

Beyond Artificial Misalignment: Detecting and Grounding Semantic-Coordinated Multimodal Manipulations

Jinjie Shen
Hefei University of Technology
Hefei, Anhui, China
shenjinnie22@gmail.com

Yaxiong Wang*
Hefei University of Technology
Hefei, Anhui, China
wangyx@hfut.edu.cn

Lechao Cheng
Hefei University of Technology
Hefei, Anhui, China
chenglc@hfut.edu.cn

Nan Pu
University of Trento
Trento, Trentino, Italy
nan.pu@unitn.it

Zhun Zhong*
Hefei University of Technology
Hefei, Anhui, China
zhunzhong007@gmail.com

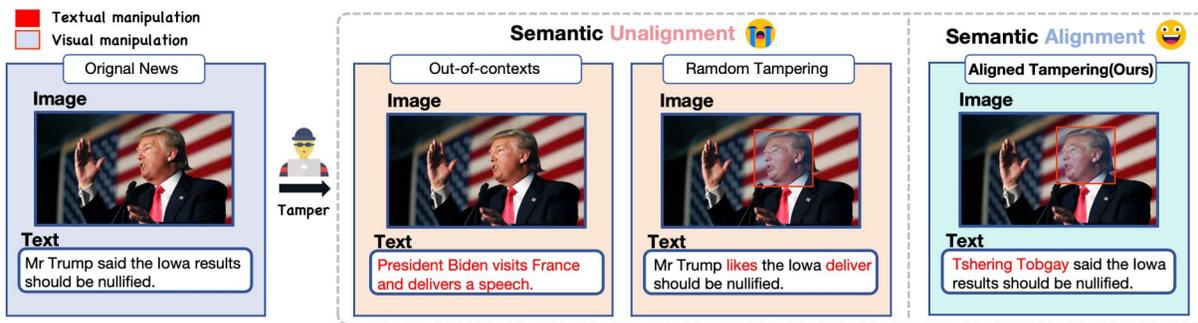


Figure 1: Unlike existing multimodal deepfake datasets where cross-modal semantic alignment is lacking, SAMM proposes semantically aligned fake news, which better reflects real-world scenarios.

Abstract

The detection and grounding of manipulated content in multimodal data has emerged as a critical challenge in media forensics. While existing benchmarks demonstrate technical progress, they suffer from misalignment artifacts that poorly reflect real-world manipulation patterns: practical attacks typically maintain semantic consistency across modalities, whereas current datasets artificially disrupt cross-modal alignment, creating easily detectable anomalies. To bridge this gap, we pioneer the detection of semantically-coordinated manipulations where visual edits are systematically paired with semantically consistent textual descriptions. Our approach begins with constructing the first Semantic-Aligned Multimodal Manipulation (SAMM) dataset, generated through a two-stage pipeline: 1) applying state-of-the-art image manipulations, followed by 2) generation of contextually-plausible textual narratives that reinforce the visual deception. Building on this foundation, we propose

a Retrieval-Augmented Manipulation Detection and Grounding (RamDG) framework. RamDG commences by harnessing external knowledge repositories to retrieve contextual evidence, which serves as the auxiliary texts and encoded together with the inputs through our image forgery grounding and deep manipulation detection modules to trace all manipulations. Extensive experiments demonstrate our framework significantly outperforms existing methods, achieving 2.06% higher detection accuracy on SAMM compared to state-of-the-art approaches. The dataset and code are publicly available at <https://github.com/shen8424/SAMM-RamDG-CAP>

CCS Concepts

• Security and privacy → Social aspects of security and privacy; • Information systems → Multimedia streaming.

Keywords

Media Manipulation Detection, DeepFake Detection, Multi-Modal Learning

ACM Reference Format:

Jinjie Shen, Yaxiong Wang*, Lechao Cheng, Nan Pu, and Zhun Zhong*. 2025. Beyond Artificial Misalignment: Detecting and Grounding Semantic-Coordinated Multimodal Manipulations. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746027.3754951>

*Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
<https://doi.org/10.1145/3746027.3754951>

1 Introduction

The rapid development of generative models has driven significant progress across domains [13, 20, 27, 56, 62]. Concurrently, this technological advancement precipitates critical societal risks, particularly through the synthesis of highly plausible yet falsified media content [18, 26, 46, 49]. Such fabricated information not only erodes public trust through fabricated misleading contents but also engenders systemic vulnerabilities in digital information ecosystems [8, 12, 44].

Many efforts have been made to recognize the fake news in social media [39, 45, 63, 65]. In recent years, different types of scenarios of manipulated multimodal news have been studied [5, 16, 33, 47]. NewsCLIPings [33] uses randomly selected news texts to form out-of-context image-text Pairs. DGM⁴ [47] employs a random modification strategy as well, which in the visual modality involve using randomly selected non-celebrity faces as substitutes, and in the textual modality involve randomly replacing certain words or segments. Despite these pioneering attempts, the focused scenarios of these works all have a severe artifact of semantic misalignment. For example, as shown in the top row of Figure 1: “*an image of Mr Trump giving a speech is paired with a caption that reads president Biden performing at a concert.*” This misalignment, on the one hand, renders the fake news detection too easy to cheat the people. On the other hand, the semantic-misaligned news fails to stimulate the practical situations, since the attackers usually maintain the consistency across modalities to deceive the public. For example, if Biden’s face is swapped with Trump’s in the visual modality, the corresponding text modality would also reflect Biden being replaced by Trump, as shown in the bottom of Figure 1.

In response to the aforementioned challenges, we focus on a more practical problem in this work: multimodal manipulation detection and grounding with semantic-alignment manipulations. To facilitate this research, we present the SAMM (Semantic-Aligned Multimodal Manipulation) - a comprehensive dataset containing 260,970 carefully crafted semantic-coordinated samples. First, we perform visual alterations through either face swapping [4, 9] or facial attribute editing [37, 54] on celebrity images, considering their heightened social impact and misinformation risks in public domains [47]. Subsequently, we generate semantically-aligned fake text descriptions that maintain logical consistency with the manipulated visual content. This two-stage manipulation pipeline ensures sophisticated alignment between visual tampering and textual fabrication, creating convincing multimodal forgeries that pose significant detection challenges.

