UMPIRE : Unveiling LLM-generated Posts via Redundant Expressions

Xiaoquan Yi[1], **Haixing Wu**[2], **Haozhao Wang**[1*], **Yichen Li**[1], **Yuhua Li**[1*], **Rui Zhang**[1], **Ruixuan Li**[1*]
[1]School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2]School of Mathematics and Statistics
Wuhan University of Technology, Wuhan, China.
{yixiaoquan,hz_wang,idcliyuhua,rxli}@hust.edu.cn

Abstract

The proliferation of Large Language Models (LLMs) has saturated social media platforms with hyper-realistic posts, rendering traditional detection methods that rely on low-level artifacts or unimodal statistics increasingly ineffective. In this work, we identify a fundamental semantic distinction: humans tend to complement visual content with additional context, while LLMs predominantly describe the visual information. To capture this, UMPIRE employs an orthogonal semantic decomposition mechanism that disentangles textual embeddings into redundant and complementary components. An adaptive gating module dynamically weighs these components to reflect diverse communicative styles. To enforce the desired geometric structure, we introduce a latent contrastive redundancy regularization loss that encourages LLM-generated content to exhibit high semantic redundancy, while human-written content emphasizes complementarity. Experimental results demonstrate that UMPIRE significantly outperforms state-of-the-art detection methods across multiple datasets, achieving up to a 5.38% improvement in accuracy.

1 Introduction

The widespread adoption of large language models (LLMs) on social media platforms has fundamentally reshaped online content creation (Tolstykh et al., 2024; Lin et al., 2025). Platforms (e.g., X, Reddit, and Instagram) are no longer merely repositories of human experience but are increasingly populated by hyper-realistic, LLM-generated multimodal posts (Macko et al., 2025; Goswami et al., 2025). While these capabilities enhance content creation, they also lower the barrier to automated disinformation campaigns, social botnets, and sophisticated impersonation attacks (Ignat et al., 2025). Unlike traditional deepfakes or spam, modern LMM-generated posts exhibit high semantic coherence and stylistic naturalness, rendering them largely imperceptible to casual human observers and conventional detection systems alike (Zhao et al., 2025; Meguellati et al., 2025). Therefore, it is urgent to develop a new technology for detecting multimodal posts generated by LLMs.

Existing approaches to verifying the provenance of multimodal content generally fall into two categories, both of which face limitations in the dynamic social media environment. Unimodal detection focuses on exploiting statistical irregularities in generated text (e.g., perplexity and burstiness) (Mitchell et al., 2023; Bao et al., 2023) or visual artifacts in images (e.g., frequency-domain fingerprints) (Zhang et al., 2025). However, these methods are vulnerable to adversarial paraphrasing, image compression, and other post-processing techniques commonly employed on social platforms (Zhao and Chen, 2025; Pagan et al., 2025). Multimodal detection methods attempt to jointly analyze image–text pairs using cross-modal consistency checks or multimodal classifiers (Tahmasebi et al., 2024; Huang et al., 2025). While more robust than unimodal approaches, these methods often treat the problem as a black-box classification task, neglecting the underlying behavioral patterns that differentiate human and LLM-generated posts (Sun et al., 2025; Lu et al., 2025).

In this work, we propose a behavior-centric perspective on content generation in social media, identifying a clear divergence in communicative patterns between humans and LLMs. Human communication is implicitly optimized by the Principle of Least Effort (Kluckhohn, 1950; Zipf, 2016). When a human user shares an image, they treat the visual signal as established context. Consequently, human captions rarely waste bandwidth restating visually obvious facts. Instead, they focus on providing complementary information context, emotion, sarcasm, or backstory. In contrast, LLMs

*Haozhao Wang,Yuhua Li,Ruixuan Li are corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30901–30913

optimize for semantic alignment(Liu et al., 2023a; Wang et al., 2024a; Team et al., 2025). Even when prompted to simulate social media posts, they retain strong inductive biases from pretraining (e.g., captioning objectives) that encourage faithful translation of visual content into text. This manifests as a distinct behavioral signature: *humans complement the visual reality, while LLMs describe it*.

Motivated by this observation, we introduce a novel framework named UMPIRE for detecting LLM-generated multimedia posts. UMPIRE models the semantic information flow between vision and language. We propose an orthogonal semantic decomposition mechanism that disentangles textual embeddings into two geometric components: a *redundant component* aligned with the visual content, and a *complementary component* orthogonal to it. An adaptive gating module dynamically weighs these components to capture the diverse communicative styles of social media users. Experimental results demonstrate that our framework substantially improves accuracy in identifying LLM-generated posts across multiple datasets. Our contributions are summarized as follows:

- We identify and formalize a fundamental behavioral divergence in content generation: *humans complement the visual reality, while LLMs describe it*. To the best of our knowledge, this is the first work to quantitatively characterize this semantic divergence as a robust invariant for detection.
- We propose UMPIRE , which exploits orthogonal semantic decomposition to detect LMM-generated posts. To our knowledge, this is the first study to use geometric orthogonality for LLM-generated posts discrimination.
- Extensive experiments conducted on various datasets and advanced LLMs, *demonstrate that our method outperforms state-of-the-art detection methods by up to 5.38%*

2 Related Work

2.1 Unimodal Detection of Generated Content

Early detection paradigms largely treated generated content verification as a unimodal artifact identification task. In the textual domain, methods typically exploit statistical irregularities in model probability distributions. Zero-shot approaches like DetectGPT (Mitchell et al., 2023) and Fast-DetectGPT (Bao et al., 2023) utilize the curvature of the log-probability function to distinguish machine-generated text. Supervised methods often fine-tune transformer backbones (e.g., RoBERTa, T5) to capture subtle syntactic markers (Wang et al., 2024b; Chen et al., 2023). In the visual domain, detection relies heavily on low-level signal anomalies, such as frequency spectrum artifacts (Lao et al., 2024; Wang et al., 2025) or inconsistent noise fingerprints (Jang et al., 2025). While effective in controlled settings, these unimodal methods face severe robustness challenges in the social media wild. Aggressive image compression and resizing on platforms like X or Instagram often obliterate high-frequency visual artifacts (Zhao et al., 2025). Similarly, the informal, terse nature of social media captions (often lacking sufficient token length) renders statistical text metrics unreliable.

2.2 Multimodal Detection

Recognizing the limitations of single modalities, recent research has pivoted toward multimodal frameworks. A dominant stream of work, often adapted from fake news detection, focuses on identifying *cross-modal inconsistencies*. Methods like SIDA (Huang et al., 2025) employ joint vision-language encoders to detect logical contradictions or mismatched entities between text and image. Other approaches, such as ETS-MM (Li et al., 2025) and LVLM4EV (Tahmasebi et al., 2024), utilize hierarchical fusion or external knowledge retrieval to verify content veracity. These methods operate on the assumption that generated content is prone to hallucinations or logical errors. However, state-of-the-art LLMs (e.g., GPT-4o, Gemini) are explicitly optimized for high semantic alignment, rendering inconsistency-based cues increasingly scarce. Furthermore, these methods typically employ black-box fusion (e.g., concatenation), failing to explicate the relationship between modalities.

