Deepfake Explainability Challenge 2026

Abhinav Dhall, *Monash University*
Shreya Ghosh, *The University of Queensland*
Muhammad Haris Khan, *MBZUAI*
Usman Tariq, *American University of Sharjah*

Challenge Description:

The rapid advancement of *Generative AI* has fundamentally reshaped how visual content is produced, circulated, and consumed. Recent diffusion and vision language models (VLMs) can now fabricate highly persuasive, photorealistic deepfakes that extend far beyond identity alterations, enabling manipulations of human actions, intent, object interactions, and scene-level semantics. Over the next 3–5 years, these high-level, contextually rich fabrications are expected to become increasingly prevalent across social media, news ecosystems, and interactive multimedia platforms.

As deepfakes evolve from simple facial swaps to rich, context-aware manipulations driven by large-scale generative models, the core challenge facing the multimedia community is no longer detection alone but *understanding why* an image is identified as manipulated. Traditional detectors often act as black boxes, flagging inconsistencies without offering human-interpretable reasoning. This lack of transparency limits trust and hinders deployment in high-stakes environments, and makes it difficult for researchers and practitioners to diagnose model failures.

The **Deepfake Explainability Challenge** addresses this crucial gap by shifting the focus from mere classification to interpretable, evidence-driven deepfake understanding. Instead of asking models to simply detect or localise manipulated regions, the challenge requires participants to generate meaningful, human-understandable explanations that pinpoint *why* particular pixels, regions, or semantic attributes indicate manipulation.

Built on top of *MultiFakeVerse [1]*, a large-scale dataset of 845,286 VLM-guided person-centric manipulations, the challenge is designed to test models on subtle, high-semantic edits involving human object interactions, roles, activities, emotional intent, co-occurring people, and contextual or narrative shifts. These are precisely the manipulations that fool both humans and state-of-the-art detectors making explainable reasoning essential.

After two years successfully hosting 1M Deepfake detection challenge at ACM Multimedia 2024, 2025 with more than 200+ teams participating the challenge, this year our focus is more into explainability part of the manipulated content along with the

detection performance. The challenge organisers intend to run this benchmarking effort over the next five years while introducing different facet of deepfakes.

Task Description

The MultiFakeVerse Challenge comprises of two sub-tasks:

a. Deepfake Detection – Given an image sample, the task is to identify if the image is a deepfake or real.

b. Deepfakes Explainability – Given an image sample, the task is to find out the explanation behind the model's decision with respect to the manipulation. The assumption here is that from the perspective of spreading misinformation.

The dataset used for the two tasks above is the recently proposed MultiFakeVerse [1]. The database contains over 845,286 person-centric images.

Outline for SOTA

Most existing deepfake datasets concentrate on person-level facial manipulations, restricting edits to identity swaps or changes in facial expressions. For example, DFFD provides 300K GAN-generated or edited facial images, while widely used datasets such as FFHQ and FakeSpotter similarly emphasise face-centric alterations. Beyond faces, some datasets explore other domains: OHImg targets aerial imagery, M3Dsynth focuses on manipulated medical images, and SIDA [7] uses vision–language models to replace objects in images via inpainting (e.g., cat → dog). However, these efforts mainly capture object- or scene-level perturbations and do not address person-centric edits that meaningfully alter an image's interpretation. SemiTruths [6] extends this direction by introducing multi-level manipulation across objects and scenes using Stable Diffusion and prompt-based LLAMA-7B. Yet, it remains object-scene oriented and does not engage with the nuanced challenges of manipulating people within images.

A person-centric focus is crucial. Studies indicate that deepfakes involving real individuals are viewed and shared up to six times more than generic content [5], and over 96% of online deepfakes are non-consensual pornography [4]. This highlights the urgent societal and ethical implications of human-targeted manipulations yet existing datasets rarely address them. Real-world deepfakes frequently involve full-body edits, multiple people, contextual shifts, and composite scene alterations. These complexities demand new benchmarks designed specifically for rich, high-semantic, person-centred manipulations to advance the next generation of deepfake detection systems.

Data Documentation

We will release link to sites containing relevant datasets to be used for objective training and evaluation of the grand challenge tasks. Full appropriate documentation on the datasets will be provided.

