HARF-News: A Multimodal Benchmark for Disentangling Authorship and Veracity in AI-Era News

Anonymous ACL submission

Abstract

The spread of false information on social media poses concrete risks to public health, democratic elections, and institutional trust. Yet existing benchmarks do not jointly capture this evolving reality; most treat AI-authored text as inherently fake and overlook authorship diversity, while ignoring the multimodal nature of real-world news. Critically, authorship and veracity are independent axes—LLMs are increasingly adopted by journalists to produce genuine content, while human-authored misinformation continues to circulate at scale. We introduce **HARF-News**, the first unified text-image benchmark for 4-way authorship-veracity classification: 241,590 headlines and 138,882 text-image pairs across Human Real (HR), Human Fake (HF), AI Real (AR), and AI Fake (AF). We further propose **HARFNet**, a multi-task framework designed as a strong baseline for joint authorship-veracity detection. HARFNet integrates stylometric, cross-modal discrepancy, and patch-level alignment signals via bilinear interaction, achieving 87.6 Macro-F1 on 4-way classification and showing competitive generalization on external multimodal misinformation benchmarks.[1]

1 Introduction

Online platforms have become a central medium for accessing news and public information, where a significant volume of content circulates before its accuracy can be verified. Empirical evidence demonstrates that false information spreads significantly faster and more broadly than truthful content on social media (Vosoughi et al., 2018), posing concrete risks to public health, democratic elections, and institutional trust (Lazer et al., 2018). Consequently, detecting unreliable news has emerged as a critical challenge in modern information ecosystems.

[1]The dataset and code are available at https://anonymous.4open.science/r/HARF-News.

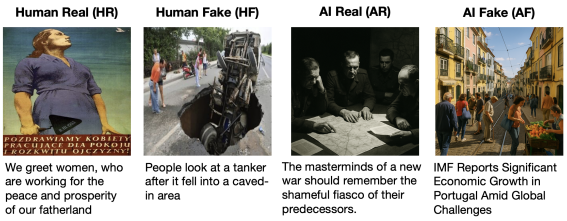


Figure 1: Example samples from the HARF-News dataset, covering four distinct classes: Human Real (HR), Human Fake (HF), AI Real (AR), and AI Fake (AF).

Simultaneously, the advent of Large Language Models (LLMs) and generative image techniques is fundamentally reshaping news production. LLMs are now capable of generating journalistically styled articles at scale—over half of UK journalists now report using AI tools on a weekly basis (Thurman et al., 2025)—and AI-generated visuals have emerged as a growing vehicle for disinformation (Feuerriegel et al., 2023). As a result, both the textual and visual components of a news item can now be independently AI-generated, giving rise to hybrid text-image news packages whose authorship is increasingly opaque.

A critical consequence of this shift is that *authorship* and *veracity* must be treated as independent axes. The prevailing assumption—that AI-generated content is inherently fake—no longer reflects reality: LLMs are now actively adopted by professional journalists for content production at scale (Thurman et al., 2025; Cools, 2024), while human-authored misinformation predates generative AI and continues to circulate at scale (Shu et al., 2020; Nakamura et al., 2020). Treating authorship and veracity as a single axis introduces two systematic failure modes: AI-written real news (AR) is misclassified as fake, while human-written fake news (HF) may evade detectors tuned to stylometric signals of AI generation (Su et al., 2023).

Fine-grained detection therefore requires distinguishing all four combinations—Human Real (HR), Human Fake (HF), AI Real (AR), and AI Fake (AF)—as illustrated in Figure 1. Yet exist-

ing benchmarks fail to address this. Traditional and synthetic datasets either lack authorship diversity or remain text-only, treating machine-generated content as inherently fake (Shu et al., 2020; Nakamura et al., 2020; Yang et al., 2024; Chen and Shu, 2024; Wang et al., 2024; Huang et al., 2024). The only work to consider both dimensions jointly operates in a text-only setting (Su et al., 2024). Moreover, existing multimodal detectors treat cross-modal inconsistency as the primary signal of manipulation (Papadopoulos et al., 2024)—yet AI Fake news may exhibit hyper-concentrated text-image alignment, the opposite of what these detectors expect, while Human Fake news produces the mismatched patterns they are designed to catch (Kang et al., 2025). To the best of our knowledge, no existing benchmark jointly covers all four multimodal categories—HR, HF, AR, and AF—in a unified text-image format.

In this paper, we introduce **HARF-News**, the first unified sample-level text-image benchmark covering all four authorship-veracity combinations, comprising 241,590 headlines and 138,882 text-image pairs. To enable fine-grained authorship-veracity analysis, we construct AI-authored samples by applying LLM-based rewriting and controlled factual manipulation to human-authored seeds, and pair each sample with an image generated via an edit-based pipeline that preserves the visual realism of the source domain. Our primary contribution is the formulation and construction of a unified multimodal benchmark for disentangling authorship from veracity, enabling controlled evaluation of whether models can separately identify who produced a news item and whether it is factually reliable. We further propose **HARFNet** as a strong task-specific baseline for this new evaluation setting, jointly modeling authorship, veracity, and cross-modal alignment through dedicated stylometric, cross-modal discrepancy, and patch-level alignment modules fused via bilinear interaction.

Experiments across multiple benchmarks show that existing detectors struggle with the joint authorship-veracity setting, while HARFNet achieves a 4-way Macro-F1 of 87.6 with competitive cross-domain generalization.

2 Related Work

2.1 Misinformation Datasets

Benchmark datasets have shaped fake news detection under different assumptions about misinformation. The Weibo dataset (Jin et al., 2017) provides early text-image pairs verified by Xinhua News Agency, while FakeNewsNet (Shu et al., 2020) adds social context from PolitiFact and GossipCop. Fakeddit (Nakamura et al., 2020) scales to over one million samples but relies on distant supervision without modeling AI-generated content. NewsCLIPpings (Luo et al., 2021) focuses on cross-modal inconsistency, and DGM^4 (Shao et al., 2023) introduces generative manipulations.

Table 1: Coverage of authorship (Human/AI) and veracity (Real/Fake) across existing datasets. HARF-News is the first unified text-image benchmark covering all four authorship-veracity categories.

Dataset	Modality	HR	HF	AR	AF
FakeNewsNet (Shu et al., 2020)	Text+Image	✓	✓	–	–
Fakeddit (Nakamura et al., 2020)	Text+Image	✓	✓	–	–
LLMFake (Chen and Shu, 2024)	Text	–	✓	–	✓
MegaFake (Wang et al., 2024)	Text	✓	✓	✓	✓
GossipCop++ (Su et al., 2024)	Text	✓	✓	✓	✓
MiRAGeNews (Huang et al., 2024)	Caption+Image	✓	–	–	✓
MFND (Zhu et al., 2025)	Text+Image	partial	partial	partial	partial
HARF-News (Ours)	**Text+Image**	✓	✓	✓	✓

Despite their impact, these datasets mainly model human-authored misinformation via manipulation or inconsistency, with limited consideration of generative authorship. Recent efforts such as LLMFake (Chen and Shu, 2024) and MegaFake (Wang et al., 2024) introduce LLM-generated fake news but remain confined to text. MiRAGeNews (Huang et al., 2024) extends to multimodal AI generation but still lacks a unified authorship-veracity formulation. MFND (Zhu et al., 2025) adopts a 4-way label space but assumes all AI-generated content to be fake, excluding the AR class in HARF-News and allowing cross-modal inconsistency where text and image may differ in authorship or veracity (see Table 1). Unlike manipulation-oriented datasets where text and image may carry different authorship or veracity states, HARF-News assigns a unified sample-level label for each text-image pair, enabling direct evaluation of the 4-way HARF taxonomy.

2.2 Multimodal Fake News Detection

Early fake news detection methods rely on single-modality signals: text-based approaches exploit linguistic patterns and pretrained language models (Potthast et al., 2018; Shu et al., 2020; Kaliyar et al., 2021), while image-based methods target manipulated or synthetic imagery (Afchar et al., 2018; Rossler et al., 2019; Wang et al., 2020, 2023) but struggle with cross-modal mismatches such as semantically inconsistent text-image pairs.

Multimodal approaches address this by inte-

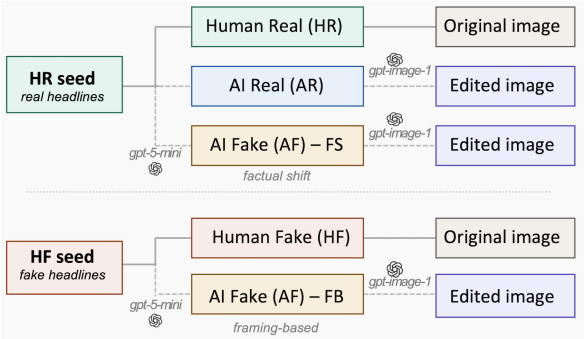


Figure 2: Overview of the HARF-News construction pipeline.

grating both modalities, ranging from early fusion via shared attention or embeddings (Jin et al., 2017; Singhal et al., 2019, 2020; Khattar et al., 2019) to late fusion that combines independently encoded representations (Su et al., 2023; Xue et al., 2021). More recent methods leverage cross-modal alignment through vision-language pretraining and contrastive learning (Qian et al., 2021; Zhou et al., 2023; Chen et al., 2022; Ying et al., 2023), achieving stronger performance. Nevertheless, these methods are designed to implicitly treat AI-generated content as inherently fake, conflating authorship with veracity. Consequently, they cannot distinguish human manipulation from AI generation, nor recognize factually accurate AI-generated content—leaving authorship and veracity entangled and limiting fine-grained detection.

