What Improves Multimodal Misinformation Detection? Answers from a Large-Scale Empirical Study

Akshit Sharma and Prashant W. Patil
CVPR Lab, MFSDSAI
Indian Institute of Technology Guwahati
akshitsharma.rs@gmail.com pwpatil@iitg.ac.in

Abstract

Multimodal misinformation is increasingly crafted to look convincing by pairing a textual claim with an image that appears to "prove" it. Yet in practice, building effective detectors often hinges on a small set of design choices that are rarely examined in a controlled way. In this paper, we conduct a large-scale study of multimodal design choices for misinformation detection with over 3,375 experiments-spanning three benchmark datasets and a broad range of pre-trained vision and language backbones. Through systematic comparisons and targeted robustness analyses, we distill practical guidance on which design choices help, when do they fail silently, and what aspects of the pipeline most strongly shape model behavior, answering 4 key Research Questions (RQs). We aim to provide a reliable foundation for designing stronger and more dependable multimodal misinformation detection systems, thus contributing to the broader research community.

1 Introduction

Misinformation on social media and online news is increasingly multimodal: the core claim is expressed through text, while an accompanying image is often interpreted as corroborating evidence. In practice, however, the image may be irrelevant, out-of-context, weakly related, or intentionally manipulated, yet still persuasive enough to mislead readers. This makes credibility assessment difficult for unimodal systems, since text-only models cannot verify visual evidence, while image-only models cannot interpret what is being claimed. As a result, multimodal misinformation classification has become a standard framing for fake-news detection, rumor detection, and image–text inconsistency analysis (Abdali et al., 2024; Comito et al., 2023; Jin et al., 2017; Luo et al., 2021). A large body of work addresses this setting, but the design space is now broader than fusion alone. Att-RNN (Jin et al., 2017) integrates textual, visual, and social-context signals; EANN (Wang et al., 2018) targets event-specific bias; SpotFake (Singhal et al., 2019) and SpotFake+ (Singhal et al., 2020) show the strength of transfer-learning pipelines built from pre-trained encoders; MVAE (Khattar et al., 2019) studies shared latent spaces; and models such as SAFE (Zhou et al., 2020), MCNN (Xue et al., 2021), and recent out-of-context image–text verification work such as NewsCLIPpings (Luo et al., 2021) and SNIFFER (Qi et al., 2024) explicitly target cross-modal mismatch. Adjacent image–text classification tasks expose similar design tensions. In multimodal meme classification, CLIP-based methods investigate explicit cross-modal interaction and lightweight adaptation of pretrained representations (Kumar and Nandakumar, 2022; Shah et al., 2024), while more recent work explores keyword-guided representation learning (Sharma and Patil, 2026a) and local–global pivotal-cue grounding with frozen CLIP features (Sharma and Patil, 2026b). Although these tasks target harm, hate, or sarcasm rather than veracity, they motivate the same broader question of how encoder choice, fusion, and modality reliance shape robust image–text classification.

Across this literature, however, several recurring design factors are typically entangled: the visual backbone, the text backbone, the level of multimodal interaction, the downstream head, and the model's dependence on each modality. Yet a basic question remains unresolved: *when it comes to multimodal misinformation classification, which of these design factors actually contribute to model performance and which don't?* Prior work often varies several factors at once, or reports only the best benchmark-specific configuration, making it difficult to isolate what actually drives performance (Abdali et al., 2024; Comito et al., 2023; Chen et al., 2023). This matters because many practical pipelines rely on frozen pre-trained en-

Proceedings of the 10th Widening NLP Workshop, pages 37–47
October 28, 2026 ©2026 Association for Computational Linguistics

coders and lightweight heads, where small design decisions can dominate the final result. In this paper, we conduct a controlled study of multimodal misinformation detection that isolates each design component and evaluate the resulting benefits, trade-offs at scale. Concretely, we define three fusion strategies based on the abstraction level being combined—*early* fusion (feature-level), *mid* fusion (score-level), and *late* fusion (decision-level)—and evaluate them as part of the design factors under a unified protocol. We then address four key research questions:

- **RQ1. Generalized Performance Trends:** Which end-to-end performance patterns are seen to remain stable across different large-scale multimodal misinformation benchmarks?
- **RQ2. Modality Complementarity:** When do text and image contribute complementary evidence, and when does one modality dominate when it comes to multimodal misinformation detection?
- **RQ3. Impact of Each Component on Overall Performance:** Which pipeline components-visual encoder, text encoder, fusion strategy, or classifier, exert the most reliable effect on the classification performance?
- **RQ4. Benchmark-Shift Generalization:** To what extent do strong in-benchmark design choices and observed performance trends generalize under transfer to unseen misinformation benchmarks?

Across the three benchmark datasets, we conduct a total of 3,375 experiments- consisting of 675 in-domain model evaluations, 1,350 modality-drop evaluations, and 1,350 cross-dataset transfer evaluations. **Contributions.**

1. **RQ-based Analysis:** We provide a unified empirical analysis of performance trends, modality dependence, component robustness, and transfer under benchmark shift, providing answers to the 4 key research questions.
2. **Evidence at scale:** 675 + 1,350 + 1,350 = 3,375 controlled experiments across three benchmark datasets a unified protocol and diverse backbone combinations, supporting reproducible and broadly applicable conclusions.
3. **Statistically grounded conclusions:** We report confidence intervals, paired hypothesis tests, and transfer analyses that separate benchmark-stable findings from dataset-contingent effects.

