WHAT A DEEPFAKE SCORE DOES NOT TELL YOU: SAME-SOURCE CONTROLS FOR FACE-FORGERY AND GENERATED-VIDEO DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real/fake performance alone does not determine how a detector responds to non-generative processing, neural re-rendering, or cross-identity conditioning. We introduce source-matched controls for face-forgery and generated-video evaluation. For 136 FaceForensics++ test excerpts and two face-swap generators, we retain each original (R) and construct a pipeline-only arm (N), a self-swap (S), and a cross-identity swap (F) with the same implemented downstream pipeline. Across 29 released checkpoints from five families, several assign high forgery scores to pipeline-only videos (N/R AUC up to 0.97; at a common 0.5 threshold two video pipelines judge 63–65% of them fake against 1% of originals). The conditional discrimination between cross-identity and self-swap outputs (F/S) varies across the 26 primary checkpoints (0.47–0.79) and is weaker than the conventional forgery-versus-original discrimination for every one of them. On a third generator, GHOST, absent from the documented detector-training methods, N/R exceeds 0.5 for all 26 primary checkpoints and F/S is below F/R for 25. On FaceForensics++ test crops at threshold 0.5, Gaussian blur ($\sigma = 4$) turns 99–100% of initially correctly classified real frames into false alarms for 16 of the 19 frame-level checkpoints, and JPEG quality 30 turns 15–96% of correctly detected NeuralTextures frames into misses for the 12 FaceForensics++-trained ones. In a controlled Xception system, pipeline-matched negatives suppress the positive pipeline response. With fixed SimSwap four-arm inputs and five epochs, changing only the self-swap target raises F/S from 0.55 to 0.97 on the training generator and from 0.56 to 0.69 on GHOST (not confirmed on InSwapper), and a shared-backbone two-head model supports both targets. A source-matched generated-video case study shows detector- and implementation-dependent responses to format conversion and same-family VAE reconstruction. Standard AUC does not resolve these operational sensitivities; we recommend reporting matched non-generative controls, conditional contrasts, and the inference and control-construction settings alongside detection results.

1 INTRODUCTION

Face-forgery detection is evaluated as binary classification: a model receives a video, emits a score, and is ranked by AUC on held-out manipulation methods (Rössler et al., 2019; Li et al., 2020; Yan et al., 2023b; 2024). Cross-method AUCs above 0.9 are common and are read as evidence that the model has learned what a manipulated face looks like; the reading skips a question the evaluation cannot answer. In the comparison actually made, an original video against a neural identity swap, three changes arrive together: the pipeline detects, aligns, warps, blends, and re-encodes the face region; a neural generator re-renders the aligned crop; and a different source identity replaces the person. A high AUC on that comparison does not say which of the three drives the score.

We build controls that hold the implemented downstream processing fixed while the processing path, the generator, or the source-identity conditioning changes (Figure 1a). For each real test video we build arms that share one face detection, one affine alignment, and (for the two primary generators) one blending mask. R is the original. N runs the complete pipeline but pastes the aligned original crop back where the generator output would go, so it contains no synthesised pixel. S runs the generator with the same person as source, so the face is re-rendered but nobody is replaced. F runs the generator

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with a fixed other person. The adjacent-arm contrasts quantify the responses to pipeline processing (N/R), to neural re-rendering (S/N), and to cross-identity rather than self-identity conditioning (F/S). Two labels make the contrasts interpretable: a broad synthesis label, positive whenever the face contains generator output, and an identity label, positive only when the person is replaced; S is positive on the first and negative on the second, so scoring S as fake is right under the synthesis contract, and F/S asks whether the score also separates a replaced identity from a re-rendered one, which the standard evaluation never asks.

We apply the protocol in three steps. First, we evaluate 29 released checkpoints on arms built with two swap generators and two encodings (Section 4). Responses to the pipeline-only arm vary across checkpoints, two video pipelines separating it from the original almost perfectly; F/S is far below the cross-method AUC for every primary checkpoint and varies by 0.1–0.2 even for the same DF40 training data. Second, we intervene in one controlled training system (Section 5): pipeline-only negatives suppress the pipeline response and leave F/S unchanged; with the SimSwap training set and five epochs fixed, changing only the target of S changes the identity ranking on the training generator and partially on GHOST (not confirmed on InSwapper), and a two-head backbone carries both axes. Third, the same construction carried to full-frame generated video (Section 6) shows detector- and implementation-dependent responses to format conversion and same-family VAE reconstruction.

Standard real/fake AUC leaves a detector's sensitivity to non-generative processing, neural reconstruction, and cross-identity generation unresolved. Our same-source controls measure these responses, and fixed-support target interventions test how supervision changes rankings and their transfer. Contributions:

- A source-matched four-arm protocol for video face forgery that distinguishes synthesis and identity labels and defines task-level and conditional contrasts for two generators and two encodings (Section 3).
- An evaluation of 29 released checkpoints (26 in the primary comparison) on two generators and a held-out third showing that conventional detection performance ensures neither low pipeline-only false-alarm rates nor strong conditional identity discrimination, and documenting false alarms and misses induced by non-generative post-processing (Sections 4.2 to 4.4), with source-held-out analyses connecting blur-induced feature displacements to the released read-outs (Appendix H.1).
- Controlled interventions in one Xception system showing how negative construction and self-swap supervision alter detector responses (Section 5), complemented by source-matched format controls for generated-video detectors (Section 6).

2 Related work

Benchmarks and evaluation protocols. FaceForensics++ (Rössler et al., 2019), Celeb-DF (Li et al., 2020), DFDC (Dolhansky et al., 2020), DeepfakeBench (Yan et al., 2023b), and DF40 (Yan et al., 2024) evaluate detectors by AUC against held-out manipulation methods or domains. For neural identity swaps in these benchmarks, F/R jointly changes processing, neural rendering, and source-identity conditioning without a matched pipeline-only arm; DF40 crosses methods with domains but does not separate processing from source. Identity-aware detectors exploit identity cues (Nirkin et al., 2022; Cozzolino et al., 2021; Dong et al., 2022); F/S tests whether any detector's score separates cross- from self-identity conditioning.

