Understanding the Limits of Vision-Language Models as Deepfake Detectors Under Realistic Settings

Simone Teglia[1], Irene Amerini[1]

[1]Sapienza University of Rome, Department of Computer, Control and Management Engineering "Antonio Ruberti"

teglia@diag.uniroma1.it, amerini@diag.uniroma1.it

In recent years, realistic synthetic media, commonly known as deepfakes, have been used to spread fake news or alter narratives of important events, therefore raising the importance of reliable deepfake detection systems. As specialized detectors continue to improve, the emergence of Vision-Language Models (VLMs) question their usability as prompt-driven deepfake detectors. Recent studies have focused on evaluating VLMs in zero-shot [9, 8] or more informative prompted settings [6, 5], demonstrating that models such as CLIP [13], Flamingo [1] and QwenVL [12] exhibit emergent sensitivity to visual inconsistencies without task-specific training. However, despite their inherent robustness allows them to transfer knowledge to new visual concepts defined purely by language, previous studies have primarily concentrated on other types of zero-shot tasks like image classification [15, 13], object detection [3] and visual question answering [10, 7, 2], and only handful of them have explored the use of VLMs as standalone deepfake detectors in true zero-shot or few-shot settings [11, 16], particularly under realistic visual conditions.

Consequently, our work addresses this gap, first by reframing deepfake detection within the context of social network imagery: instead of evaluating on controlled, high-resolution facial datasets such as FaceForensics++ [14] or Celeb-DF [17], we use SID-Set [4], a dataset designed to mirror the heterogeneous nature of images commonly found on modern social media platforms. We evaluate recent State-Of-The-Art VLMs in zero-shot and one-shot settings, exploring a range of prompting strategies and observing that more structured or information-dense prompts do not always lead to improved performance. As part of this exploration, we also introduce dynamic prompting, an image-tailored approach that encourages the model to consider visual cues inspired by classical computer vision techniques, such as edges, color distributions, lighting, and depth. Our observations suggest that, while dynamic prompting can influence the model's analytical process, its effectiveness is not uniform.

Finally, extending our focus on social network imagery, we evaluate the models on compressed versions of the dataset to better reflect real-world social-network conditions. Results show that compression artifacts further challenge VLM performance, suggesting that additional work is needed before prompt-driven deepfake detectors can operate reliably in the wild.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[2] Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Zero-shot visual question answering with language model feedback. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9268–9281, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Yongchao Feng, Yajie Liu, Shuai Yang, Wenrui Cai, Jinqing Zhang, Qiqi Zhan, Ziyue Huang, Hongxi Yan, Qiao Wan, Chenguang Liu, Junzhe Wang, Jiahui Lv, Ziqi Liu, Tengyuan Shi, Qingjie Liu, and Yunhong Wang. Vision-language model for object detection and segmentation: A review and evaluation, 2025.

[4] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model, 2025.

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

[6] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 12:193057–193075, 2024.

[8] XIAOSONG MA, Jie ZHANG, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 65252–65264. Curran Associates, Inc., 2023.

[9] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

[10] Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis, 2024.

[11] Viacheslav Pirogov. Visual language models as zero-shot deepfake detectors, 2025.

[12] Qwen, :, 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, Jiaxi 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[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, 2021.

[14] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 01 2019.

[15] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17542–17552, 2024.

[16] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection, 2025.

[17] Pu Sun Honggang Qi Yuezun Li, Xin Yang and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.