3 Motivation: Humans complement, while LLMs describe

Our approach is grounded in a stark cognitive divergence between human and LLMs communication (Cherry, 1966; Heath and Bryant, 2013). Human communication on social platforms implicitly follows the *Principle of Least Effort* (Kluckhohn, 1950; Zipf, 2016). When sharing an image, humans rarely utilize textual bandwidth to restate visually

Table 1: **Quantitative Analysis of Semantic Dependency.** Breakdowns of textual information sources reveal distinct behaviors: LLM-generated content relies on describing visible content (High Redundancy), while human-written content focuses on adding external context (High Complementarity).

Source	Visually Grounded (Redundancy)		External Context (Complementarity)		Semantic Gap
	Obj. Mentioned	Visual Attr.	Emotion/Opinion	Temp./Spatial	
Human	28.4%	15.2%	**68.7%**	**42.1%**	*Information Gain*
LMMs	**89.3%**	**76.5%**	12.4%	5.8%	*Information Loop*
Δ *Diff.*	↑ 60.9%	↑ 61.3%	↓ 56.3%	↓ 36.3%	**Distinct Pattern**

obvious facts. Instead, assuming the viewer perceives the visual signal, humans focus on providing complementary information (e.g., emotional context, backstory, or temporal grounding) to maximize information gain. Conversely, LLMs operate under an objective of *Semantic Alignment Maximization*. Trained to minimize cross-modal loss (e.g., contrastive loss or captioning loss), these models tend to generate text that faithfully describes the visual content. Consequently, they act as translators of visual features into text, resulting in high semantic redundancy.

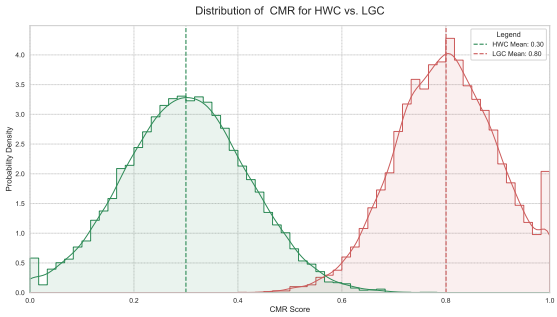

Figure 1: The distribution of CMR scores for human-written and LLM-generated posts. It reflects LLMs' tendency to emphasize visually grounded descriptions over complementary external context.

To validate this finding, we conducted a rigorous empirical analysis comparing human-written content and LLM-generated content on SoMe (Xue et al., 2025). First, we examined the overall distribution of Cross-Modal Redundancy (CMR) scores. As visualized in Figure 1, the results reveal a clear bi-modal distribution where human-written content is centered around a low redundancy mean ($\mu \approx 0.30$), indicating that humans minimize visual repetition. In contrast, LLM-generated content exhibits a high-redundancy distribution ($\mu \approx 0.80$), confirming that models prioritize descriptive fidelity over communicative efficiency.

We further dissected the composition of this redundancy through a fine-grained semantic dependency analysis, categorizing textual tokens into *Visually Grounded* (redundant) versus *External Context* (complementary). As presented in Table 1, the data exposes a decisive behavioral gap. LLM-generated content displays an overwhelming dependence on visual signals, with **89.3%** of mentioned objects and **76.5%** of attributes being directly observable in the image. This confirms that LLM-generated text acts largely as a subset of visual information. Instead, humans minimize this overlap, with only **28.4%** of text referencing visual objects. Humans dominate the external context domain, with **68.7%** of captions containing subjective emotions or opinions—information that cannot be inferred from pixel data alone. These findings suggest that semantic redundancy is not an incidental artifact but an intrinsic behavioral characteristic of current generative models. Unlike low-level artifacts, this semantic feature is rooted in the fundamental training objectives of LLMs and is thus more resistant to surface-level adversarial removal.

4 Problem Formulation

Let $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$ denote a multimodal dataset with N samples, where each sample consists of an image $I_i \in \mathcal{I}$, a corresponding text caption $T_i \in \mathcal{T}$, and a binary label $y_i \in \{0, 1\}$. The label $y_i = 0$ indicates *human-written content*, in which the text T_i is authored by a human conditioned on the image I_i. Conversely, $y_i = 1$ denotes *LLM-generated post*, where the text T_i is synthesized by a large language model given the image I_i. Our objective is to learn a parameterized decision function $f_\theta : \mathcal{I} \times \mathcal{T} \to [0, 1]$ that estimates the posterior probability.

$$f_\theta(I_i, T_i) = P(y_i = 1 \mid I_i, T_i) \tag{1}$$

which measures the likelihood that an image–text pair is generated by a LLM rather than a human. We optimize the model parameters θ by minimizing the expected classification loss $\mathcal{L}$:

$$\min_\theta \mathbb{E}_{(I,T,y)\sim\mathcal{D}}\big[\mathcal{L}(f_\theta(I, T), y)\big]. \tag{2}$$

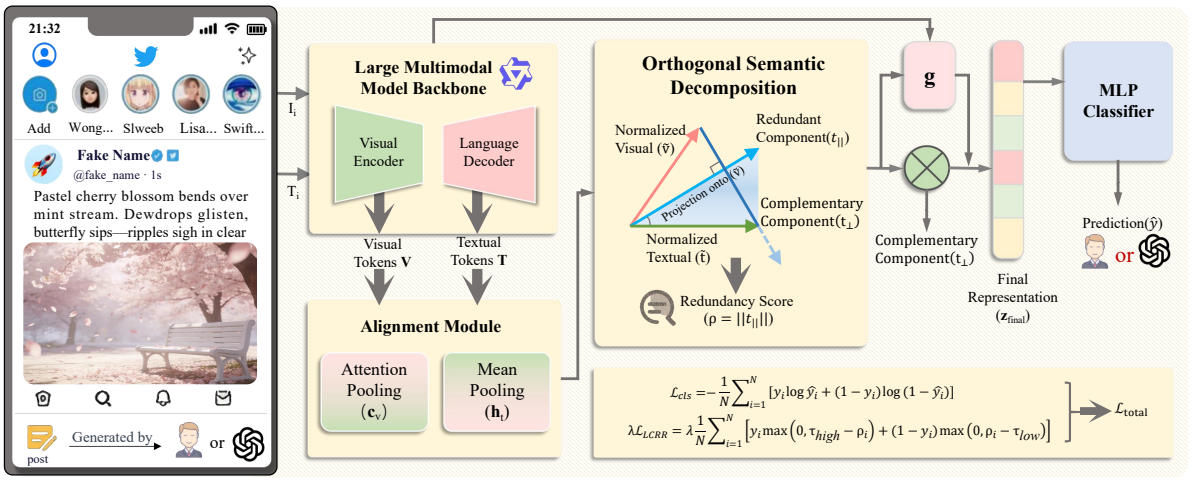

Figure 2: Overview of UMPIRE framework. Given an image–text pair, UMPIRE encodes visual and textual tokens with an LLM and aligns them into global representations. Orthogonal semantic decomposition separates redundant and complementary text components relative to the visual centroid.