3 HARF-News Dataset

3.1 Design Goals

HARF-News provides the first unified sample-level text-image benchmark that jointly covers all four authorship-veracity combinations: Human Real (HR), Human Fake (HF), AI Real (AR), and AI Fake (AF), with representative examples shown in Figure 1. Our dataset construction follows three principles: (i) full 2×2 coverage over Human/AI and Real/Fake; (ii) preservation of realistic headline-image alignment as observed in real online news; and (iii) controlled AI generation that distinguishes fact-preserving from fact-distorting transformations. Overall, HARF-News contains 241,590 headlines for unimodal (text-only) evaluation, of which 138,882 are paired with images to support multimodal evaluation. Figure 2 provides an overview of the construction pipeline.

3.2 Human-Written Headlines (HR/HF)

We sample human-written headlines from the Fakeddit corpus (Nakamura et al., 2020), which contains over 1 million posts across 22 subreddits. After preprocessing (see Appendix A.1 for full details and statistics), we retain 84,954 HR and 56,636 HF headlines.

3.3 AI Real Headlines (AR)

AI Real (AR) consists of factually accurate news rewritten by LLMs—a class absent from prior benchmarks that conflate AI-generated content with fakeness, thereby distorting authorship evaluation. To construct AR, we sample 60,000 HR headlines via stratified sampling over subreddits and instruct `gpt-5-mini` to rewrite each headline while preserving all factual content and introducing only stylistic variation in wording and structure under a Reddit-style news-headline tone.

3.4 AI Fake Headlines (AF)

AI Fake (AF) headlines are generated using two distinct seed types, each reflecting a different real-world mechanism of AI-assisted misinformation. The first captures AI-driven factual distortion of true news (Feuerriegel et al., 2023), while the second reflects the recirculation of existing misinformation through LLM paraphrasing—a pattern shown to evade detection by altering surface form while preserving deceptive intent (Chen and Shu, 2024; Su et al., 2023). Full prompt templates are provided in Appendix A.2: (1) factual corruption of Human Real (HR) headlines, and (2) reframing of Human Fake (HF) headlines to simulate AI-assisted recirculation of misinformation. We design six prompt templates in total: three factual-corruption templates (FS-1–3) applied to HR seeds, and three framing templates (FB-1–3) applied to HF seeds. Each template contributes approximately 6,667 headlines, yielding a total of 40,000 AF samples. All AF generations use `gpt-5-mini`.

FS Family: Factual Shift (HR → AF)

The FS (Factual Shift) family is motivated by the observation that AI-assisted misinformation frequently operates through direct factual distortion of true content—such as inverting outcomes, altering numerical values, or substituting key entities—reflecting the *information manipulation* category identified in prior work on LLM-generated misinformation (Chen and Shu, 2024). The three templates correspond to progressively weaker levels of semantic distortion:

- **FS-1 (Inversion):** Reverses the event's stance

or main outcome (e.g., "wins"→ "loses"), representing the strongest factual shift.

- **FS-2 (Large shift):** Modifies multiple factual elements such as numbers, locations, actors, or event scale.
- **FS-3 (Small shift):** Alters only one or two minor factual elements to introduce subtle but meaningful inconsistencies.

FB Family: Framing-Based (HF → AF)

The FB (Framing-Based) family is motivated by the threat of AI-assisted recirculation of existing misinformation through surface-level transformation—a pattern shown to evade detection while preserving deceptive intent (Su et al., 2023; Wang et al., 2024). By altering style or sentence structure without changing the underlying false claim, FB samples simulate how LLMs are exploited to launder and redistribute human-authored fake news. We define three templates:

- **FB-1 (Paraphrasing):** Rewrite the headline in clearly different words while preserving meaning and tone.
- **FB-2 (Expand):** Adds a short phrase of background or framing detail to make the same false claim appear more concrete.
- **FB-3 (Disguise):** Shifts sentence structure and tone (e.g., toward a formal news-reporting style) to make the rewritten headline harder to align with the original.

3.5 Multimodal Images

A central design choice in HARF-News is how to construct images for AI-authored classes (AR/AF). Previous synthetic-news image pipelines often generate images *from scratch* from text prompts using diffusion models (Huang et al., 2024). However, from-scratch generation produces images stylistically detached from real user-uploaded content, risking trivially solvable AI/Human visual distinction. To address this, we adopt an *image-edit-based* construction strategy.

For each AR/AF sample, we take the original Reddit-sourced image of its HR or HF seed and edit it conditioned on the AI-modified headline (`ai_text`). Concretely, we use the `gpt-image-1` edit API with the prompt template *"Edit this image to visually reflect the following news headline: {ai_text}"*, generating 1024×1024 images at medium quality. By starting from the original photograph rather than a blank canvas, the edited image inherits the visual realism and amateur photographic style of the source domain while embedding visual cues consistent with the AI-modified text.

For the Human classes (HR/HF), we directly inherit the original user-provided images from Fakeddit, retaining 79,794 images for HR and 54,088 for HF after preprocessing. For the AI classes, the image-edit pipeline is applied to a stratified subset, producing 3,000 AR and 2,000 AF images, with proportions matched across the six AF strategies. We accordingly prioritize coverage across all four classes and all six AF strategies over per-class scale. In total, HARF-News provides 138,882 multimodal text-image pairs, supporting both unimodal and multimodal evaluation under a unified authorship-veracity setup. We empirically validate that edit-based images preserve the visual realism of source photographs and are less visually distinguishable from real images than from-scratch generated counterparts (see Appendix A.3). Although image-only models can still detect AI-edited images, HARF-News is designed not as a pure AI-image detection benchmark but as a joint authorship-veracity reasoning task.

3.6 Quality Validation

To validate dataset quality, three graduate students with backgrounds in natural language processing independently evaluated samples from each generation strategy without prior knowledge of each other's judgments, using a three-point scale (*Yes / No / Unsure*). Text samples (randomly selected, 200 per strategy) were assessed on whether the intended modification was correctly applied, whether unintended factual changes were introduced, and whether the result remained fluent and natural as a news headline. Image samples (randomly selected, 100 per modality) were assessed on semantic consistency with the modified headline and visual naturalism as a news photograph.

Text samples achieved Pass Rates of 87.0% for AR, 92.3% for FS, and 93.8% for FB. The lower AR rate reflects borderline cases where minimal stylistic variation was deemed insufficiently distinct from the original, rather than factual errors in the generated content. Image samples achieved 93.3%, confirming that edit-based images are semantically coherent with their modified headlines and visually natural. Agreement rates ranged from

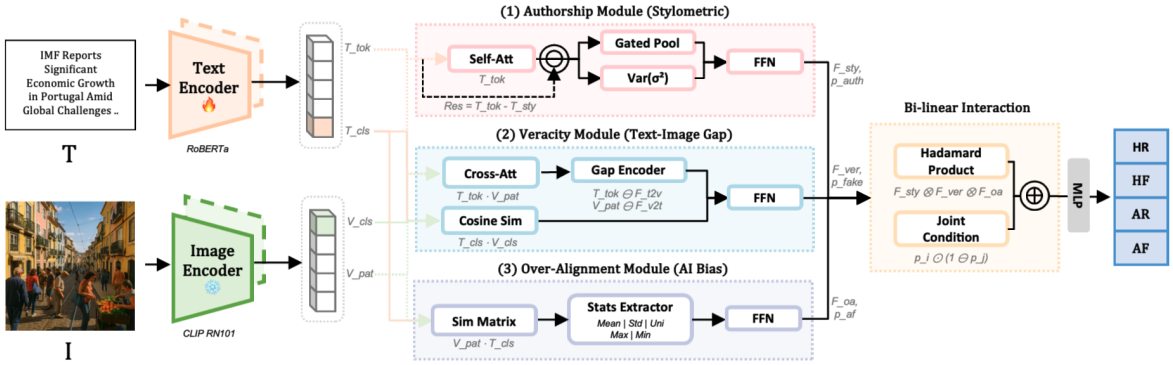


Figure 3: HARFNet architecture. The Authorship Module (AM), Veracity Module (VM), and Over-Alignment Module (OAM) extract stylometric, cross-modal discrepancy, and patch-level alignment signals, respectively, fused via Bilinear Interaction and Joint Condition reasoning for 4-way classification into HR, HF, AR, and AF.

Table 2: Statistics of the HARF-News dataset. #Images denotes the number of valid text-image pairs available for multimodal evaluation.

Label	#Texts	#Images	Source
HR	84,954	79,794	Fakeddit
HF	56,636	54,088	Fakeddit
AR	60,000	3,000	gpt-5-mini / gpt-image-1
AF	40,000	2,000	gpt-5-mini / gpt-image-1
Total	**241,590**	**138,882**	—

77.2% to 82.8% for text and reached 80.0% for images, consistent with annotation studies on generated text quality. We use agreement rates rather than adjudicated factuality labels because the validation aims to assess construction quality rather than provide an independent fact-checking benchmark. Cases marked as 'Unsure' were treated conservatively as non-pass cases in the reported pass rates. Detailed criteria and full results are provided in Appendix A.4.

3.7 Dataset Summary

Table 2 summarizes the final composition of HARF-News. Headlines average 11.2 words ($\sigma = 3.4$), following a Reddit-style news format. The dataset spans 21 subreddits covering diverse topics including general interest (*mildlyinteresting*), news (*neutralnews*, *nottheonion*), and satirical content (*fakehistoryporn*); detailed subreddit distributions are provided in Appendix A.1.