The information on MultiFakeVerse (data, baseline code and ACM Multimedia 2025 Paper) are accessible at https://github.com/Parul-Gupta/MultiFakeVerse

Along with this, we will provide a sophisticated webpage for 2026 version of the challenge (similar to our previous iterations https://github.com/ControlNet/AV-Deepfake1M).

Evaluation and Submission Platform

The evaluation scripts, baseline models and code will be made available challenge website (similar to our previous iterations https://github.com/ControlNet/AV-Deepfake1M). The participants will submit the labels for evaluation and the top three teams need to share their code base for evaluation purposes. Each team will be invited to submit a paper describing their method.

A commitment to publish and maintain a website

The organisers are committed to the problem of deepfake detection as also reflected by prior works in deepfakes detection. The data, code information and later meta-information regarding the findings from the challenge will be maintained on the public Github link for at least the next 3 years.

Work with ACM Multimedia Conference organizers to publicize the Grand Challenge tasks to researchers for participation.

The organisers have previously organised events at ACM Multimedia (two iterations of One Million Deepfakes on 2025 and 2025, MRAC workshop series-Ghosh et al 2023, 2024, 2025). The organisers will work with the ACM Multimedia 2026 publicity and grand challenge chairs for dissemination of the call for participation and media releases for the different stages of the challenge.

Contact information of at least two organizers

1. Abhinav Dhall, Abhinav.dhall@monash.edu
2. Shreya Ghosh, shreya.ghosh@uq.edu.au

Multi year Commitment:

The organisers are planning for a series of challenges in deepfake detection. This proposal is the third iteration of deepfake challenge.

In the past too, the organisers (Dhall and Ghosh) have been successfully organising the Emotion Recognition in the Wild Challenge series.

Bio of the Organisers

Abhinav Dhall is an Associate Professor at Monash University, Australia. He received a PhD in computer science from the Australian Naonal University. He has been on the

organisation teams of ACM Multimedia 2024, ACII 2024, ICVGIP 2023, ICMI 2023 and Emotion Recognition in the Wild challenge. He is an Associate Editor of IEEE Transactions on Multimedia.

Shreya Ghosh is a Lecturer (Assistant Professor) at The University of Queensland, Australia. She pursued her postdoc and PhD at Monash University, Australia. Her research interests are in Affective Computing and deep learning. Over past few years, she has been working on deepfake. She is an Associate Editor of IEEE Transactions on Affective Computing.

Muhammad Haris Khan is an Assistant Professor at MBZUAI, UAE. His research interests include domain adaptation, domain generalization, few-shot / zero-shot learning, active learning. He has served as an Area Chair at CVPR 2024-25, WACV 2024-25, ICML2025, Neurips 2024 and BMVC 2024. He is acting as Assoicate Editor at IET Computer Vision journal and a regular program committee member at top conferences. He is an organizer of workshop at ACCV 2022, a competition at ACM MM Grand Challenge 2024, a special issue at IJCV, and a workshop at CVPR 2025.

Usman Tariq is an Associate Professor of Electrical Engineering at the American University of Sharjah. He received his PhD from the University of Illinois at Urbana Champaign. His research interests are in computer vision and deepfakes analysis.

References

[1] Gupta, Parul, Shreya Ghosh, Tom Gedeon, Thanh-Toan Do, and Abhinav Dhall. "Multiverse Through Deepfakes: The MultiFakeVerse Dataset of Person-Centric Visual and Conceptual Manipulations." In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13258-13265. 2025.

[2] Cai, Zhixi, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset." In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7414-7423. 2024.

[3] Cai, Zhixi, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. "Av-deepfake1m++: A large-scale audio-visual deepfake benchmark with real-world perturbations." In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13686-13691. 2025.

[4] Claire Wardle. 2019. The Disturbing World of Deepfake Pornography. WIRED (October 2019). https://www.wired.com/story/deepfakes-pornography Accessed: 2024-05-30.

[5] Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The spread of true and false news online." science 359.6380 (2018): 1146-1151.

[6] Pal, Anisha, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. "Semi-Truths: A Large-Scale Dataset of AI-Augmented Images for Evaluating Robustness of AI-Generated Image detectors." Advances in Neural Information Processing Systems 37 (2024): 118025-118051.

[7] Huang, Zhenglin, 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." In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 28831-28841. 2025.