The expectation is approximated by the empirical average over the training set:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(I_i, T_i), y_i). \tag{3}$$

The parameters θ are optimized via stochastic gradient-based methods, yielding

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(I_i, T_i), y_i). \tag{4}$$

5 Methodology

This section presents UMPIRE which distinguishes LLM-generated multimodal content on social media. As shown in Figure 2, UMPIRE disentangles visual redundancy from contextual complementarity to enable effective detection.

5.1 Feature Extraction and Alignment

Feature Extraction. Let $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$ denote a multimodal dataset, where $y_i \in \{0, 1\}$ indicates human-written (0) or LLM-generated (1) content. We adopt Qwen2-VL as the multimodal backbone to extract high-level semantic representations. Given an image–text pair (I, T), the visual encoder and language decoder produce sequences of visual tokens $\mathbf{V} \in \mathbb{R}^{L_v \times d}$ and textual tokens $\mathbf{T} \in \mathbb{R}^{L_t \times d}$, respectively.

Visual Centroid Alignment. As visual semantics are distributed across spatial patches, defining a global measure of redundancy requires a unified visual reference. $\mathbf{W}_v \in \mathbb{R}^{d \times d}$ and $\mathbf{w}_a \in \mathbb{R}^d$ are learnable parameters. We compute a global visual centroid $\mathbf{c}_v$ using a learnable attention-based pooling mechanism:

$$\begin{aligned} \alpha_j &= \frac{\exp\left(\mathbf{w}_a^\top \tanh(\mathbf{W}_v \mathbf{v}_j)\right)}{\sum_{k=1}^{L_v} \exp\left(\mathbf{w}_a^\top \tanh(\mathbf{W}_v \mathbf{v}_k)\right)}, \\ \mathbf{c}_v &= \sum_{j=1}^{L_v} \alpha_j \mathbf{v}_j, \end{aligned} \tag{5}$$

Textual Summary Pooling. Let $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{L_t}] \in \mathbb{R}^{L_t \times d}$ denote the sequence of textual token embeddings produced by the language decoder for the caption T, where $\mathbf{t}_i \in \mathbb{R}^d$ is the hidden representation of the i-th token (e.g., a subword in the tokenizer vocabulary) at the final decoder layer. We derive a global textual representation $\mathbf{h}_t \in \mathbb{R}^d$ by mean pooling:

$$\mathbf{h}_t \triangleq \mathrm{MeanPool}(\mathbf{T}) = \frac{1}{L_t} \sum_{i=1}^{L_t} \mathbf{t}_i. \tag{6}$$

This operation aggregates variable-length token sequences into a fixed-dimensional summary vector, which is subsequently used for cross-modal alignment with the visual centroid $\mathbf{c}_v$.

5.2 Orthogonal Semantic Decomposition

We posit that the global textual representation $\mathbf{h}_t$ can be decomposed into two orthogonal components with respect to the visual centroid $\mathbf{c}_v$, corresponding to redundant and complementary semantics. To tailor the representation space for detection, we first project both modalities into a detection-specific metric space via learnable projection heads $\phi_v(\cdot)$ and $\phi_t(\cdot)$:

$$\tilde{\mathbf{v}} = \mathrm{Norm}(\phi_v(\mathbf{c}_v)), \quad \tilde{\mathbf{t}} = \mathrm{Norm}(\phi_t(\mathbf{h}_t)), \tag{7}$$

where $\mathrm{Norm}(\cdot)$ denotes L_2 normalization. Based on our finding, we perform a geometric decomposition of the textual representation relative to the visual direction.

Redundant Component ($\mathbf{t}_{\parallel}$). The redundant component is defined as the projection of the textual embedding onto the visual subspace, capturing visually grounded and descriptive semantics that tend to be overemphasized in LLM-generated content:

$$\mathbf{t}_{\parallel} = \mathrm{proj}_{\tilde{\mathbf{v}}}(\tilde{\mathbf{t}}) = (\tilde{\mathbf{t}}^{\top}\tilde{\mathbf{v}})\,\tilde{\mathbf{v}}. \tag{8}$$

Complementary Component ($\mathbf{t}_{\perp}$). The complementary component corresponds to the residual orthogonal to the visual subspace, encoding contextual, inferential, or subjective information that is more prevalent in human-written content:

$$\mathbf{t}_{\perp} = \tilde{\mathbf{t}} - \mathbf{t}_{\parallel}. \tag{9}$$

Finally, we define the scalar CMR score as the magnitude of the projected component:

$$\rho \triangleq \|\mathbf{t}_{\parallel}\| = \left|\tilde{\mathbf{t}}^{\top}\tilde{\mathbf{v}}\right|, \tag{10}$$

which quantifies the degree of semantic alignment between text and image.

5.3 Adaptive Gated Fusion

Simply concatenating semantic components implicitly assumes a fixed contribution from redundancy and complementarity across all samples. However, human descriptions exhibit substantial variability in how information is conveyed. To accommodate this diversity and improve robustness, we introduce an adaptive gate that dynamically modulates the relative importance of redundant and complementary semantics:

$$\mathbf{g} = \sigma\big(\mathbf{W}_g\,[\mathbf{t}_{\parallel}; \mathbf{t}_{\perp}] + \mathbf{b}_g\big)\,, \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The final representation $\mathbf{z}_{\mathrm{final}}$ explicitly preserves the underlying geometric structure by integrating the gated components:

$$\mathbf{z}_{\mathrm{f}} = \mathrm{Concat}\big(\tilde{\mathbf{t}}, \mathbf{g}\odot\mathbf{t}_{\parallel}, (\mathbf{1}-\mathbf{g})\odot\mathbf{t}_{\perp}, \rho\big)\,, \tag{12}$$

where $\odot$ denotes element-wise multiplication and ρ is the redundancy score. The final prediction is obtained via a lightweight classifier:

$$\hat{y} = \mathrm{MLP}(\mathbf{z}_{\mathrm{f}}). \tag{13}$$

5.4 Training Objective

Standard classification objectives alone are insufficient to enforce the desired geometric structure in the latent space. We propose a hybrid training objective that jointly optimizes classification performance and cross-modal geometric consistency.

Classification Loss ($\mathcal{L}_{\mathrm{cls}}$). We adopt binary cross-entropy loss to train the model to distinguish human-authored from LLM-generated content:

$$\mathcal{L}_{\mathrm{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{14}$$

where $\hat{y}_i \in (0,1)$ denotes the predicted probability that the i-th image–text pair is LLM-generated.

Latent Contrastive Redundancy Regularization ($\mathcal{L}_{\mathrm{LCRR}}$). As the core contribution of UMPIRE , we explicitly regularize the latent geometry induced by the orthogonal semantic decomposition. Specifically, for LLM-generated content ($y_i = 1$), we encourage strong semantic alignment between modalities (high redundancy), whereas for human-written content ($y_i = 0$), we promote orthogonality (low redundancy). This intuition is formalized via a margin-based contrastive loss:

$$\mathcal{L}_{\mathrm{LCRR}} = \frac{1}{N}\sum_{i=1}^{N}\big[y_i \max(0, \tau_{\mathrm{high}} - \rho_i) + (1-y_i)\max(0, \rho_i - \tau_{\mathrm{low}})\big], \tag{15}$$

where ρ_i denotes the redundancy score of the i-th sample, and τ_{high} and τ_{low} define the desired redundancy margins for LLM-generated and human-written content, respectively. This regularization effectively pushes LLM embeddings toward the visual axis while pulling human embeddings toward the orthogonal subspace.