2 Methodology

2.1 Problem Formulation

Let $\mathcal{D} = \{(x_i^{\text{img}}, x_i^{\text{text}}, y_i)\}_{i=1}^N$ be a multimodal misinformation dataset, where x_i^{img} is an image, x_i^{text} is the associated text, and $y_i \in \{1, \ldots, C\}$ is the class label. We learn a classifier

$$f : \mathcal{X}_{\text{img}} \times \mathcal{X}_{\text{text}} \to \{1, \ldots, C\}$$

using a modular pipeline with a vision encoder $\phi_v : \mathcal{X}_{\text{img}} \to \mathbb{R}^{d_v}$ and a text encoder $\phi_t : \mathcal{X}_{\text{text}} \to \mathbb{R}^{d_t}$:

$$z^{\text{img}} = \phi_v(x^{\text{img}}) \in \mathbb{R}^{d_v},$$
$$z^{\text{text}} = \phi_t(x^{\text{text}}) \in \mathbb{R}^{d_t}.$$

Fusion combines $(z^{\text{img}}, z^{\text{text}})$ into a joint representation, which is then passed to the final classifier.

2.2 Encoders and Classifiers Used

2.2.1 Vision Encoders

We instantiate ϕ_v using five pre-trained vision backbones that span both convolutional and transformer-based families: ResNet-50 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2021), CLIP ViT-B/32 (Radford et al., 2021), SigLIP (Zhai et al., 2023), and VGG-16 (Simonyan and Zisserman, 2015). This set covers standard CNNs, vanilla vision transformers, and image encoders trained with image–text pretraining objectives.

This selection is included to make the visual encoder a controlled experimental factor rather than a single default backbone. VGG-16 and ResNet-50 represent canonical convolutional architectures, with VGG-16 providing a plain deep CNN baseline and ResNet-50 representing residual CNNs. ViT-B/16 introduces a transformer-based visual encoder that operates over image patches rather than convolutional feature maps. CLIP ViT-B/32 and SigLIP add vision towers trained through paired image–text supervision, which is especially relevant for misinformation detection because the visual evidence must be interpreted relative to a textual claim. Holding the downstream pipeline fixed while varying these encoders allows us to test whether performance is driven mainly by convolutional versus transformer architecture, by unimodal versus

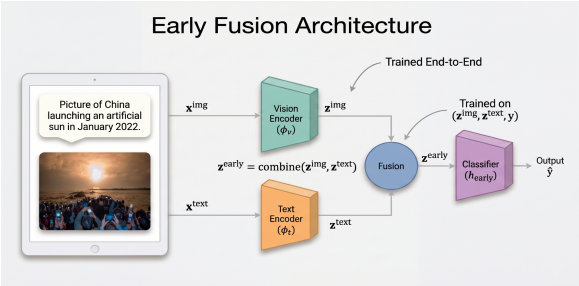

Figure 1: Figure showcasing the early fusion architecture used as part of our experiments

multimodal pretraining, or by the specific visual representation learned by each backbone.

2.2.2 Text Encoders

We instantiate ϕ_t using five pre-trained language backbones: BERT-base (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), SBERT (Reimers and Gurevych, 2019), CLIP-Text (Radford et al., 2021), and SigLIP-Text (Zhai et al., 2023). Together, these models cover standard bidirectional transformers, distilled encoders, sentence-oriented encoders, and text towers trained jointly with images.

This set is included to make the text encoder a separate controlled axis of the study. BERT-base provides a standard bidirectional Transformer baseline for contextual token-level language representations, while DistilBERT tests whether a smaller distilled encoder preserves enough semantic signal for the same downstream task. SBERT represents sentence-level representation learning, which is useful because misinformation examples are classified from complete claims or captions rather than isolated tokens. CLIP-Text and SigLIP-Text represent text towers trained under image–text alignment objectives, making them natural counterparts to multimodal vision encoders. Comparing these choices allows us to distinguish gains from general language understanding, sentence-level semantic embedding quality, and multimodal alignment.

2.2.3 Classifiers

We use three lightweight heads on top of frozen multimodal representations: logistic regression (Cox, 1958), random forest (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016). This set provides a clean comparison between a standardized linear classifier, a bagging-based ensemble, and a boosting-based ensemble. Including logistic regression is important in our setting because frozen pre-trained representations can often be evaluated effectively with simple linear probes (Kornblith et al., 2019), making a strong linear baseline difficult to beat.

These classifiers are included because the goal is to evaluate the representations and fusion strategies rather than to hide their effects behind a large trainable prediction head. Logistic regression provides the lowest-capacity linear probe: if it performs well, the frozen multimodal representation already exposes task-relevant information in a nearly linearly separable form. Random forest adds non-linear decision boundaries through bagged decision trees, while XGBoost provides a stronger boosting-based alternative that can capture feature interactions more aggressively. Using these three lightweight heads therefore gives an interpretable range of classifier capacity while keeping the main comparison focused on the encoders, fusion level, and modality dependence.

2.2.4 Fusion Architectures

Early fusion performs **feature-level** fusion by concatenating the image and text embeddings:

$$z^{\text{early}} = [z^{\text{img}}; z^{\text{text}}] \in \mathbb{R}^{d_v+d_t},$$
$$\hat{y} = h_{\text{early}}(z^{\text{early}}).$$

This is a single-stage scheme with no intermediate unimodal decision layer.

