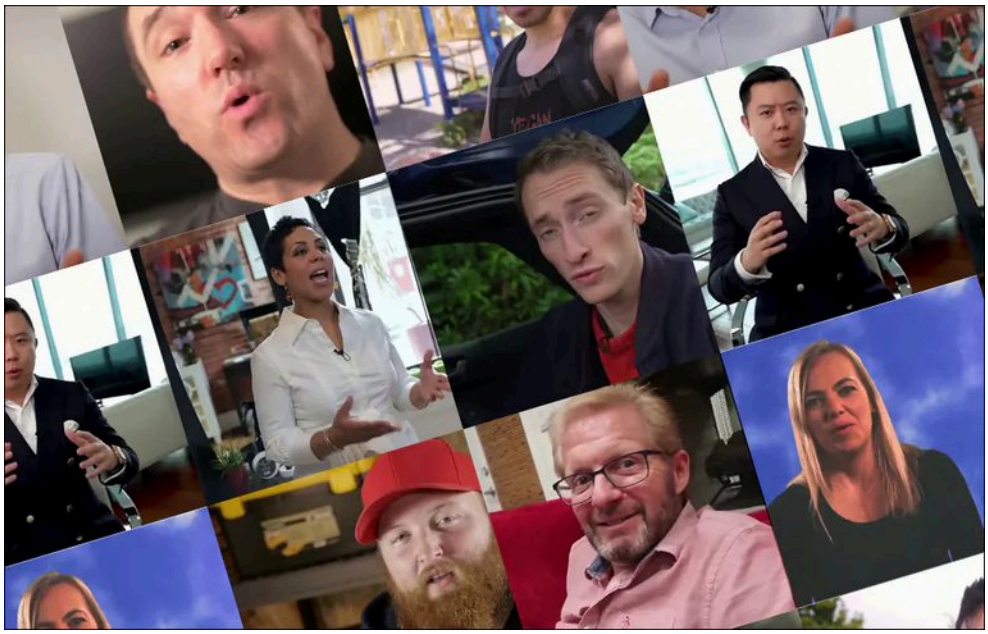
Deepfake Detection Accuracy Falls 49% in a Year of AI Video Progress

Originally published 14 September 2026, at https://www.unite.ai/deepfake-detection-accuracy-falls-49-in-a-year-of-ai-video-progress/



I note with interest a new Ukraine-led paper indicating that the older generation of deepfake video detectors perform very poorly on current state-of-the-art genAI video, whether those generators are open source, local AI installs such as Alibaba's Wan 2.2 or Hunyuan Video; or API-only, closed-source models such as Grok Imagine or Kling.

CLICK TO PLAY VIDEO ON SITE

Click to play (AUDIO CONTENT). *Three examples from the 'Direct to camera casual' category in the DF26 dataset.* Source

Back in 2024, another paper had already established that human ability to detect AI video was not that far above chance, at 57%. What the new work (titled *DF26: We Cannot Tell Fake From Real Anymore*) adds, is empirical evidence that automated deepfake detectors have fallen dramatically in efficacy since then – from 94% to 48%, when faced with the very latest genAI video output.

From the new dataset generated for the work, examples of image-to-video and text-to-video generated videos (below), compared to real videos on the left. The corpus concentrates on 'persuasive' and credible head-and-shoulders setups, which are among the most concerning genAI domains at the moment. Source

The steepest decline in the tests undertaken for the work represents a drop of 49%. Since the earlier benchmark hails from 2025, and the new work tests a range of the latest 2026-generation video models, this figure represents nearly a year of deterioration in AI-video detection efficacy.

Generator	Creator	Released	Elo
Closed-source			
Wan 2.6	Alibaba	Dec 2025	1,185
Veo 3.1	Google	Jan 2026	1,213
Grok Imagine 1.0	SpaceXAI	Jan 2026	1,221
Kling 3.0	KlingAI	Feb 2026	1,212
Open-source			
Wan 2.2 A14B	Alibaba	Jul 2025	1,106
HunyuanVideo 1.5	Tencent	Nov 2025	1,016
LTX 2.3	Lightricks	Mar 2026	1,120

Details of the models used to generate the videos against which the detectors were tested.