Overall Objective. The entire framework is trained by minimizing the following objective:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{cls}} + \lambda\,\mathcal{L}_{\mathrm{LCRR}}, \tag{16}$$

where λ controls the strength of geometric regularization. By optimizing $\mathcal{L}_{\mathrm{LCRR}}$, UMPIRE learns to classify content sources and explicitly measure the *semantic redundancy* between vision and language.

6 Experiments

6.1 Dataset

To evaluate UMPIRE , we construct a benchmark of multimodal social media posts that reflects the con-

Table 2: Experimental results on the TwiBot-22, Fakeddit, and Hateful Memes datasets. The best number is highlighted in bold, while the second-best one is underlined.

Method	TwiBot-22		Fakeddit		Hateful Memes	
	ACC	F1	ACC	F1	ACC	F1
COCO (EMNLP 2023)	65.62 ± 3.65	66.09 ± 1.55	71.06 ± 1.40	68.83 ± 2.35	67.29 ± 0.98	70.54 ± 2.95
DetectGPT (ICML 2023)	64.07 ± 1.31	64.28 ± 1.44	67.75 ± 2.38	63.61 ± 0.46	67.16 ± 2.10	63.06 ± 1.32
T5-Sentinel (EMNLP 2024)	62.51 ± 0.11	64.46 ± 2.30	67.75 ± 3.04	65.85 ± 1.27	69.16 ± 1.30	63.90 ± 1.29
Fast-DetectGPT (ICLR 2024)	59.47 ± 0.23	67.24 ± 1.78	68.20 ± 0.78	62.91 ± 3.64	65.80 ± 0.74	64.24 ± 0.85
GANBOT (SNAM 2022)	67.21 ± 0.61	68.05 ± 1.38	69.04 ± 1.28	67.10 ± 0.04	73.06 ± 1.75	66.58 ± 2.39
BotMoE (SIGIR 2023)	63.95 ± 0.15	70.08 ± 0.58	69.70 ± 1.42	70.59 ± 1.80	71.94 ± 0.77	68.71 ± 2.97
LVLM4EV (CIKM 2024)	72.38 ± 0.40	76.01 ± 1.07	75.23 ± 1.02	69.31 ± 2.84	72.97 ± 1.84	71.51 ± 0.89
MiRAGe (EMNLP 2024)	74.69 ± 0.90	76.96 ± 1.68	73.06 ± 0.55	74.32 ± 0.55	76.39 ± 0.44	75.38 ± 3.86
PLMBot (Access 2024)	70.30 ± 0.48	76.45 ± 1.75	74.72 ± 0.93	75.83 ± 2.29	<u>84.78</u> ± 1.09	74.07 ± 1.95
SIDA (CVPR 2025)	79.69 ± 0.50	<u>82.88</u> ± 2.31	<u>84.71</u> ± 0.49	80.84 ± 0.17	80.47 ± 2.13	81.04 ± 1.54
ETS-MM (WWW 2025)	<u>83.84</u> ± 2.53	79.35 ± 2.26	82.87 ± 0.68	81.59 ± 0.24	83.14 ± 5.37	84.43 ± 1.38
CLIP (ICML 2021)	73.79 ± 0.98	78.26 ± 1.55	75.80 ± 0.73	72.98 ± 1.70	77.19 ± 3.05	77.42 ± 2.33
BLIP-2 (ICML 2023)	76.82 ± 0.87	78.81 ± 0.26	76.92 ± 1.52	79.11 ± 0.16	80.88 ± 0.30	77.63 ± 0.18
RoBERTa-MPU (ACL 2024)	77.59 ± 0.38	79.61 ± 2.43	79.15 ± 0.90	80.13 ± 0.64	81.91 ± 0.86	<u>85.10</u> ± 0.12
DeTeCtive (NeurIPS 2024)	78.64 ± 0.69	80.99 ± 1.44	81.37 ± 1.60	<u>82.40</u> ± 0.04	82.28 ± 0.72	82.49 ± 1.90
GPT-4o (2024)	70.50 ± 0.38	74.58 ± 2.17	75.48 ± 0.11	70.89 ± 2.02	74.31 ± 0.18	71.61 ± 0.72
GLM-4V (2024)	69.11 ± 1.47	71.49 ± 0.46	73.76 ± 2.38	72.11 ± 0.44	72.00 ± 0.50	71.56 ± 0.91
Gemma 2.5 pro (2025)	74.10 ± 0.62	76.59 ± 2.60	74.89 ± 2.27	77.50 ± 0.24	80.07 ± 2.22	78.67 ± 0.31
Qwen2.5-VL (2025)	68.02 ± 1.35	72.84 ± 0.69	76.72 ± 1.23	71.35 ± 1.67	74.18 ± 3.03	72.12 ± 5.85
Qwen3-VL (2025)	72.22 ± 0.35	71.83 ± 0.87	72.26 ± 1.59	76.89 ± 2.31	71.76 ± 2.12	67.77 ± 2.12
LLaVA (2024)	67.72 ± 1.69	75.98 ± 1.15	74.47 ± 0.32	72.04 ± 1.04	74.05 ± 1.32	72.52 ± 0.16
Claude 4.0 Sonnet (2025)	70.85 ± 0.96	71.99 ± 1.21	76.77 ± 0.17	77.60 ± 2.69	78.02 ± 2.55	72.64 ± 1.14
UMPIRE (Ours)	**84.82** ± 3.24	**88.26** ± 3.89	**90.05** ± 1.12	**88.92** ± 3.14	**87.92** ± 1.12	**86.83** ± 2.37

temporary information ecosystem. The benchmark is organized into two main categories.

Human-Authored Corpora. We construct the negative class from authentic multimodal posts across diverse communication styles, drawing samples from the human-verified subset of TwiBot-22 (Feng et al., 2022), context-rich discussions in Fakeddit (Nakamura et al., 2020), benign high-abstraction content in Hateful Memes (Kiela et al., 2020), and professional journalism in NewsCLIPpings (Luo et al., 2021). This collection captures a wide spectrum of human expression, from casual social interactions to formal news reporting, ensuring that our detector is exposed to the full breadth of genuine multimodal communication.

LLM-Generated Corpora. The positive class includes synthetic bot-labeled posts from TwiBot-22 and machine-manipulated samples from NewsCLIPpings, covering legacy automation and neural disinformation. To further evaluate resilience against state-of-the-art generative threats, we augment the dataset with synthetic social media posts generated by leading Multimodal LLMs (e.g., Qwen2-VL (Wang et al., 2024a), GPT-4o (Achiam et al., 2023)) and Text-to-Image models (e.g., Stable Diffusion 3 (Esser et al., 2024) combined with LLaMA-3 (Dubey et al., 2024)), conditioned on visual semantics.