Shortcuts in generated-image detection. Dataset bias (Torralba & Efros, 2011) and shortcut learning (Geirhos et al., 2020) are documented for still-image forensics: Chai et al. (2020) localise the patches that make GAN images detectable; Gragnaniello et al. (2021) and Corvi et al. (2023) show that resizing and compression mismatches between real and generated sets inflate accuracy; Grommelt et al. (2024) align JPEG and size statistics between the classes; Rajan et al. (2025) make the latent-diffusion autoencoder reconstruction of each real image the positive class; Wang et al. (2020) and Ojha et al. (2023) show that CNN and CLIP features detect generator families. Inpainting Exchange (Nebioglu et al., 2026) restores the pixels outside the inpainted region, drops inpainting detectors towards chance, and traces their cue to the VAE reconstruction.

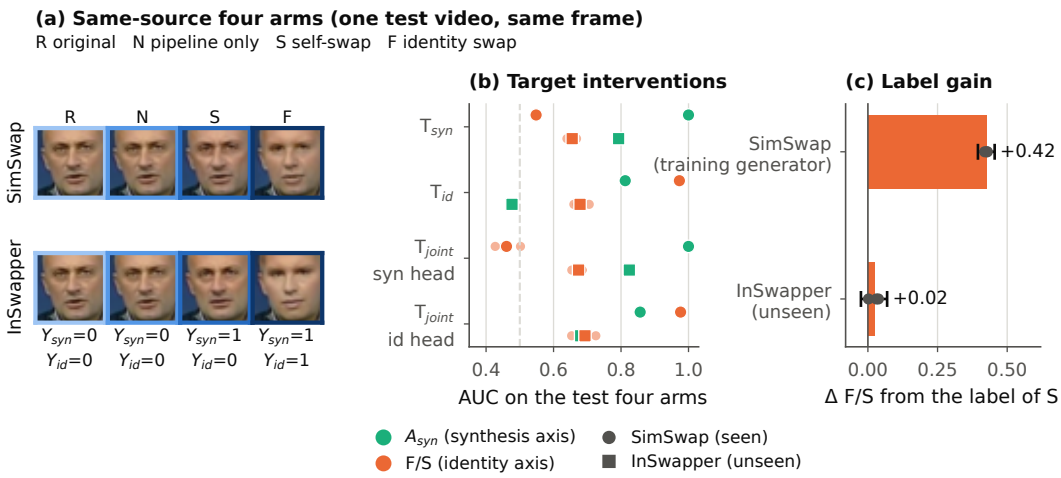

Figure 1: **Same-source arms and the fixed-support target intervention.** (a) One test video, one frame, four arms sharing detection, alignment, and blending; S is positive on the synthesis axis and negative on the identity axis, so the standard comparison (R vs. F) moves both at once. (b) With the training four arms fixed, T_{syn} and T_{id} differ only in the target of S (positive vs. negative); T_{joint} trains both targets as two heads on one backbone, read per head. Green: A_{syn}; orange: F/S. Circles: SimSwap (training generator); squares: InSwapper (unseen); small markers: seeds. (c) The label change raises F/S by $+0.42$ on SimSwap and $+0.02$ on InSwapper (bars: three-seed means; dots: seeds; error bars: 95% source bootstrap of the three-seed mean with one index set shared by the seeds, i.e. source uncertainty with the models fixed); the joint identity head's InSwapper improvement comes from F/R and F/N (Appendix K).

Compositing shortcuts in face forgery. Closest to this paper is the alpha-blending hypothesis of Yermakov et al. (2026), under which frame-level detectors act as blending searchers: four FaceForensics++-trained detectors reach video-level AUROC of 90.6–97.6 when the fakes are replaced by self-blended images they never saw, real faces pasted back onto their own frames with a brightness shift trigger them, and adding self-blended images to the real class lowers cross-dataset AUROC. Self-blended images (Shiohara & Yamasaki, 2022) are a training design whose positive class is blended real faces. Our N arm makes the non-generative point with the generator's own pipeline rather than a synthetic perturbation, and the protocol adds what a blending control cannot: a neural self-reconstruction S that separates the synthesis and identity labels, a fixed-support intervention on that label, native video-pipeline inference with source-paired intervals, and format controls for full-frame generated video, where there is no compositing.

Controls for AI-generated video. VidAudit (Cakiroglu et al., 2026) proposes six protocol-level controls for generated-video detection (canonical re-encoding, frame-count leakage, real-versus-real, matched harness, seeds, cross-dataset) and shows that they change rankings; Michels et al. (2026) show that motion-based detectors collapse when a motion bias is removed. We claim no priority for these controls; Section 6 adds a same-source pipeline-only arm built from the released generators' own output format, so that G/R can be compared directly with N/R.

3 Same-source four-arm protocol

3.1 Arms

Let v be a real video and g a face-swap generator that takes an aligned target crop x_t and a source identity embedding e_s and returns a rendered crop $g(x_t, e_s)$. A swap pipeline computes per frame a detection and five-point landmarks, an affine map M to the generator's canonical frame, the rendered crop, and a paste-back with a soft mask m: $y = m \odot \text{warp}_M^{-1}(g(x_t, e_s)) + (1-m) \odot x$. For SimSwap

and InSwapper we define four arms that share the detection, M, and m of every frame:

$$\mathsf{R} = x, \qquad \mathsf{N} = m \odot \mathrm{warp}_M^{-1}(x_t) + (1-m) \odot x,$$
$$\mathsf{S} = m \odot \mathrm{warp}_M^{-1}(g(x_t, e_v)) + (1-m) \odot x, \quad \mathsf{F} = m \odot \mathrm{warp}_M^{-1}(g(x_t, e_{d(v)})) + (1-m) \odot x,$$

where e_v is the identity embedding of the same video's first detectable frame and $d(v)$ is a fixed donor video. N contains the warps, the mask erosion and feathering, and the re-encoding of the pipeline, but no synthesised pixel; S contains synthesis without replacement, preserving identity up to the generator's own reconstruction error (Appendix B); F contains both. The contrasts N/R, S/N, and F/S are operational: each changes the processing path, the generator, or the source conditioning while the implemented pipeline stays fixed. F/S measures discrimination between cross-identity and self-identity outputs of the same generator; it does not isolate an abstract notion of identity from the synthesis differences that a changed source conditioning brings.

Label contract. We use a broad synthesis label Y_{syn} that is positive whenever the visible face contains generator output, covering self-reconstruction, expression, lip, and attribute edits as well as swaps, and an identity label Y_{id} that is positive only when the person is replaced. The arms carry $(Y_{\mathrm{syn}}, Y_{\mathrm{id}}) = (0,0), (0,0), (1,0), (1,1)$ for R,N,S,F. Under this contract N judged fake is a false alarm, S judged fake is correct, and F/S is a conditional diagnostic of the identity sub-problem rather than a score on the main task. Detectors trained with a different contract, for example the blended-face positives of SBI, are evaluated under ours and read accordingly.