The closed-source, commercial models used to generate this detector-defeating output were Grok 1.0; Kling 3.0; Veo 3.1; and Wan 2.6. Only text-to-video examples were generated for these, since using attempting image-to-video (I2V) generations often triggered filters in regard to 'deepfake generation', or else would have violated the platform's TOS.

Look Into My Eyes

These triggers occurred, likely, because the researchers concentrated on the most potentially powerful, and arguably sinister form of video-based persuasion – diverse variations on 'head and shoulder' and 'talking head' videos, with subjects talking to camera, as they would in YouTube and TikTok influencer videos, or in a news reporting scenario (including interviewees over satellite connections, etc.).

The authors contend that these scenarios are prime targets for deepfake activity, not least because they represent 'influencer' and various other 'informational' contexts – contexts where an assumption of a prior relationship of trust exists, and where that relationship therefore has great potential for abuse (naturally this applies at least equally to deepfake video calls, though the researchers do not deal with this).

From the zenith of the autoencoder deepfake age, manipulation of real Richard Nixon footage was able to simulate his announcement of the death of the Apollo 9 astronauts in 1969 – an event that never arose, but for which the real-world script was used to impose a deepfaked voice and lip movements.

Besides focusing on this influencer scenario, the new benchmark – and the videos created for it – concentrate on evaluating the *entire frame* of the video, in contrast to historical deepfake detection strategies.

Those older strategies enacted facial or expression manipulation, or facial (or at best, full head) substitution within real videos – which was indeed the only reasonable approach, until diffusion-based models became effective enough to supplant them over the last 18-24 months.

Face Away

That's not how its done any more; the older autoencoder-based method dates back to the initial advent of deepfakes in late 2017; and this technique, which replaces only the area within the outer lineaments of the face, was completely exhausted by 2022.

The prime real estate for deepfake activity, 2017-2023 – but for the new generation of genAI video platforms, there's no such boundary to concentrate the attentior

Nonetheless, since the scientific research sector loves a long-term constant – where the target is fixed, and where an evolving range of approaches can be thrown against it over the years/decades – deepfake detection methods based on 'inner-face substitution/amendment' have persisted in the literature, and in commercial detection offerings, far beyond the validity of the underlying technology being targeted.

Besides the three closed-source models tested for the new work, three very capable consumer-usable open source models were also trialed: Wan 2.2 A14B; HunyuanVideo 1.5; and LTX 2.3, which has been dazzling the FOSS genAI community lately. The LTX range is also capable of generating speech:

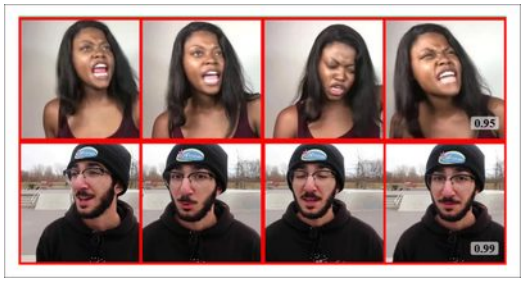
CLICK TO PLAY VIDEO ON SITE

***Click to play**: experiments with the LTX series at r/stablediffusion, whose community deals only with FOSS models.* Source

Data Design

The authors intend their new dataset, DF26, as a 'hold-out' set for evaluation of unseen material. The collection consists of 2,691 videos across three public-speaking scenarios: direct-to-camera or casual addresses; official statements; and studio interviews.

All the AI-generated video clips were based on 271 real-world clips curated from OpenVid; TalkingCelebs; and MAVOS-DD.