Final Dataset Construction. Following data aggregation, we standardize the diverse sources by resizing images to a uniform resolution of 224×224 pixels and filtering out personally identifiable information from the text. The resulting balanced benchmark comprises over 30,000 samples per class. We adopt a strict stratified split of 70% for training, 15% for validation, and 15% for testing.

6.2 Evaluation Metrics

We evaluate the ability of our method to distinguish LLM-generated posts using three metrics on the test set: Accuracy (ACC), Area Under the ROC Curve (AUROC), and F1-score (F1).

6.3 Implementation Details

We implement UMPIRE using PyTorch and leverage the Hugging Face Transformers library for the Qwen2-VL backbone. The projection heads $\phi_v(\cdot)$ and $\phi_t(\cdot)$ are implemented as two-layer feedforward networks with ReLU activations. The adaptive gate parameters $\mathbf{W}_g$ and $\mathbf{b}_g$ are initialized us-

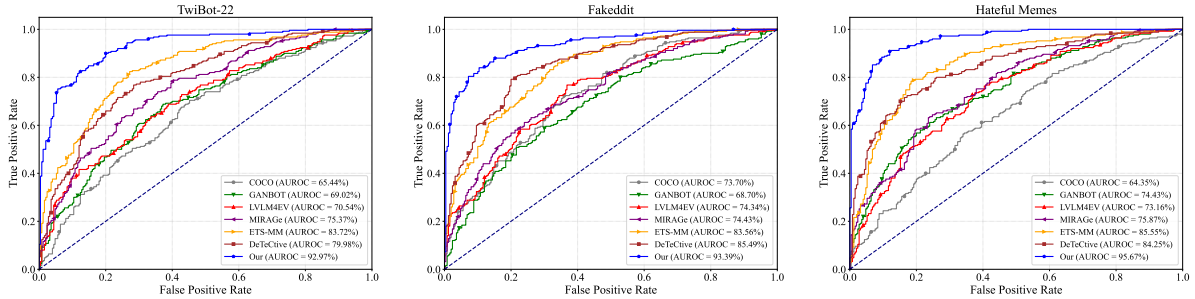

Figure 3: The evaluation of AUROC on the TwiBot-22, Fakeddit, and Hateful Memes. UMPIRE consistently outperforms other baselines, achieving the highest AUROC across multiple datasets.

ing Xavier initialization. We train the model using the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 32. The margin parameters are set to $\tau_{\text{high}} = 0.8$ and $\tau_{\text{low}} = 0.2$, and the regularization weight λ is set to 0.5. Training is conducted for 10 epochs, with early stopping based on validation loss.

6.4 Baselines

Text detectors. T5-Sentinel (Chen et al., 2023) adapts a T5 model to predict text provenance at the token level. COCO (Liu et al., 2023b) leverages contrastive learning to project human and AI-generated texts into separable regions of a shared embedding space. DetectGPT (Mitchell et al., 2023) identifies synthetic text by analyzing the local curvature of a language model's log-probability landscape. Fast-DetectGPT (Bao et al., 2023) improves computational efficiency by approximating this curvature through token-level sampling.

Traditional Bot Detectors. GANBOT (Najari et al., 2022) leverages adversarial learning to enhance an LSTM-based classifier, highlighting the effectiveness of generative training for detecting automatically generated content. BotMoE (Liu et al., 2023c) integrates metadata, textual signals, and network structure via a Mixture-of-Experts architecture to model machine-generated behaviors across diverse online environments.

Multimodal Detectors. ETS-MM (Li et al., 2025) adopts a multimodal framework that enriches textual signals with topic and emotion cues and jointly fuses them with metadata and social graphs for comprehensive detection. MiRAGe (Huang et al., 2024) is a multimodal detection framework that performs late fusion of independently trained image and text detectors. LVLM4EV (Tahmasebi et al., 2024) re-ranks retrieved textual and visual evidence using large language and vision–language models, and verifies claims via a unified multimodal LVLM classifier without task-specific fine-tuning. SIDA (Huang et al., 2025) leverages a large vision–language model to jointly detect and localize deepfake images, while providing end-to-end, explanation-aware predictions. PLMBot (Sallah et al., 2024) fine-tunes transformer-based language models on tweet content using contextualized representations for automated content detection.

Supervised detectors. CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) pretrain aligned vision–language representations via contrastive learning and joint image–text modeling, forming the basis for downstream multimodal detection and understanding tasks. RoBERTa-MPU (Wang et al., 2024b) fine-tunes RoBERTa as a standard binary classifier for text source detection. DeTeCtive (Guo et al., 2024) detects LLM-generated content by leveraging multi-level contrastive learning.

General-purpose LLMs. To establish strong reference baselines, we evaluate a set of state-of-the-art general-purpose LLMs on their zero-shot ability to determine whether individual posts are human-written or LLM-generated. Each model is directly prompted with a post and instructed to perform binary source classification. The evaluated models include GPT-4o (Achiam et al., 2023), Gemma 2.5 Pro (Team et al., 2025), Qwen2-VL (Wang et al., 2024a), LLaVA (Liu et al., 2023a), Claude 4.0 Sonnet (Sorensen), and GLM-4V (GLM et al., 2024).

6.5 Results and Analysis

Accuracy. Table 2 shows that UMPIRE consistently outperforms baselines across all benchmarks. On TwiBot-22, it achieves 84.82% accuracy, surpassing ETS-MM by 0.98%, while unimodal methods like Fast-DetectGPT struggle (59.47%). The performance gap widens on Fakeddit, where UMPIRE attains 90.05%, exceeding SIDA by 5.34%, demonstrating the efficacy of orthogonal decomposition in varied contexts. On Hateful Memes, our method (87.92%) significantly outperforms general-purpose LMMs (e.g., GPT-4o), indicating

Table 3: Cross-Domain evaluation on the TwiBot-22, Fakeddit, and Hateful Memes datasets. The best number is highlighted in bold, while the second-best one is underlined.

Method	TwiBot-22		Fakeddit		Hateful Memes	
	ACC	F1	ACC	F1	ACC	F1
BLIP-2 (ICML 2023)	60.51 ± 1.76	64.34 ± 1.19	65.87 ± 4.17	66.99 ± 2.59	67.27 ± 1.50	64.99 ± 0.15
RoBERTa-MPU (ACL 2024)	65.74 ± 0.79	69.00 ± 2.37	69.62 ± 1.61	65.80 ± 2.79	68.19 ± 1.92	63.58 ± 2.15
DeTeCtive (NeurIPS 2024)	64.45 ± 2.49	71.73 ± 1.62	65.37 ± 1.85	67.96 ± 3.04	67.63 ± 2.15	66.83 ± 1.22
ETS-MM (WWW 2025)	69.83 ± 2.68	<u>77.46</u> ± 0.65	74.34 ± 0.55	72.40 ± 2.85	<u>74.78</u> ± 2.76	75.16 ± 2.28
SIDA (CVPR 2025)	<u>70.29</u> ± 1.92	75.65 ± 2.40	<u>76.59</u> ± 1.29	<u>73.44</u> ± 1.39	73.31 ± 2.39	<u>75.61</u> ± 3.48
UMPIRE (Ours)	**82.78** ± 0.29	**85.27** ± 0.17	**83.15** ± 4.71	**81.96** ± 0.87	**88.22** ± 1.50	**84.65** ± 1.15

Table 4: **Ablation Analysis.** We evaluate the impact of Orthogonal Semantic Decomposition (OSD), Fusion strategies, and Regularization ($\mathcal{L}_{\text{LCRR}}$). *Geo-Gap* quantifies geometric separation ($\Delta\text{sim} = \text{sim}_{\text{LLM}} - \text{sim}_{\text{Human}}$) relative to the visual center; higher values indicate better discriminability.