Semantic-coordinated manipulations reflect real-world cases but pose greater detection challenges, as prior methods fail in such scenarios. Notably, human usually check the confused information by conducting the cross-verification with external knowledge, such as using the fact that “Messi is a great football player” to identify the fake news claiming “Messi won the Nobel Prize in Literature.” Inspired by this, we propose the Retrieval-augmented manipulation Detection and Grounding (RamDG) framework. First, by integrating a large-scale external knowledge base Celeb Attributes Portfolio (CAP) containing information on celebrities from various domains, we design the Celebrity-News Contrastive Learning (CNCL) mechanism to facilitate RamDG in leveraging external

knowledge for semantic-level fake news detection akin to human capabilities. Furthermore, we introduce the Fine-grained Visual Refinement Mechanism (FVRM) module to enhance the model’s ability to accurately localize visually manipulated regions.

Our main contributions include:

- Introducing SAMM, a more realistic deepfake dataset featuring multi-modal semantic alignment and purposeful tampering, with a large scale and rich fine-grained annotations to meet training or evaluation needs;
- Proposing CAP, an external knowledge base containing multi-domain celebrity information, enabling logical fake news detection through simple “string matching” integration into existing datasets;
- Presenting RamDG, which outperforms in binary classification and excels in fine-grained tampering localization compared to all current models.

2 Related work

DeepFake Detection. Historically, deepfake detection has primarily focused on single modalities, such as text [34, 61, 64] or visual [3, 50, 59]. Within visual modalities, methods are categorized into those based on the spatial domain [25, 31, 35] and those based on the frequency domain [7, 17, 40]. With the advancement of multimodal techniques [20, 22, 41, 51, 53, 57, 58], news content is increasingly presented in multimodal formats, leading to the emergence of recent multimodal detection approaches, including methods based on modality fusion [22, 47, 48] and those leveraging Vision Language Large Model [16, 29, 38]. Modality fusion methods struggle with small datasets due to limited external knowledge, whereas large language models, despite their rich internal knowledge, face challenges in fine-grained tampering localization. To address these limitations, we propose a hybrid approach that integrates CAP-derived external knowledge for detection and enhances fine-grained localization accuracy.

DeepFake Datasets. Existing DeepFake datasets primarily consist of single-modality data focusing on either visual [32, 36] or textual content [30, 55]. While some multimodal datasets exist [16, 33], they typically adopt either contextually irrelevant pairings (out-of-context pairs) [33] or rely entirely on synthetic data produced by generative models [16]. DGM⁴ [47] addresses these issues to some extent by providing over 230k samples with fine-grained annotations, constructed through modifications of authentic news articles. However, All the datasets mentioned above exhibit two critical limitations: 1) Cross-modal semantic inconsistency; 2) Purposeless tampering. To address these limitations, we propose the **SAMM Dataset**, a large-scale multimodal dataset with comprehensive fine-grained annotations that better aligns with real-world fake news distribution patterns.

3 SAMM Dataset

Unlike existing benchmarks that randomly manipulate the multimedia, which result in misaligned semantics in fabricated multimodal media. [33, 47]. In response, we construct SAMM dataset to remedy this weakness. The construction of the dataset comprises three steps: 1) Source data collection to filter out data for subsequent tampering operations and build an external knowledge base; 2)

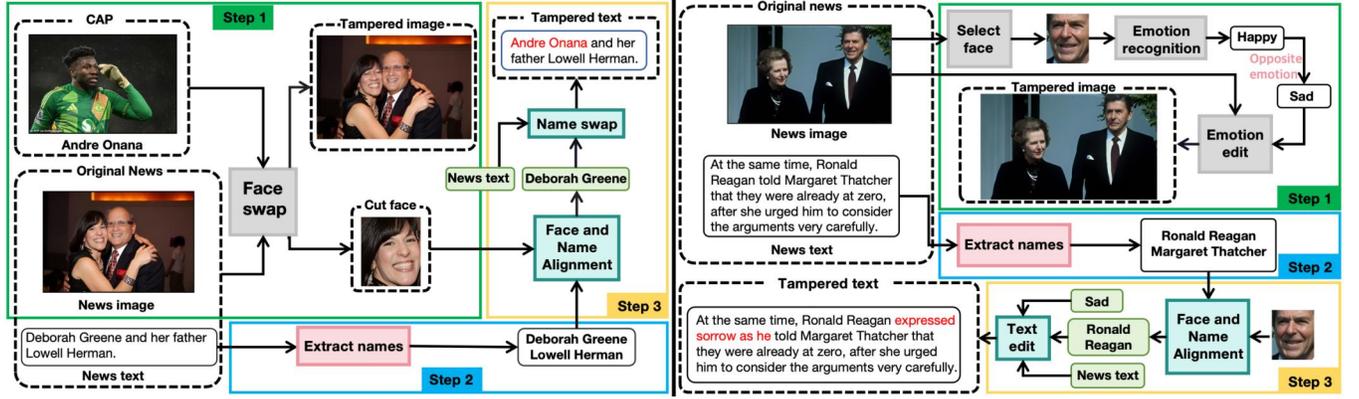


Figure 2: The process of swap manipulation is shown on the left, while the process of attribute manipulation is shown on the right. We use image tampering models and the Qwen series of models to carry out manipulation.

Multi-modal manipulation, which details the processes of swap manipulation and attribute manipulation. The construction steps of the dataset are as follows:

3.1 Source Data Collection

Given the social impact and risks of multimodal news manipulation, we align with DGM⁴ [47] to adopt human-centered news data as raw material. Our dataset is built upon **VisualNews** [28] and **GoodNews** [2]. To ensure human-centeredness and diversity, the following strategies are adopted to filter the raw data: (1) We select news that at least contains one individual. (2) We encode the news images using CLIP [41], compute their similarity scores with all other images in the dataset, and remove those with high cumulative similarity scores. The filtered dataset $S = \{P_s \mid P_s = (I_s, T_s)\}$ forms the basis for subsequent dataset construction.