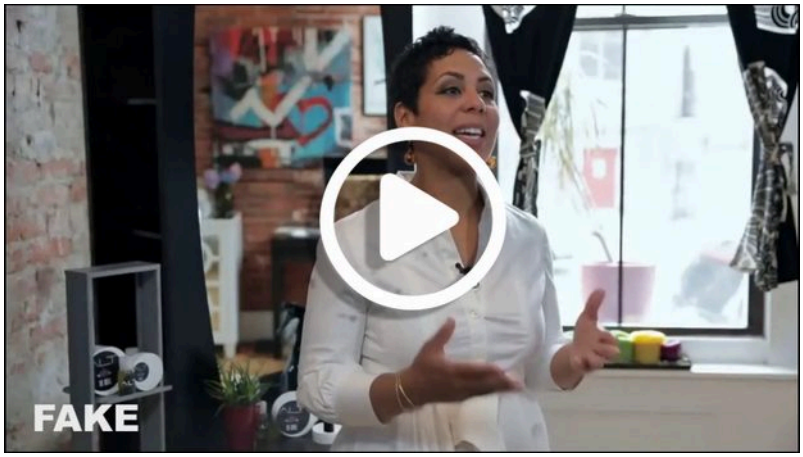
Sample from the OpenVid collection, from which 271 real-world videos were sourced as root material for the AI-generated videos in the DF26 collection. Two other datasets also contributed to this non-fake component of DF26. Source

The 271 real videos were used to generate the AI videos by deriving matched scene prompts from their frames, then feeding those prompts – and the first frame itself, where supported – into the seven video-generation models, to produce a total of 2,420 synthetic clips.

The authors ensured that no real-world frames featuring text overlays (such as announcements as to who is talking, etc.) were allowed into the generation workflows, since these would likely have encouraged shortcuts and assumptions. Candidate clips were evaluated by Gemini 2.5.

Additionally, segments featuring more than one face were rejected, to enforce a 'single speaker' setting.

CONTINUED ON NEXT PAGE



CLICK TO PLAY VIDEO ON SITE

Click to play [NO SOUND]. *From the dataset associated with the paper, examples of state-of-the-art video generations across seven of the latest and most capable generative models, both open and closed source. The models used for the dataset include those also capable of generating speech (see examples earlier in this article).* Source

The researchers used the popular commercial Higgsfield AI platform, which makes available a wide variety of closed and open source models, with diverse levels of gatekeeping and guardrails.

With a NVIDIA H200 (141GB of VRAM) as a base GPU model for an internal research cluster, it took about 440 GPU hours to generate 1,626 videos for the collection.

Besides ensuring that no text was visible in the output, it was also necessary to obfuscate or in general avoid the depiction of AI watermarking, since this too would represent a potential 'shortcut' to an evaluator or evaluation system.

Tests

The primary metric used for a closing round of tests was Area Under Receiver Operating Characteristic Curve (AUC / AUROC), with Equal Error Rate (EER) as a complementary measure.

Frame-based detectors (which assess still frames independently) were tested on 32 evenly spaced frames from each video, with the scores then averaged into a single result. Temporal detectors (which can use information from successive frames) instead analyzed the video sequence itself.

Results for the temporal detectors DFD-FCG and PwTF-DVD can be seen below, together with results for ForAda; Effort; FSFM; GenD-CLIP and GenD-DINO; and DFD-HR. All were trained on the venerable FaceForensics++ dataset.

Detector	Venue	Type	CDFv3		DF26		DF26 – CDFv3	
			AUROC ↑	EER ↓	AUROC ↑	EER ↓	ΔAUROC	ΔEER
DFD-FCG [10]	CVPR'25	Temporal	**94.3**	**12.3**	48.2	51.9	−46.1	+39.6
PwTF-DVD [12]	ICCV'25	Temporal	92.3	15.7	61.6	40.8	−30.7	+25.1
ForAda [5]	CVPR'25	Frame	75.6	32.1	56.6	45.8	−19.0	**+13.7**
Effort [27]	ICML'25	Frame	78.7	29.8	44.0	55.7	−34.7	+25.9
FSFM [24]	CVPR'25	Frame	79.6	27.5	52.6	48.1	−27.0	+20.6
GenD-CLIP [28]	WACV'26	Frame	85.2	23.0	54.2	47.7	−31.0	+24.7
GenD-PE [28]	WACV'26	Frame	89.9	16.9	**69.7**	**35.9**	−20.2	+19.0
GenD-DINO [28]	WACV'26	Frame	82.6	24.5	54.7	47.3	−27.9	+22.8
DFD-HR [21]	CVPR'26	Frame	81.6	27.0	63.5	40.7	**−18.1**	**+13.7**

Test results comparing deepfake detectors on CelebDF++ and DF26. All detectors were trained on FaceForensics++, with higher AUROC and lower EER indicating better performance.