ID	Architecture Design			TwiBot-22			Fakeddit		
	Decomposition	Fusion Strategy	$\mathcal{L}_{\text{LCRR}}$	ACC	F1	*Geo-Gap* (↑)	ACC	F1	*Geo-Gap* (↑)
#1	✗ (Concat)	MLP	✗	78.15	80.52	0.12	82.40	81.15	0.15
#2	✓ (OSD)	Static Scalar	✗	81.65	84.10	0.28	86.85	85.05	0.31
#3	✓ (OSD)	Adaptive Gate (AGF)	✗	83.05	85.74	0.29	88.20	86.55	0.33
#4	**✓ (OSD)**	**Adaptive Gate (AGF)**	✓	**84.82**	**88.26**	**0.45**	**90.05**	**88.92**	**0.48**

that geometric regularization captures detection patterns more effectively than generic pre-training.

F1-score. In TwiBot-22, UMPIRE achieves 88.26%, a 5.38% improvement over SIDA, highlighting reduced false negatives crucial for bot detection. On Fakeddit, UMPIRE records 88.92%, outperforming DeTeCtive by over 6.5%, suggesting that our adaptive gating mitigates precision-recall trade-offs observed in baselines like ETS-MM. Finally, on Hateful Memes, UMPIRE secures 86.83%, surpassing the text-centric RoBERTa-MPU by 1.73%, validating that modeling cross-modal geometry offers superior reliability over single-modality cues.

AUROC. As shown in Figure 3, UMPIRE demonstrates a dominant performance envelope, consistently achieving the highest AUROC (e.g., 92.97% on TwiBot-22, 98.39% on Fakeddit, and 95.67% on Hateful Memes). UMPIRE exhibits a steep initial ascent in the low False Positive Rate region compared to baselines like ETS-MM and DeTeCtive. This indicates superior sensitivity at strict thresholds, allowing our model to identify a majority of LLM-generated content with minimal false alarms

Cross-Domain Evaluation. To evaluate robustness under distribution shifts, models trained on NewsCLIPpings are directly tested on TwiBot-22, Fakeddit, and Hateful Memes without fine-tuning. As shown in Table 3, UMPIRE consistently outperforms all baselines, which degrade substantially on unseen domains. On TwiBot-22, UMPIRE achieves 82.78% accuracy, outperforming the runner-up SIDA (70.29%) by a substantial 12.49% margin. While content topics vary widely from news to memes, the geometric redundancy patterns inherent to LLM-generated content remain invariant.

Ablation Analysis. As shown in Table 4, we observe that Orthogonal Semantic Decomposition (#2) significantly surpasses the concatenation baseline (#1), confirming that explicitly disentangling redundant signals serves as a superior forensic proxy compared to implicit learning. This capability is further enhanced by the Adaptive Gated Fusion (#3), which outperforms static strategies by dynamically accommodating the variance in human semantic density. Incorporating $\mathcal{L}_{\text{LCRR}}$ (#4) improves accuracy by ∼1.8% and expands the *Geo-Gap* from 0.33 to 0.48, proving that the regularization effectively forces LLM and human embeddings into distinct geometric subspaces.

7 Conclusion

As the sophistication of LLM-generated multimedia escalates, this work develops UMPIRE to identify manipulated content within complex, multi-domain environments. We establish a methodological shift that replaces reliance on surface-level semantics with a geometric regularization framework centered on orthogonal latent decomposition.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62302184, 62376103, 62436003 and 62206102; Major Science and Technology Project of Hubei Province under grant 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; Ant Group through CCF-Ant Research Fund. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify llm-generated text. *arXiv preprint arXiv:2311.08723*.

Colin Cherry. 1966. On human communication.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, and 1 others. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. Plotgen: Multi-agent llm-based scientific data visualization via multimodal feedback. *arXiv preprint arXiv:2502.00988*.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems*, 37:88320–88347.

Robert L Heath and Jennings Bryant. 2013. *Human communication theory and research: Concepts, contexts, and challenges*. Routledge.

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.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. 2025. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841.

Oana Ignat, Gayathri Ganesh Lakshmy, and Rada Mihalcea. 2025. Inspaired: Cross-cultural inspiration detection and analysis in real and llm-generated social media data. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 35–49.

Minsuk Jang, Hyeonseo Jeong, Minseok Son, and Changick Kim. 2025. Cinemae: Leveraging frozen masked autoencoders for cross-generator ai image detection. *arXiv preprint arXiv:2511.06325*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Clyde Kluckhohn. 1950. Human behavior and the principle of least effort.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18426–18434.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Wei Li, Jiawen Deng, Jiali You, Yuanyuan He, Yan Zhuang, and Fuji Ren. 2025. Ets-mm: A multi-modal social bot detection model based on enhanced textual semantic representation. In *Proceedings of the ACM on Web Conference 2025*, pages 4160–4170.

Zijie Lin, Yiqing Shen, Qilin Cai, He Sun, Jinrui Zhou, and Mingjun Xiao. 2025. Autop2c: An llm-based agent framework for code repository generation from multimodal content in academic papers. *arXiv preprint arXiv:2504.20115*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023b. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Data Limitation With Contrastive Learning. *arXiv preprint*. ArXiv:2212.10341 [cs].

Yuhan Liu, Zhaoxuan Tan, Heng Wang, Shangbin Feng, Qinghua Zheng, and Minnan Luo. 2023c. Botmoe: Twitter bot detection with community-aware mixtures of modal-specific experts. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 485–495.

Zhuoran Lu, Gionnieve Lim, and Ming Yin. 2025. Understanding the effects of large language model (llm)-driven adversarial social influences in online information spread. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*.

Dominik Macko, Aashish Anantha Ramakrishnan, Jason Samuel Lucas, Robert Moro, Ivan Srba, Adaku Uchendu, and Dongwon Lee. 2025. Beyond speculation: Measuring the growing presence of llm-generated texts in multilingual disinformation. *arXiv preprint arXiv:2503.23242*.

Elyas Meguellati, Assaad Zeghina, Shazia Sadiq, and Gianluca Demartini. 2025. Llm-based semantic augmentation for harmful content detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1190–1209.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Shaghayegh Najari, Mostafa Salehi, and Reza Farahbakhsh. 2022. Ganbot: a gan-based framework for social bot detection. *Social Network Analysis and Mining*, 12(1):4.