3.2 Construction and verification

We take 136 of the 140 real test videos of the FaceForensics++ c23 split (Rössler et al., 2019) as 101-frame excerpts at the original spatial extent (about four seconds at 25 fps) and build the arms with SimSwap-224 (Chen et al., 2020) and InSwapper-128 (InsightFace contributors, 2023); GHOST (Groshev et al., 2022), absent from the documented detector-training methods of every checkpoint, is held out as a third generator with its released output-adaptive paste-back, so its S/N and F/S include changes in the blending mask (Figure 2, right; Appendix G). Each generator's released alignment and blending is used (Appendices B and G; no watermark, no JPEG intermediates). The donor map $d(\cdot)$ is a fixed permutation of the 136 videos with self-mapping excluded, so each video serves once as target and once as donor. Arms are written as lossless PNG (C0) and additionally re-encoded with identical ffmpeg settings (libx264, CRF 23, yuv420p, 25 fps; C23). We verify intended identity preservation and substitution with an embedding check from the same face-analysis stack; InSwapper conditions on that embedding, so the check shares its representation with one of the generators. On the shared landmarks, the mean cosine similarity to R over the 136 C0 test videos is 0.76 (SimSwap) and 0.82 (InSwapper) for S and 0.19 and 0.06 for F, which is close to its donor (0.56 and 0.89; distributions in Appendix B). Expression and gaze are not verified. The mean absolute difference between N and R over the full frame is 0.10 and 0.16 grey levels on the 0–255 scale; by construction it is confined to the mask's support. Frame-level detectors receive the same square crop (1.3× the original bounding box, resized to the model's input) from all four arms, a common track that lets checkpoints be compared; full-frame pipelines receive the whole excerpt under each checkpoint's released preprocessing, a native track (Appendix A).

3.3 Metrics and uncertainty

A video score is the mean logit or probability over 32 uniformly spaced frames, or the pipeline's own aggregate. For an ordered pair of arms (A, B) we report the video-level AUC of A against B over the 136 sources, written A/B. A value below 0.5 is a reversed ranking, not an absent response; the AUCs of different contrasts are not shares of a mechanism and are never subtracted. Intervals are pointwise 95% paired bootstrap over sources (2,000 resamples; the four arms of a source are resampled together, dependence through the donor side of F is not modelled, so the intervals are source-level; a donor-endpoint-disjoint check is in Appendix F). We summarise the axes by $A_{\mathrm{syn}} = (\mathsf{S/R} + \mathsf{S/N} + \mathsf{F/R} + \mathsf{F/N})/4$ and $A_{\mathrm{id}} = (\mathsf{F/R} + \mathsf{F/N} + \mathsf{F/S})/3$ but always report the pairwise values (Appendix C). The paired win rate W_{id} is the fraction of sources whose F outscores their own S, ties counting one half. Judged-fake rates use a fixed threshold, 0.5 on the model's probability (0 on a logit), without calibration, so a rate is read within a checkpoint, not as a matched operating point across checkpoints; released thresholds are in Appendix C.

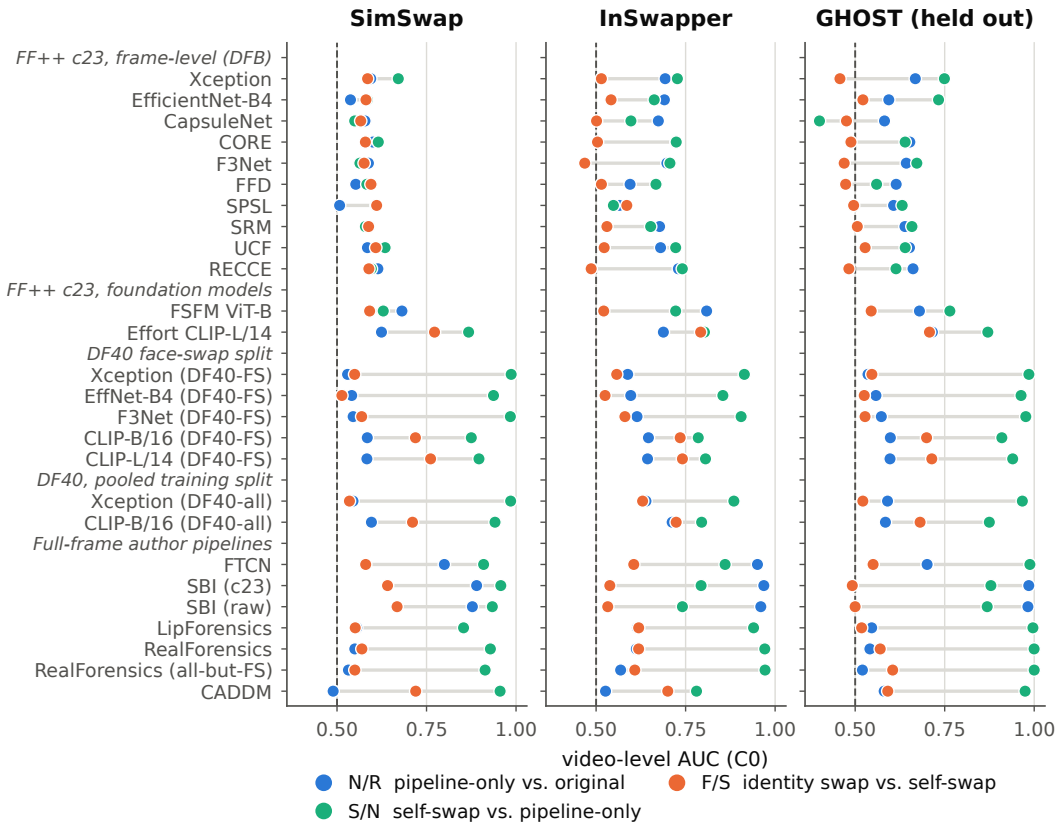

Figure 2: **The 26 primary checkpoints on the same-source arms of three generators (C0, video-level AUC, 136 sources).** Blue: N/R, pipeline without generator against original. Green: S/N, self-swap against pipeline. Orange: F/S, identity swap against self-swap. Pipeline responses vary and several are substantial; the DF40 checkpoints and the video pipelines respond strongly to re-rendering; F/S is below 0.70 for all but CADDM and the four CLIP-backbone models and never exceeds 0.79 on SimSwap or InSwapper. GHOST (right) is held out from the documented detector-training methods and uses its released adaptive paste-back; N/R exceeds 0.5 for all 26 and F/S is at most 0.71. Intervals, W_{id}, and judged-fake rates for SimSwap and InSwapper are in Appendix D; GHOST AUC point estimates and pipeline-only judged-fake rates are in Appendix G. Both appendices list the separately reported checkpoints. Gray segments connect contrasts, not intervals.

