Synthetic Images at MediaEval 2026: Advancing detection of generative AI in real-world online images

Olga Papadopoulou[1,*,†], Dimitrios Karageorgiou[1,3], Christos Koutlis[1], Efstratios Gavves[3], Hannes Mareen[4] and Symeon Papadopoulos[1]

[1]*Information Technologies Institute (ITI) @ CERTH, Greece*
[3]*University of Amsterdam, The Netherlands*
[4]*IDLab, Ghent University — imec, Belgium*

Abstract

The task "Synthetic Images: Advancing Detection of Generative AI in Real-World Online Images" at MediaEval 2026 continues the 2025 edition, focusing on AI methods for (i) synthetic image detection and (ii) manipulated region localization. The same tasks are retained to enable broader participation and benchmarking. Methods are evaluated under real-world transformations such as compression, resizing, and cropping. While prior results show strong performance in constrained settings, generalizable detection and precise localization—especially for fully AI-generated images—remain challenging. The 2026 edition also emphasizes "insight-driven" participation, encouraging Quest for Insight papers that analyze dataset characteristics, methodological strengths and weaknesses, and evaluation protocols.

1. Introduction

Synthetic media generation has rapidly advanced with the widespread adoption of powerful generative models such as diffusion and multimodal foundation models, enabling the creation of highly realistic and diverse visual content at scale [1]. While these technologies benefit creative industries and digital communication, they also amplify risks related to misinformation, manipulation, and erosion of trust in visual media. As a result, robust synthetic image detection has become critical for safeguarding information integrity and supporting trustworthy AI ecosystems.

Recent approaches to synthetic image detection have demonstrated promising performance under controlled conditions [2, 3, 4]; however, their effectiveness often degrades in the wild [5, 6, 7]. The continuous evolution of generative models, combined with post-processing operations such as compression, resizing, and platform-specific transformations, significantly obscures forensic traces. Moreover, modern generative tools enable not only fully synthetic image creation but also fine-grained, text-guided manipulations (e.g., inpainting and object insertion), further blurring the line between authentic and manipulated content [8, 9, 10].

A key challenge is generalization [11, 12]: models trained on specific datasets or generation techniques frequently fail to transfer effectively to unseen models, domains, or real-world distributions. This limitation is exacerbated by the gap between curated training datasets and in-the-wild data [6, 13], where variations in quality, diversity, and content characteristics play a crucial role in detection performance. Despite encouraging progress, current methods still exhibit limited robustness and struggle to maintain consistent performance across different manipulation types and realistic conditions.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
olgapapa@iti.gr (O. Papadopoulou); dkarageo@iti.gr (D. Karageorgiou); ckoutlis@iti.gr (C. Koutlis); e.gavves@uva.nl (E. Gavves); hannes.mareen@ugent.be (H. Mareen); papadop@iti.gr (S. Papadopoulos)

In this context, the 2026 edition of the task continues to explore both image-level classification and manipulation localization, reflecting the complex nature of synthetic media detection. The challenge aims to foster the development of methods that go beyond benchmark performance, encouraging solutions that are robust, generalizable, and capable of providing meaningful insights into the evolving landscape of AI-generated media detection.

2. Task Description

The task is organized into two subtasks, following the same structure as in the 2025 edition [14].

2.1. Subtask A: Real vs. Synthetic Image Detection

The objective of this subtask is to develop methods that classify whether a given image is real or synthetic. To ensure direct comparability with last year's results, the same training, validation, and evaluation datasets are reused. Maintaining identical data splits enables consistent benchmarking across editions and supports longitudinal analysis of progress in synthetic image detection.

Participants are required to submit at least one constrained run and one open run. The constrained run must rely exclusively on the provided training data, while the open run may incorporate additional external data sources, excluding the evaluation data. In addition to reporting results, participants are expected to provide a comprehensive analysis comparing these runs. This analysis should examine how variations in training data, model design, and data curation strategies impact system behaviour and performance.

The dataset comprises three parts: a training set (5,000 real and 5,000 synthetic in-the-wild images [15, 4]), a validation set for development (10,000 balanced labelled real/synthetic images), and a test set of 10,000 images without labels, used by the organizers for the final evaluation.

Evaluation metrics. For evaluating synthetic image detection, we use the SIDBench framework [16], which provides a comprehensive set of metrics to assess model robustness in real-world scenarios. While multiple indicators are reported (Accuracy, Precision, Recall, F1-Score, AUC, AP, and EER), F1-Score is used as the main metric for ranking, as it best reflects the balance between false alarms and missed detections, which is crucial for reliable deployment.

2.2. Subtask B: Manipulated Region Localization

Subtask B focuses on the localization of AI-generated image manipulations, covering both spliced and fully-regenerated images. With the rapid progress of generative AI, modern tools enable complex, highly realistic image edits with simple text prompts. These methods support a wide range of manipulations, from subtle object insertions or replacements to large-scale modifications and full scene regeneration, often preserving strong semantic coherence with the original content.