The authors note that most of the detectors perform well on Celeb-DF++ (CDFv3), but fall sharply on the more challenging new DF26 collection.

> *'[Multiple] state-of-the-art detectors achieve strong performance on the CelebDF++ [16] (CDFv3) benchmark, with temporal methods reaching an AUROC of 94.3 and 92.3.*
>
> *'However, their performance drops substantially on DF26, to 48.2 and 61.6 AUROC, respectively.*
>
> *'Most methods degrade substantially, remaining near chance; the highest AUROC of 69.7 is achieved by GenD-PE.*
>
> *'This shows that DF26 is a more challenging benchmark for state-of-the-art detectors.'*

Image-to-video, they observe, proved more difficult to detect than text-to-video, for both people and automated detectors, with PwTF-DVD performing best on I2V material, and GenD-PE leading on T2V.

Performance varied dramatically depending on which generator produced the fake video, suggesting that detectors were often learning generator-specific traces rather than a general signature of synthetic video. PwTF-DVD, for example, reached 92.9 AUROC on text-to-video output from Wan 2.2, but fell to roughly chance performance on HunyuanVideo 1.5 – even though both belonged to the same modern generation landscape:

Detector	Commercial generators				Open-source generators						Mean
	Grok 1.0	Kling 3.0	Veo 3.1	Wan 2.6	HV 1.5		Wan 2.2		LTX 2.3		
	T2V	T2V	T2V	T2V	I2V	T2V	I2V	T2V	I2V	T2V	
DFD-FCG [10]	20.5	27.4	29.6	60.3	41.1	76.5	61.3	58.4	61.1	45.9	48.2
PwTF-DVD [12]	34.5	60.4	26.7	71.6	41.9	56.7	**87.7**	**92.9**	79.5	64.0	61.6
ForAda [5]	38.9	53.4	39.9	63.8	50.8	78.0	62.6	48.4	70.8	57.7	56.4
Effort [27]	13.6	32.3	21.3	29.0	44.3	68.8	61.7	30.4	61.6	45.5	40.9
FSFM [24]	36.3	49.7	38.2	69.2	46.3	75.4	62.7	55.0	65.2	41.3	53.9
GenD-CLIP [28]	47.7	62.5	39.6	62.6	52.2	73.9	66.5	35.5	72.2	43.5	55.6
GenD-PE [28]	**66.5**	71.0	**76.6**	**88.3**	**64.8**	82.7	66.1	58.6	81.5	**71.4**	**72.7**
GenD-DINO [28]	35.4	70.5	42.4	76.6	55.6	73.0	63.0	34.7	72.0	46.5	57.0
DFD-HR [21]	43.9	**72.1**	49.9	64.4	61.3	**84.5**	73.1	48.3	**83.8**	56.5	63.8

Detector performance across commercial and open-source video generators. The wide variation between columns shows that detection reliability depends strongly on which model produced the video, with GenD-PE achieving the highest overall mean AUROC.

Below we see that retraining GenD-PE on newer DF26 material substantially improved its ability to detect videos from generators that it had not seen during training:

Train set	Commercial generators				Open-source generators						Mean
	Grok 1.0	Kling 3.0	Veo 3.1	Wan 2.6	HV 1.5		Wan 2.2		LTX 2.3		
	T2V	T2V	T2V	T2V	I2V	T2V	I2V	T2V	I2V	T2V	
FF++	66.5	71.0	76.6	88.3	68.5	**88.7**	69.1	65.2	**86.5**	79.8	76.0
HV I2V	**96.8**	**69.9**	**95.6**	**93.1**	**83.2**	85.3	65.8	84.1	67.0	82.4	**82.3**
Wan I2V	88.5	50.6	80.5	87.8	67.3	78.1	**78.4**	**83.6**	70.6	80.3	76.6
LTX I2V	74.7	65.5	73.0	74.3	50.6	80.3	60.4	71.3	78.5	**94.0**	72.3

Retraining results for GenD-PE across unseen video generators. Training on newer DF26 material generally improved cross-generator performance compared with the original FaceForensics++ training.