Celeb Attributes Portfolio. Before diving into the construction of dataset, we first prepare a Celeb Attributes Portfolio (CAP) to aid the building of SAMM dataset and provide external knowledge. We have collected and curated multimodal data for celebrities from the internet using the Google Search API [10], encompassing visual modality information (images) and textual modality information (gender, birth year, occupation, main achievements). CAP covers celebrities featured in datasets such as VisualNews[28], GoodNews [2], DGM⁴ [47], and SAMM. Furthermore, a celebrity’s information card can be acquired by simple name matching.

3.2 Multi-Modal Manipulation

To ensure the alignment between the manipulated text and image, the image tampering is performed first, and the text fabrication follows. We adopt two types of image manipulation: **Swap Manipulation** and **Attribute Manipulation**. Specifically, Swap Manipulation includes face replacement in the visual modality and corresponding name replacement in the textual modality, while Attribute Manipulation involves emotion manipulation in the visual modality and corresponding emotion-related vocabulary manipulation in the textual modality. Combined with unaltered original image-text pairs, the dataset comprises three data categories.

Swap Manipulation. As shown in Figure 2, given (I, T) , we employ the existing face-swapping model SimSwap [4] and InfoSwap [9] for visual manipulation and large language models [53, 60] for textual name replacement.

► **Face swap.** We randomly select a face I_f and corresponding name N_f from CAP, randomly apply SimSwap [4] or InfoSwap [9] to I to generate the tampered image I_m , and record the bounding box coordinates (x_1, y_1, x_2, y_2) of the swapped face region F_m in I_m . The region correspond to the bbox coordinates in I is denoted as F .

► **Text forgery.** To generate text that is semantically aligned with the swapped image, we need to know the name of the face that has been swapped out. To acquire this information, we utilize the large language model Qwen2.5 [60] to extract all names $N = \{N^i \mid i = 1, 2, \dots\}$ from T (e.g., “Joe Biden”, “Vladimir Putin”). These names and F are fed into the multimodal model Qwen2-VL [53] to align F with the corresponding name $N^t \in N$. We perform manual sampling verification on the name extraction and face matching results generated by the large model to ensure accuracy. (implementation details and accuracy validation are provided in the appendix). Finally, all N^t in T are replaced with N_f to produce the manipulated text T_m , accompanied by a one-hot vector $label_t$ to indicate whether the i -th word is tampered.

Attribute Manipulation. As shown in figure 2, given (I, T) , we employ the HFGI [54] and StyleCLIP [37] for visual emotion editing and large language models [53, 60] for textual emotion manipulation.

► **Face attribute edit.** We first utilize the DSFD [24] model for face detection to randomly select a target face F in I , recording its bounding box coordinates (x_1, y_1, x_2, y_2) . The Qwen2-VL model [53] predicts the emotional state of F , which is then randomly fed into HFGI [54] or StyleCLIP [37] to generate an opposite emotional manipulation (e.g., modifying “happy” to “sad”), resulting in F_m . We then replace F in I with F_m , yielding I_m .

► **Text distortion.** Following the same operation as in the swap manipulation, we utilize Qwen2.5 [60] and Qwen2-vl [53] to complete the matching of the name N^t and the face F . To achieve diverse emotional expressions, we collect multiple emotions and provide various expressions for each emotion option $\mathcal{E} = \{e_i \mid e_i =$

equip the celebrities knowledge to the raw image feature:

$$V_f = \text{Attn}(Q = V, K = V_{cb}, V = V_{cb}|\Theta_1) \quad (1)$$

where $V_f = \{V_f^{cls}, V_f^{pat}\}$ is the resultant comprehensive features. $\text{Attn}(\cdot)$ is the multi-head attention function and Θ_1 represents the parameters in the $\text{Attn}(\cdot)$.

Text Encoding with Celebrity Notes. For text input, as $\{T_j\}$ have high information density and no redundancy, we directly append the celebrity meta text(s) to the text, thereby directly enriching the textual context. Subsequently, the enhanced text is fed into text encoder E_t , thus the compressive representative for text is obtained, marked as $L_f = E_t(T_{all}) = \{L_f^{cls}, L_f^{tok}\}$. The pure text feature $L = \{L^{cls}, L^{tok}\}$ and celebrity textual feature $L_f = \{L_f^{cls}, L_f^{tok}\}$ are also obtained by feeding the text and the concatenated celebrity notes to text encoder.

4.2 Celebrity-News Contrastive Learning.

Human usually doubts a piece of news when they found conflicts between the news and the information they know. Motivated by this consideration, Celebrity-News Contrastive Learning (CNCL) conducts a contrastive learning [14] procedure between the multimodal news and auxiliary celebrity information, aiming to endow the network a human-like reasoning ability for fake news detection. In specific, we adopt a contrastive learning mechanism to simulate human logical reasoning : by aligning the semantics of untampered celebrity information with the news, we enhance the model's detection ability:

$$\mathcal{L}_{v2v}(I_j, I, \mathcal{I}) = -\log\left(\frac{\exp(s(I_j, I^+)/\tau)}{\sum_{I_k \in \mathcal{I}} \exp(s(I_j, I_k)/\tau)}\right) \quad (2)$$

where $s(\cdot)$ is the cosine similarity function, τ is the learnable parameter controlling the temperature, I_j is the image of the j -th celebrity appearing in the news (I, T) , I is the corresponding positive sample news image, and \mathcal{I} is the set of a positive sample news image I and multiple negative samples that do not include the celebrity represented by I_j . We map the [CLS] token through a projection layer and incorporate it into the cosine similarity calculation as follows:

$$s(I_j, I) = [P_v(V_j^{cls})]^T \hat{P}_v(\hat{V}^{cls}) \quad (3)$$

Where \hat{V}^{cls} is the [CLS] token obtained by encoding I with momentum encoder \hat{E}_v [14], P_v and \hat{P}_v are the mapping layers. Similarly, the contrastive learning $\mathcal{L}_{v2t}(I_j, T, \mathcal{T})$ between external knowledge images and news text can be performed. In a analogous fashion, we further augment the contrastive learning on the text side, introducing $\mathcal{L}_{t2v}(T_j, I, \mathcal{I})$ and $\mathcal{L}_{t2t}(T_j, T, \mathcal{T})$. (The specific expressions for \mathcal{L}_{v2t} , \mathcal{L}_{t2v} , \mathcal{L}_{t2t} can be found in the appendix).