4 Method

4.1 HARFNet Pipeline Overview

We propose **HARFNet**, a hierarchical framework that decomposes detection into three factors: (i) text authorship, (ii) veracity, and (iii) AI-specific over-alignment artifacts. As illustrated in Figure 3, each factor is modeled by a dedicated module producing feature representations and probabilistic signals. The resulting representations are then fused through bilinear interaction and joint-condition reasoning to perform 4-way classification.

4.2 Parallel Modules with Hierarchical Semantics

Given a text-image pair (I, T), we extract multimodal representations using a frozen CLIP encoder and a partially fine-tuned RoBERTa. This yields token-level and patch-level representations together with global representations, enabling hierarchical semantic reasoning across modalities.

4.2.1 Authorship Module (AM)

The Authorship Module (AM) determines whether the text is human-written by contextualizing the token sequence $\mathbf{T}_{\text{tok}}$ via self-attention. Following DetectGPT (Mitchell et al., 2023), LLM-generated text exhibits more uniform residual patterns, whereas human-written text shows greater local variability. AM first applies self-attention to $\mathbf{T}_{\text{tok}}$ to obtain $\mathbf{T}_{\text{sty}}$, and computes the stylometric residual:

$$\mathbf{R}_{\text{sty}} = \mathbf{T}_{\text{tok}} - \mathbf{T}_{\text{sty}}. \tag{1}$$

Two branches then pool $\mathbf{R}_{\text{sty}}$: (1) *Var(σ^2) Pool* computes token-level variance and projects it via an FFN,

$$\mathbf{f}_{\text{var}} = \text{FFN}\left(\text{Var}(\mathbf{R}_{\text{sty}})\right), \tag{2}$$

and (2) *Gated Pool* selectively aggregates residual tokens weighted by a learned gate $g_i = \text{softmax}(\mathbf{w}_g^\top \mathbf{R}_{\text{sty}}[i])$,

$$\mathbf{f}_{\text{gate}} = \sum_{i=1}^{L} g_i \cdot \mathbf{R}_{\text{sty}}[i]. \tag{3}$$

The pooled features are concatenated and passed through an FFN with layer normalization, producing a stylometric representation $\mathbf{F}_{\text{sty}} \in \mathbb{R}^{768}$ and an AI-authorship probability p_{auth}.

4.2.2 Veracity Module (VM)

The Veracity Module (VM) assesses whether the image semantically supports the headline via *cross-attention*, *Gap Encoder*, and *global similarity*.

In the text-to-image cross attention, tokens $\mathbf{T}_{\text{tok}}$ attend to image patches $\mathbf{V}_{\text{pat}}$ to produce grounded features $\hat{\mathbf{T}}$, and the discrepancy $\text{gap}_t = \text{mean}(\mathbf{T}_{\text{tok}} - \hat{\mathbf{T}})$ captures deviation from visual grounding. The reverse direction similarly produces gap_v. These bidirectional discrepancy signals are processed through the *Gap Encoder*, while a global semantic similarity is computed as $\text{sim}_g = \cos(\mathbf{T}_{\text{cls}}, \mathbf{V}_{\text{cls}})$. In general, HF samples tend to exhibit large cross-modal gaps, while AF samples show small gaps but unusually high semantic similarity. We fuse these signals to obtain a veracity feature $\mathbf{F}_{\text{ver}} \in \mathbb{R}^{768}$ and a fake probability p_{fake}.

4.2.3 Over-Alignment Module (OAM)

Although VM captures text-image consistency, semantic alignment alone cannot distinguish genuine photographs from AI-generated images tailored to the text. Prior work has shown that AI-generated images exhibit stronger text alignment than real photographs, which often contain visual content not fully captured by the caption (Kang et al., 2025). Motivated by this observation, OAM introduces patch-level alignment concentration (*over-alignment*) as an auxiliary signal complementary to VM's cross-modal discrepancy signals. We treat over-alignment as an auxiliary cue rather than a sufficient condition for AI-generated fake news, since low-saliency or weakly edited images may exhibit alignment patterns similar to real news.

To quantify this behavior, we compute a *Similarity Matrix* between image patch embeddings $\mathbf{V}_{\text{pat}}$ and the global text representation $\mathbf{T}_{\text{cls}}$, yielding per-patch alignment scores: $s_n = \cos(\mathbf{V}_{\text{pat}}[n], \mathbf{T}_{\text{cls}})$. From these scores, we derive three statistics:

$$\mu_{\text{oa}} = \frac{1}{P}\sum_{n=1}^{P} s_n, \tag{4}$$

$$\sigma_{\text{oa}} = \text{std}(\{s_n\}), \tag{5}$$

$$u_{\text{oa}} = \frac{\mu_{\text{oa}}}{s_{\max} - s_{\min} + \varepsilon}. \tag{6}$$

Here, μ_{oa}, σ_{oa}, and u_{oa} measure alignment strength, patch-level variability, and uniformity, respectively. AF samples tend to exhibit high μ_{oa} and u_{oa}, whereas HF/HR samples show larger σ_{oa} due to natural background diversity. The extracted statistics are processed by the *Stats Extractor* and FFN, and combined with text-guided weighted patch pooling via $\text{softmax}(s_n)$ to produce $\mathbf{F}_{\text{oa}} \in \mathbb{R}^{768}$ and the AF probability p_{af}.

4.3 Bilinear Interaction (BM) and Classification

The three modules produce complementary yet individually insufficient signals. p_{auth} cannot separate AR from AF, p_{fake} conflates HF and AF, and p_{af} is unreliable in isolation. Crucially, AF is the only class where all three signals are simultaneously elevated. To explicitly capture this dependency, we apply a *Bilinear Interaction*. We first project each representation from $\mathbb{R}^{768}$ to $\mathbb{R}^{128}$. We then compute their interaction using a low-rank Hadamard product. The Hadamard product is large only when all three inputs are active, yielding strong responses for AF and suppressing others. To further separate AR from AF, the *Joint Condition* module augments the interaction features with scalar probability terms such as $p_{\text{auth}}p_{\text{af}}$ and $p_{\text{auth}}(1-p_{\text{af}})$, encoded through a lightweight MLP. Finally, the classifier concatenates all four representations into $\mathbf{H} \in \mathbb{R}^{3072}$ and applies a 2-layer MLP head:

$$\begin{aligned} \mathbf{H} &= [\mathbf{F}_{\text{sty}}; \mathbf{F}_{\text{ver}}; \mathbf{F}_{\text{oa}}; \mathbf{F}_{\text{bi}}], \\ \hat{y} &= \mathbf{W}_2\, \text{GELU}(\mathbf{W}_1\mathbf{H}). \end{aligned} \tag{7}$$

4.4 Distribution-Aware Multi-Task Training Objectives

Training is governed by five complementary losses:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{focal}} &+ \lambda_1\, \mathcal{L}_{\text{ver}} + \lambda_2\, \mathcal{L}_{\text{auth}} \\ &+ \lambda_3\, \mathcal{L}_{\text{oa}} + \lambda_4\, \mathcal{L}_{\text{con}}, \end{aligned} \tag{8}$$

where λ_1=0.30, λ_2=0.15, λ_3=0.30, λ_4=0.05. $\mathcal{L}_{\text{focal}}$ is a 4-class focal loss that mitigates extreme class imbalance (HR : HF : AR : AF $\approx$ 40 : 27 : 1.5 : 1). $\mathcal{L}_{\text{ver}}$ and $\mathcal{L}_{\text{auth}}$ are binary cross-entropy terms that directly supervise p_{fake} and p_{auth}, anchoring each module's output to its intended semantic role independently of the final classifier. $\mathcal{L}_{\text{con}}$ is a supervised contrastive loss on VM embeddings $\mathbf{z}_{\text{ver}}$ that sharpens the Real/Fake decision boundary in representation space.

$$\mathcal{L}_{\text{oa}} = \|\mu_{\text{oa}} - \mu^*_{\text{oa}}\|_2^2 + \|\sigma_{\text{oa}} - \sigma^*_{\text{oa}}\|_2^2 + \|u_{\text{oa}} - u^*_{\text{oa}}\|_2^2\,. \tag{9}$$

Meanwhile, $\mathcal{L}_{\text{oa}}$ penalizes deviations of the mean μ_{oa}, standard deviation σ_{oa}, and uniformity u_{oa} of patch-alignment scores from class-specific targets (μ^*, σ^*, u^*) via MSE. This distributional constraint captures authorship-dependent alignment

Type	Model	H/A		R/F		4-way		Class-wise F1				
		Acc	F1	Acc	F1	Acc	F1	HR	HF	AR	AF	Macro Avg
Text-only	BERT	97.9	83.1	84.0	83.3	82.7	68.1	86.0	80.2	57.4	49.0	68.1
	RoBERTa	99.2	93.9	84.5	83.8	84.3	78.2	87.0	80.9	89.3	55.7	78.2
	DeBERTa-v3	99.3	94.4	85.7	84.9	85.5	79.9	88.1	82.0	**90.9**	58.6	79.9
Image-only	ResNet-50	99.5	96.3	74.4	72.8	74.2	61.5	80.0	66.9	58.7	40.5	61.5
	ViT-B/16	99.6	97.0	68.6	66.6	68.4	58.3	75.0	59.1	59.2	40.0	58.3
	CLIP (Image-only)	99.8	98.5	82.2	81.3	82.1	69.3	86.2	77.9	66.8	46.4	69.3
Multimodal	Qwen2.5-VL-7B	94.6	49.4	59.5	59.4	56.3	29.3	56.7	59.1	1.3	0.3	29.3
	GPT-5-mini	91.5	50.9	70.8	69.9	64.7	34.5	73.4	58.8	1.1	4.7	34.5
	CAFE	97.9	83.9	76.7	75.4	75.4	58.4	80.7	69.9	52.0	31.1	58.4
	SpotFake+	98.6	89.7	75.8	73.6	75.1	65.1	81.1	65.8	68.4	45.3	65.2
	FND-CLIP	99.8	98.6	88.4	88.0	88.3	77.0	90.8	**86.6**	73.1	57.7	77.0
	HAMMER	99.8	98.8	87.1	86.4	87.0	76.7	90.0	84.0	75.2	57.5	76.7
	MiRAGe	99.7	98.1	87.3	86.7	87.3	76.1	90.1	84.5	77.5	52.1	76.1
	HARFNet	**100.0**	**100.0**	**89.0**	**88.5**	**89.0**	**87.6**	**90.9**	86.2	89.6	**83.7**	**87.6**

Table 3: Performance comparison on the HARF-News dataset.

Model	MiRAGeNews															DGM^4		
	NYT+MJ (ID)			BBC+DE			CNN+DE			BBC+SX			CNN+SX					
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
CAFE	88.3	88.3	95.1	66.2	66.1	73.1	75.6	75.2	83.6	70.4	70.4	78.3	79.5	79.4	87.5	50.2	50.2	59.9
SpotFake+	91.2	91.2	97.2	73.4	73.3	81.1	76.8	76.7	84.9	73.6	73.5	81.3	80.0	80.0	87.0	62.4	60.7	68.1
FND-CLIP	99.4	99.4	**100.0**	80.1	80.1	85.7	82.4	82.3	90.9	86.7	86.6	92.9	90.3	90.3	96.3	77.6	75.7	85.6
HAMMER	99.4	99.4	100.0	78.1	77.7	88.0	78.9	78.0	88.7	**87.4**	**87.3**	94.7	**90.7**	**90.6**	**96.5**	79.6	**84.5**	87.0
MiRAGe	98.2	98.2	76.8	76.8	81.8	81.8	81.3	81.2	85.7	85.7	85.7	85.6	85.7	85.7	85.7	65.9	64.6	74.0
HARFNet	**99.6**	**99.6**	**100.0**	**87.1**	**87.0**	**95.4**	**89.3**	**89.3**	**96.1**	85.8	85.7	**95.2**	89.3	89.3	**96.5**	**81.2**	79.6	**89.8**

Table 4: Performance comparison on the MiRAGeNews and DGM^4 datasets. DE and SX denote DALL-E and SDXL, respectively.

patterns across classes, providing a stable signal complementary to the veracity-oriented $\mathcal{L}_{\text{ver}}$. The class-specific target statistics are summarized in Appendix B.1.

5 Experiments

5.1 Experimental Setup

We evaluate our method on HARF-News and further assess its generalization on MiRAGeNews (Huang et al., 2024) and DGM^4 (Shao et al., 2023). MiRAGeNews is a multimodal binary benchmark with 15K image-caption pairs, with both in-distribution and out-of-distribution splits across different image generators and news sources. DGM^4 is a large-scale multimodal manipulation benchmark containing 281K text-image news pairs with pristine and manipulated samples, including Face Swap (FS), Face Attribute (FA), Text Swap (TS), Text Attribute (TA), and their mixed variants. For HARF-News, we use only the multimodal news subset. Although these benchmarks do not share the 4-way HARF label space, they are used to evaluate cross-domain generalization of the proposed multimodal signals.

We compare HARFNet with representative unimodal, VLM, and multimodal baselines, including

Model	HARF-News		MiRAGeNews		DGM^4	
	Acc	F1	Acc	F1	Acc	F1
HARFNet (Full)	**89.0**	**88.5**	**90.2**	**90.2**	**81.2**	**79.6**
w/o BM	89.0	88.6	85.5	85.2	81.3	79.6
w/o BM+OAM	89.1	88.6	90.3	90.2	80.2	78.4
w/o BM+OAM+AM	89.0	88.6	87.5	87.4	80.2	78.5
w/o BM+OAM+VM	82.7	82.0	79.7	78.3	51.7	51.6

Table 5: Ablation study on Real/Fake binary classification across three benchmarks. MiRAGeNews results are averaged over five subsets.

CAFE (Chen et al., 2022), SpotFake+ (Singhal et al., 2020), FND-CLIP (Zhou et al., 2023), HAMMER (Shao et al., 2023), and MiRAGe (Huang et al., 2024), as well as text-only and image-only baselines. VLM baselines are evaluated zero-shot without task-specific fine-tuning or access to dataset-specific generation templates. Because HARF-News is highly imbalanced, particularly in AR and AF, we emphasize Macro-F1 and class-wise F1 alongside accuracy. We report Accuracy, F1, and AUC on MiRAGeNews and DGM^4, and additionally evaluate H/A, R/F, and 4-way classification on HARF-News. Dataset and implementation details are provided in Appendix C.3.

5.2 Results on HARF-News

Table 3 presents the results on HARF-News. HARFNet achieves strong performance across all tasks, obtaining 100.0 F1 on H/A classification, 88.5 F1 on R/F classification, and 87.6 F1 on 4-way classification. Compared with the best prior multimodal baseline, FND-CLIP, HARFNet improves the 4-way F1 from 77.0 to 87.6 (+10.6), with especially large gains on the AI-related AR and AF classes. Among unimodal models, DeBERTa-v3 (He et al., 2021) achieves the strongest text-only performance, reaching 79.9 4-way F1 and the highest AR F1 (90.9) among non-HARFNet models, suggesting that it captures linguistic artifacts in AI-generated text. In contrast, image-only models achieve near-perfect H/A performance but substantially lower 4-way results. For example, CLIP (image-only) obtains 98.5 H/A F1 but only 69.3 4-way Macro-F1, with particularly weak performance on AR and AF. This indicates that visual artifacts alone are insufficient for determining whether AI- or human-authored content is factually reliable. Overall, the strongest supervised multimodal detectors outperform unimodal baselines on 4-way classification, highlighting the importance of joint multimodal reasoning.

5.3 Results on DGM^4 and MiRAGeNews

HARFNet achieves strong overall performance on both DGM^4 and MiRAGeNews (Table 4), obtaining the best Accuracy and AUC on DGM^4 and stronger cross-domain generalization on MiRAGeNews. In particular, it surpasses FND-CLIP by +7.0 Acc on BBC+DALL-E and +6.9 Acc on CNN+DALL-E, demonstrating stronger adaptation to unseen generative distributions. These results suggest that the proposed approach generalizes beyond HARF-News, which is specifically designed for 4-way classification, to other fake-news benchmarks defined under different settings and distributions.

5.4 Ablation Study

To identify which modules are primarily responsible for this robustness, we conducted ablation studies across benchmarks. Results are summarized in Table 5. On **HARF-News**, the Veracity Module (VM) is the most influential component. Removing VM together with BM and OAM causes a substantial drop (-6.3 Acc, -6.5 F1), whereas removing BM, OAM, or AM individually causes only minor changes. This reflects the nature of HARF-News, where fake samples often exhibit semantic inconsistencies between image and text that VM is designed to capture.

On **MiRAGeNews**, both VM and AM contribute substantially. Removing AM (*w/o BM+OAM+AM*) reduces accuracy by 2.7 points, while additionally removing VM (*w/o BM+OAM+VM*) causes a further 7.8-point decline. This matches the OOD setting of MiRAGeNews, where unseen generators such as DALL-E and SDXL are paired with authentic news text. AM provides generator-agnostic authenticity cues, while VM maintains cross-modal consistency.

For **DGM^4**, VM removal causes a catastrophic drop (-29.5 Acc, -28.0 F1), far larger than any other ablation. Since DGM^4 focuses on manipulation-based fake news, manipulated regions often create strong semantic conflicts between text and image. VM is therefore particularly effective in this setting, whereas the remaining modules contribute comparatively little.

Overall, VM is the primary driver of Real/Fake discrimination, while AM, OAM, and BM provide complementary benefits depending on dataset characteristics. This is expected because the three benchmarks emphasize different detection conditions: HARF-News requires disentangling authorship and veracity, **MiRAGeNews** emphasizes generator shift, and **DGM^4** emphasizes localized multimodal manipulation. We therefore view HARFNet not as a universally optimal architecture, but as a modular baseline for analyzing useful signals across different authorship-veracity conditions.

6 Conclusion

In this paper, we introduce HARF-News, a multimodal benchmark that jointly models authorship and veracity with four classes: Human Real, Human Fake, AI Real, and AI Fake, enabling fine-grained analysis of multimodal inconsistencies in human- and AI-generated content. We further propose HARFNet, a multi-task framework for jointly learning veracity, authorship, and cross-modal alignment. Experiments show that existing multimodal misinformation detectors struggle in this setting, while HARFNet provides a strong baseline with competitive cross-domain generalization. These results highlight the importance of jointly modeling authorship and veracity under controlled cross-modal alignment.

Limitations

Despite its contributions, this study has several limitations. First, HARF-News exhibits severe class imbalance (HR:HF:AR:AF = 40:27:1.5:1), resulting in relatively few AR and AF samples per test fold. Consequently, evaluation on AI-related classes remains limited, and fine-grained analysis across AF manipulation strategies is currently infeasible. This limitation primarily stems from the high cost of large-scale image generation using the GPT-image-1 editing API.