4 Released checkpoints under the protocol

4.1 Checkpoints

We evaluate 29 released checkpoints without retraining; 26 form the primary comparison and three low-discrimination DeepfakeBench checkpoints (Meso4, MesoInception, CNN-Aug), which assign near-constant fake probability to our crops, are reported separately in Appendix D. The 26 are ten DeepfakeBench models trained on FaceForensics++ c23 (Xception, EfficientNet-B4, CapsuleNet, CORE, F3Net, FFD, SPSL, SRM, UCF, RECCE), and FSFM ViT-B and Effort CLIP-L/14 fine-tuned on the same data. Seven DF40 models have training data that contain both SimSwap and InSwapper outputs (Xception, EfficientNet-B4, F3Net, CLIP-B/16, and CLIP-L/14 trained on the face-swap split; Xception and CLIP-B/16 trained on the pooled DF40 training split). Seven author-released full-frame pipelines are evaluated by us (FTCN; SBI c23 and raw; LipForensics; RealForensics trained on FaceForensics++ and on all-but-FaceSwap; CADDM EfficientNet-B3). The two DF40-all models judge 68–71% of the originals fake at 0.5 but rank them normally (Appendix A). As a within-dataset sanity check on our crops, the FaceForensics++-trained frame-level checkpoints reach video AUC 0.81–0.99 per manipulation (Appendix A).

Table 1: **Full-frame author-released pipelines on the same-source arms.** Video-level AUC over 136 sources at C0, with the C23 value of N/R in parentheses; W_{id} is the paired win rate of F over its own S; "N fake" is the fraction of pipeline-only videos judged fake at threshold 0.5 (originals: 0–7% for every model). Brackets: 95% paired bootstrap over sources for F/S. Conditional identity discrimination (F/S, shaded) is 0.53–0.67 except for CADDM.

	SimSwap					InSwapper				
Model	N/R (C23)	S/N	F/S	W_{id}	N fake	N/R (C23)	S/N	F/S	W_{id}	N fake
FTCN	0.80 (0.65)	0.91	0.58 [.56,.61]	0.74	16%	0.95 (0.84)	0.86	0.61 [.58,.64]	0.76	65%
SBI (c23)	0.89 (0.67)	0.96	0.64 [.60,.69]	0.73	26%	0.97 (0.79)	0.79	0.54 [.49,.58]	0.58	63%
SBI (raw)	0.88 (0.65)	0.93	0.67 [.62,.71]	0.76	28%	0.96 (0.76)	0.74	0.53 [.49,.58]	0.60	63%
LipForensics	0.56 (0.53)	0.85	0.55 [.51,.59]	0.62	0%	0.62 (0.57)	0.94	0.62 [.58,.66]	0.64	1%
RealForensics	0.55 (0.56)	0.93	0.57 [.53,.61]	0.65	9%	0.61 (0.63)	0.97	0.62 [.58,.66]	0.66	11%
RealForensics (all-but-FS)	0.53 (0.54)	0.91	0.55 [.52,.59]	0.57	1%	0.57 (0.59)	0.97	0.61 [.57,.65]	0.74	1%
CADDM	0.49 (0.53)	0.96	0.72 [.67,.77]	0.80	1%	0.53 (0.57)	0.78	0.70 [.65,.75]	0.75	2%

4.2 The pipeline arm is scored as forged

Figure 2 and Table 1 give the three contrasts. On InSwapper N/R exceeds 0.5 for all 26 primary checkpoints, on SimSwap for 25 of them (CADDM: 0.49). It is 0.51–0.68 for the frame-level FaceForensics++ models on SimSwap and 0.56–0.81 on InSwapper, consistent with its larger paste-back footprint (Appendix B), and it reaches 0.80–0.97 for FTCN and SBI. At the common 0.5 threshold used throughout, FTCN calls 16% of SimSwap and 65% of InSwapper pipeline-only videos fake, SBI 26% and 63%, against 1% of the originals for both models (held-out GHOST: 76–81% for SBI). This agrees with the alpha-blending hypothesis; SBI's training positives are blended real faces, so a boundary response is consistent with its design. LipForensics and RealForensics, which model mouth and audio-visual motion, have small pipeline responses at C0 (N/R 0.53–0.62) and among the largest re-rendering responses (S/N 0.85–0.97). Under a common H.264 re-encoding of all four arms (C23), the pipeline response of FTCN and SBI shrinks (N/R 0.80 to 0.65 and 0.89 to 0.67 on SimSwap; pipeline-only judged-fake rates 4–24%) while the other four pipelines change by at most 0.05, and S/N on SimSwap falls by 0.04–0.27 for six of the seven (Appendix D); what one compression pass removes is a candidate explanation, not an identified one.

4.3 Conditional identity discrimination varies across checkpoints

Conditional identity discrimination is lower than the conventional F/R discrimination for all 26 primary checkpoints on SimSwap and InSwapper and for 25 on GHOST (CapsuleNet: F/R 0.46, F/S 0.48; Figure 2). On SimSwap and InSwapper, F/R is 0.68–1.00 while F/S is 0.47–0.61 for the ten frame-level FaceForensics++ models and FSFM, 0.51–0.63 for the four DF40 CNN and frequency models that were trained on outputs of both generators, and 0.53–0.67 for FTCN, SBI, LipForensics, and RealForensics. A high F/R does not ensure strong conditional identity discrimination: the DF40 CNN checkpoints separate S from N with AUC 0.85–0.99, yet their F/S remains 0.51–0.63. Five checkpoints have F/S of 0.70–0.79: CADDM (0.72 and 0.70), whose training suppresses identity leakage, Effort (0.77 and 0.79), and the DF40 CLIP-B/16 and CLIP-L/14 checkpoints (0.71–0.76); on GHOST, the CLIP-based checkpoints remain the highest (0.68–0.71) and CADDM falls to 0.59. CLIP-based checkpoints achieve higher F/S than the CNN checkpoints released for the same DF40 training subsets, so exposure to the two generators alone does not explain the differences; architecture, pre-training, and adaptation remain confounded. The paired win rate agrees: W_{id} is 0.46–0.69 for the ten FaceForensics++ frame-level models and 0.57–0.76 for the video pipelines other than CADDM, and even the five checkpoints at or above 0.70 reach a paired, tie-adjusted win rate of only 0.75–0.85. Re-centring every score on its source baseline $(s(\mathsf{R}) + s(\mathsf{N}))/2$ changes F/S by less than 0.04 for the FaceForensics++ CNN and frequency checkpoints, FSFM, the DF40 face-swap CNN and frequency models, LipForensics, and RealForensics and lowers it by up to 0.22 for the rest (Appendix E).

4.4 Non-generative post-processing moves decisions in both directions

We test whether non-generative post-processing changes detector decisions while leaving the synthesis label unchanged. On the FaceForensics++ test crops, Gaussian blur ($\sigma = 4$) flips 99–100% of the real frames judged real before processing to fake for 16 of the 19 frame-level checkpoints (the

Table 2: **Training interventions in one Xception system (three seeds).** Top (last epoch, no model selection): identical four-arm minibatches (16 source frames × R,N,S,F), identical initialisation, data order, and flip sampling per seed; T_{syn} and T_{id} differ only in the target of S; T_{joint} has two linear heads on one backbone and is read per head. Bottom (epoch chosen by validation AUC): same positives F, negatives changed. Cells: mean ± sample standard deviation over seeds; F/S columns shaded. 95% paired bootstrap over the 136 sources for F/S (seed 303; SimSwap / InSwapper): T_{syn} [.51,.58] / [.60,.68], T_{id} [.96,.99] / [.62,.72], T_{joint} id head [.96,.99] / [.60,.70]. Video AUC on the four FaceForensics++ manipulations (official labels, seed means): 0.42–0.57 (fixed-support rows); 0.49–0.59, 0.44–0.52, 0.46–0.55 (negatives recipes).

	SimSwap (training generator)				InSwapper (unseen)			
Recipe	N/R	S/R	A_{syn}	F/S	N/R	S/R	A_{syn}	F/S
T_{syn} (S positive), syn head	0.45±0.03	1.00±0.00	1.00±0.00	0.55±0.00	0.57±0.08	0.76±0.08	0.79±0.05	0.66±0.01
T_{id} (S negative), id head	0.43±0.01	0.61±0.04	0.81±0.02	0.97±0.01	0.41±0.02	0.32±0.03	0.48±0.03	0.68±0.02
T_{joint}, syn head	0.48±0.02	1.00±0.00	1.00±0.00	0.46±0.04	0.66±0.02	0.84±0.04	0.82±0.02	0.67±0.02
T_{joint}, id head	0.45±0.03	0.70±0.04	0.86±0.02	0.98±0.00	0.56±0.04	0.63±0.10	0.68±0.06	0.69±0.04
U_{base} (neg. R)	0.71±0.04	1.00±0.00	–	0.65±0.02	0.91±0.04	0.98±0.01	–	0.66±0.02
U_{matched} (neg. N)	0.43±0.01	1.00±0.00	–	0.66±0.01	0.55±0.02	0.79±0.02	–	0.67±0.01
U_{both} (neg. R,N)	0.50±0.05	1.00±0.00	–	0.65±0.02	0.62±0.11	0.78±0.11	–	0.65±0.03

three CLIP models: 61–77%), and $\sigma = 2$ already flips 21–84%. JPEG quality 30 flips 15–78% of the NeuralTextures fake frames judged fake before processing to real for the ten FaceForensics++ CNN and frequency models, 96% for FSFM and 48% for Effort; the DF40 models, trained on compressed generator outputs, flip 0–5%. The full-frame pipelines (FTCN, both SBI models, the FaceForensics++-trained RealForensics) behave alike at video level: blur $\sigma = 2$ raises the fraction of real videos judged fake from 1–7% to 64–90% and JPEG 30 lowers the fraction of identity swaps judged fake from 84–99% to 9–51%. Under the synthesis contract the blur flips are false alarms and the JPEG flips are misses, since neither adds nor removes synthesised content (per-checkpoint flips and foreign real sets, where an SBI model judges 90% of YouTube reals fake, are in Appendix H). Across sources, mean within-video score–sharpness correlations on real crops are negative for all 19 primary frame-level checkpoints; in six prespecified checkpoints, inner-region blur produces the largest mean score increase at $\sigma = 2$ and 4, and one blur-derived direction nearly retains F/R AUC; on SimSwap, so does the leading natural-change direction for five of six checkpoints (Appendix H.1).

5 Training interventions

We now intervene in one controlled training system: SimSwap arms of the 706 training and 128 validation reals of the FaceForensics++ split (16 frames per video, 299×299 crops with the rule of Section 3), Xception-299, five epochs, AdamW (Loshchilov & Hutter, 2019) at 10^{-4}, three seeds; evaluation on the test four arms of all three generators. Causal statements below are about this system. Table 2 reports seed means with sample standard deviations and source-level paired-bootstrap intervals.

5.1 Fixed support, only the target changes

Every minibatch holds 16 source frames with all four arms; initialisation, data order, and flip sampling are identical across recipes for a given seed. T_{syn} $(\mathsf{R}, \mathsf{N}, \mathsf{S}, \mathsf{F}) \mapsto (0, 0, 1, 1)$ and T_{id} $(0, 0, 0, 1)$ differ only in the target of S, which also halves the positive fraction, so their comparison is the total effect of relabelling S; T_{joint} instead trains both targets with two linear heads on a shared backbone. We report the last epoch (epoch selection changes F/S by at most 0.025, Appendix K). With the input support and training procedure fixed, the target assigned to S is sufficient to change the identity ranking on the training generator: T_{id} raises F/S from 0.548 to 0.973 (seed means; per-seed difference 0.425 ± 0.005; source-level intervals in Table 2) and W_{id} from 0.58 to 0.99, while T_{syn} keeps A_{syn} at 1.00 with F/S at 0.55. The two-head model's synthesis head reaches $A_{\text{syn}} = 1.00$ and its identity head F/S 0.98, the same as the single-target models: the two abilities coexist. All three recipes see N as a negative, and on the training generator the positive pipeline response is suppressed (0–2% of N judged fake); N/R of 0.43–0.48 shows that the pipeline difference remains visible to the network with reversed sign, not that it disappeared.

Table 3: **Source-matched generated-video controls under the primary inference and encoding settings (AIGVDBench, 300 sources, video-level AUC).** N/R: format-and-temporal-resampling control against original; V/N: VAE round trip against the control; G/R, G/N: released video-to-video output (image-to-video for SVD and Pyramid-Flow) against original and control. Each detector runs under its authors' released inference code; paired-bootstrap half-widths are at most 0.04; italics mark displayed AUC estimates below 0.50 (reversed score ranking, not statistical significance); judged-fake rates, contract variants, and the kernel and CRF sensitivity of N are in Appendix J.

	CogVideoX1.5				LTX-Video				SVD			Pyramid-Flow		
Detector	N/R	V/N	G/R	G/N	N/R	V/N	G/R	G/N	N/R	G/R	G/N	N/R	G/R	G/N
DeCoF	0.82	0.56	0.81	0.51	0.80	0.61	0.84	0.57	0.89	0.97	0.77	0.86	0.88	0.54
WaveRep	0.52	0.94	0.98	0.98	0.58	0.99	1.00	0.99	0.64	1.00	1.00	0.51	0.93	0.94
D3	*0.38*	0.50	0.63	0.73	*0.34*	*0.48*	*0.49*	0.66	*0.31*	*0.37*	0.59	*0.40*	0.52	0.63

5.2 The identity gain is generator-conditional

For the original five-epoch models, the intervention moves F/S on the unseen InSwapper from 0.655 to 0.679 (per-seed difference 0.023 ± 0.019), and the joint identity head reaches 0.69; on GHOST, which no training run saw, it raises F/S from 0.56 to 0.69 (gain $+0.14$, interval $[+0.09, +0.18]$). In exploratory runs trained on both SimSwap and InSwapper, T_{id} reaches F/S 0.98 on InSwapper but 0.77 on GHOST (0.59 for T_{syn}); on GHOST this gain ($+0.18$) shows no confirmed increase over an update-matched SimSwap-only control ($+0.17$; Appendix K.1). The label gain thus transfers partially and unevenly across the tested unseen generators.

5.3 What "real" is in training decides the pipeline response

The bottom block of Table 2 varies only the negatives (epoch chosen by validation AUROC; all five epochs in Table 29): U_{base} (negatives R, the standard construction) and U_{matched} (negatives N) are exposure-matched, sharing the positives F, the training procedure, and the number of updates (every seed selects epoch 1, 353 updates); U_{both} (negatives R and N) takes about 1.5 times as many updates per epoch and is supplementary. Replacing R by N lowers N/R from 0.71 ± 0.04 to 0.43 ± 0.01 on the training generator and from 0.91 ± 0.04 to 0.55 ± 0.02 on the unseen InSwapper, with U_{both} in between. S/R stays at 1.00 for all three recipes on SimSwap, so every recipe still ranks self-swaps above originals, and F/S stays at 0.65–0.67, so suppressing the positive pipeline response does not create an identity response. On InSwapper, U_{matched} also loses part of its cross-generator score (S/R 0.98 to 0.79, F/R 0.99 to 0.89): part of the base model's transfer to the unseen generator depended on responses that pipeline-matched negatives suppress; which features carry them is not localised.

6 Beyond faces: controls for generated-video detectors

The generated-video pipelines have no face paste-back stage, but a released generator still fixes a resolution, frame rate, frame count, and encoder, and we build the control from those settings. The sources are 300 real videos of the AIGVDBench test split (Ma et al., 2026) (1920×1080, 30 fps, 82–113 frames) whose released generated versions share the source video or prompt. N resamples R to the released resolution, frame rate, and frame count (bicubic, uniform temporal sampling, so the playback speed changes as in the released files) and re-encodes it with the libx264 settings read from the released files (preset medium, CRF 25; CRF 23 for SVD). V applies the family's released VAE (version match with the released videos not verified) to the same uncompressed frames and encodes them like N. G are the released outputs: video-to-video (CogVideoX1.5, LTX), image-to-video from the first frame (LTX, SVD, Pyramid-Flow), and text-to-video from the caption (CogVideoX1.5, LTX). They are not content-preserving counterfactuals (G may change motion, camera, and objects), so the comparisons are G/R and G/N. Three released detectors score every arm under the inference contract of their authors' released code (Appendix J): DeCoF (Ma et al., 2024), the CLIP-L/14 plus temporal-transformer detector released with the benchmark; WaveRep (Corvi et al., 2025), a DINOv2 detector trained with forensic-oriented augmentation (every frame, mean logit); and D3 (Zheng et al., 2025), a training-free statistic (the temporal standard deviation of the first differences of consecutive-embedding distances) over XCLIP embeddings of 16 frames at 8 fps.

Table 3 shows three different responses, two of which also depend on how the control or the detector's contract is implemented. DeCoF scores the format control almost as it scores generated video: N/R is 0.80–0.89, so its separation of the released outputs from the originals (G/R 0.81–0.97) is matched by a separation of resampled and re-encoded reals, and G/N falls to 0.51–0.57 for three families (SVD 0.77); applied one at a time and followed by the released encoding, spatial rescaling gives 0.79–0.83 and re-encoding alone or temporal resampling 0.51–0.52 (Appendix J.1). DeCoF separates the VAE round trip from the format control only weakly (V/N 0.56–0.61), and its frame-rule and JPEG variants move no contrast by more than 0.05. WaveRep separates the VAE round trip (0.94, 0.99) and the released outputs (0.94–1.00) from the format control and responds less to the control at the released encoder settings (N/R 0.51–0.64); that response grows to 0.69–0.80 at CRF 18 across the tested kernels (Appendix J.1), so it is partly a property of the control's implementation. Under D3's author-released extraction and operator, the format control ranks below the originals (N/R 0.31–0.40). Temporal resampling followed by encoding reproduces this direction (0.36–0.39), whereas re-encoding alone or rescaling followed by encoding yields 0.48–0.50. The strength and direction of G/R depend on the task (0.37–0.63), while G/N (0.59–0.73) exceeds G/R for all four families. The toolkit variant produces different rankings on the same videos (Appendix J). An author-aligned vision-language detector (Skyra, Appendix J.2) judges 94–97% of the CogVideoX1.5 and LTX format controls fake against 6% of the originals and shows little detection sensitivity on the other two families. Format-control results depend on both the control's construction and the detector's inference contract; both must therefore accompany the reported contrasts, with the released settings as the default. These measurements characterise operational responses, not internal mechanisms.

7 Discussion

What should change in evaluation. The four-arm construction needs no annotated forgery masks, only the generator's own alignment and blending, two generator passes per real video (Appendix L), and an identity check. Reports should add to the cross-method AUC the N/R and S/N contrasts of each generator, F/S with its paired interval when identity is the claim, and, for generated video, a format control at the released settings and contract.

Limitations. The face protocol covers face swapping with three generators, one of them held out, one source dataset, and four-second excerpts; expression, lip, and attribute edits would need their own S construction. The identity axis is an intended assignment with an imperfect check: same-stack embeddings support it for all but a few sources (Appendix B); all outputs are kept. The paired bootstrap resamples sources and does not model dependence through donors; a donor-endpoint-disjoint subset moves the contrasts by at most 0.09 and a second donor assignment moves the released frame-level checkpoints' F/S by at most 0.03 (Appendix F). The original protocols were specified before scoring; the donor stratification, the N grid with its single-step variants, the additional detectors, the GHOST evaluation, and the sharpness and feature-space analyses were exploratory additions after inspection of the initial results and internal feedback (including AI feedback; see the AI use statement), and the bootstrap intervals do not account for this selection. The interventions establish effects within the stated Xception training procedures, not mechanisms in the released checkpoints; the CLIP observation is a checkpoint-level association. The generated-video study uses format-and-temporal-resampling controls, same-family VAEs whose versions may differ from the released videos', and generation tasks that do not preserve content; its primary comparison uses three detectors and one released encoder setting per family (Appendix J).

8 Conclusion

For neural face swaps, the original-versus-swap comparison (F/R) changes processing, synthesis, and identity together; same-source arms measure the operational responses that its AUC leaves unresolved. Several released checkpoints score pipeline-only videos as forged, and F/S is below F/R for all 26 primary checkpoints on SimSwap and InSwapper and for 25 on GHOST. In one controlled Xception system, negative construction and the self-swap label alter the responses, but the original label gain is partial on GHOST and not confirmed on InSwapper; on generated video the responses depend on the control's construction and the inference contract. Detection reports should therefore pair conventional AUC with source-matched controls, explicit label definitions, and inference contracts.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. AI-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 460–470, 2024. doi: 10.1007/978-981-97-8792-0_32. arXiv:2403.16638.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017.

Mert Onur Cakiroglu, Zhihe Lu, Mehmet Dalkilic, and Hasan Kurban. Auditing generalization in AI-generated video detection: A six-control protocol and the VidAudit toolkit. *arXiv preprint arXiv:2606.31004*, 2026.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, 2022.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *ECCV*, 2020.

Chirui Chang, Jiahui Liu, Zhengzhe Liu, Xiaoyang Lyu, Yi-Hua Huang, Xin Tao, Pengfei Wan, Di Zhang, and Xiaojuan Qi. How far are AI-generated videos from simulating the 3D visual world: A learned 3D evaluation approach. In *ICCV*, 2025.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. SimSwap: An efficient framework for high fidelity face swapping. In *ACM Multimedia*, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *CVPR Workshops*, 2023.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable AI-generated video detection with forensic-oriented augmentation. In *NeurIPS*, 2025.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. ID-Reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15088–15097, 2021. doi: 10.1109/ICCV48922.2021.01483.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *CVPR*, 2020.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, 2023.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from DeepFake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9458–9468, 2022. doi: 10.1109/CVPR52688.2022.00925.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *ICME*, 2021.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? Revealing common biases in generated image detection datasets. In *ECCV Workshops*, 2024.

Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. GHOST—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022. doi: 10.1109/ACCESS.2022.3196668.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, 2021.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, 2022.

InsightFace contributors. InsightFace: 2d and 3d face analysis project, software release with the `inswapper_128.onnx` face-swapping model and the `buffalo_l` model pack. `https://github.com/deepinsight/insightface`, model release of 2023, 2023.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-Fake: A unified explainable framework and benchmark for image and video AIGC detection. In *ICMR*, 2026. arXiv:2506.00979.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: AI-generated video detection via grounded artifact reasoning. In *CVPR*, pp. 4482–4493, 2026.

Yuezun Li and Siwei Lyu. Exposing DeepFake videos by detecting face warping artifacts. In *CVPR Workshops*, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *CVPR*, 2020.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting AI-generated video via frame consistency. *arXiv preprint arXiv:2402.02085*, 2024.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for AI-generated video detection. *arXiv preprint arXiv:2601.11035*, 2026.

Joren Michels, Lode Jorissen, and Nick Michiels. Dataset biases and shortcut learning in motion-based AI-generated video detection. *arXiv preprint arXiv:2607.00948*, 2026.

Elif Nebioglu, Emirhan Bilgiç, and Adrian Popescu. AI-generated image detectors overrely on global artifacts: Evidence from inpainting exchange. *arXiv preprint arXiv:2602.00192*, 2026.

Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP*, 2019.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: Consistent representation learning for face forgery detection. In *CVPR Workshops*, 2022.

Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2022. doi: 10.1109/TPAMI.2021.3093446.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *ICLR*, 2025.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, 2022.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. FSFM: A generalizable face security foundation model via self-supervised facial representation learning. In *CVPR*, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *ICCV*, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *NeurIPS Datasets and Benchmarks Track*, 2023b.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward next-generation deepfake detection. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *ICML*, 2025.

Andrii Yermakov, Jan Cech, Mario Fritz, and Jiri Matas. The alpha blending hypothesis: Compositing shortcut in deepfake detection. *arXiv preprint arXiv:2605.10334*, 2026.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D^3: Training-free AI-generated video detection using second-order features. In *ICCV*, 2025.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, 2021.

REPRODUCIBILITY STATEMENT

Every evaluation metric in the paper is computed from stored per-video or per-frame predictions by the analysis scripts (the feature-space analyses of Appendix H.1 additionally use stored penultimate features, released classifier parameters, and source-split records); construction checks, embedding checks, and timings come from separate validation and execution records. Appendix A lists the 29 checkpoints with their publishers, training data, and the inference contract each one was run under; Appendices B and G give the arm construction (generators, alignment and blending, the fixed donor permutation, the two encodings, and the embedding checks); Appendix C defines every contrast and the paired bootstrap; Appendix K gives the training protocol and seeds of the interventions;

Appendix J gives the construction of the generated-video arms, the encoder settings read from the released files, and the released inference code of each detector; Appendix L gives the compute. The arm-construction scripts, the donor permutation, the scoring harness with the inference contracts, the per-video and per-frame prediction files of all checkpoints and detectors, and the scripts that produce every table and figure will be released after the review period; no repository is provided at submission. FaceForensics++ and AIGVDBench are obtained from their distributors under their terms and are not redistributed; the arms are rebuilt from them with the released scripts.

