SynthForensics

Benchmarking and Evaluating People-Centric Synthetic Video Deepfakes

Roberto Leotta[1,*] Salvatore Alfio Sambataro[2,*]
Claudio Vittorio Ragaglia[2] Mirko Casu[2] Yuri Petralia[1]
Francesco Guarnera[2] Luca Guarnera[2] Sebastiano Battiato[2]

[1]iCTLab s.r.l. [2]University of Catania [*]Equal contribution

40th Conference on Neural Information Processing Systems

NeurIPS 2026

Track on Evaluations and Datasets

Synthetic people and the evaluation gap

State-of-the-art open-source T2V/I2V generators synthesize people increasingly hard to distinguish from authentic footage. Unlike manipulation-based deepfakes, which locally alter existing recordings, purely synthetic generators can fabricate entirely novel individuals from a text prompt or animate a single photograph of a real person through I2V conditioning. This dramatically lowers the barrier for identity fraud, non-consensual content, disinformation, and impersonation; threats that fundamentally require realistic human depiction to succeed.

Current evaluation suites lag: legacy benchmarks such as FF++ [1] target manipulation-based forgeries, and recent synthetic-video benchmarks prioritize scale over realistic human depiction.

We introduce **SynthForensics**, a people-centric benchmark of 20,445 videos from 8 T2V and 7 I2V open-source generators, paired-source from FF++ and DFD [2] reals, two-stage human-validated, in four compression versions with full generation metadata.



T2V: new synthetic person, same scene and context. **I2V**: same real person, synthetically re-animated.

[1] A. Rössler et al., FaceForensics++, ICCV 2019. [2] N. Dufour, A. Gully, Contributing Data to Deepfake Detection Research, Google AI Blog, 2019.

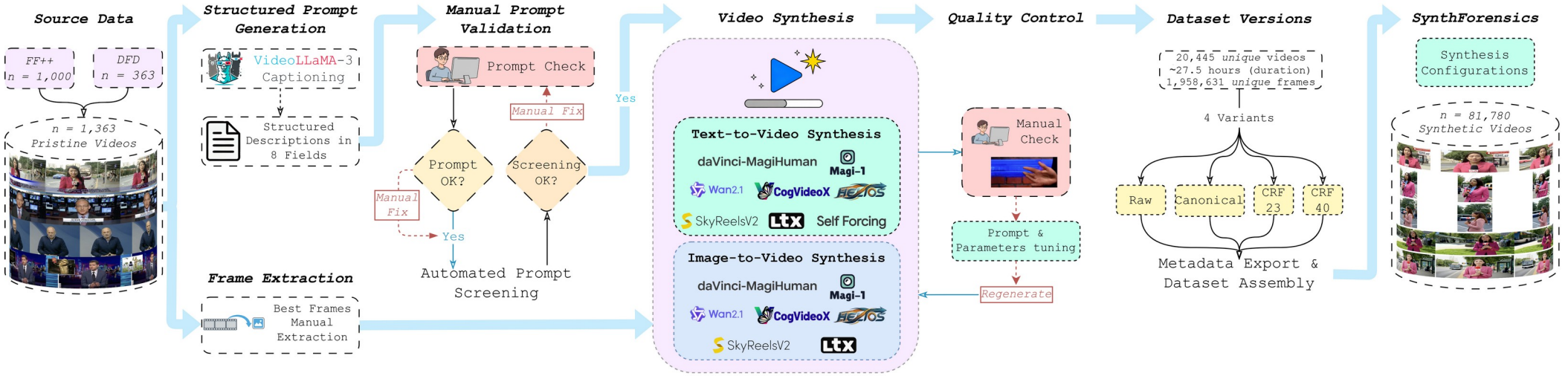
Paired-source protocol

SynthForensics employs a paired-source protocol using 1,363 pristine videos from FF++ (1,000) and DFD (363). Each pristine video is captioned and the resulting caption is used as the positive prompt to anchor synthetic counterparts in both T2V and I2V modes, controlling high-level semantic variables (scene, subjects, actions) to isolate low-level generative artifacts from semantic confounds. T2V videos are generated from the caption alone and their subjects are entirely synthetic, whereas I2V videos also receive a reference frame of the source video and animate its real subject.



Frames of three real sources (centre) and of their T2V (left) and I2V (right) counterparts, each labelled with its generator. The eight generators are Wan2.1, CogVideoX, SkyReels-V2, Self-Forcing, MAGI-1, LTX-2.3, Helios and daVinci-MagiHuman; all but Self-Forcing are also used in I2V mode.

Construction pipeline



Prompts and reference frames. VideoLLaMA 3 [3] describes each source in eight structured fields, from which model-specific positive and negative prompts are built. For I2V, human annotators select a reference frame per source, prioritizing a large, centered and sharp face in a neutral, frontal pose.

Caption validation. Human annotators validated every caption field by field, for consistency with the source video and for sensitive or harmful content. An LLM [4] then flagged references to violence and warfare, weapons, vulnerable individuals, political figures and institutions, geopolitical references, national symbols and identifiable real persons; flagged captions were rewritten by the annotators.

Synthesis and quality control. Generators run with their authors' settings, refined per generator and per video. Human annotators blind to the generator inspected every video for anatomical, rendering and temporal artifacts, semantic coherence with the prompt and ethical compliance; rejected videos were regenerated with refined prompts and parameters.

[3] B. Zhang et al., VideoLLaMA 3, arXiv:2501.13106, 2025. [4] A. Grattafiori et al., The Llama 3 Herd of Models, arXiv:2407.21783, 2024.

Dataset statistics and release

Metric	Value
Unique videos (T2V)	10,904
Unique videos (I2V)	9,541
Total unique synthetic videos	20,445
Total video files (4 compression versions)	81,780
Total unique frames	1,958,631
Total unique duration	~27.5 hours
Resolution range	384×640 to 1088×1920
Frame rate range (FPS)	8 to 25
Duration range	4 to 6 s
Landscape / Portrait videos	16,349 / 4,096

Compression versions. We provide four versions of each video for robustness analysis: the Raw generator output; a Canonical version, re-encoded with uniform parameters (H.264 CRF 0, YUV420p, BT.709) to neutralize format-specific confounds; and its progressively compressed CRF23 and CRF40 versions, which simulate real-world distribution.

Identities. The benchmark depicts 11,815 individuals: 10,904 novel synthetic identities created by T2V and 911 real people animated by I2V.

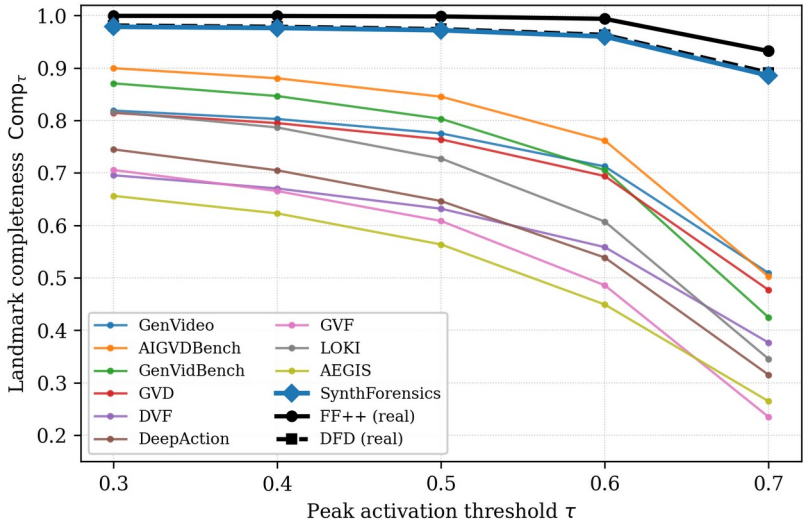
Released metadata. For each video we release a JSON sidecar with the generator identification (model, version, Hugging Face commit, weight checksums), the source-data references (structured caption and, for I2V, the human-selected reference frame), the prompts as fed to the model and the seed, the complete inference configuration and the environment, providing full reproducibility metadata for every video.


```
{
  "generator": "LTX-2.3", "mode": "i2v",
  "model_version": "ltx-2.3-22b-dev",
  "prompt": "A man in traditional white attire sitting [...]",
  "negative_prompt": "blurry, out of focus, overexposed, [...]",
  "seed": 1364115978, "num_inference_steps": 30,
  "width": 1536, "height": 1024, "num_frames": 121, [...]
  "source_caption": "000.json", "source_image": "000.png",
  "environment": {
    "repo_commit": "9e8a28e17ac4dd9e49695223d50753a1ebda36fe",
    "model_sha256": { [...] },
    "gpu_name": "NVIDIA A100 80GB PCIe", [...]
  },
  "generation_time_seconds": 264.0
}
```

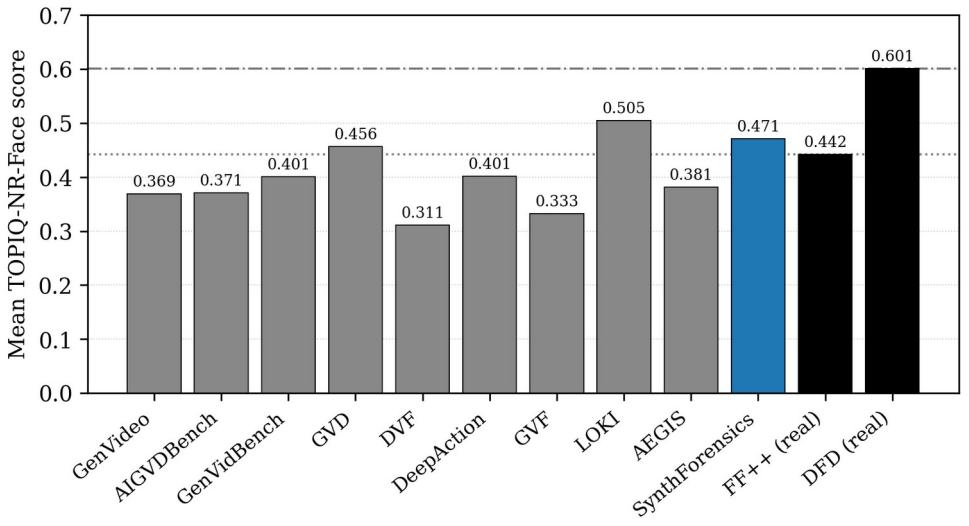

Excerpt of the JSON sidecar of an I2V video generated by LTX-2.3 from FF++ source video #000 (example from the paper's appendix; the released JSONs contain the complete fields).

Landmark stability and face quality

We compare SynthForensics with nine publicly available synthetic-video benchmarks, with FF++ and DFD pristine videos as real-video baselines, on the face-positive subset of each. Landmark stability uses the 2D Face Alignment Network [5]; face quality uses TOPIQ-NR-Face [6]; both on 32 frames uniformly sampled per video.



Landmark completeness. The curves of FF++, DFD and SynthForensics remain within five percentage points of one another across the entire range of τ. At τ = 0.7 the values are 0.932, 0.891 and 0.886, whereas the best benchmark (GenVideo) reaches only 0.509: a gap of more than 37 points.

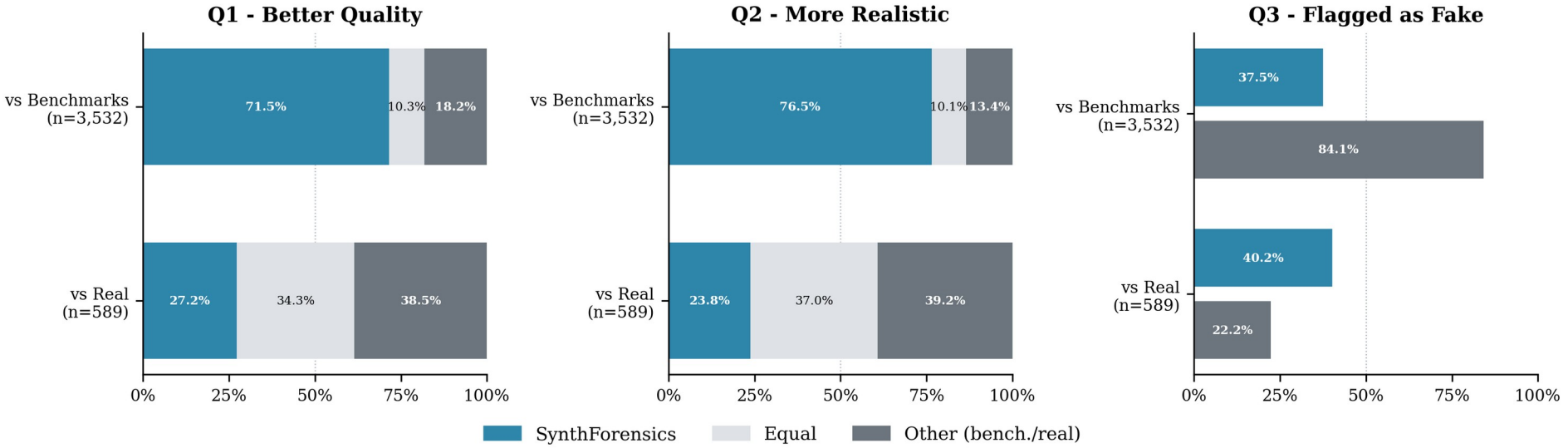


Face quality. SynthForensics scores 0.471, between the real-video baselines FF++ (0.442) and DFD (0.601). Seven benchmarks fall below both real baselines; two more (GVD at 0.456, LOKI at 0.505) sit above FF++ but below DFD.

[5] A. Bulat, G. Tzimiropoulos, How Far are We from Solving the 2D & 3D Face Alignment Problem?, ICCV 2017. [6] C. Chen et al., TOPIQ: A Top-Down Approach from Semantics to Distortions for Image Quality Assessment, IEEE TIP 2024.

Human perceptual study

Following ITU-T P.910 [7] and T2V-HE [8], we ran a paired-comparison study on a pool of 110 video pairs: 100 pairing SynthForensics with one of the nine benchmarks and 10 with a real FF++ or DFD video. Of 159 recruited participants, 118 passed screening, yielding 4,121 ratings per question. For each pair they chose the video of better overall quality (Q1), the more realistic one (Q2), and the one that could be a fake (Q3).

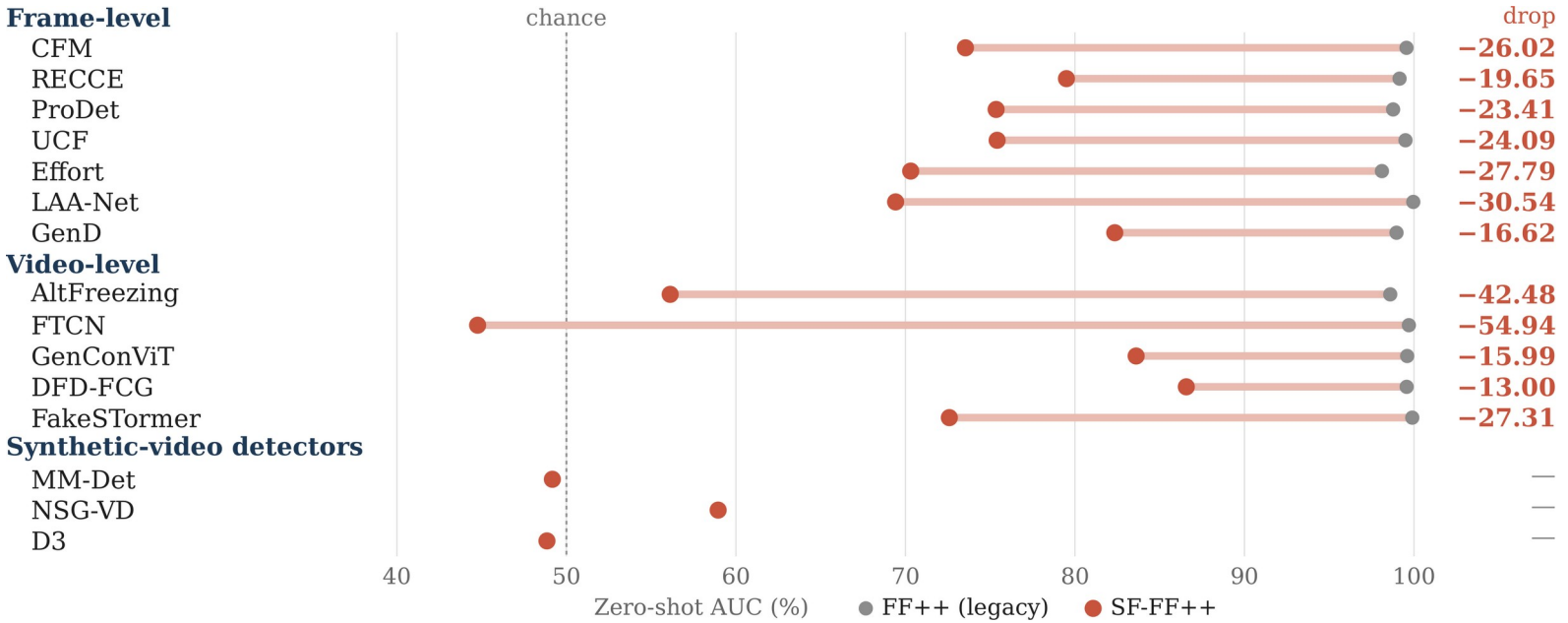


Against existing benchmarks, SynthForensics is preferred on quality (71.5%) and realism (76.5%) and is flagged as fake less than half as often (37.5% vs 84.1%). Against real videos it is still preferred in 27.2% and 23.8% of the comparisons, and its 40.2% fake-flag rate indicates that participants struggle to reliably mark it as synthetic.

[7] ITU-T P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, 2023. [8] T. Zhang et al., Rethinking Human Evaluation Protocol for Text-to-Video Models, NeurIPS 2024.

Zero-shot evaluation

We evaluate 12 face-based detectors (official implementations or DeepfakeBench [9]) and 3 detectors purpose-built for synthetic video, by video-level AUC. SF-FF++ pairs the reals of the FF++ official test split with their SynthForensics counterparts; SF-DFD does the same with all 363 DFD reals, and SF-CDF combines Celeb-DF [10] test reals with unpaired SynthForensics videos.

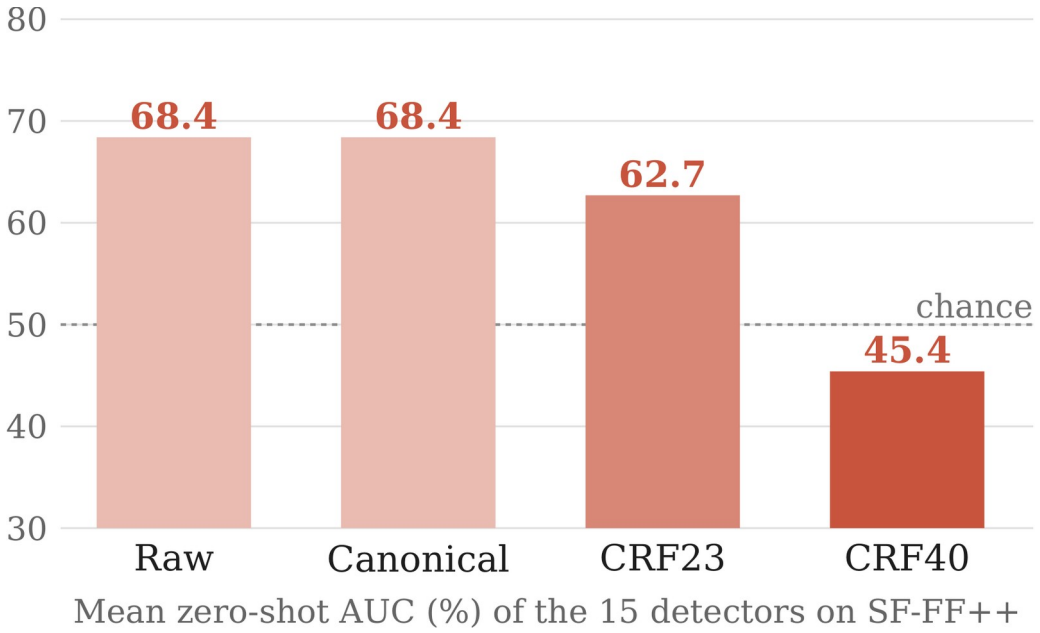


Face-based detectors lose on average 27 AUC points (13–55) from FF++ to SF-FF++, with FTCN and AltFreezing collapsing by over 40 points; the gaps widen further on SF-DFD and SF-CDF, with LAA-Net falling to 47% on SF-CDF. The three synthetic-video detectors reach only 49–59% on SF-FF++ and 41–71% across the three test sets (both extremes on SF-CDF), frequently at or below chance.

[9] Z. Yan et al., DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection, NeurIPS 2023. [10] Y. Li et al., Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics, CVPR 2020.

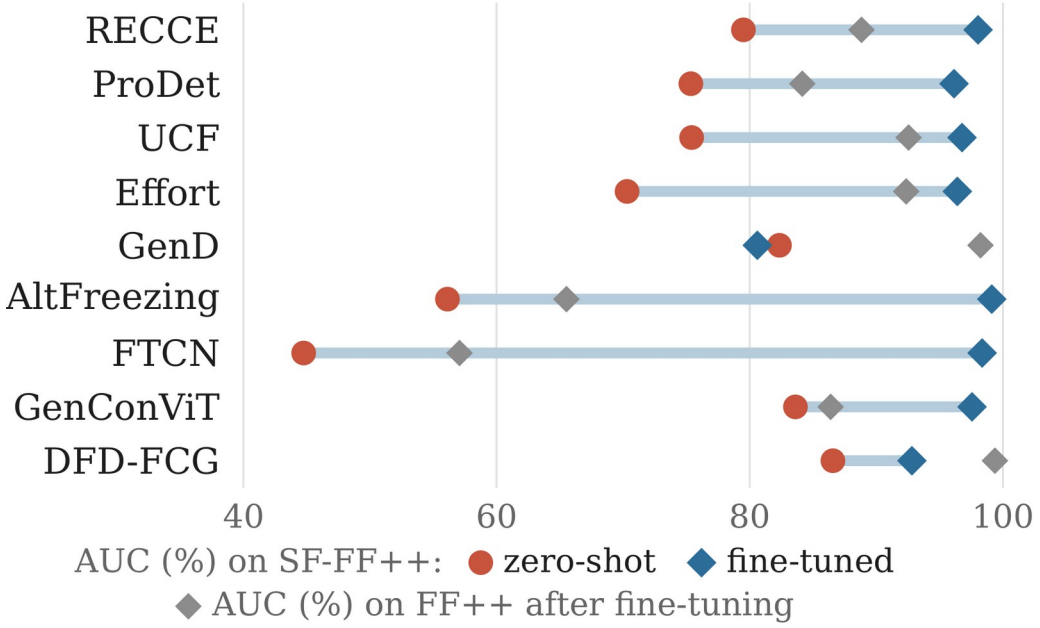
Compression and fine-tuning

Compression. Real and synthetic videos of every test set are re-encoded with the same pipeline, giving four matched versions.



The mean AUC of the 15 detectors on SF-FF++ falls from 68% (Raw and Canonical) to 63% at CRF23 and 45% at CRF40.

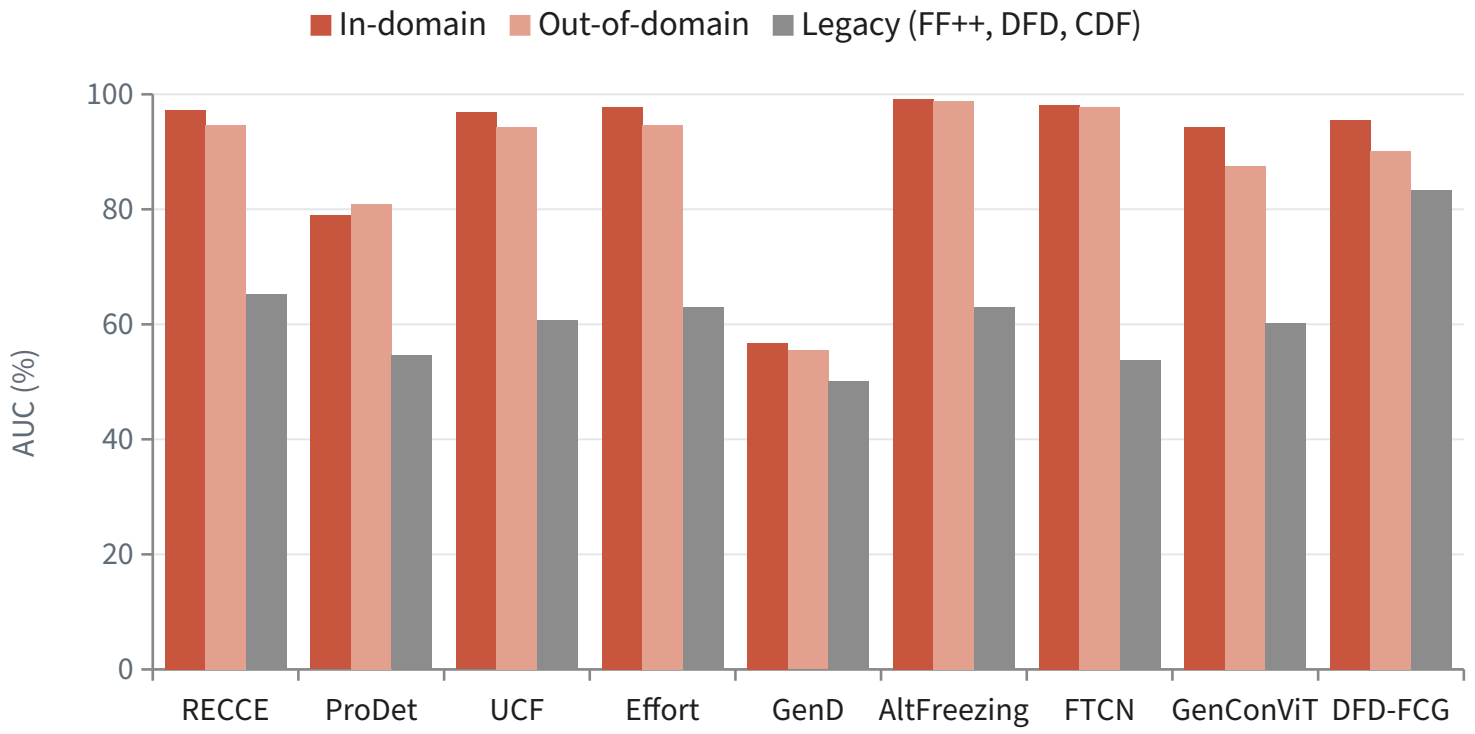
Fine-tuning. Nine face-based detectors are fine-tuned on SynthForensics and tested on SF-FF++ and on FF++.



Fine-tuning raises the mean AUC of the 9 detectors on SF-FF++ from 73% to 95%, but lowers it on FF++ from 99% to 85%; video-level AltFreezing and FTCN pay the steepest cost, at 65.5% and 57.1%.

Training from scratch

To simulate the emergence of new generators, we train from scratch the nine face-based detectors of the fine-tuning protocol on a partition of the generator pool. The in-domain pool, used for training, comprises CogVideoX, Wan2.1, LTX-2.3 and MAGI-1 in both T2V and I2V; the out-of-domain pool, never seen during training, comprises SkyReels-V2, Self-Forcing, Helios and daVinci-MagiHuman. Both are tested on the SF-FF++ test split, and legacy transfer on FF++, DFD and CDF.



Training from scratch closes the in-domain and out-of-domain gap for seven detectors, all reaching 87–99% AUC, with AltFreezing and FTCN holding within 0.5 points across the two pools; ProDet and GenD lag behind.

Legacy transfer collapses to 50–65%, with DFD-FCG the only exception (83.4%): the features learned on fully synthetic videos do not generalize to traditional manipulation benchmarks.

Takeaways

A people-centric benchmark. 20,445 videos from 8 T2V and 7 I2V open-source generators, paired-source from FF++/DFD reals, two-stage human-validated, in four compression versions with full reproducibility metadata.

Ahead of existing benchmarks. Above each of the nine existing synthetic-video benchmarks on landmark stability and in human paired comparisons (win rates above 71%), with face quality in the FF++/DFD baseline range.

A hard test for detectors. Face-based detectors drop 13–55 AUC points (mean 27) zero-shot and 23 more under CRF40; fine-tuning recovers the gap at a backward cost on legacy benchmarks, and training from scratch shows that synthetic and manipulation features barely overlap for most detectors.

Dataset & code







rorobertostring.github.io/
SynthForensics.github.io

Contact
Roberto Leotta · roberto.leotta@ictlab.srl