The 2026 edition follows the same task design as in 2025. To support continuity while remaining up-to-date with recent generative models, the 2025 evaluation set [8] is incorporated into the 2026 training and validation data [9, 10]. A new evaluation set has been created for 2026 following the same generation protocol, updated to include content produced by more recent generators. This setup enables fair comparison across editions while reflecting the evolving capabilities of state-of-the-art generative systems.

Evaluation metrics. For subtask B, we report performance for both image-level detection and manipulated region localization.

Image-level detection is evaluated using the F1 score and the Area Under the ROC Curve (AUC), which respectively measure threshold-dependent classification performance and threshold-agnostic discriminative ability of a detector. For localization, the Intersection over Union (IoU) metric is employed to evaluate the overlap between predicted and ground-truth manipulated regions. We compute the IoU both with the predicted mask and with its inverted version, and then pick the higher of the two. This adjustment accounts for some localization methods that correctly separate the two areas, but misclassify the manipulated and pristine areas. Together, these metrics enable comprehensive evaluation of approaches' ability to both detect and precisely localize AI-generated manipulations.

Details about the datasets of both subtasks can be found in the 2025 edition of the task [14], as well as in the official task repository available at: https://github.com/mever-team/mediaeval2026-sid.

2.3. Quest for Insights

Beyond quantitative evaluation, the task places strong emphasis on insight generation. Participants are encouraged to move beyond reporting performance metrics and to critically analyse the behaviour of their methods, the characteristics of the data, and the broader challenges of synthetic image detection in practical settings. Understanding why systems succeed or fail is essential for advancing the field, particularly given the rapidly evolving nature of generative models and the persistent gap between curated datasets and in-the-wild content.

As a starting point, participants are invited to reflect on the following research questions:

- Are the synthetic images or manipulated regions that are difficult for automatic methods also challenging for human observers to identify?
- What are the characteristics of false positives (i.e., real images or regions misclassified as synthetic)? Can common patterns be identified in terms of visual quality, artifacts, or semantic content?
- Are there specific characteristics (e.g., semantic content, context, framing) that distinguish in-the-wild examples from curated datasets?
- How does the choice of training data influence model performance and behaviour?
- What are the benefits of incorporating additional or newly collected datasets (e.g., from fact-checking platforms or social media)? Which types of content remain underrepresented?
- To what extent can collected datasets approximate the distribution of real-world synthetic media, and what strategies can help reduce the gap between curated training data and in-the-wild content?
- Do the proposed methods exhibit biases or fairness issues (e.g., across different content types, demographic attributes, geographic contexts, or image sources)? Are certain groups, scenes, or styles more prone to misclassification, and what factors may contribute to these disparities?

Findings from both quantitative and qualitative analyses contribute to a deeper understanding of the limitations and opportunities in synthetic media detection.

3. Results and Analysis

The approaches submitted in the 2025 edition of the Synthetic Image Detection (SID) task were primarily based on deep learning architectures, with most methods relying on Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Common backbone models included widely used architectures such as EfficientNet, ResNet, and U-Net, typically employed as pretrained models and adapted through transfer learning or fine-tuning.

In addition to these standard approaches, several teams explored more original directions. These included contrastive learning and knowledge distillation to enhance generalization, multi-channel architectures designed to capture complementary features, and methods leveraging frequency-domain representations or noise residuals to better expose artifacts introduced by generative models.

For Task A, the approaches proposed in [17, 18] outperform the baseline methods BFree [11] and RINE [4], establishing a new performance baseline exceeding 0.8 F1 score for the constrained run. In the open run setting, the approaches in [19, 18, 20] further improve upon the RINE-TWIGMA baseline, increasing the F1 score from 0.8 to 0.89. These results establish updated reference baselines for the 2026 SID task. Table 1 summarises the results for the constrained and open run settings, reporting only methods that outperform the baselines.

Task B remained highly challenging, attracting limited participation and yielding relatively low performance overall. Only one team exceeded [21] the DeCLIP baseline [22], while none reached the performance of the TruFor baseline [23], with the majority of approaches falling below both. Table 2 reports only methods surpassing the baselines. A notable insight is that several teams cited limited computational and development resources as a primary reason for not participating in this subtask.

Table 1
Performance of teams for constrained and open runs (F1-score). Dashes (–) indicate that no experiment was conducted for the corresponding setting. Only teams outperforming the baselines are reported.

Team	Approach	Constrained F1	Open F1
MICC-UNIFI [19]	CLIP for feature extraction	0.05	**0.89**
SYN-CHK [17]	ViT, EfficientNet, ResNet & VGG19	**0.86**	0.66
CodingSoft-REC [18]	EfficientNet-B0 (Constrained), EfficientNet-B4 with CIFAKE dataset (Open)	0.85	0.83
CVG-IBA [20]	ResNet-based architectures	0.42	0.81
RINE-TWIGMA (baseline)	Trained on TWIGMA dataset [24]	–	0.80
BFree (baseline) [11]	Bias-free training paradigm with diffusion images	–	0.74
RINE (baseline) [4]	Intermediate ViT (CLIP) layers, fine-tuned on provided training data	0.67	–

Table 2
Performance of teams for the manipulated region localization task (IoU). Only teams outperforming the baselines are reported.