Mid fusion is a **score-level** two-stage scheme. We first train unimodal classifiers h^{img} and h^{text} and extract unimodal score representations $\ell^{\text{img}}, \ell^{\text{text}} \in \mathbb{R}^C$. These are concatenated and passed to a final fusion classifier:

$$z^{\text{mid}} = [\ell^{\text{img}}; \ell^{\text{text}}] \in \mathbb{R}^{2C}, \qquad \hat{y} = h_{\text{mid}}(z^{\text{mid}}).$$

Late fusion is a **decision-level** two-stage scheme. We first obtain unimodal probabilities $p^{\text{img}}, p^{\text{text}} \in [0,1]^C$, concatenate them, and train a final classifier on the resulting vector:

$$z^{\text{late}} = [p^{\text{img}}; p^{\text{text}}] \in \mathbb{R}^{2C}, \qquad \hat{y} = h_{\text{late}}(z^{\text{late}}).$$

2.3 Benchmark Setup

We evaluate on three large-scale multimodal misinformation benchmarks: MMFakeBench, MiRAGeNews, and VLDBench (Liu et al., 2025; Huang et al., 2024; Raza et al., 2026). MMFakeBench contains 11,000 image–text pairs, MiRAGeNews contains 15,000 pairs across its training, validation, and test partitions, and VLDBench contains 31,339 article–image pairs. Together, they cover mixed-source content, AI-generated multimodal news, and vision–language disinformation.

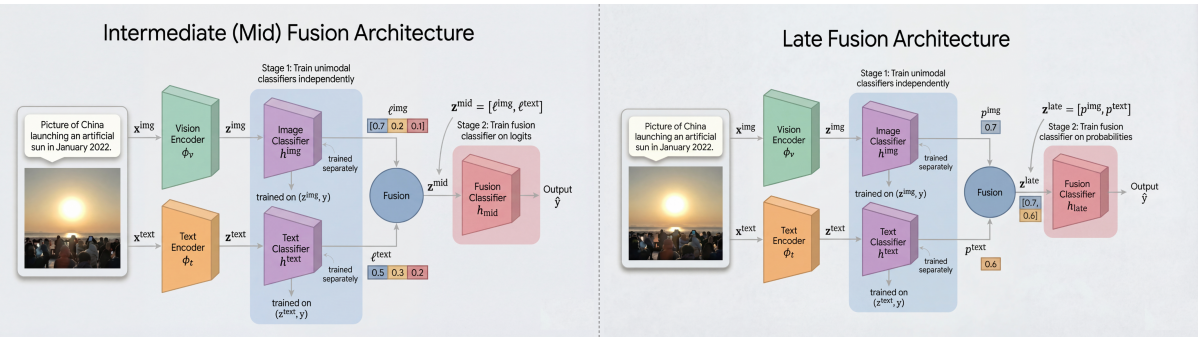

Figure 2: Figure showcasing the mid and late fusion architecture used as part of our experiments

For consistent supervised evaluation, we construct 80/20 train–test partitions for all three datasets, resulting in 8,800/2,200 examples for MMFakeBench, 12,000/3,000 for MiRAGeNews, and 25,071/6,268 for VLDBench. For MiRAGeNews, we remove 841 records with missing images—673 from the training partition and 168 from the test partition—leaving 14,159 instances (11,327/2,832). The final study comprises 56,498 examples: 11,000 from MMFakeBench, 14,159 from MiRAGeNews, and 31,339 from VLDBench. For cross-dataset transfer, we train on the constructed source training partition and evaluate on the target test partition.

2.4 Implementation Details

All encoders are frozen, and only the downstream classifier heads are trained. Logistic regression uses standardized inputs, the `lbfgs` solver, and `max_iter`=2000; random forest uses 300 trees; and XGBoost uses 300 trees, a 0.05 learning rate, maximum depth 6, row and column subsampling of 0.8, histogram-based tree construction, and a binary logistic objective. Feature extraction uses batch sizes of 16 and 32 for vision and text, respectively, with a maximum text length of 64. All experiments run on an NVIDIA RTX A4000 GPU with 16 GB VRAM.

Mid and late fusion use leakage-free stacking. For each outer training split, the unimodal branches generate out-of-fold predictions or features using shuffled 5-fold `StratifiedKFold` with a fixed random seed. The fusion classifier is trained only on these outputs, ensuring that no example both trains a unimodal model and contributes its in-sample prediction to the fusion model. The unimodal branches are then retrained on the full outer training split to generate outputs for the held-out test split, which is never used for training, calibration, or model selection. Cross-dataset transfer follows the same protocol: out-of-fold features and the fusion model are learned only from the source training set, and the frozen source-trained pipeline is evaluated directly on the target test set.

Statistical analysis. For all four research questions, we report Macro-Precision, AUROC, and Accuracy with 95% confidence intervals. Macro-Precision is the primary metric because false-positive credibility flags can be costly, while AUROC and Accuracy assess ranking quality and operating-point behavior. Aggregate comparisons over matched configurations use bootstrap confidence intervals. Paired comparisons use Wilcoxon signed-rank tests, with Holm correction for multiple pairwise tests; multi-level component comparisons use Friedman tests followed by Holm-corrected Wilcoxon post-hoc tests. We report Wilson score intervals for proportions, such as the frequency with which early fusion wins, and use exact McNemar tests for selected instance-level comparisons.

3 Results and Discussion

We organize the results by the four research questions. Unless otherwise noted, all table entries are mean percentages over matched configurations, and comparative claims such as "best," "strongest," or "default" refer to Macro-Precision, with AUROC and Accuracy reported as secondary diagnostics.

3.1 RQ1: Generalized Performance Trends

Based on Table 2, one in-domain trend is exceptionally stable: **early fusion is the only strategy that is best on average across all three datasets and metrics**. Over 225 matched in-domain configurations, it achieves 76.02% Macro-Precision (95% CI [74.58, 77.50]), 82.92% AUROC ([81.46, 84.39]), and 76.87% Accuracy ([75.51, 78.26]), compared with 74.23 / 81.64 / 76.04 for mid fusion and 74.09 / 81.37 / 75.91 for late fusion. Its Macro-Precision advantage is significant over both mid and late fusion ($p = 1.00 \times 10^{-8}$ and 5.45×10^{-11}), and

it wins 136/225 matched blocks, or 60.4% with Wilson interval [53.9, 66.6].

Classifier rankings support exclusion rather than a single universal winner. **Random forest is the weakest head across all three datasets and metrics**, while logistic regression and XGBoost form the competitive pair in Table 1. Logistic regression is nominally best in Macro-Precision at 76.42%, but is tied with XGBoost at 76.34% ($p = 0.160$). XGBoost achieves the highest AUROC at 84.07%, significantly above logistic regression at 83.33% ($p = 0.0022$), whereas logistic regression has the highest Accuracy at 78.08%, with no significant difference from XGBoost's 77.93% ($p = 0.175$). XGBoost leads on MMFakeBench and VLDBench, while logistic regression leads on MiRAGeNews. Thus, random forest should be avoided, and the final choice should be between logistic regression and XGBoost depending on whether Macro-Precision or AUROC is prioritized.

The single-dataset evidence is strongest on VLDBench and MMFakeBench, where early fusion improves Macro-Precision by roughly two to three points over the best alternative, while preserving the same ordering on MiRAGeNews. This benchmark-consistent ranking shows that the result is not merely a small average effect across runs. When the evaluation target is unknown, early fusion is therefore the safest in-domain default for Macro-Precision, with AUROC and Accuracy providing consistent support.

3.2 RQ2: Modality Complementarity

The modality-drop analysis still supports complementarity, but the dominant pattern is not symmetric synergy; it is **image-heavy reliance**. Table 3 shows the full multimodal result together with both single-modality drop tests for every dataset–fusion–classifier block. Across all 675 matched comparisons, removing text reduces Macro-Precision by 4.83 points (95% CI [4.18, 5.53]), AUROC by 3.93 ([3.65, 4.23]), and Accuracy by 5.78 ([5.28, 6.31]), whereas removing the image stream reduces them by 13.42 [12.20, 14.69], 12.08 [11.48, 12.66], and 16.93 [16.10, 17.79] points, respectively. Full multimodal models significantly outperform both text-drop and image-drop variants overall on all three metrics (all $p < 10^{-78}$), and the difference between image-drop and text-drop degradation is also highly significant (Macro-Precision: $p = 8.99 \times 10^{-46}$; AUROC: $p = 2.26 \times 10^{-66}$; Accuracy: $p = 1.31 \times 10^{-68}$).The asymmetry is strongly dataset dependent, but the direction is now uniform across all three benchmarks. MiRAGeNews is by far the most image-dependent benchmark: removing images causes mean losses of 21.68 points in Macro-Precision, 17.05 in AUROC, and 26.62 in Accuracy, far larger than the corresponding text-removal losses of 6.48, 3.27, and 8.52 points. MMFakeBench is the most balanced benchmark, although image removal is still worse on average (9.62 / 10.60 / 10.01 versus 6.61 / 6.42 / 5.33). VLDBench shows the smallest text losses and remains image-dominated as well (8.96 / 8.60 / 14.17 versus 1.39 / 2.11 / 3.49). Over all configurations, image removal hurts more than text removal in 76.3% [72.9, 79.4] of Macro-Precision comparisons, 80.0% [76.8, 82.8] of AUROC comparisons, and 82.4% [79.3, 85.1] of Accuracy comparisons. The practical conclusion is therefore not merely that multimodal models help, but that their gains are real, asymmetric, and usually driven by harder-to-replace visual evidence.

3.3 RQ3: Component Impact Analysis

The component-level picture remains stable and practically actionable under the revised metric protocol. All four component families exhibit significant overall effects under Friedman testing on Macro-Precision, AUROC, and Accuracy, but their magnitudes differ substantially. Vision encoder choice is dominant on all three metrics (all $p < 8 \times 10^{-94}$), with average per-dataset spreads of 7.46 points on Macro-Precision, 7.34 on AUROC, and 6.95 on Accuracy. Classifier choice is the next largest source of stable variation (5.32 / 5.75 / 5.68 points), whereas text encoder choice (1.99 / 2.33 / 1.70) and fusion strategy (1.93 / 1.55 / 0.97) are materially smaller. Put differently, once a reasonable text encoder is in place, most of the robust variance comes from the image tower and the downstream head.

Two trends are especially stable. First, **SigLIP is the strongest vision encoder on all three reported metrics**, reaching 79.19% Macro-Precision, 86.01% AUROC, and 80.37% Accuracy overall, with CLIP ViT-B/32 ranking second at 77.43 / 84.93 / 78.67. Post-hoc testing confirms that SigLIP significantly exceeds CLIP ViT-B/32 after Holm correction on all three metrics ($p_{\text{Holm}} = 7.40 \times 10^{-15}$, 2.53×10^{-20}, and 5.67×10^{-15}). Second, **text encoders matter less and separate much less cleanly**. DistilBERT has the strongest overall in-domain text means at 75.17 / 82.57 /

Classifier	Metric	MMFakeBench	MiRAGeNews	VLDBench
Logistic Regression	Prec	73.93 (± 1.07)	91.65 (± 0.95)	63.68 (± 0.48)
	AUROC	83.55 (± 1.18)	96.80 (± 0.54)	69.63 (± 0.67)
	Acc	77.87 (± 0.87)	91.65 (± 0.95)	64.73 (± 0.45)
XGBoost	Prec	74.51 (± 0.74)	90.00 (± 0.83)	64.52 (± 0.32)
	AUROC	85.45 (± 0.82)	96.18 (± 0.48)	70.60 (± 0.42)
	Acc	78.29 (± 0.59)	89.99 (± 0.83)	65.53 (± 0.30)
Random Forest	Prec	71.06 (± 1.31)	83.34 (± 1.29)	60.32 (± 0.68)
	AUROC	79.79 (± 1.27)	90.94 (± 1.02)	64.86 (± 0.92)
	Acc	73.74 (± 0.65)	83.32 (± 1.29)	61.37 (± 0.67)

Table 1: Classifier-wise in-domain averages (%) over all encoder pairs and fusion strategies. Values are mean ± 95% bootstrap CI half-width across matched configurations.

Fusion	Metric	MMFakeBench	MiRAGeNews	VLDBench
Early	Prec	75.20 (± 1.07)	88.80 (± 1.15)	64.06 (± 0.41)
	AUROC	83.73 (± 1.20)	95.10 (± 0.76)	69.93 (± 0.59)
	Acc	76.79 (± 0.85)	88.77 (± 1.16)	65.06 (± 0.40)
Mid	Prec	72.22 (± 1.07)	88.20 (± 1.41)	62.27 (± 0.71)
	AUROC	82.78 (± 1.26)	94.51 (± 1.02)	67.63 (± 0.98)
	Acc	76.62 (± 0.86)	88.19 (± 1.41)	63.31 (± 0.70)
Late	Prec	72.09 (± 1.07)	88.00 (± 1.39)	62.20 (± 0.71)
	AUROC	82.28 (± 1.22)	94.31 (± 0.99)	67.52 (± 0.96)
	Acc	76.48 (± 0.85)	87.99 (± 1.38)	63.25 (± 0.69)

Table 2: Fusion-wise in-domain averages (%) over all encoder pairs and classifier heads. Values are mean ± 95% bootstrap CI half-width across matched configurations.

76.67, but the top four text encoders are all within 0.32 Macro-Precision points of one another, and DistilBERT does not significantly exceed BERT-base or CLIP-Text after Holm correction. This is exactly the pattern one would expect from a component family that matters, but matters less than the vision side.

Classifier conclusions mirror RQ1. Logistic regression and XGBoost again form the leading pair: logistic regression has the highest overall Macro-Precision and Accuracy means (76.42 and 78.08), but both are statistically tied with XGBoost ($p = 0.160$ and 0.175), whereas XGBoost has the strongest AUROC at 84.07%, significantly above logistic regression ($p = 0.0022$). Random forest is substantially behind on every metric. Figure 3 reinforces the same conclusion visually: variation across vision rows is larger and more systematic than variation across most text columns. Taken together, the component analysis suggests a clear priority order for future system design: get the vision encoder right first, choose between logistic regression and XGBoost as the serious downstream heads, then refine the fusion strategy, and only then expect modest gains from swapping among strong text encoders such as DistilBERT, BERT-base, CLIP-Text, and SBERT.

3.4 RQ4: Benchmark-Shift Generalization

Benchmark shift is the most challenging setting. As Table 4 shows, all six source–target transfers degrade substantially relative to their target in-domain baselines. Macro-Precision gaps range from 8.33 points for MMFakeBench→VLDBench to 36.30 for VLDBench→MiRAGeNews; the corresponding AUROC and Accuracy gaps range from 11.76 to 41.57 and from 6.45 to 36.93 points, respectively. The highest transfer Macro-Precision is achieved by MMFakeBench→MiRAGeNews at 63.25% (±0.59), followed by MiRAGeNews→MMFakeBench at 62.56% (±0.53), whereas VLDBench→MiRAGeNews exhibits the largest overall collapse. Every transfer direction is significantly worse than its target in-domain baseline on all three metrics (paired Wilcoxon tests, all $p \leq 1.15 \times 10^{-38}$).

Although transfer rankings are flatter than in-domain rankings, early fusion remains the strongest precision-first default. Across 450 matched transfer blocks, it achieves 56.16% Macro-Precision, significantly exceeding mid fusion (55.51%, $p = 1.73 \times 10^{-5}$) and late fusion (55.45%, $p = 1.47 \times 10^{-5}$). It also leads in AUROC at 58.78%, while Accuracy is nearly tied across fusion strategies. Direction-specific reversals occur—MiRAGeNews→MMFakeBench slightly favors

Dataset	Fusion	Classifier	Full	Text drop	Δ_{text}	Image drop	Δ_{img}
MMFakeBench	Early	Logistic Reg.	72.2/82.0/76.6	64.4/75.2/68.2	7.8/6.8/8.5	62.8/71.0/65.7	9.4/11.0/11.0
		XGBoost	76.2/86.2/79.0	71.0/79.2/72.5	5.2/7.1/6.5	63.4/72.9/65.7	12.9/13.3/13.3
		Random Forest	77.2/83.0/74.7	74.2/76.8/72.8	3.0/6.1/1.9	58.4/72.3/71.0	18.8/10.7/3.8
	Mid	Logistic Reg.	74.9/84.9/78.6	66.4/78.0/72.3	8.4/6.9/6.3	69.2/77.7/73.2	5.7/7.2/5.4
		XGBoost	73.8/85.2/78.0	69.5/81.0/73.6	4.3/4.2/4.4	67.2/76.2/61.9	6.6/9.0/16.2
		Random Forest	68.0/78.2/73.2	62.3/70.4/68.5	5.7/7.8/4.8	58.0/65.1/64.2	10.0/13.1/9.0
	Late	Logistic Reg.	74.7/83.7/78.4	59.2/78.0/72.4	15.5/5.7/6.0	68.4/77.7/72.9	6.3/6.0/5.5
		XGBoost	73.5/84.9/77.8	69.5/79.6/72.9	4.0/5.3/4.9	66.4/73.0/60.7	7.1/12.0/17.1
		Random Forest	68.0/78.2/73.2	62.4/70.4/68.5	5.6/7.8/4.7	58.1/65.1/64.3	9.8/13.1/8.9
MiRAGeNews	Early	Logistic Reg.	91.6/96.9/91.6	86.0/93.8/83.4	5.6/3.1/8.3	77.1/82.7/61.9	14.6/14.2/29.7
		XGBoost	89.8/96.1/89.7	84.5/92.4/81.9	5.3/3.7/7.9	67.4/79.1/64.5	22.3/17.0/25.2
		Random Forest	85.0/92.3/84.9	79.9/87.8/73.4	5.1/4.5/11.5	67.4/76.2/62.4	17.6/16.1/22.5
	Mid	Logistic Reg.	91.9/97.0/91.9	82.9/93.5/75.0	9.0/3.5/16.8	78.8/87.3/64.6	13.1/9.7/27.3
		XGBoost	90.2/96.3/90.2	86.5/93.9/84.8	3.7/2.4/5.4	66.6/80.2/65.0	23.6/16.1/25.1
		Random Forest	82.5/90.3/82.5	79.2/86.9/78.1	3.4/3.4/4.5	64.1/64.0/55.2	18.4/26.3/27.3
	Late	Logistic Reg.	91.4/96.5/91.4	72.1/93.5/78.6	19.4/3.0/12.8	47.9/87.3/60.7	43.6/9.2/30.8
		XGBoost	90.0/96.2/90.0	86.5/93.7/85.0	3.6/2.5/5.1	66.5/77.7/65.6	23.6/18.5/24.4
		Random Forest	82.5/90.3/82.5	79.1/86.9/78.0	3.4/3.4/4.5	64.1/63.9/55.3	18.4/26.3/27.2
VLDBench	Early	Logistic Reg.	62.9/68.4/63.9	62.8/67.3/58.6	0.1/1.1/5.4	55.8/58.5/48.7	7.1/9.9/15.3
		XGBoost	65.0/71.3/66.0	63.7/69.3/63.3	1.4/2.0/2.7	59.8/59.4/46.3	5.2/11.9/19.7
		Random Forest	64.2/70.1/65.2	62.6/67.7/62.7	1.6/2.4/2.5	45.3/59.2/46.6	18.9/10.9/18.6
	Mid	Logistic Reg.	64.1/70.2/65.1	63.1/68.3/59.6	1.0/1.9/5.5	51.1/64.5/52.0	13.0/5.7/13.1
		XGBoost	64.4/70.4/65.4	64.1/69.6/63.9	0.3/0.9/1.5	59.4/61.7/52.3	4.9/8.8/13.1
		Random Forest	58.4/62.2/59.4	56.3/59.3/56.8	2.1/2.9/2.7	54.9/56.2/49.8	3.5/6.1/9.6
	Late	Logistic Reg.	64.1/70.3/65.1	61.6/68.3/60.1	2.5/2.0/5.0	46.8/64.5/51.4	17.3/5.8/13.7
		XGBoost	64.2/70.0/65.2	62.7/67.4/61.7	1.5/2.7/3.4	57.0/57.9/50.3	7.1/12.2/14.8
		Random Forest	58.3/62.2/59.4	56.3/59.3/56.8	2.0/3.0/2.6	54.8/56.1/49.8	3.5/6.1/9.7

Table 3: Modality-drop study. Each cell reports Macro-Precision/AUROC/Accuracy (%). "Full" is the complete multimodal model, and Δ_{text} and Δ_{img} are performance drops in percentage points when one modality is removed at test time. Detailed entries are rounded to one decimal for readability.

Train → Test	Macro-Precision	AUROC	Accuracy
MMFakeBench → MiRAGeNews	63.25 (± 0.59)	67.70 (± 0.70)	57.84 (± 0.57)
MMFakeBench → VLDBench	54.51 (± 0.33)	56.60 (± 0.25)	57.42 (± 0.14)
MiRAGeNews → MMFakeBench	62.56 (± 0.53)	69.26 (± 0.70)	55.44 (± 0.70)
MiRAGeNews → VLDBench	49.25 (± 0.15)	48.88 (± 0.22)	44.85 (± 0.19)
VLDBench → MMFakeBench	52.64 (± 0.26)	54.69 (± 0.46)	48.68 (± 0.39)
VLDBench → MiRAGeNews	52.03 (± 0.35)	53.07 (± 0.52)	51.39 (± 0.27)

Table 4: Cross-dataset transfer by source–target direction. Values are mean ± 95% bootstrap CI half-width (%).

mid fusion, and VLDBench→MiRAGeNews favors late fusion—but do not alter the aggregate ranking. Table 5 shows that other component rankings also flatten under transfer. XGBoost has the highest Macro-Precision at 56.31% but is statistically tied with logistic regression at 56.21% ($p = 0.264$). SigLIP leads the vision encoders at 57.73% but is tied with CLIP ViT-B/32 at 57.49% ($p = 0.819$), while BERT-base leads the less-stable text encoders at 56.00%, only 0.19 points above SBERT.

Overall, the safest precision-first transfer configuration uses early fusion, SigLIP or CLIP ViT-B/32, BERT-base or SBERT, and either XGBoost or logistic regression. Source-specific effects may change the finer ranking, but strong visual representations and early fusion remain the most dependable choices. Cross-dataset evaluation should therefore be treated as a first-class requirement, as even strong in-domain models can lose more than 30 points under benchmark shift.

3.5 SHAP-Based Interpretation

We also use SHAP to test whether the modality conclusions are visible in feature attributions rather than only in modality-drop ablations. Across 675 in-domain SHAP analyses, image features account for a mean 61.57% of total attribution mass (95% bootstrap CI [60.52, 62.61]). Image attribution exceeds text attribution in 81.6% of configurations,

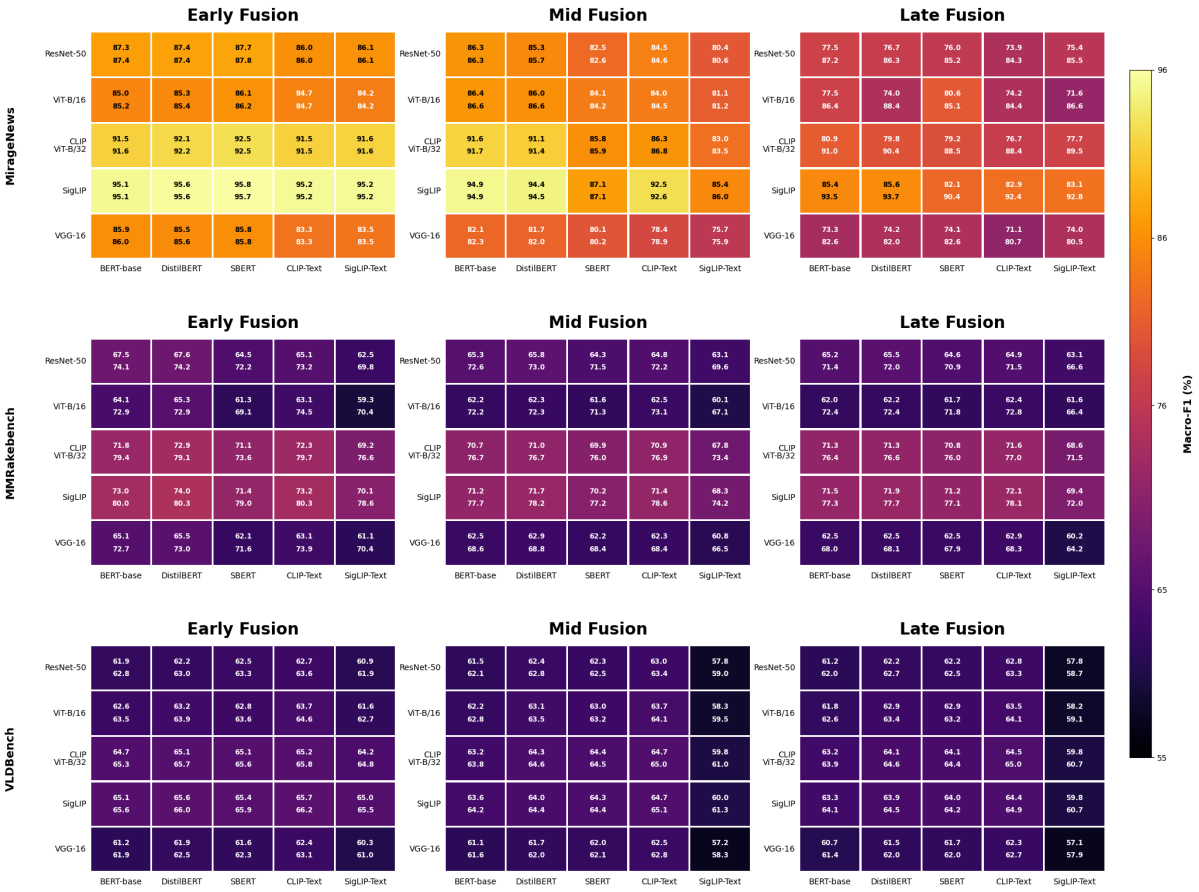

Figure 3: Main encoder-pair heatmap across the three datasets. Each cell reports mean Macro-Precision / Accuracy (%) for a fixed vision–text encoder pair, averaged over the three classifier heads under the given fusion strategy; color encodes AUROC.

and the paired image-vs-text difference is highly significant ($p = 7.24 \times 10^{-50}$). This pattern is not a fusion-specific artifact: the mean image share is essentially identical for early, mid, and late fusion (all about 61.6%).

The dataset breakdown mirrors RQ2 almost exactly. MMFakeBench is the most balanced benchmark at 53.8% mean image share, MiRAGeNews is the most visually dominated at 67.1%, and VLDBench remains clearly image-leaning at 63.8%. In the best-performing SHAP-analyzed configurations for each dataset, the top individual contributors are overwhelmingly SigLIP dimensions, which independently reinforces the RQ3 conclusion that visual encoder quality—and especially strong contrastive vision backbones—explains a large share of stable multimodal performance.

3.6 Practical Recommendations

The aggregate results point to two concrete default approaches for future multimodal misinformation studies. If the goal is strong in-domain performance under frozen representations, a robust Macro-Precision-oriented starting point is SigLIP on the visual side, with CLIP ViT-B/32 as the main alternative, DistilBERT on the text side, early fusion, and either logistic regression or XGBoost as the downstream head. This recommendation is supported by the in-domain averages: early fusion is the strongest fusion strategy overall on Macro-Precision (76.02%), SigLIP is the strongest vision encoder (79.19%), DistilBERT has the strongest text-encoder mean (75.17%), and logistic regression and XGBoost form the leading classifier pair, with logistic regression nominally highest on Macro-Precision (76.42%) and XGBoost essentially tied (76.34%).

If the focus is stronger OOD performance, the default should be chosen for transfer stability rather than in-domain dominance. Our results suggest early fusion with a strong contrastive vision encoder, especially SigLIP or CLIP ViT-B/32, together with a strong generic text encoder such as BERT-base or SBERT, and either XGBoost or logistic regression as the final head. This recommendation follows the transfer-wide Macro-Precision ranking: early fusion is best overall (56.16%), SigLIP is the strongest vision encoder (57.73%)

(a) Classifier-wise transfer averages by target dataset				
Classifier	Metric	MMFakeBench	MiRAGeNews	VLDBench
Logistic Regression	Prec	58.22 (± 0.97)	58.77 (± 1.02)	51.64 (± 0.52)
	AUROC	63.10 (± 1.46)	62.02 (± 1.32)	52.80 (± 0.59)
	Acc	52.38 (± 0.85)	55.21 (± 0.73)	51.00 (± 1.05)
XGBoost	Prec	58.22 (± 1.01)	58.45 (± 1.12)	52.26 (± 0.55)
	AUROC	62.62 (± 1.39)	61.89 (± 1.53)	52.82 (± 0.74)
	Acc	51.89 (± 1.02)	55.99 (± 0.84)	51.24 (± 1.07)
Random Forest	Prec	56.36 (± 0.87)	55.70 (± 1.05)	51.73 (± 0.51)
	AUROC	60.19 (± 1.28)	57.24 (± 1.22)	52.60 (± 0.69)
	Acc	51.92 (± 0.79)	52.64 (± 0.58)	51.17 (± 0.93)

(b) Fusion-wise transfer averages by target dataset				
Fusion	Metric	MMFakeBench	MiRAGeNews	VLDBench
Early	Prec	57.69 (± 0.92)	58.74 (± 1.21)	52.04 (± 0.57)
	AUROC	62.28 (± 1.33)	60.96 (± 1.49)	53.09 (± 0.70)
	Acc	52.55 (± 0.92)	53.92 (± 0.73)	51.50 (± 0.99)
Mid	Prec	57.59 (± 0.96)	57.11 (± 1.00)	51.82 (± 0.49)
	AUROC	61.90 (± 1.38)	60.19 (± 1.35)	52.58 (± 0.66)
	Acc	51.74 (± 0.86)	54.98 (± 0.77)	50.91 (± 1.02)
Late	Prec	57.51 (± 0.94)	57.07 (± 0.99)	51.78 (± 0.52)
	AUROC	61.74 (± 1.35)	60.01 (± 1.33)	52.55 (± 0.68)
	Acc	51.90 (± 0.84)	54.95 (± 0.77)	50.99 (± 1.02)

Table 5: Cross-dataset transfer averages by component family and target dataset. Values are mean ± 95% bootstrap CI half-width (%).

but statistically tied with CLIP ViT-B/32 (57.49%), BERT-base is the strongest text encoder (56.00%) with SBERT close behind (55.81%), and XGBoost is nominally best among classifiers (56.31%) but statistically tied with logistic regression (56.21%). The main shift from the in-domain recommendation is therefore on the text side: DistilBERT is the strongest in-domain text choice, whereas under benchmark shift the more reliable text defaults become BERT-base and SBERT.

4 Conclusion

In this paper, we presented a large-scale empirical study of multimodal misinformation detection, covering 3,375 controlled experiments across three benchmarks, multiple vision and text encoders, fusion strategies, and classifiers. Our findings show that early fusion is relatively the most reliable in-domain strategy, visual evidence often contributes more strongly than text, and vision encoder and classifier choices have the largest impact on performance. We also find that cross-dataset generalization remains challenging, highlighting the need for stronger transfer evaluation. We hope that these findings and protocols will serve as a useful reference for the broader research community.

Limitations

Our study intentionally focuses on frozen encoders and simple downstream heads. This choice improves comparability and isolates component effects, but it does not cover fully trainable cross-attention architectures or parameter-efficient adaptation methods. Finally, statistical significance should not be confused with practical magnitude: because we aggregate many matched configurations, some small gaps are estimated precisely without necessarily justifying additional system complexity. Future work should therefore extend the benchmark scope and pair significance testing with stronger effect-size, calibration, and more in-depth analyses.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):76:1–76:29.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Jinyin Chen, Chengyu Jia, Haibin Zheng, Ruoxi Chen, and Chenbo Fu. 2023. Is multi-modal necessarily better? robustness evaluation of multi-modal fake news

detection. *IEEE Transactions on Network Science and Engineering*, 10:3144–3158.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Carmela Comito, Luciano Caroprese, and Ester Zumpano. 2023. Multimodal fake news detection on social media: A survey of deep learning techniques. *Social Network Analysis and Mining*, 13(1):101.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal realistic AI-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448, Miami, Florida, USA. Association for Computational Linguistics.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 795–816.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. 2019. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xuannan Liu, Zekun Li, Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. MMFakeBench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In *International Conference on Learning Representations*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13062.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and Mubarak Shah. 2026. VLDBench evaluating multimodal disinformation with regulatory alignment. *Information Fusion*, 130:104092.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Akshit Sharma and Prashant W. Patil. 2026a. MemeTAG: Keyword-driven meme classification through tag embedding reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7679–7688.

Akshit Sharma and Prashant W. Patil. 2026b. Small cues, big consequences: Learning pivotal cues for multimodal meme classification. *Preprint*, arXiv:2609.26907. Accepted to Findings of the Association for Computational Linguistics: EMNLP 2026.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. SpotFake+: A multimodal framework for fake news detection via transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13915–13916.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 849–857.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining*, volume 12085 of *Lecture Notes in Computer Science*, pages 354–367. Springer, Cham.