In summary, the overall loss function of Celebrity-News Contrastive Learning mechanism is:

$$\mathcal{L}_{cncl} = \mathcal{L}_{v2v} + \mathcal{L}_{v2t} + \mathcal{L}_{t2v} + \mathcal{L}_{t2t} \quad (4)$$

4.3 Image Forgery Grounding via FVRM

To integrate the knowledge of both modalities, we first fuse the information from two modalities to obtain a hybrid multimodal

representation with comprehensive contexts, where the text L_f serves as the query to collect clues from image via attention:

$$M_f = \text{Attn}(Q = V_f, K = L_f, V = L_f|\Theta_2) \quad (5)$$

where $M_f = \{M_f^{cls}, M_f^{tok}\}$. Since visual manipulation is small-scale and localized, we need to extract local semantic information related to visual tampering. To achieve this, we adopt the Fine-grained Visual Refinement Mechanism (FVRM).

Fine-grained Visual Refinement Mechanism. Patches in V^{pat} from tampered regions differ semantically from those in unaltered regions, revealing local tampering traces. Based on this observation, we add a classification head (three linear layers) after V^{pat} to predict patch manipulation. The loss \mathcal{L}_{pat} is computed using cross-entropy. Specifically, the loss function is defined as:

$$\mathcal{L}_{pat} = -\sum_{i=1}^C [y_{pat} \log(P_{pat})] \quad (6)$$

where y_{pat} is the label converted from bbox coordinates to indicate whether a patch is manipulated. P_{pat} represents the probability, predicted by the model, that the patch has been tampered with. C is the number of patches in an image. Under the supervision of y_{pat} , V^{pat} is mapped through the first two linear layers of classification head to obtain \tilde{V}^{pat} , which contains visual manipulation traces. M_f^{cls} contains semantic information related to the global examination of the detected news pair. By fusing \tilde{V}^{pat} and M_f^{cls} , the resulting \tilde{M}_f captures both global and local information. The process is as follows:

$$\tilde{M}_f = \text{Attn}(Q = M_f^{cls}, K = \tilde{V}^{pat}, V = \tilde{V}^{pat}|\Theta_3) \quad (7)$$

Let \tilde{M}_f perform attention computation with a learnable vector Q [23] to capture semantic information at different scales. The fusion of external knowledge in M_f dilutes the local details of the original image. To address this, we perform residual connection [15], as follows:

$$M_{fv} = M_f + V \quad (8)$$

Finally, cross-attention is performed between \tilde{M}_f and M_{fv} to achieve fine-grained visual refinement, denoted as \tilde{M}_{fv} . The specific process is as follows:

$$\tilde{M}_f = \text{Attn}(Q = Q, K = \tilde{M}_f, V = \tilde{M}_f|\Theta_4) \quad (9)$$

$$\tilde{M}_{fv} = \text{Attn}(Q = \tilde{M}_f, K = M_{fv}, V = M_{fv}|\Theta_5) \quad (10)$$

We use the obtained \tilde{M}_{fv} for bounding box prediction. The L_1 loss and IoU loss [43] between the predicted boxes and the bounding box coordinates are then computed:

$$\mathcal{L}_{bbox} = -\|P_{gro} - \text{Sigmoid}(y_{bbox})\|_1 + \mathcal{L}_{IoU}(P_{gro}, \text{Sigmoid}(y_{bbox})) \quad (11)$$

where $\mathcal{L}_{IoU}(\cdot)$ is the IoU loss function, P_{gro} represents the predicted bounding box coordinates from the model.