Training on HunyuanVideo 1.5 raised AUROC to at least 93.1 on Grok Imagine 1.0, Veo 3.1 and Wan 2.6, supporting the paper's argument that detectors trained only on older datasets such as FaceForensics++ are poorly matched to current generative video.

Perhaps predictably, the two temporal detectors had a much easier time with open-source video than with clips from commercial models. DFD-FCG dropped from 57.4 AUROC on open-source material, compared to 34.5 on closed-source video; while PwTF-DVD fell from 70.5 to 48.3:

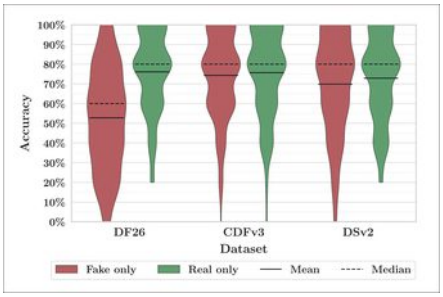
Group	Subset	DFD-FCG [10]		PwTF-DVD [12]	
		AUROC ↑	EER ↓	AUROC ↑	EER ↓
Generator source type					
Source	Open-source	57.4	45.2	70.5	33.8
Source	Closed-source	34.5	61.7	48.3	51.3
Source	**Closed − Open**	−22.9	+16.5	−22.2	+17.5
Semantic class					
Semantic	Direct-to-camera	50.6	50.1	60.1	43.9
Semantic	Official statement	51.7	53.2	72.3	34.9
Semantic	Studio interview	50.5	48.1	66.1	39.1

Results comparing two temporal detectors across generator source and speaking scenario. Both performed substantially worse on closed-source video, while differences between scene types were smaller.

However, the paper stops short of declaring that commercial models are simply harder to detect, since differences in model family, post-processing and visual quality could also be factors.

The *kind* of scene mattered much less: DFD-FCG stayed close to chance across direct-to-camera clips, official statements and studio interviews, while PwTF-DVD did best on official statements. According to the paper, the generator itself seems to matter more than the presentation style.

A human study was also run to see whether people struggled with DF26 in the same way as the automated detectors. Across 232 labeling sessions, participants judged short clips from DF26, CelebDF++ and DeepSpeak v2 as real or fake, without being told how many examples of each class they would see.



Human accuracy across DF26, CelebDF++ and DeepSpeak v2. Participants identified real videos at similar rates across all three datasets, while accuracy on fake DF26 videos fell close to chance.

Performance on real videos was similar across all three datasets, at 76.0% for DF26, 75.5% for CelebDF++ and 72.8% for DeepSpeak v2. The difference appeared with the fake videos, where accuracy fell to 52.6% on DF26, compared with 74.5% and 69.8% on the two older datasets.

Therefore, according to the paper, though participants were not generally confused by DF26, they struggled to find *specific* visible evidence that its synthetic videos were fake.

Conclusion

Despite a reported 16x increase in deepfake frequency over the last two years, the actual reality of maleficent deepfakes is largely absent from our common experience, unless we are ourselves targeted by them.

In the case of victims of sexual deepfakes, this is entirely understandable – but since deepfakes now have an increasing impact on the crime of fraud, it might be useful if more of that material could be shared with the public, so that our context is updated from the 'golden age' of autoencoder deepfakes into the far more incisive and deceptive era of diffusion-based deepfakes.

Of course, most material relating to output from the models in the new study, as well as other models, surfaces without adequate context in social media platforms whose overseers either may not be able to distinguish AI from real, or who just don't care.

One additional metric that we can all apply to videos that we may doubt is reasonable credulity – but this faculty varies so much across individuals as to be unreliable. It therefore may be that for a transitional period, as we acclimatize to the impact of this technology and its ever-increasing capabilities, we simply must defer judgement.

First published Monday, September 14, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai