The Coherence Trap: When MLLM-Crafted Narratives Exploit Manipulated Visual Contexts

Supplementary Material

Contents

. Related Work 1

. Prompt Paradigm 2
. Prompt for AMD . . . 2
. Prompt for General-purpose Model . . . 2

. Experimental Setup 2
. Implementation Details . . . 2
. Baselines . . . 2
. Evaluation Metrics . . . 3

. Ethics Statements 3

1. Related Work

Deepfake Detection. The rapid progress of generative models and the surge in synthetic content have accelerated advances in Deepfake detection. Existing work spans unimodal and multimodal approaches. Unimodal, including image-based [8, 17] and text-based [4, 16] approaches, already achieve strong results. With the rise of Multimodal Large Language Models (MLLMs), multimodal Deepfake detection has received increasing attention [9, 10, 14]. Regarding datasets, Shao *et al.* [14] introduced the pioneering DGM^4 benchmark for multimodal manipulation detection and grounding. However, its manipulations are rule-based, leading to semantically fragmented image-text discrepancies that do not accurately reflect real-world misinformation. MMFakeBench [10] recognized this limitation and proposed generating semantically aligned news images using text-to-image models. Yet such semantically matched samples constitute only 30% of its fake subset, and the dataset contains merely 11k samples, limiting its utility for training robust detectors. Existing Deepfake datasets also fail to consider the risk of semantically coherent but misleading text generated by modern MLLMs. On the methodological side, HAMMER [14] integrates contrastive learning to build a detector capable of classifying manipulation types and grounding manipulated regions, but it does not address cross-domain robustness. Beyond conventional multimodal detectors, FKA-Owl [9] employs a 7B-scale MLLM with several architectural modifications to enhance generalization. However, it is trained on DGM^4, where text manipulations follow fixed editing rules rather than being synthesized by MLLMs, making it unsuitable for detecting more subtle, semantically aligned misinformation produced by modern models. Moreover, FKA-Owl performs only binary real/fake classification without fine-grained manipulation type prediction or grounding, and its large backbone and heavy architectural design result in substantially increased computational cost and slower inference.

Multi-Modal Large Language Model. In recent years, Multi-Modal Large Language Models have emerged as a crucial technology for understanding and reasoning across multiple modalities, particularly text and images. By extending the capabilities of Large Language Models (LLMs) to incorporate visual inputs, these models have demonstrated outstanding performance in tasks such as image captioning and visual question answering. CLIP [13] and ALIGN [5] leveraged contrastive learning to align visual and textual representations, enabling efficient zero-shot vision-language understanding. Subsequently, models such as Flamingo [1] and BLIP-2 [7] have introduced vision-language transformers, integrating pre-trained LLMs with vision encoders to enhance cross-modal reasoning and generative capabilities. More recently, GPT-4V [12] and Florence-2 [2] have significantly enhanced the potential of MLLMs in tackling complex multi-modal tasks by leveraging a more efficient framework and larger-scale pre-training data. A key advantage of MLLMs is their acquisition of extensive world knowledge through large-scale pretraining, which substantially strengthens their reasoning abilities in downstream tasks. Such knowledge not only enhances cross-modal understanding but also proves essential for handling challenging problems,

including misinformation detection.

2. Prompt Paradigm

2.1. Prompt for AMD

The details of the heuristic question-answer prompts in AMD are as follows:

###Human:

```
  The following are multiple choice
questions about fake news detection.
The text caption of the news is:
<Text>.  The identity and emotion
of the face, and the semantic and
sentiment of the text should not be
manipulated.  Question:  Is there any
face swap/attribute or text fabrication
in the news?
  A. No.
  B. Only face swap.
  C. Only face attribute.
  D. Only text swap.
  E. Both face swap and text
fabrication.
  F. Both face attribute and text
fabrication.
  If there is manipulation of a face,
locate the one most likely manipulated
face in the image and append the
results to your selected option.
  The answer is:
```

###Assistant:

```
  <Option>[Manipulated face:
<loc_x1><loc_y1><loc_x2><loc_y2>]
```

Where $<$ Text $>$ refers to the textual narratives paired with the input image, $<$ Option $>$ represents the correct answer option for this sample, such as *E. Both face swap and text fabrication.* And $<$ loc_ $>$ is added to the vocabulary as a special token representing coordinates. Fig. 6 shows two kinds of prompts.

2.2. Prompt for General-purpose Model

To ensure fairer testing and more credible results for general-purpose models (Tab.2 in the main paper), we enhanced the invocation of general-purpose models by adding more detailed descriptions to the AMD prompt, as follows:

###Human:

```
  <Same as AMD>
  If face manipulation, use rectangular
box coordinates in the format of
[x1,y1,x2,y2], where the top-left
vertex of the image is defined as (0,0)
and the bottom-right vertex as (1,1)
for relative positioning, and append
the results to the option you have
selected.
  DO NOT output analysis.  ONLY output
final answer in format:  [Option +
Coordinates (if applicable).]
```

3. Experimental Setup

3.1. Implementation Details

All experiments are conducted on 4 NVIDIA GeForce RTX 4090 GPUs using **Distributed Data Parallel (DDP)** training in PyTorch. The image encoder $\mathcal{E}_v$ is based on DaViT [3], with Florence-2-B [2] serving as the backbone. The APE $\mathcal{E}_m^p$ is based on the Florence-2 encoder and remains frozen during training. Thus, in the APE stage, only the artifact token, classifier head, and attention pooling module are jointly trained. The classifiers and bounding box (bbox) detector consist of two Multi-Layer Perceptron layers, with output dimensions of 2 and 4, respectively. For manipulation detection guidance, the **Dual-Branch Manipulation** shares a common classifier.

The training images are resized to 224×224 and undergo random horizontal flipping. The batch size per GPU is set to 6, and the model is trained for 12 epochs. We use the AdamW optimizer [11] with an initial learning rate of $1e^{-7}$ and a weight decay of 0.01. A cosine learning rate scheduler with a warm-up phase is applied, gradually increasing the learning rate to $1e^{-6}$ in the first 1000 steps, and then decaying it to $1e^{-7}$ throughout training. Our code will be released to provide further implementation details.

3.2. Baselines

We adapt four state-of-the-art multi-modal methods to the MDSM setting for comparison, including three multi-modal manipulation detection models and one multi-modal learning approach:

- **HAMMER** [14] is a pioneering model for the multi-modal manipulation detection and grounding. It employs two unimodal encoders to extract visual and textual forgery features, which are then aligned through contrastive learning. Following this, a multi-branch transformer architecture with two specialized decoders is utilized for manipulation detection and grounding.
- **HAMMER++** [15] is a more powerful model that builds upon HAMMER by integrating contrastive learning from both global and local perspectives.
- **FKA-Owl** [9] is another pioneering model designed for large vision-language models to perform multi-modal fake news detection, and it demonstrates outstanding cross-domain performance. Since FKA-Owl does not support fine-grained classification tasks, we fine-tuned it using the same prompts as those used for AMD.
- **ViLT** [6], for the multi-modal learning approach, is a

Samples with coordinates

###Human:
The following are multiple choice questions about fake news detection. The text caption of the news is: Benjamin Netanyahu and Elon Musk discuss global innovation and future collaborations in high-level meeting. The identity and emotion of the face, and the semantic and sentiment of the text should not be manipulated. Question: Is there any face swap/attribute or text fabrication in the news?

A. *No.*
B. *Only face swap.*
C. *Only face attribute.*
D. *Only text swap.*
E. *Both face swap and text fabrication.*
F. *Both face attribute and text fabrication.*
If there is manipulation of a face, locate the one most likely manipulated face in the image and append the results to your selected option.
The answer is:

###Assistant:
E. Both face swap and text fabrication. Manipulated face: <loc_39><loc_30><loc_58><loc_72>

Samples without coordinates

###Human:
The following are multiple choice questions about fake news detection. The text caption of the news is: A comedian threw stacks of money at Sepp Blatter during a meeting of FIFA's executive committee in Zurich. The identity and emotion of the face, and the semantic and sentiment of the text should not be manipulated. Question: Is there any face swap/attribute or text fabrication in the news?

A. *No.*
B. *Only face swap.*
C. *Only face attribute.*
D. *Only text swap.*
E. *Both face swap and text fabrication.*
F. *Both face attribute and text fabrication.*
If there is manipulation of a face, locate the one most likely manipulated face in the image and append the results to your selected option.
The answer is:

###Assistant:
A. No.

Figure 6. Examples of Image-Prompt pairs in AMD.

representative single-stream method where cross-modal interaction layers operate on the concatenation of image and text inputs. For adaptation, We add classification and detection heads to the corresponding outputs of the model.

3.3. Evaluation Metrics

To comprehensively evaluate our proposed MDSM, we follow the rigorous evaluation protocols and metrics outlined in [14] for all manipulation detection and grounding tasks. The detailed evaluation setup is organized as follows:

- **Binary Classification. Accuracy (ACC)** is adopted as the evaluation metric to measure the correctness of real/fake news classification results.
- **Multi-Label Classification.** For multi-label classification tasks, we employ the **mean Average Precision (mAP)**, which measures the per-class average precision and then takes the arithmetic mean across all manipulation types. This macro-averaged mAP provides a comprehensive evaluation of the model's overall performance across different manipulation types.
- **Manipulated Image Bounding Box Grounding.** To evaluate the precision of predicted manipulated bounding boxes, we calculate the **mean Intersection over Union (mIoU)** between the ground-truth and predicted coordinates for all testing samples. This metric quantifies the spatial overlap between detected regions and actual manipulated areas, reflecting the localization accuracy of the model.
- **Manipulated Text Token Grounding.** In the DGM^4 benchmark, an additional task of manipulated text token grounding is included. For this task, **Precision** is used as the evaluation metric to measure the accuracy of identifying manipulated text tokens within input sequences.

This standardized evaluation framework ensures a systematic and comparative assessment of MDSM across diverse manipulation scenarios, aligning with both general detection tasks and benchmark-specific requirements.

4. Ethics Statements

The MDSM dataset and associated analyses were created solely to support research on detection and mitigation of modern MLLM-driven multimodal misinformation. We recognize that assembling realistic, semantically coherent synthetic examples entails dual-use risks: the same materials and procedures could be misused to produce deceptive content. To minimize harm, we adopt a harm-minimizing, controlled-release approach: we will not publish the generation pipeline, detailed prompts, or prompt–response pairs to prevent their exploitation by adversaries for generating harmful content; public distribution is limited to vetted, research-only access under a signed Data Usage Agreement (DUA); distributed images will carry conspicuous visual watermarks and standardized metadata tags; high-fidelity originals and sensitive metadata will be withheld; images of minors and clearly sensitive contemporary conflict content have been excluded; and reserve the right to revoke access on evidence of misuse. Full technical and procedural details of these safeguards are documented in the Appendix and in the dataset README.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Xiao Bin, Wu Haiping, Xu Weijian, Dai Xiyang, Hu Houdong, Lu Yumao, Zeng Michael, Liu Ce, and Yuan Lu. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2023. 1, 2

[3] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *ECCV*, 2022. 2

[4] Kung-Hsiang Huang, McKeown Kathleen, Nakov Preslav, Choi Yejin, and Ji Heng. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, 2023. 1

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[6] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[8] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 1

[9] Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *ACM MM*, 2024. 1, 2

[10] Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *ICLR*, 2025. 1

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2017. 2

[12] OpenAI. Gpt-4 technical report. *CoRR*, 2023. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[14] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *CVPR*, 2023. 1, 2, 3

[15] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *TPAMI*, 46(8):5556–5574, 2024. 2

[16] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 1

[17] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, 2021. 1