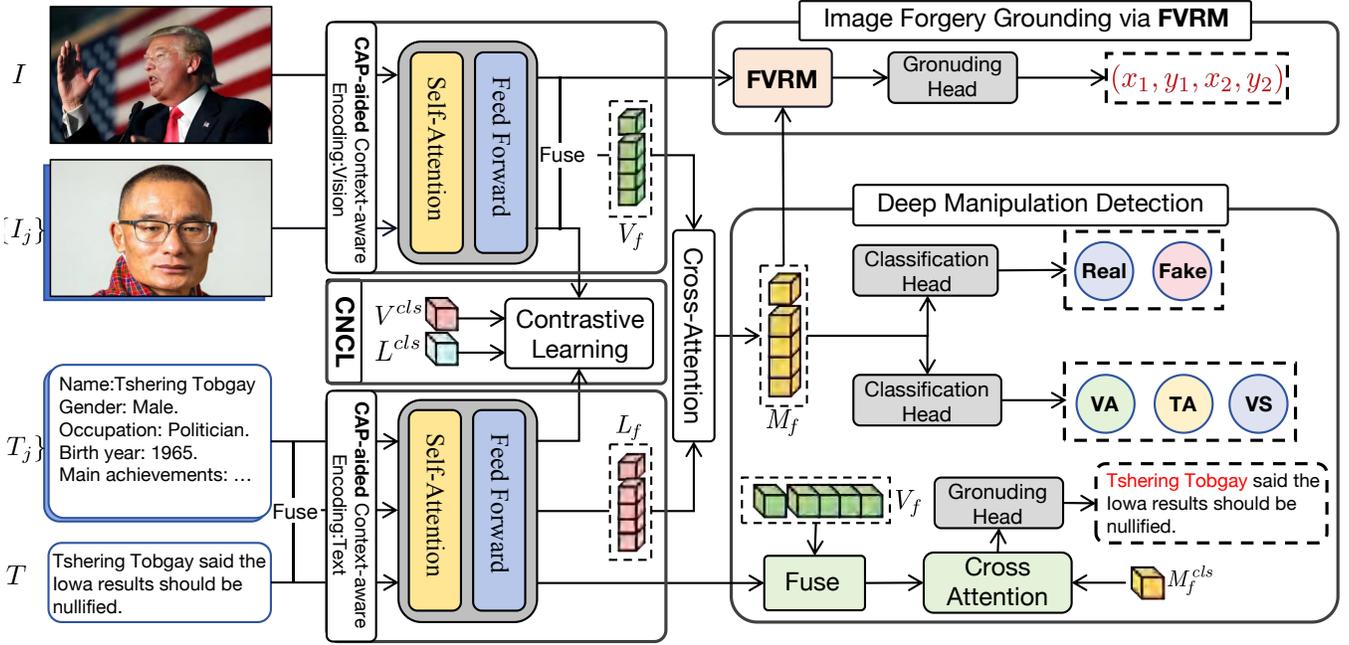


Figure 4: The architecture of our proposed RamDG. It processes image-text pairs and CAP-retrieved knowledge using uni-Encoders and the CNCL module, then fuses knowledge-enriched embeddings for visual/textual manipulation localization, fake news recognition, and manipulation type prediction.

4.4 Deep Manipulation Detection

Text Manipulation Localization. This section achieves fine-grained text manipulation localization: predicting whether each word in T is replaced or added. We first fuse the textual representation L^{tok} with the visual semantic feature V_f through cross-attention to generate a fine-grained cross-modal feature L_v^{tok} :

$$L_v^{tok} = \text{Attn}(Q = L^{tok}, K = V_f, V = V_f | \Theta_5) \quad (12)$$

M_f^{cls} can locate tokens in L_v^{tok} with abnormal logical relationships due to its global semantic information, which are often added or modified. Therefore, we further allow M_f^{cls} to extract information from L_v^{tok} . The above process is as follows:

$$\tilde{L}_v^{tok} = \text{Attn}(Q = L_v^{tok}, K = M_f^{cls}, V = M_f^{cls} | \Theta_6) \quad (13)$$

Finally, we use \tilde{L}_v^{tok} to predict the probability of manipulation for each token, and construct the text manipulation localization loss function \mathcal{L}_{tok} based on the cross-entropy loss function:

$$\mathcal{L}_{tok} = - \sum_i [y_{tok}^i \log(P_{tok}^i)] \quad (14)$$

where y_{tok}^i is the label indicating whether i -th token is manipulated. P_{tok}^i represents the probability, as output by the model, that the i -th token has been tampered.

Fake News Recognition. This section is used to determine whether (I, T) has been tampered. We utilize M_f^{cls} for recognition, as it integrates multi-modal news data and external knowledge from CAP to provide a comprehensive summary. We compute the loss using

a cross-entropy function, as detailed below:

$$\mathcal{L}_{2cls} = -y_{2cls} \log(P_{2cls}) \quad (15)$$

where y_{2cls} is the binary classification label and P_{2cls} represents the probability, predicted by the model, that (I, T) has been tampered with.

Manipulation Type Recognition. SAMM incorporates three types of tampering traces, including Visual Swap(VS) manipulation, Visual Attribute(VA) manipulation, Textual Attribute(TA) manipulation (including name swapping, and the addition or alteration of emotions). We predict the specific tampering method used in the given news. Consistent with Fake News Recognition, we use M_f^{cls} to predict and compute the loss using a cross-entropy function, as detailed below:

$$\mathcal{L}_{mcls} = - \sum_{i=1}^4 [y_{mcls}^i \log(P_{mcls}^i)] \quad (16)$$

where y_{mcls}^i denotes the label of the i -th tampering method used in the given news. P_{mcls}^i represents the probability, predicted by the model, that the i -th tampering method has been used.

4.5 Overall Loss Function

The overall loss function for the training process is as follows:

$$\mathcal{L} = \mathcal{L}_{cncl} + \mathcal{L}_{2cls} + \mathcal{L}_{mcls} + \mathcal{L}_{pat} + \mathcal{L}_{bbox} + \mathcal{L}_{tok} \quad (17)$$

5 Experiments

Please refer to the supplementary material for implementation details and evaluation metrics.

Table 2: Comparison of multi-modal learning methods for SAMM. We conduct a comprehensive evaluation of models' performance across four key tasks: binary classification, multi-label classification, image grounding, and text grounding.

Methods	Binary Cls(BC)			Multi-Label Cls(MLC)			Image Grounding(IG)			Text Grounding(TG)		
	AUC	EER	ACC	mAP	CF1	OF1	IoUmean	IoU50	IoU75	Precision	Recall	F1
VILT [22]	96.10	11.02	88.83	96.03	90.21	89.84	65.38	71.91	54.49	77.42	69.78	73.40
HAMMER [47]	97.85	7.80	92.43	97.98	93.77	93.44	77.68	84.41	78.44	85.94	82.74	84.31
HAMMER++ [48]	97.60	7.99	92.26	97.72	93.70	93.34	77.66	84.12	78.62	85.86	82.89	84.35
FKA-Owl [29]	98.09	7.19	92.60	2.53	13.97	13.84	66.40	73.54	54.82	19.16	49.71	27.66
Qwen2.5VL-72b [1]	76.67	44.93	55.06	—	—	—	—	—	—	—	—	—
RamDG(Ours)	98.79	5.42	94.66	98.86	95.52	95.33	80.90	87.56	82.00	86.16	83.54	84.83

Table 3: Comparison on the SAMM 20000 training set.

Methods	BC	MLC	IG	TG
	ACC	mAP	IoUmean	F1
VILT [22]	81.97	87.31	39.95	51.15
HAMMER [47]	85.74	92.98	52.65	71.65
HAMMER++ [48]	84.03	92.72	58.61	72.10
FKA-Owl [29]	87.91	1.94	49.48	9.79
Qwen2.5VL-72b [1]	55.06	—	—	—
RamDG(Ours)	88.40	95.32	64.30	73.69

Table 4: Additional VLMs' performance on the SAMM.

Methods (BC/ACC)	Zero-Shot	Finetuned
LLaMA-3.2-Vision-90B [42]	60.4	—
Gemini-3-27B [21]	59.7	—
SeedVL-1.5 [11]	64.1	—
Qwen2.5VL-3B [1]	—	82.0
RamDG(Ours)	—	94.66

Comparison Methods. We selected three modality fusion-based methods – VILT [22], HAMMER [47] and HAMMER++ [48], along with the state-of-the-art Visual-Language Large Models (VLLMs) – FKA-Owl [29] and Qwen2.5VL-72b [1] as baselines for performance comparison with RamDG on the SAMM. Implementation details for the methods can be found in the Appendix.

5.1 Quantitative Results

Performance Comparison. Table 2 shows the performance of all the aforementioned baselines on the SAMM dataset. To simulate real-world scenarios with scarce training samples, we train these models on randomly selected subsets of 20,000 and 50,000 samples. We then evaluate their performance on the complete test set, as shown in Table 3 and Table 5. Experimental results showed in tables prove that our method achieved state-of-the-art performance across various tasks on the SAMM dataset. Notably, under conditions of limited training data, RamDG demonstrated significant advantages

Table 5: Comparison on the SAMM 50000 training set.

Methods	BC	MLC	IG	TG
	ACC	mAP	IoUmean	F1
VILT [22]	85.18	92.73	55.50	64.91
HAMMER [47]	88.16	95.58	65.97	76.02
HAMMER++ [48]	87.99	95.29	68.15	78.48
FKA-Owl [29]	90.36	1.19	63.28	27.66
Qwen2.5VL-72b [1]	55.06	—	—	—
RamDG(Ours)	91.07	97.18	73.65	79.10

Table 6: Generalization to unseen entities.

Methods	ACC	mAP	IoUmean	F1
HAMMER [47]	92.0	97.0	77.6	83.8
FKA-Owl [29]	92.3	4.0	68.1	28.1
RamDG(Ours)	94.1	97.3	78.7	83.9

over baseline models, particularly in the precision of visual tampering region localization, fully demonstrating the effectiveness and superiority of our proposed method.

While FKA-Owl slightly outperforms HAMMER in binary classification (+0.24%), it fails at fine-grained tampering localization. Though VLLMs leverage rich knowledge for fake news judgment, they lack fine-grained extraction capability. In contrast, our RamDG: 1) retrieves CAP knowledge, 2) integrates it via CNCL, and 3) enhances visual localization with FVRM – achieving superior performance across all tasks.

More VLMs' performance. We conduct evaluations on additional VLMs under both zero-shot and fine-tuned settings, as shown in the table 4.

Generalization to Unseen Entities. To evaluate new entities absent from CAP, we select a sub-test set whose entities absent from the training set and directly input them into RamDG without retrieving information from CAP. As shown in the table 6, our RamDG still outperforms comparison methods.

5.2 Ablation study

External knowledge from CAP. To investigate the impact of different celebrity information in CAP on model performance, we

Table 7: Ablation study for external knowledge from CAP. For each task, we present the most representative metrics: ACC, mAP, IoUmean, and F1.

External Knowledge From CAP					BC	MLC	IG	TG
Gender	Birth Year	Occupation	Main Achievements	Images	ACC	mAP	IoUmean	F1
				✓	91.00	96.88	75.77	83.49
				✓	91.17	97.01	76.79	83.26
✓	✓	✓	✓	✓	93.73	98.08	79.49	84.84
	✓	✓	✓	✓	94.32	98.21	79.84	84.10
✓		✓	✓	✓	94.57	98.35	80.59	84.50
✓	✓		✓	✓	93.25	98.15	77.96	84.12
✓	✓	✓		✓	93.48	98.34	79.46	84.69
✓	✓	✓	✓	✓	94.66	98.86	80.90	84.83

Table 8: Ablation study for CNCL and FVRM.

Module		BC	MLC	IG	TG
CNCL	FVRM	ACC	mAP	IoUmean	F1
	✓	93.24	98.18	79.32	84.23
✓		94.79	98.88	78.01	85.28
✓	✓	94.66	98.86	80.90	84.83

conducted a series of ablation experiments. The results, as shown in Table 7, reveal several key observations:

1) Without leveraging CAP-derived external knowledge, model performance across tasks drops by an average of 3%. Single-modal external knowledge alone is insufficient: textual knowledge is indispensable, with its absence causing significant declines in fake news detection (4.11% drop) and even visual localization. In contrast, visual knowledge provides minimal improvement (0.78% average gain) due to image redundancy.

2) Analysis shows textual knowledge components affect performance variably. Occupation information contributes the most (1.44% average gain), reflecting its role in providing contextual and social cues for human verification. For example, knowing Messi’s occupation helps debunk false claims like his Nobel Prize win.

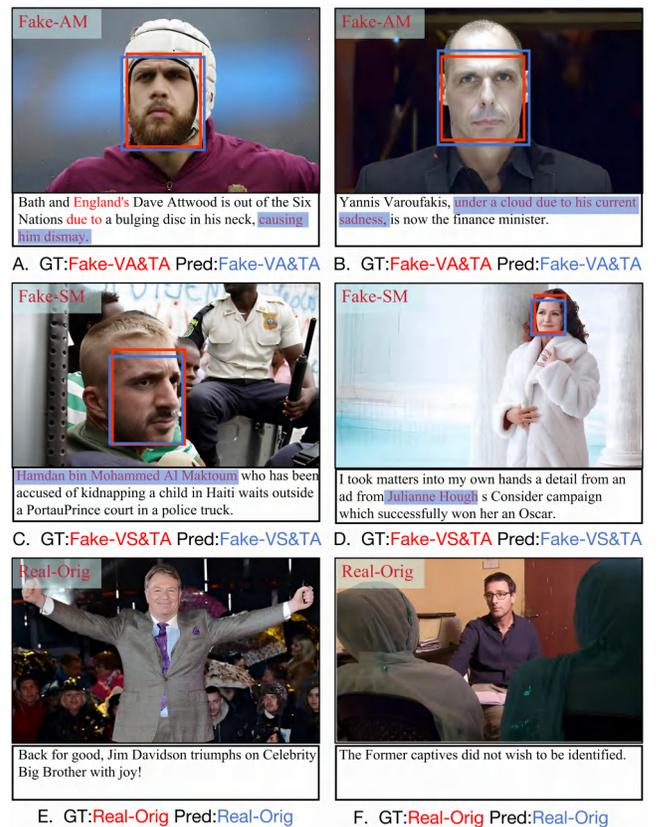
Framework Component Ablation. As shown in Table 8, we investigated the impact of CNCL and FVRM on model performance. After removing the FVRM module, we directly used M_{fv}^{cls} for visual tampering localization. The results reveal several key insights:

1) Removing CNCL reduces performance across all tasks (avg. -1.07%), confirming its role in enhancing external knowledge understanding for multimodal tampering detection.

2) FVRM specifically boosts visual tampering localization (+2.89%) with minimal impact on other tasks, demonstrating its fine-grained visual tampering capture capability.

5.3 Visualized results

Figure 5 presents results for six cases: Examples A–B (Attribute Manipulation, AM) involve Visual Attribute (VA) and Textual Attribute (TA) manipulations; C–D (Swap Manipulation, SM) feature Visual Swap (VS) and TA manipulations; E–F represent original, unmanipulated news.

**Figure 5: Visualized results. Red and blue regions denote the ground truth and predicted outputs, respectively.**

6 Conclusion

In this paper, we propose a realistic research scenario: detecting and grounding semantic-coordinated multimodal manipulations, and introduce a new dataset SAMM. To address this challenge, we design the RamDG framework, proposing a novel approach for detecting fake news by leveraging external knowledge, which consists of two core components: CNCL and FVRM. Extensive experimental results demonstrate the effectiveness of our approach.

Acknowledgments

The paper is supported by the Fundamental Research Funds for the Central Universities with No. JZ2024HG7B0261 and the NSFC project under grant No. 62302140. The computation is completed on the HPC Platform of Hefei University of Technology.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *CoRR* abs/2502.13923 (2025). doi:10.48550/ARXIV.2502.13923 arXiv:2502.13923
- [2] Ali Furkan Biten, Lluís Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context driven entity-aware captioning for news images. In *CVPR*. 12466–12475.
- [3] Ali Borji. 2023. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vision Comput* 137 (2023).
- [4] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework for High Fidelity Face Swapping. In *ACM MM*. 2003–2011.
- [5] Federico Cocchi, Lorenzo Baraldi, Samuele Poppi, and Marcella Cornia. 2023. Unveiling the Impact of Image Transformations on Deepfake Detection: An Experimental Analysis. In *ICIAP*. 345–356.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.
- [7] Tarik Dzanic, Karan Shah, and Freddie Witherden. 2020. Fourier spectrum discrepancies in deep network generated images. In *NeurIPS*.
- [8] Sheera Frenkel and Cecilia Kang. 2021. How Facebook Failed to Stop Misinformation in India. *The New York Times* (October 2021).
- [9] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information Bottleneck Disentanglement for Identity Swapping. In *CVPR*. 3404–3413.
- [10] Google. 2024. Google Search API. <https://developers.google.com/custom-search/v1/overview>.
- [11] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, Jingji Chen, Jingjia Huang, Kang Lei, Liping Yuan, Lishu Luo, Pengfei Liu, Qinghao Ye, Rui Qian, Shen Yan, Shixiong Zhao, Shuai Peng, Shuangye Li, Sihang Yuan, Sijin Wu, Tianheng Cheng, Weiwei Liu, Wenqian Wang, Xianhan Zeng, Xiao Liu, Xiaobo Qin, Xiaohan Ding, Xiaojun Xiao, Xiaoying Zhang, Xuanwei Zhang, Xuehan Xiong, Yanghua Peng, Yangrui Chen, Yanwei Li, Yanxu Hu, Yi Lin, Yiyuan Hu, Yiyuan Zhang, Youbin Wu, Yu Li, Yudong Liu, Yue Ling, Yujia Qin, Zhanbo Wang, Zhiwu He, Aoxue Zhang, Bairen Yi, Bencheng Liao, Can Huang, Can Zhang, Chaorui Deng, Chaoyi Deng, Cheng Lin, Cheng Yuan, Chenggang Li, Chenhui Gou, Chenwei Lou, Chengzhi Wei, Chundian Liu, Chunyuan Li, Deyao Zhu, Donghong Zhong, Feng Li, Feng Zhang, Gang Wu, Guodong Li, Guohong Xiao, Haibin Lin, Haihua Yang, Haoming Wang, Heng Ji, Hongxiang Hao, Hui Shen, Huixia Li, Jiahao Li, Jialong Wu, Jianhua Zhu, Jianpeng Jiao, Jiashi Feng, Jiaye Chen, Jianhui Duan, Jihao Liu, Jin Zeng, Jingqun Tang, Jingyu Sun, Jia Chen, Jun Long, Junda Feng, Junfeng Zhan, Junjie Fang, Junting Lu, Kai Hua, Kai Liu, Kai Shen, Kaiyuan Zhang, and Ke Shen. 2025. Seed1.5-VL Technical Report. *CoRR* abs/2505.07062 (2025). doi:10.48550/ARXIV.2505.07062 arXiv:2505.07062
- [12] Drew Harwell. 2019. A 'Deepfake' President Trump Mocked Facebook and Mark Zuckerberg in a Video That Went Viral. *The Washington Post* (June 2019).
- [13] Kilichbek Haydarov, Aashiq Muhamed, Xiaoqian Shen, Jovana Lazarevic, Ivan Skorokhodov, Chamuditha Jayanga Galappaththige, and Mohamed Elhoseiny. 2024. Adversarial Text to Continuous Image Generation. In *CVPR*. 6316–6326.
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. 9726–9735.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.
- [16] Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRA-GeNews: Multimodal Realistic AI-Generated News Detection. In *EMNLP*. 16436–16448.
- [17] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In *CVPR*. 6458–6467.
- [18] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *CVPR*. 2889–2898.
- [19] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *CVPR*. 2889–2898.
- [20] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. 2024. CaLa: Complementary Association Learning for Augmenting Composed Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*, Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (Eds.). ACM, 2177–2187. doi:10.1145/3626772.3657823
- [21] Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Novan Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Ji-aming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesh Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreenpathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Jan Balantyne, Idan Szepes, and Ivan Nardini. 2025. Gemma 3 Technical Report. *CoRR* abs/2503.19786 (2025). doi:10.48550/ARXIV.2503.19786 arXiv:2503.19786
- [22] Wonjae Kim, Bokyoung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.
- [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*. 19730–19742.
- [24] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: Dual Shot Face Detector. In *CVPR*. 5060–5069.
- [25] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-ray for More General Face Forgery Detection. In *CVPR*. 5001–5010.
- [26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *CVPR*. 3207–3216.
- [27] Zhengqi Li, Richard Tucker, Noah Snively, and Aleksander Holynski. 2024. Generative Image Dynamics. In *CVPR*. 24142–24153.
- [28] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual News: Benchmark and Challenges in News Image Captioning. In *EMNLP*. 6761–6771.
- [29] Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. FKA-Owl: Advancing Multimodal Fake News Detection through Knowledge-Augmented LLMs. In *ACM MM*. 10154–10163.
- [30] Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023. COCO: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning. In *EMNLP*. 16167–16188.
- [31] Peter Lorenz, Ricard L. Durall, and Janis Keuper. 2023. Detecting Images Generated by Deep Diffusion Models using their Local Intrinsic Dimensionality. In *ICCV*. 448–459.
- [32] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: benchmarking human and model perception of AI-generated images. In *NeurIPS*.
- [33] Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLippings: Automatic Generation of Out-of-Context Multimodal Media. In *EMNLP*. 6801–6817.
- [34] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *ICML*. 24950–24962.
- [35] Minh-Quang Nguyen, Khanh-Duy Ho, Hoang-Minh Nguyen, Canh-Minh Tu, Minh-Triet Tran, and Trong-Le Do. 2023. Unmasking The Artist: Discriminating Human-Drawn And AI-Generated Human Face Art Through Facial Feature Analysis. 1–6.
- [36] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *CVPR*. 24480–24489.
- [37] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*. 2085–2094.
- [38] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. In *CVPR*. 13052–13062.

- [39] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection.. In *SIGIR*. 153–162.
- [40] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2022. Detecting and re-covering sequential deepfake manipulation.. In *ECCV*. 86–103.
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Sandhini Agarwal, Gabriel Goh, Girish Sastry, Amanda Askell, Pamela Mishkin, Gretchen Krueger, Jack Clark, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.
- [42] Meta AI Research. 2025. Llama 3.2 Model. <https://llama.meta.com/llama3.2>.
- [43] Hamid Rezaatofghi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*. 658–666.
- [44] Adam Satariano and Davey Alba. 2020. How Facebook and YouTube Spread False 'Plandemic' Video. *The New York Times* (May 2020).
- [45] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. 2020. Regularized fine-grained meta face anti-spoofing.. In *AAAI*. 11974–11981.
- [46] Rui Shao, Tianxing Wu, and Ziwei Liu. 2022. Detecting and re-covering sequential deepfake manipulation.. In *ECCV*.
- [47] Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and Grounding Multi-Modal Media Manipulation. In *CVPR*. 6904–6913.
- [48] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. 2024. Detecting and Grounding Multi-Modal Media Manipulation and Beyond. *IEEE TPAMI* 46, 8 (2024), 5556–5574.
- [49] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective.. In *KDD*. 22–36.
- [50] Sergey Sinitisa and Ohad Fried. 2024. Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis. In *WACV*. 4067–4076.
- [51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023). doi:10.48550/ARXIV.2307.09288 arXiv:2307.09288
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.
- [53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *CoRR* abs/2409.12191 (2024). doi:10.48550/ARXIV.2409.12191 arXiv:2409.12191
- [54] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-Fidelity GAN Inversion for Image Attribute Editing. In *CVPR*. 11379–11388.
- [55] William Yang Wang. 2017. Liar, Liar Pants on Fire : A New Benchmark Dataset for Fake News Detection. In *ACL*. 422–426.
- [56] Yaxiong Wang, Lianwei Wu, Lechao Cheng, Zhun Zhong, Yujiao Wu, and Meng Wang. 2025. Beyond General Alignment: Fine-Grained Entity-Centric Image-Text Matching with Multimodal Attentive Experts. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 792–802. doi:10.1145/3726302.3729902
- [57] Yaxiong Wang, Hao Yang, Xiuxiu Bai, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2021. PFAN++: Bi-Directional Image-Text Retrieval With Position Focused Attention Network. *IEEE Trans. Multim.* 23 (2021), 3362–3376. doi:10.1109/TMM.2020.3024822
- [58] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position Focused Attention Network for Image-Text Matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 3792–3798. doi:10.24963/IJCAI.2019/526
- [59] Zhenqiang Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. In *ICCV*. 22445–22455.
- [60] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *CoRR* abs/2412.15115 (2024). doi:10.48550/ARXIV.2412.15115 arXiv:2412.15115
- [61] KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust Multi-bit Natural Language Watermarking through Invariant Features. In *ACL*. 2092–2115.
- [62] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. 2024. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. In *CVPR*. 25669–25680.
- [63] Zhenxing Zhang, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Dan Guo, and Meng Wang. 2025. ASAP: Advancing Semantic Alignment Promotes Multi-Modal Manipulation Detecting and Grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 4005–4014. https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_ASAP_Advancing_Semantic_Alignment_Promotes_Multi-Modal_Manipulation_Detecting_and_Grounding_CVPR_2025_paper.html
- [64] Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022. Distillation-Resistant Watermarking for Model Protection in NLP. In *EMNLP*. 5044–5055.
- [65] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. 2021. Face Forensics in the Wild. In *CVPR*. 5778–5788.