodal fake news detection,” in *Proc. 31st ACM International Conference on Multimedia*, 2023, pp. 5696–5704.

[27] S. Y. Lin, Y. C. Chen, Y. H. Chang, S. H. Lo, and K. M. Chao, “Text–image multimodal fusion model for enhanced fake news detection,” *Science Progress*, vol. 107, no. 4, p. 00368504241292685, 2024.

[28] S. Kumar, S. Sai, V. Chamola, A. Gaur, C. Agarwal, K. Huang, and A. Hussain, “Peeping into the future: Understanding and combating generative ai-based fake news,” *Cognitive Computation*, vol. 17, no. 3, p. 103, 2025.