Team	Approach	IoU
TruFor (baseline) [23]	Noise map (Noiseprint++), non-finetuned official model	**0.627**
HCMUS-Aqua [21]	FR classification: LoRA on CLIP-ViT; Localization: Frequency-based Edge Detector & SegFormer	0.515
DeCLIP (baseline) [22]	CLIP-based features, non-finetuned official model	0.470

4. Acknowledgments

The task organization is supported by the Horizon Europe AI-CODE project (Grant Agreement no. 101135437) that focuses on the development of AI tools for supporting media professionals in their verification and fact-checking activities.

Declaration on Generative AI

ChatGPT 5 was used for Grammar and spelling checks of this paper. The authors reviewed and edited the resulting content as needed and take full responsibility for the publication's content.

References

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. Yu, L. Sun, A survey of ai-generated content (aigc), ACM Computing Surveys 57 (2025) 1–38.

[2] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, E. Gavves, Any-resolution ai-generated image detection by spectral learning, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18706–18717.

[3] D. Cozzolino, G. Poggi, M. Nießner, L. Verdoliva, Zero-shot detection of ai-generated images, in: European conference on computer vision, Springer, 2024, pp. 54–72.

[4] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: European Conference on Computer Vision (ECCV), Springer, 2024, pp. 394–411.

[5] D. Karageogiou, Q. Bammey, V. Porcellini, B. Goupil, D. Teyssou, S. Papadopoulos, Evolution of detection performance throughout the online lifespan of synthetic images, in: European Conference on Computer Vision (ECCV) Workshops, Springer, 2024, pp. 400–417.

[6] D. Konstantinidou, D. Karageorgiou, C. Koutlis, O. Papadopoulou, E. Schinas, S. Papadopoulos, Navigating the challenges of ai-generated image detection in the wild: What truly matters?, in: 5th ACM International Workshop on Multimedia AI Against Disinformation (MAD'26), 2026.

[7] C. Li, X. Wang, M. Li, B. Miao, P. Sun, Y. Zhang, X. Ji, Y. Zhu, Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 20379–20389.

[8] P. Giakoumoglou, D. Karageorgiou, S. Papadopoulos, P. C. Petrantonakis, SAGI: Semantically aligned and uncertainty guided ai image inpainting, arXiv preprint arXiv:2502.06593 (2025).

[9] H. Mareen, D. Karageorgiou, G. Van Wallendael, P. Lambert, S. Papadopoulos, TGIF: Text-guided inpainting forgery dataset, in: Int. Workshop on Information Forensics and Security (WIFS), 2024.

[10] H. Mareen, D. Karageorgiou, P. Giakoumoglou, P. Lambert, S. Papadopoulos, G. Van Wallendael, TGIF2: Extended text-guided inpainting forgery dataset & benchmark, Journal on Information Security (2026).

[11] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A Bias-Free Training Paradigm for More General AI-generated Image Detection, in: IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 18685–18694.

[12] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, S. Hu, Detecting multimedia generated by large ai models: A survey, arXiv preprint arXiv:2402.00045 (2024).

[13] T. Pantsios, D. Karageorgiou, C. Koutlis, G. Karantaidis, O. Papadopoulou, S. Papadopoulos, Automated in-the-wild data collection for continual ai-generated image detection, in: 5th ACM International Workshop on Multimedia AI Against Disinformation (MAD'26), 2026.

[14] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos, Synthetic images at mediaeval 2025: Advancing detection of generative ai in real-world online images, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025, pp. 25–26.

[15] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic

images generated by diffusion models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
[16] M. Schinas, S. Papadopoulos, SIDBench: A python framework for reliably assessing synthetic image detection methods, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, 2024, pp. 55–64.
[17] N. ul Huda, A. Fayyaz, U. Asad, Y. S. Afridi, Synthetic vs real image detection using vision transformers and cnn-based architectures, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[18] S. M. T. Mariappan, M. Ramasamy, B. Arul, An efficientnet framework: Methods and results for synthetic image detection and manipulation localization, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[19] Q. Li, A. Ciamarra, R. Caldelli, S. Berretti, A clip-based approach for synthetic image detection under distribution shift, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[20] H. M. O. Raza, M. A. Tahir, R. A. Khan, Real versus synthetic classification using resnet, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[21] M.-H. Le, M.-K. Le-Phan, K.-N. Vu-Nguyen, M.-T. Tran, T.-L. Do, A resolution-agnostic three-stage framework for image forgery detection and localization, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
[22] S. Smeu, E. Oneata, D. Oneata, DeCLIP: Decoding clip representations for deepfake localization, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 149–159.
[23] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20606–20615.
[24] Y. Chen, J. Y. Zou, TWIGMA: A dataset of AI-Generated Images with Metadata From Twitter, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023, pp. 37748–37760.