Second, all AI-generated images are produced using a single GPT-image-1-based editing pipeline. Although this ensures construction consistency, it may introduce generator-specific biases. HARF-News should therefore be viewed as an initial benchmark for authorship-veracity disentanglement rather than a complete representation of all AI-assisted news production pipelines. Our cross-dataset experiments on MiRAGeNews and DGM^4 provide preliminary evidence that HARFNet generalizes beyond GPT-image-1, but future versions of HARF-News should incorporate multiple text and image generators.

Third, although the proposed over-alignment hypothesis is effective for many AF samples, some low-saliency examples exhibit alignment patterns similar to real news, suggesting that alignment statistics alone may be insufficient. Nevertheless, OAM remains a useful auxiliary signal in settings involving text-conditioned image generation, suggesting that over-alignment may reflect a broader characteristic of text-conditioned image generation rather than a GPT-image-1-specific artifact.

Beyond these limitations, this work also raises broader societal and ethical concerns. HARF-News includes AI-generated multimodal news samples and associated prompting strategies, which could be misused to generate realistic misinformation or deceptive media content. Although the dataset is intended solely for AI-generated fake news detection research, its dual-use risk cannot be fully eliminated. To mitigate this risk, we release HARF-News and HARFNet exclusively for detection, robustness, and safety research, while withholding generation pipelines that directly optimize deceptive realism. Note that the released dataset includes labels, splits, and evaluation scripts, but does not include automated scripts for scaling misinformation generation or prompts optimized to increase deceptive realism. Future work should build on the released framework to develop more robust and generalizable multimodal misinformation detectors.

References

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

Hannes Cools. 2024. Uses of generative ai in the newsroom. *Journalism Practice*, 18(4):411–428.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Stefan Feuerriegel, Renée DiResta, Josh A. Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Prollochs. 2023. Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, 7:1818–1821.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. Miragenews: Multimodal realistic AI-generated news detection. In *Findings of EMNLP*.

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.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Ju Yeon Kang, Jaehong Park, Semin Kim, Ji Won Yoon, and Nam Soo Kim. 2025. Semantic-aware reconstruction error for detecting ai-generated images. *arXiv preprint arXiv:2508.09487*.

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.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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.

Eric Mitchell, Yoonho Lee, Rajat Khanna, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 6149–6157.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2024. Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(4).

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 231–240.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

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 *International Conference on Machine Learning (ICML)*.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

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 (student abstract). 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. IEEE.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1473–1490, Mexico City, Mexico. Association for Computational Linguistics.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Neil Thurman, Sina Thäsler-Kordonouri, and Richard Fletcher. 2025. AI adoption by UK journalists and their newsrooms: Surveying applications, approaches, and attitudes. Technical report, Reuters Institute for the Study of Journalism, Oxford, UK.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Lionel Z Wang, Yiming Ma, Renfei Gao, Beichen Guo, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024. Megafake: A theory-driven dataset of fake news generated by large language models. *arXiv preprint arXiv:2408.11871*.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

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.

Migyeong Yang, Chaewon Park, Jiwon Kang, Daeun Lee, Daejin Choi, and Jinyoung Han. 2024. Fighting against fake news on newly-emerging crisis: A case study of covid-19. In *Companion Proceedings of the ACM Web Conference 2024*, pages 718–721.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 5384–5392.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pages 2825–2830. IEEE.

Ye Zhu, Yunan Wang, and Zitong Yu. 2025. Multimodal fake news detection: Mfnd dataset and shallow-deep multitask learning. *arXiv preprint arXiv:2505.06796*.

A Dataset Details

A.1 Fakeddit Preprocessing Details

The original Fakeddit corpus contains 682,661 samples across three splits (564,000 train / 59,342 validation / 59,319 test) (Nakamura et al., 2020), of which 268,908 are labelled *real* and 413,753 are labelled *fake*. We apply three sequential filtering steps before sampling. The resulting sample counts at each stage are summarized in Table 6.

1. **Image filtering.** We retain only samples that have both a non-empty `image_url` and `hasImage=True`, removing 1,863 samples (1,307 real, 556 fake) that correspond to deleted posts or broken URLs. This leaves 680,798 samples.

2. **Headline length filtering.** We compute the character length of each `clean_title` and apply IQR-based outlier removal ($Q_1 = 19$, $Q_3 = 57$, IQR $= 38$; bounds: $[-38, 114]$), followed by removal of headlines shorter than the post-IQR mean length (38.7 characters). We additionally exclude the `propagandaposters` subreddit (14,642 samples, all fake), whose text-image veracity labels are systematically inconsistent with the rest of the corpus. In total, 405,688 samples are removed (102,647 real, 303,041 fake), leaving 275,110 samples (164,954 real, 110,156 fake).

3. **Sampling.** From the filtered pool we allocate 80,000 samples as AI-generation sources (AR: 60,000 from real; AF: 20,000 from real + 20,000 from fake) and retain the remainder as human-authored samples. To maintain an overall 6:4 real-to-fake ratio across all four labels, we subsample the human-fake pool to 56,636 items. The final human-authored subset contains 84,954 HR and 56,636 HF samples.

Stage	Real	Fake	Total
Original Fakeddit	268,908	413,753	682,661
After image filtering	267,601	413,197	680,798
Removed	1,307	556	1,863
After length filtering	164,954	110,156	275,110
Removed	102,647	303,041	405,688
Final HR / HF (human)	84,954	56,636	141,590
AR / AF (AI–generated)	60,000	40,000	100,000
HARF-News total	**144,954**	**96,636**	**241,590**

Table 6: Sample counts at each Fakeddit preprocessing stage. "Real" and "Fake" refer to the original Fakeddit 2-way labels.

Subreddit	HR	HF	AR	AF
mildlyinteresting	33,760	–	23,571	7,800
nottheonion	18,588	–	13,044	4,407
upliftingnews	12,164	–	8,690	2,858
photoshopbattles	9,040	–	6,540	2,149
fakehistoryporn	–	14,905	–	5,288
psbattle_artwork	–	10,879	–	3,840
pareidolia	–	9,211	–	3,231
subredditsimulator	–	5,387	–	1,882
neutralnews	4,168	–	2,987	1,014
savedyouaclick	–	4,179	–	1,530
Others	7,234	12,075	5,168	6,001
Total	**84,954**	**56,636**	**60,000**	**40,000**

Table 7: Subreddit distribution across HARF-News labels. HR and AR are sampled from real-labelled subreddits; HF from fake-labelled subreddits; AF from both.

A.2 Dataset Construction Prompts

This section provides the full prompt templates used to generate AI Real (AR) and AI Fake (AF) headlines in HARF-News. All generations use `gpt-5-mini`. The AR template preserves factual content while introducing only stylistic variation. The six AF templates are divided into two families: Factual Shift (FS-1–3), applied to Human Real (HR) seeds, and Framing-Based (FB-1–3), applied to Human Fake (HF) seeds, as described in Section 3.4.

AI Real (AR) Prompt

The AR prompt instructs the model to rewrite a Human Real headline while preserving all factual content and introducing only stylistic variation in wording and structure under a Reddit-style news-headline tone.

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Rewrite the headline so it reads more natural and easy to understand as a Reddit-style news title, while keeping all underlying facts exactly the same. **Writing rules:** • Keep every person, organization, location, date, and number from the original headline unchanged. • Do not change the event itself or its outcome; only rephrase the wording and style. • Make small but meaningful changes to wording and word order so the rewritten headline is not identical to the original. • Try to keep the rewritten headline roughly similar in length to the original. • Do not introduce any new facts, claims, or background information; use only what is in the original headline. **Original Reddit headline:** {title} **Rewritten Reddit headline (AI assistant):**

Table 8: Prompt template for generating AI Real (AR) headlines.

FS Family: Factual Shift Prompts (FS-1, FS-2, FS-3)

The FS family applies factual corruption to HR seeds, producing AF headlines whose underlying claims are no longer true. The three templates correspond to progressively weaker levels of semantic distortion: FS-1 (Inversion), FS-2 (Large shift), and FS-3 (Small shift).

FS-1 (Inversion). The model reverses the event's main outcome, decision, or stance (e.g., "wins" → "loses"), while keeping the same actors and overall situation.

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Rewrite the headline by reversing the main outcome, decision, or stance, while keeping the situation otherwise consistent and realistic. **Writing rules:** • Maintain the same underlying real-world event and general topic category as the original headline. You may adjust details if necessary for coherence, but do not introduce a different event. • Keep the same main actors, organizations, and overall situation as in the original headline. • Reverse the central result, decision, or stated position (e.g., approval → rejection, increase → decrease, win → loss). • Ensure all details in the rewritten headline are internally consistent with each other. • The rewritten headline should be similar in length to the original (stay within ±10–15%). • Ensure the headline reads as a natural, plausible, single-line Reddit-style news headline. • Do not introduce extreme, sensational, or unrelated details beyond the scope of the original headline. **Original Reddit headline:** {title} **Rewritten Reddit headline (outcome-flip, AI assistant):**

Table 9: FS-1 prompt template (Inversion).

FS-2 (Large shift). The model modifies multiple factual elements such as numbers, locations, actors, or event scale, so the situation becomes clearly different but remains realistic.

FS-3 (Small shift). The model alters only one or two minor factual elements to introduce subtle but meaningful inconsistencies, producing a headline that remains highly similar to the original.

FB Family: Framing-Based Prompts (FB-1, FB-2, FB-3)

The FB family rewrites HF headlines to preserve their deceptive intent while altering surface form, mimicking recirculation of misinformation through different stylistic expressions. The three templates are FB-1 (Paraphrase), FB-2 (Expand), and FB-3 (Disguise).