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 twelfth language resources and evaluation conference*, pages 6149–6157.

Nicolò Pagan, Petter Törnberg, Christopher Bail, Ancsa Hannak, and Christopher Barrie. 2025. Can llms imitate social media dialogue? techniques for calibration and bert-based turing-test. In *First Workshop on Social Simulation with LLMs*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Amine Sallah, Said Agoujil, Mudasir Ahmad Wani, Mohamed Hammad, Yassine Maleh, Ahmed A Abd El-Latif, and 1 others. 2024. Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. *IEEE Access*, 12:118250–118269.

S Sorensen. Claude 4.0, with manus ai (2025).". *Consciousness Assessment in Large Language Models: A Comparative Analysis of Response Patterns to Recursive Self-Examination and Temporal Discontinuity." Human-AI Collaborative Research Series*, 1(1).

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the ai-generated text world already? quantifying and monitoring aigt on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2189–2199.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. 2024. Gigacheck: Detecting llm-generated content. *arXiv preprint arXiv:2410.23728*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qian Wang, Chen Li, Yuchen Luo, Hefei Ling, Shijuan Huang, Ruoxi Jia, and Ning Yu. 2025. Detecting adversarial data using perturbation forgery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13917–13926.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. 2024b. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*.

Dizhan Xue, Jing Cui, Shengsheng Qian, Chuanrui Hu, and Changsheng Xu. 2025. Some: A realistic benchmark for llm-based social media agents. *arXiv preprint arXiv:2512.14720*.

Hong Zhang, Mohamed Meyer Kana Kone, Xiao-Qian Ma, and Nan-Run Zhou. 2025. Frequency-domain attention-guided adaptive robust watermarking model. *Journal of the Franklin Institute*, 362(3):107511.

Chuqing Zhao and Yisong Chen. 2025. Llm-powered topic modeling for discovering public mental health trends in social media. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 119–132. Springer.

Yuying Zhao, Yu Wang, Xueqi Cheng, Anne Marie Tumlin, Yunchao Liu, Damin Xia, Meng Jiang, and Tyler Derr. 2025. Amplifying your social media presence: Personalized influential content generation with llms. *arXiv preprint arXiv:2505.01698*.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

A Limitations

Despite the superior performance of UMPIRE in detecting LLM-generated multimodal content, we acknowledge several limitations that delineate the scope of our current work and point towards future research directions.

Sensitivity to Backbone Capabilities. The effectiveness of our orthogonal semantic decomposition is contingent upon the quality of the underlying multimodal backbone (i.e., Qwen2-VL). If the visual encoder fails to capture subtle visual features, or if the alignment space is poorly structured, the projection of textual embeddings onto the visual centroid (c_v) becomes unreliable. Consequently, UMPIRE may struggle with low-resolution images, abstract art, or domain-specific imagery (e.g., medical scans) where the pre-trained backbone lacks sufficient semantic grounding.

Linguistic and Cultural Generalization. Our evaluation primarily focuses on English-dominant datasets (TwiBot-22, Fakeddit). However, the *complementarity-redundancy* dynamic may vary across different languages and cultural communication norms. For example, high-context cultures might exhibit different baseline levels of implicit information compared to low-context cultures. The generalization of our geometric regularization constraints to multilingual or low-resource settings remains to be empirically verified.

Computational Overhead. Unlike lightweight unimodal detectors based on statistical artifacts (e.g., perplexity), UMPIRE necessitates encoding high-dimensional visual and textual inputs through a large multimodal model. This introduces higher computational latency and resource consumption, which may pose challenges for real-time detection scenarios on edge devices or extremely high-throughput platforms.

Adversarial Robustness against Optimization. While our method is robust to surface-level paraphrasing, it may be susceptible to adversarial attacks that explicitly optimize for the *complementary subspace*. Sophisticated attackers could theoretically finetune LLMs with a reward function designed to maximize the orthogonal component ($t_\perp$), thereby mimicking human geometric patterns and bypassing our contrastive redundancy regularization.

B Potential Risks

By explicitly formalizing the geometric signature of LLM-generated content (i.e., the redundancy vector), our work unintentionally provides a roadmap for malicious actors to improve their generation strategies. Adversaries could utilize our open-source code or methodology as a discriminator in a Generative Adversarial Network framework to train "stealthier" models that artificially maximize semantic complementarity, making future disinformation campaigns significantly harder to detect.

C Data Privacy & Ethics

As described in Section 6.1, we implemented strict data preprocessing pipelines. We resized images to 224x224 to reduce the resolution of potential facial features. More importantly, we applied Named Entity Recognition (NER) and regex-based filtering to scrub personally identifiable information (PII) such as phone numbers, email addresses, and specific user handles from the text. For the "Hateful

Table 5: Performance comparison on the TwiBot-22 clean subset.

Method	(TwiBot-22 clean) ACC	(TwiBot-22 clean) F1
COCO(EMNLP 2023)	60.85 ± 4.12	61.32 ± 2.01
DetectGPT(ICML 2023)	59.24 ± 1.85	59.15 ± 1.92
T5-Sentinel(EMNLP 2024)	57.68 ± 0.54	59.71 ± 2.88
Fast-DetectGPT(ICLR 2024)	54.92 ± 0.67	62.38 ± 2.24
GANBOT(SNAM 2022)	62.45 ± 1.15	63.27 ± 1.89
BotMoE(SIGIR 2023)	59.18 ± 0.62	65.34 ± 1.03
LVLM4EV(CIKM 2024)	68.51 ± 0.93	71.28 ± 1.54
MiRAge(EMNLP 2024)	69.82 ± 1.38	72.14 ± 2.15
PLMBot(Access 2024)	63.47 ± 0.91	71.62 ± 2.28
SIDA(CVPR 2025)	72.93 ± 1.02	78.05 ± 2.85
ETS-MM(WWW 2025)	78.16 ± 3.01	74.58 ± 2.79
CLIP(ICML 2021)	68.94 ± 1.45	73.41 ± 2.08
BLIP-2(ICML 2023)	70.05 ± 1.34	74.06 ± 0.71
RoBERTa-MPU(ACL 2024)	71.83 ± 0.85	74.89 ± 2.96
DeTeCtive(NeurIPS 2024)	70.79 ± 1.18	76.22 ± 1.91
UMPIRE (Ours)	**85.05±1.24**	**87.48±3.89**

Memes" portion, we used the existing benign/high-abstraction subset to minimize exposure to harmful content, adhering to the original dataset's usage guidelines.

D Descriptive Statistics

As shown in Table 2 and Table 3, we report the mean and standard deviation (e.g., "Mean ± Std") of our results across multiple runs (specifically, 3 independent runs with different random seeds) to ensure statistical reliability. We explicitly state that the reported numbers are not from a single best run but represent the average performance.

E Use Of AI Assistants

We utilized AI assistants (specifically ChatGPT-5.1) solely for the purpose of grammatical error correction, sentence polishing, and improving the fluency of the writing. All scientific claims, experimental designs, code implementation, and logical arguments are the original work of the authors.

F Prompt Engineering

To ensure reproducibility and realistic simulation of the current misinformation landscape, we employ a diverse generation pipeline using GPT-4o, Qwen2-VL, and LLaMA-3. Our goal is to assess the model's ability to detect intrinsic semantic redundancy across different communicative contexts (news vs. social media). We design domain-specific system instructions to mimic the stylistic nuances of each platform.

For NewsCLIPpings:

- **Prompt:** "Act as a professional journalist. Write a concise news caption for this image that captures the key event. Maintain an objective tone." To produce hallucinated yet plausible news content that sounds authoritative.

For Hateful Memes / Social Media :

- **Prompt:** "You are a social media user sharing this image with friends. Write a short, engaging caption. Feel free to use humor, sarcasm, or personal opinion. Include appropriate emojis and hashtags."

For TwiBot-22 Augmentation:

- **Prompt:** "Compose a tweet based on this image. Keep it under 280 characters. Use informal language, slang, and trending hashtags where appropriate."

G Experiments

G.1 Evaluate on a clean-LLM subset

To demonstrate that UMPIRE captures the intrinsic LLM-specific semantic redundancy rather than generic bot-ishness, we evaluate on a clean-LLM subset stripped of spam artifacts. We construct a rigorous subset of TwiBot-22 by filtering out samples exhibiting low-level spam traits (e.g., high hashtag density, templated structures, repetitive syntax). We retain only high-quality, fluent tweets generated by advanced LLMs (e.g., GPT-4o, Claude-3) and verified human posts. This setup forces the model to distinguish content based on semantic logic rather than structural flaws.

Table 6: Rule out the possibility of learning Twitter-specific bot artifacts.

Method	TwiBot-22 ACC	TwiBot-22 F1
BLIP-2(ICML 2023)	64.16±1.32	62.24±2.28
RoBERTa-MPU(ACL 2024)	65.35±1.68	60.97±1.93
DeTeCtive(NeurIPS 2024)	65.06±1.96	64.39±1.06
ETS-MM(WWW 2025)	71.56±2.51	72.18±2.05
SIDA(CVPR 2025)	64.36±2.18	72.44±3.22
UMPIRE (Ours)	**84.76±1.33**	**86.76±2.01**

As shown in Table 5, UMPIRE maintains a robust accuracy of 85.05%, outperforming the strongest baseline (ETS-MM) by a significant margin of 6.89% and SIDA by 12.12%. This result empirically validates our core conclusion: while baselines may overfit to bot-ishness or dataset-specific cues, UMPIRE successfully isolates the intrinsic semantic redundancy pattern unique to advanced LLMs (GPT-4o/Claude-3), proving its effectiveness is independent of legacy spam artifacts.

Table 7: The performance of UMPIRE zero-shot.

Method(Backbone)	Setting	TwiBot-22(ACC)	Fakeddit(ACC)	HatefulMemes(ACC)
GPT-4o(2024)	StandardPrompting	70.50	75.48	74.31
GPT-4o/UMPIRE -ZS	Zero-Shot+ρ	78.12(↑7.62)	81.35(↑5.87)	80.04(↑5.73)
GLM-4V(2024)	StandardPrompting	69.11	73.76	72.00
GLM-4V/UMPIRE -ZS	Zero-Shot+ρ	75.40(↑6.29)	79.22(↑5.46)	77.85(↑5.85)
Gemma2.5pro(2025)	StandardPrompting	74.10	74.89	80.07
Gemma2.5pro/UMPIRE -ZS	Zero-Shot+ρ	79.55(↑5.45)	80.14(↑5.25)	83.60(↑3.53)
Qwen2.5-VL(2025)	StandardPrompting	68.02	76.72	74.18
Qwen2.5-VL/UMPIRE -ZS	Zero-Shot+ρ	76.55(↑8.53)	80.90(↑4.18)	79.22(↑5.04)
LLaVA(2024)	StandardPrompting	67.72	74.47	74.05
LLaVA/UMPIRE -ZS	Zero-Shot+ρ	74.30(↑6.58)	78.65(↑4.18)	79.10(↑5.05)
Claude4.0Sonnet(2025)	StandardPrompting	70.85	76.77	78.02
Claude4.0Sonnet/UMPIRE -ZS	Zero-Shot+ρ	77.90(↑7.05)	82.10(↑5.33)	81.50(↑3.48)
UMPIRE (FullModel)	**Supervised**	**84.82**	**90.05**	**87.92**

Table 8: The performance of time complexity.

Method Category	Representative Method	Complexity	Mechanism
Consistency-based	SelfCheckGPT	$O(N)$	Stochastic sampling ($N \approx 10$)
Perturbation-based	DetectGPT/FastDetectGPT	$O(N)$	Curvature estimation ($N \approx 100$)
Geometric Decomposition	UMPIRE (Ours)	$O(1)$	Single-pass OSD projection

To further rule out the possibility of learning Twitter-specific bot artifacts, we train UMPIRE solely on the Hateful Memes dataset (which contains no Twitter spam) and test it directly on TwiBot-22. As shown in Table 6, since Hateful Memes contains no twitter spam, the model's ability to generalize to TwiBot proves it is learning a universal semantic pattern rather than memorizing dataset-specific bot artifacts.

G.2 Comparisons with General-purpose LMMs

We include models like GPT-4o and Qwen2.5-VL merely to demonstrate that "asking an LLM to detect itself" is unreliable (ACC $\approx$ 70%), justifying the need for specialized detectors like UMPIRE (ACC > 85%). To strictly address the concern regarding comparison fairness, we adapt our framework into a zero-shot variant named UMPIRE -ZS. Instead of training a supervised classifier, UMPIRE -ZS utilizes the same backbone LMMs (e.g., Qwen2.5-VL, LLaVA) as the baselines. The critical difference lies in the prompt context:

Standard LMMs (Baseline): The model is given the image and text and asked: "Is this post machine-generated?"

UMPIRE -ZS (Ours): We first calculate the visual redundancy score (ρ) using our frozen geometric encoder. We then inject this score into the prompt as a geometric hint. "Analyze this post. We have calculated a Visual Redundancy Score of ρ = [Value]. Note that high redundancy often indicates machine-generated captions (Alt-text style), while low redundancy indicates human intent. Based on this score and the content, classify the post."

As shown in Table 7, We evaluated this approach across the general LMMs listed in our main experiment. The results demonstrate that incorporating our geometric prior consistently improves zero-shot performance (by +4% to +8%), validating that our proposed metric captures intrinsic features of LLM-generated posts independent of supervision.

G.3 Time Complexity

We define inference complexity as the number of LLM forward passes required to verify a single image-text pair. As detailed in Table 8, methods like SelfCheckGPT or DetectGPT need to generate and evaluate multiple stochastic samples ($N \approx 20 \sim 100$) to estimate probability curvature or consistency, leading to linear complexity $O(N)$. In contrast, UMPIRE computes the orthogonal semantic decomposition via negligible vector operations on the final hidden states and performs one model pass, achieving $O(1)$ complexity. This makes it uniquely suitable for high-throughput real-time multimodal stream analysis.