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Rewrite the headline by changing multiple key factual details, so that the situation becomes clearly different but remains realistic. **Writing rules:** • Maintain the same underlying real-world event and general topic category as the original headline. You may alter factual details, but do not introduce a different event. • Change **at least three** key factual elements, such as numbers, dates, locations, people, organizations, or the scale of the event. • Ensure all details in the rewritten headline are internally consistent with each other. • The rewritten headline should be similar in length to the original (stay within ±10–15%). • Ensure the headline reads as a natural, plausible, single-line Reddit-style news headline. • Do not introduce extreme, sensational, or unrelated details beyond the scope of the original headline. **Original Reddit headline:** {title} **Rewritten Reddit headline (large-shift, AI assistant):**

Table 10: FS-2 prompt template (Large shift).

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Rewrite the headline by changing a small number of factual details, so that the situation becomes slightly different but remains highly similar and realistic. **Writing rules:** • Maintain the same underlying real-world event and general topic category as the original headline. You may alter factual details, but do not introduce a different event. • Change **only one or two** key factual elements, such as numbers, dates, locations, people, organizations, or the scale of the event. • Ensure all details in the rewritten headline are internally consistent with each other. • The rewritten headline should be similar in length to the original (stay within ±10–15%). • Ensure the headline reads as a natural, plausible, single-line Reddit-style news headline. • Do not introduce extreme, sensational, or unrelated details beyond the scope of the original headline. **Original Reddit headline:** {title} **Rewritten Reddit headline (small-shift, AI assistant):**

Table 11: FS-3 prompt template (Small shift).

FB-1 (Paraphrase). The model rewrites the headline in clearly different wording while preserving its meaning, claim, and all factual details.

FB-2 (Expand). The model adds a short phrase of background or framing detail to make the same false claim appear more concrete, without altering the core claim.

FB-3 (Disguise). The model shifts sentence structure and tone—for example, toward a formal news-reporting style—to make the rewritten headline harder to align with the original, while preserving the core claim and all named entities.

A.3 Image-Edit Pipeline Validation

To empirically verify that our image-edit pipeline does not introduce generator-specific visual artifacts, we compared the visual similarity between original HR photographs and their AR counterparts generated by edit-based and from-scratch approaches using three metrics: SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and CLIP-based image similarity (CLIP-I) (Radford et al., 2021).

Experimental Setup. We randomly sampled 100 matched HR-AR pairs from the HARF-News dataset. For the **edit-based** approach, we used GPT-image-1 in inpainting mode, providing the original HR photograph as the base image and the AI-rewritten AR headline as the editing prompt. The prompt followed the template:

> *"You are editing a news photo. Edit this image to visually reflect the following news headline: '{AR headline}'. Photo style: taken by bystander with consumer camera."*

For the **from-scratch** approach, the same AR headline was used as a text-to-image generation prompt without any source image. All images were resized to 256×256 pixels prior to metric computation.

Metrics.

- **SSIM** was computed on grayscale images using a data range of 255.
- **LPIPS** used a VGG backbone; images were normalized to $[-1, 1]$ before computation (↓ lower is better).
- **CLIP-I** was computed as the cosine similarity between ViT-L/14 image embeddings (↑ higher is better).

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Rewrite the headline in different words while keeping the semantics of the original headline unchanged. **Writing rules:** • Keep all people, organizations, locations, dates, and numbers from the original headline unchanged. • Keep the same underlying event and claim; do not add, remove, or alter any facts. • Use different phrasing and word order so the rewritten headline is clearly paraphrased, not a trivial edit of the original. • Keep the rewritten headline roughly similar in length and style to the original. **Original Reddit headline:** {title} **Rewritten Reddit headline (AI assistant):**

Table 12: FB-1 prompt template (Paraphrase).

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Expand the headline by adding a short phrase of realistic framing or contextual detail, while keeping the semantics of the original headline unchanged. **Writing rules:** • Preserve the core claim and what is being asserted; do not change or weaken it. • Add one short phrase of background or framing detail (for example, a brief reason, time, place, or consequence) that makes the claim sound more concrete or specific. • Do not introduce entirely new, unrelated claims. • Keep the result as a single headline-style sentence (not a full paragraph). • Keep the expanded headline only slightly longer than the original (within about 120% of the original length), and avoid making it excessively long or verbose. • Ensure the expanded headline still reads like a natural and plausible Reddit-style news title. **Original Reddit headline:** {title} **Expanded Reddit headline (AI assistant):**

Table 13: FB-2 prompt template (Expand).

As shown in Table 15, edit-based AR images consistently achieved higher similarity to the original HR photographs across all three metrics compared to from-scratch generated images (SSIM: 0.246 vs. 0.159, LPIPS: 0.646 vs. 0.753, CLIP-I: 0.711 vs. 0.574), confirming that our edit-based pipeline preserves the visual realism of the source domain while incorporating AI-modified content. This reduces the likelihood that downstream models rely on generator-specific artifacts and encourages them to focus on semantic inconsistencies introduced by manipulation.

A.4 Human Quality Validation

A.4.1 Text Quality Criteria

We evaluated generated headlines using six manipulation strategies: three Fact Shift (FS) strategies and three Fact Blend (FB) strategies.

FS strategies assess whether key factual elements are intentionally modified while preserving the original topic and event category.

- **FS-1 (Outcome Reversal).** Annotators evaluated whether the generated headline preserved the original situation and context while reversing the primary outcome, decision, or stance.
- **FS-2 (Substantial Factual Modification).** Annotators evaluated whether at least three key factual elements were modified while preserving the same event category or topic. Key factual elements include entities, actions, numerical values, time, and locations.
- **FS-3 (Limited Factual Modification).** Annotators evaluated whether exactly one or two key factual elements were modified while preserving the original event category or topic.

FB strategies assess whether factual content is preserved under paraphrasing, contextual augmentation, or stylistic variation.

- **FB-1 (Paraphrasing).** Annotators evaluated whether the generated headline preserved the original claim and key factual information while modifying wording or sentence order. Minor grammatical corrections and synonym substitutions were considered acceptable.
- **FB-2 (Contextual Augmentation).** Annotators evaluated whether short contextual information (e.g., reason, time, location, or consequence) was appropriately added without altering the original factual direction.
- **FB-3 (Stylistic Variation).** Annotators evaluated whether the generated headline preserved the original factual claim while changing sentence structure or tone.

Annotators also considered whether generated

Role	Content
System	*You are an AI writing assistant helping a Reddit user refine a news headline they wrote.*
User	**Task:** Disguise the headline by changing its surface wording, sentence structure, and tone, while keeping the semantics of the original headline unchanged. **Writing rules:** • Preserve the same core claim and overall meaning; do not change or replace it with a different claim. • Do not change the main people, organizations, locations, or events involved. • You may change sentence structure, level of formality, and tone (for example, to a more neutral, formal, news-reporting style) so that the new headline feels stylistically different from the original and is harder to match to the original text. • Use clearly different wording and style so the rewritten headline is not a trivial edit of the original. • Keep the rewritten headline relatively short and natural, as if it could realistically appear online (e.g., on a news site or Reddit). **Original Reddit headline:** {title} **Disguised Reddit headline (AI assistant):**

Table 14: FB-3 prompt template (Disguise).

Table 15: Visual similarity between HR photographs and AR images generated by edit-based (ours) vs. from-scratch approaches. Results are averaged over 100 matched pairs. ↑: higher is better; ↓: lower is better.

Comparison	SSIM ↑	LPIPS ↓	CLIP-I ↑
HR vs. AR (edit-based, GPT-image-1)	**0.246**	**0.646**	**0.711**
HR vs. AR (from-scratch)	0.159	0.753	0.574

headlines remained fluent and stylistically appropriate as news headlines.

A.4.2 Image Validation Criteria

For image samples, annotators evaluated whether the edited image was semantically consistent with the modified headline while remaining visually natural as a news image.

Annotators considered whether unnecessary or unrealistic visual changes were introduced during editing, and whether newly added elements appeared coherent with the overall event context.

Images containing unrealistic, inappropriate, or visually implausible artifacts were marked as invalid.

A.4.3 Annotation Protocol

Following the procedure described in Section 3.6, all samples were independently evaluated by three graduate student annotators using a three–point scale (*Yes / No / Unsure*).

For reporting quality statistics, only *Yes* responses were counted as passes. Cases marked as *Unsure* were treated conservatively as non–pass cases and counted together with *No* responses when computing pass rates.

Agreement was measured using strict percentage agreement, defined as the proportion of samples for which all three annotators assigned the same label (*Yes*, *No*, or *Unsure*). We report percentage agreement rather than adjudicated labels because the goal of this validation is to assess generation quality and instruction adherence rather than establish an independent factuality benchmark.

A.4.4 Validation Results

Tables 16 and 17 report per-strategy pass rates and percentage agreement. Pass rates were computed by treating *Unsure* responses as non-pass cases, while agreement corresponds to the percentage of samples for which all three annotators assigned the same label. Raw evaluation results for each sample are available in our public repository. Overall, the observed agreement rates indicate reasonably consistent judgments across annotators.

Table 16: Quality validation results for text samples.

Category	n	Pass Rate (%)	Agreement (%)
AR	200	87.0	70.0
FS-1	200	94.2	83.5
FS-2	200	90.7	72.0
FS-3	200	92.0	76.0
FS (all)	600	92.3	77.2
FB-1	200	96.0	88.0
FB-2	200	89.3	72.0
FB-3	200	96.2	88.5
FB (all)	600	93.8	82.8

Table 17: Quality validation results for image samples.

Category	n	Pass Rate (%)	Agreement (%)
AR	100	98.0	94.0
FS-1	16	89.6	68.8
FS-2	17	90.2	70.6
FS-3	17	100.0	100.0
FS (all)	50	93.3	80.0
FB-1	17	96.1	88.2
FB-2	17	92.2	76.5
FB-3	16	91.7	75.0
FB (all)	50	93.3	80.0

B Model Details

B.1 Over-Alignment Target Statistics

	HR	HF	AR	AF
μ^*_{oa}	0.10	0.02	0.65	0.95
σ^*_{oa}	0.12	0.05	0.04	0.01
u^*_{oa}	0.2	0.1	5.0	**10.0**

Table 18: Class-specific over-alignment targets used in OAM supervision. Values are estimated from validation statistics and clipped to ensure stable optimization.

To stabilize training of the Over-Alignment Module (OAM), we define class-specific target statistics for the alignment distribution. These targets are derived from validation-set empirical statistics and clipped for numerical stability.(see Table 18)

C Experimental Details

C.1 MiRAGeNews

The MIRAGeNews dataset details are shown in Table 19.

Table 19: MiRAGeNews dataset statistics.

Split	Real	Fake	Total
Train	5,000	5,000	10,000
Val	1,250	1,250	2,500
Test (ID)	250	250	500
Test (OOD)	1,000	1,000	2,000
Total	7,500	7,500	15,000

C.2 DGM^4

The DGM^4 dataset details are shown in Table 20.

C.3 Implementation Details

All models are implemented in PyTorch and trained for 10 epochs using AdamW (LR=1×10^{-4}, weight decay=1×10^{-4}, batch size=16) with early stopping (patience=4) based on validation macro-F1. Following Jacob Devlin et al. (Devlin et al., 2019), only BERT and RoBERTa are fine-tuned with a learning rate of 2×10^{-5}. We fine-tune only the last two RoBERTa layers while freezing CLIP RN101. Class imbalance is handled using WeightedRandomSampler ($\alpha = 0.5$). For OA, the alignment targets are set to $\mu_{\text{oa}} = 0.6$ for AR and $\sigma_{\text{oa}} = 0.15$ for HR samples. For evaluation, HARF-News and MiRAGeNews results are averaged over five stratified folds (k=5) to account for dataset-specific characteristics, while DGM4 is evaluated on a single run following the official repository guidelines. For zero-shot vlm baselines, we additionally evaluate two large vision–language models (VLMs) in a *zero-shot* setting without task-specific fine-tuning, providing reference points for prompt-only multimodal inference. **Qwen2.5-VL-7B-Instruct** is loaded in `bfloat16` precision with Flash Attention 2 and automatic multi-GPU distribution; inference uses greedy decoding (`do_sample=False`, $T = 0$) with a batch size of 4. **GPT-5-mini** (`gpt-5-mini`) is queried through the OpenAI Vision API with up to 8 concurrent asynchronous requests and structured JSON output enforced via `response_format`. Both models share the same structured system prompt: *"You are an expert multimodal news analyst specializing in detecting AI-generated content and misinformation. Analyze the news headline and accompanying image jointly to classify the sample into one of four categories: HR, HF, AR, or AF, based on authorship (Human vs. AI) and veracity (Real vs. Fake). Consider textual style, factual consistency, image authenticity, visual artifacts, and cross-modal alignment. Respond only with a JSON object containing the predicted label, confidence score, and brief reasoning."* Both models also receive images resized and center-cropped to 224×224 pixels (identical to the CLIP RN101 preprocessing) and are evaluated on the same stratified 20% test split (seed$= 42$) used throughout the paper.

Table 20: DGM^4 dataset statistics.

Split	Pristine	Manipulated	Total
Train	104,092	104,092	208,184
Val	7,473	14,653	22,126
Test	16,876	33,829	50,705
Total	128,441	152,574	281,015

D Additional Results

D.1 Per-Strategy Detection Performance Analysis

Table 21 summarizes per-strategy detection performance on AR and AF subsets.

AI Real (AR). AR achieves the highest performance among AI-generated categories (Acc = 91.1, F1 = 89.6), mainly driven by the Authorship Module (AM), which captures the consistently low token-level variance of LLM rewriting. The near-perfect AUC (99.9) suggests that stylistic regularities introduced by gpt-5-mini are robustly encoded

Strategy	Acc	F1	AUC
AR	91.1	89.6	99.9
FB-1	80.7	59.6	99.8
FB-2	90.0	65.7	99.9
FB-3	63.8	51.5	99.7
FS-1	82.9	64.1	99.8
FS-2	76.2	61.7	99.8
FS-3	94.1	71.1	99.9

Table 21: Per-strategy detection performance on AR and AF subsets.

in the learned representations.

Factual Shift strategies (FS-1, FS-2, FS-3). FS-3 (Small shift) achieves the highest performance (Acc = 94.1, F1 = 71.1), while FS-2 (Large shift) performs worst (Acc = 76.2, F1 = 61.7), with FS-1 (Inversion) in between (Acc = 82.9, F1 = 64.1). Small localized edits in FS-3 create sharper text-image inconsistencies that VM can exploit more reliably, whereas FS-2's broader rewriting may preserve thematic coherence. Uniformly high AUC values ($\geq$ 99.7) indicate that lower F1 mainly reflects thresholding difficulties under class imbalance rather than weak discrimination.

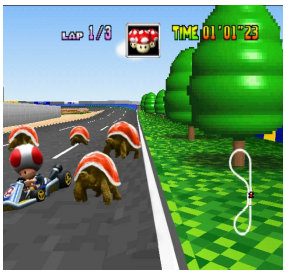
Framing-Based strategies (FB-1, FB-2, FB-3). The FB family underperforms relative to FS, with average Accuracy/F1 of 78.2/58.9 versus 84.4/65.7 for FS. FB-3 (Disguise) is the hardest setting (Acc = 63.8, F1 = 51.5) because its formal news-reporting tone makes AI-written headlines stylistically resemble human journalism, weakening AM. Additionally, rewriting headlines to be "harder to match" to the original text reduces VM's ability to capture text-image inconsistencies. Since FB samples retain unedited human photographs, OAM provides little additional signal, making FB-3 the most challenging scenario in HARF-News.

D.2 Case Study

We also conducted a case study on correctly and misclassified samples in HARF-News, revealing distinct signal patterns across classes. HF shows collapsed overlap statistics, AF exhibits highly concentrated text-image alignment, HR maintains weak and uneven alignment, and AR lies in an intermediate regime. Misclassified samples tend to shift toward other class-specific patterns, suggesting that HARFNet distinguishes classes primarily through the strength and uniformity of patch-level semantic alignment. Detailed examples are provided below.

(a) HR — *Human Real*

"this year old bottle of isopropyl alcohol from a defunct pharmacy i found in my basement"



(b) HF — *Human Fake*

"some major hazards on the race track for mario and friends"

(c) AR — *AI Real*

"Volunteer at Festival of Trees Sent Away for Wearing Pants"

(d) AF — *AI Fake*

"the shirt looks like a shar pei lying down, evoking a playful comparison to the beloved dog breed"

Cat.	Conf.	p_auth	p_fake	p_af	oa_mean	oa_uni
HR	0.780	0.000	0.007	0.002	0.299	0.440
HF	0.900	0.000	0.994	0.027	0.005	0.020
AR	0.938	1.000	0.005	0.001	0.617	2.616
AF	0.897	1.000	0.994	0.995	0.958	10.00

Figure 4: Representative correctly classified samples from each HARF class. HR and HF exhibit weak or collapsed alignment patterns, whereas AR and AF show progressively stronger semantic alignment. AF particularly demonstrates hyper-concentrated overlap statistics, reflecting strong text-image corroboration.

D.3 Case study of correctly classified samples

Human Real (HR). The HR sample—a photograph of an isopropyl alcohol bottle with a humorous caption—yields **p_auth** $= 0.000$ and **p_fake** $= 0.007$, correctly indicating human-generated, real content. The moderate **oa_mean** (0.299) with low **oa_uni** (0.440) reflects the naturally uneven patch-level alignment typical of unmanipulated photographs.

Human Fake (HF). The HF sample—a Mario Kart screenshot composited with real-world tortoises—is identified as AI-manipulated by **p_fake** $= 0.994$, the highest among all cases. Critically, **oa_mean** collapses to 0.005 and **oa_uni**

to 0.020: the OAM interprets this near-zero patch alignment as evidence of a globally incoherent composition, where the image and text share no meaningful semantic overlap despite being AI-paired.

AI Real (AR). The AR sample—a news headline overlaid on a real photograph— achieves **p_auth** $= 1.000$, reflecting the VM's detection of authentic image provenance. The elevated **oa_mean** (0.617) and **oa_uni** (2.616) indicate a structured but non-extreme patch alignment: the image genuinely depicts the scene described in the headline, yet lacks the hyper-concentrated alignment signature of the AF class.

AI Fake (AF). The AF sample—Notre-Dame Cathedral during post-fire reconstruction—exhibits the most distinctive signal profile. Both **oa_mean** (0.960) and **oa_uni** (10.00) reach their clipped maximum (`clamp(max=10)` in OAM), indicating that every image patch is almost uniformly and strongly aligned with the caption text. This hyper-concentrated over-alignment is the defining signature of the AF class, arising when an AI system deliberately selects or generates an image to maximally corroborate a factual claim. Combined with **p_auth** $= 1.000$ and **p_af** $= 0.990$, the model assigns AF at confidence 0.877.

D.4 Case study of misclassified samples

Figure 5 presents 4 failure cases that reveal the boundary conditions of our model's signal space.

HR misclassified as HF (a). The caption *"and now the really big coal plants begin to close"* is paired with a real photograph, yet the model predicts HF at low confidence (0.531). The borderline **p_fake** (0.500) and unusually low **oa_mean** (0.039) and **oa_uni** (0.044) cause the sample to mimic the HF alignment pattern. We attribute this to the abstract, weakly visual caption, which produces poor patch-level alignment despite the genuine image.

HF misclassified as HR (b). The caption describes a coincidental logo match at a grocery store outlet, and the model predicts HR (0.757). Despite the manipulated image, **p_fake** $= 0.013$ fails to detect falsity, while **oa_mean** (0.269) and **oa_uni** (0.138) remain within the HR range.

AR misclassified as AF (c). The caption *"the spot where you can literally stand on your butt"* is paired with a real photograph, yet **p_af** $= 0.994$ and **oa_uni** $= 10.00$ trigger a confident AF prediction (0.889). Highly salient visual regions strongly align with the caption tokens, inadvertently reproducing the AF-style over-alignment signature.

(a) HR→HF — *Human Real*

"and now the really big coal plants begin to close"

(b) HF→HR — *Human Fake*

"this outlet at the grocery store i work at a restaurant called tonight and their logos matched up perfectly"

(c) AR→AF — *AI Real*

"the spot where you can literally stand on your butt"

(d) AF→AR — *AI Fake*

"Wildflowers sprouting in the gutters of this abandoned cabin"

True	Pred.	Conf.	p_auth	p_fake	p_af	oa_mean	oa_uni
HR	HF	0.531	0.000	0.500	0.004	0.039	0.044
HF	HR	0.757	0.000	0.013	0.002	0.269	0.138
AR	AF	0.889	1.000	0.000	0.994	0.958	10.00
AF	AR	0.965	1.000	0.000	0.001	0.632	4.253

Figure 5: Case study of misclassified samples. Arrow (→) denotes true→predicted label. **p_auth/p_fake** denote authenticity scores, **p_af** denotes AF probability, **oa_mean/oa_uni** denote overlap statistics.

AF misclassified as AR (d). The caption *"Wildflowers sprouting in the gutters of this abandoned cabin"* describes a visually plausible scene. Although **p_auth** $= 1.000$ correctly detects AI authorship, **p_af** collapses to 0.001 because **oa_mean** (0.632) and **oa_uni** (4.253) remain below the extreme alignment pattern typical of AF. This suggests that naturalistic, low-salience AF samples are harder to detect through patch-level alignment.

D.5 Class-wise ablation study Analysis

Table 22 and Table 23 report the per-class F1 scores for each ablation variant. Overall, AI-related

Model	4-way F1	H/A F1	R/F F1
HARFNet	**87.6**	**100.0**	88.5
w/o BM	87.4	100.0	**88.6**
w/o BM + OAM	87.3	100.0	88.6
w/o AM	87.2	100.0	88.3
w/o VM	85.8	100.0	85.6

Table 22: HARFNet without module study on the HARF-News.

Model	HR	HF	AR	AF	Macro Avg
Concat Baseline	88.9	83.5	89.1	82.9	86.1
w/o BM	90.8	86.0	89.1	83.0	87.4
w/o OAM	90.8	86.1	89.4	83.0	87.3
w/o AM	90.5	85.7	88.2	81.8	87.2
w/o VM	89.2	84.4	86.1	77.2	85.8
w/o BM+OAM	90.6	85.9	88.7	82.1	87.3
w/o BM+OAM+AM	90.9	86.4	89.1	82.7	87.3
w/o BM+OAM+VM	85.3	78.3	89.2	83.1	84.0
HARFNet (Full)	**90.9**	**86.2**	**89.6**	**83.7**	**87.6**

Table 23: Class-wise F1 score on HARF-News.

classes (AR, AF) are consistently more sensitive to component removal than human-authored classes (HR, HF), indicating that AI-generated samples rely more heavily on fine-grained multimodal reasoning.

Bilinear Interaction(BM). Removing the BM causes the smallest overall degradation, but the impact is already asymmetric: AR and AF decrease more noticeably than HR and HF. This suggests that the BM mainly strengthens the joint activation of authorship, veracity, and over-alignment signals, which is particularly important for AF detection.

Over-Alignment Module (OAM). Further removing OAM primarily affects AF and AR, while the human-authored classes remain relatively stable. This reflects OAM's role in modeling patch-level alignment concentration, a characteristic signal of AI-generated images. Without OAM, AF becomes substantially harder to distinguish from AR.

Authorship Module (AM). Removing AM causes the largest non-visual degradation on AR, indicating that stylometric residual patterns are especially important for identifying AI-generated real news. In contrast, AF relies more heavily on visual and cross-modal inconsistencies than on textual style alone.

Veracity Module (VM). Removing VM produces by far the largest degradation across all classes, particularly AF. This confirms that cross-modal discrepancy signals are the primary cue for separating AI-generated fake news from semantically aligned but truthful content.

Overall, the results demonstrate that each module captures a distinct aspect of the authorship-veracity space. VM is most critical for AF detection, AM is especially important for AR, and the human-authored classes remain relatively robust across most ablation settings.

D.6 Ablation Study on benchmarks

The detailed ablation study results are shown in Table 24.

D.7 Class balanced test evaluation

Tables 25 and 26 report overall and class-wise performance on the class-balanced test set.

Table 25: HARFNet overall task performance on class balanced test set.

Metric	**Human vs AI**	**Real vs Fake**	**4-way**
Accuracy	1.0000	0.8688	0.8688
Macro-F1	1.0000	0.8684	0.8683

Table 26: HARFNet class-wise F1 scores for 4-way classification.

Metric	**HR**	**HF**	**AR**	**AF**	**Macro Avg**
F1 Score	0.8848	0.8801	0.8657	0.8424	0.8683

Since the HARF-News dataset is class-imbalanced, with HR and HF substantially outnumbering AR and AF, we additionally evaluate on a class-balanced test subset (400 samples per class, randomly under-sampled from the original 20% test split) to verify that the reported performance is not biased by class imbalance.

Results show that the model remains robust under balanced conditions. The Human vs. AI task achieves perfect Accuracy and Macro-F1 (1.0000), indicating highly discriminative authorship cues. For the Real vs. Fake and 4-way tasks, both Accuracy and Macro-F1 remain stable at approximately 0.87, suggesting that the original performance was not artificially inflated by the skewed class distribution.

At the class level, HR and HF achieve the highest F1 scores (0.8848 and 0.8801), while AR and AF are slightly lower (0.8657 and 0.8424). In particular, AF shows the lowest recall (0.7750), indicating that AI-generated fake content with strong text-image alignment remains the most challenging case. Consistent with the confusion matrix, AF samples are frequently misclassified as AR, suggesting that subtle over-alignment patterns are

Model	MiRAGeNews															DGM[4]		
	NYT+MJ (ID)			BBC+DE			CNN+DE			BBC+SX			CNN+SX					
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Baseline (Concat)	99.4	99.4	100.0	80.0	79.7	85.7	81.8	81.7	90.0	78.9	78.5	83.6	81.7	81.6	90.2	73.3	71.8	82.4
w/o BM	99.3	99.3	**100.0**	80.7	80.1	94.1	83.3	83.1	95.5	79.7	79.1	92.7	84.5	84.4	95.2	81.3	79.6	90.0
w/o OAM	99.6	99.6	**100.0**	86.5	86.4	95.3	**89.3**	**89.3**	**96.3**	85.0	84.9	94.0	**89.5**	**89.5**	96.5	80.3	78.8	88.8
w/o AM	99.6	99.6	**100.0**	85.3	85.2	94.0	87.2	87.2	95.4	84.2	84.0	94.4	88.3	88.3	96.7	**81.7**	**80.0**	**90.3**
w/o VM	99.5	99.5	**100.0**	86.5	86.4	94.3	87.8	87.8	95.6	**87.0**	**87.0**	94.9	89.4	89.4	**96.8**	78.8	77.8	89.1
w/o BM+OAM	99.5	99.5	**100.0**	86.3	86.2	95.2	89.0	89.0	96.0	86.1	85.9	94.7	**90.5**	**90.5**	96.7	80.2	78.4	88.4
w/o BM+OAM+AM	**99.7**	**99.7**	**100.0**	84.2	84.1	94.2	84.8	84.7	95.0	83.0	82.8	93.2	85.9	85.7	94.9	80.2	78.5	88.6
w/o BM+OAM+VM	92.7	92.7	97.7	68.5	65.4	93.0	84.5	84.2	94.9	68.1	64.7	93.3	84.8	84.6	94.7	51.7	51.6	63.0
HARFNet (Full)	99.6	99.6	**100.0**	**87.1**	**87.0**	**95.4**	**89.3**	**89.3**	96.1	85.8	85.7	**95.2**	89.3	89.3	96.5	81.2	79.6	90.0

Table 24: Ablation study on MiRAGeNews and DGM[4]. HARFNet demonstrates strong cross-domain robustness. In MiRAGeNews, the prefix (e.g., NYT) denotes the news source, while the suffix (e.g., MJ) denotes the image generator. DE and SX represent DALL-E and SDXL, respectively.

still difficult to distinguish in visually plausible AI-generated images.

E Disclosure of AI Usage

The authors acknowledge the use of AI assistants, including ChatGPT[2], Gemini[3], and Claude[4], to support the writing and coding processes. All AI-generated content was verified for correctness by one or more of the authors.

[2] https://chatgpt.com/
[3] https://gemini.google.com
[4] https://claude.ai/