ETHICS STATEMENT

The face-swap arms are built only from the YouTube-sourced videos of FaceForensics++, accessed under the distributor's terms, and the donors are subjects of the same dataset; we recruited no participants and collected no footage, and dataset access is not treated as evidence of individual consent to every downstream use. The outputs exist to evaluate detectors: we do not distribute complete derived videos, and the paper shows only a few labelled illustrative frames (Figure 1). The generated-video arms use the released AIGVDBench videos under their licence. The protocol runs released generators unchanged and adds no generation capability; diagnostics of detector failures can nevertheless lower the effort of evading detection, a dual-use risk we accept because the same diagnostics are what an evaluation needs, and we release no adversarial tooling.

AI USE STATEMENT

In this work we used generative AI tools (large language models with code execution) to implement methods (the arm-construction, scoring, and analysis code, including the inference harnesses that wrap the released detector code), to design and provide feedback on the experiments (model-generated reviews of drafts proposed robustness checks, among them the donor-dependence analyses, the N-implementation grid with its single-step variants, and the additional generated-video detectors, which we then specified and ran), to propose and refine hypotheses about what the contrasts measure, to clean and reformat the data (frame extraction, crops, score files), to interpret results, and to draft and edit the text. Generative models also produced the synthetic evaluation data: the self-swap, cross-identity, and VAE arms are outputs of the released face-swap generators and VAEs that are the objects of study and are cited in the text, and the generated-video arms are the released benchmark outputs; no language-model-generated examples or labels were used as evaluation data. AI assistance also covered the conceptual framing and the formulation and checking of the diagnostic definitions and statistical arguments; the paper presents no standalone theoretical proofs. Translation and qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools for coding, figure generation, and literature search. We have reviewed all AI-assisted work: the within-dataset AUCs of the FaceForensics++-trained checkpoints on our crops serve as a sanity check of the harnesses (Appendix A), the inference contracts of the main-text detectors were checked line by line against the released scripts (the wrapper-based detectors of Appendix J.2 are labelled as such), detection metrics are regenerated from the stored per-video or per-frame predictions by the analysis scripts, the feature-space analyses additionally use stored penultimate features, released classifier parameters, and source-split records, and the reference list was checked against the bibliographic records of the cited works. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

A CHECKPOINT INVENTORY AND INFERENCE PROTOCOLS

Table 4 lists the 29 checkpoints, the publisher of the weights, the training data stated by the publisher, and the inference implementation we ran. Three groups of checkpoints follow three inference contracts.

Frame-level checkpoints on the common crop track. The ten DeepfakeBench models, FSFM (Wang et al., 2025), Effort (Yan et al., 2025), and the seven DF40 models receive the same face crops: for every frame of every arm we take the detection box of the original frame (the four arms of a source share one detection, Section 3), enlarge it to a square of 1.3 times the longer side around the box centre, clip it to the frame, and resize it to the model's input size (299 for Xception-type models, 224 for the CLIP-based models, the publisher's size otherwise) with bilinear interpolation. The DeepfakeBench models are run with the DeepfakeBench inference code and the publisher's normalisation; the DF40 CLIP checkpoints are run with the DF40 CLIP detector definition (a Hugging Face CLIP vision tower with a linear head on the pooled token, mean and standard deviation 0.5); FSFM and Effort are run with the authors' code. Each model returns one fake probability per frame; the video score is the mean over the 32 uniformly spaced frames that the test crops contain, and the judged-fake rate uses the threshold 0.5 on that mean. The FaceForensics++ column of Table 4 gives the video-level AUC of each retained checkpoint on the four FaceForensics++ c23 test manipulations under this crop rule; these within-dataset AUCs provide a sanity check of the common-crop implementation, not an exact reproduction of each publisher's evaluation protocol. The DF40 face-swap checkpoints score lower on the FaceForensics++ reenactment classes (Face2Face 0.63–0.84, NeuralTextures 0.58–0.64), as expected for models trained on swaps only; the two DF40-all checkpoints (released as trained on the pooled DF40 training split, which contains its known methods but not its nine held-out ones) score 0.70–0.96 and judge 68–71% of the FaceForensics++ originals fake at threshold 0.5, so their rates in Table 6 are read against that offset.

Full-frame pipelines on the native track. FTCN (Zheng et al., 2021), SBI (Shiohara & Yamasaki, 2022), LipForensics (Haliassos et al., 2021), RealForensics (Haliassos et al., 2022), and CADDM (Dong et al., 2023) receive the whole 101-frame excerpt and run the authors' own preprocessing: FTCN uses the authors' detector, alignment, and aggregation (rates at its released threshold are in Appendix C); SBI uses RetinaFace (Deng et al., 2020) crops with the authors' margins; LipForensics and RealForensics use FAN landmarks (Bulat & Tzimiropoulos, 2017) and the authors' mouth crops; CADDM uses the authors' five-point alignment. Every pipeline returns its own video-level probability, and we report AUC on that probability and judged-fake rates at 0.5. Because these pipelines were released as end-to-end scripts, this track confirms the effect of each arm under each publisher's harness rather than under ours. No checkpoint was run on both tracks, so the tracks are not cross-calibrated; what can be said is that the pipeline response is largest on the native track (FTCN, SBI); the tracks hold different checkpoints, so no per-model consistency between them is claimed.

Separately reported checkpoints. Meso4, MesoInception, and CNN-Aug assign near-constant fake probabilities to our crops (their FaceForensics++ within-dataset AUCs under the crop rule are 0.51–0.84, far below the published numbers for the first two), so their pairwise AUCs sit at 0.49–0.78 with $\mathsf{F/S}$ at 0.49–0.56 and carry little information about the arms. They appear in Table 6 for completeness and in no summary statistic.

B ARM CONSTRUCTION AND VERIFICATION

Sources. We use the official FaceForensics++ split of 720/140/140 videos (360/70/70 manipulation pairs). Thirty videos (15 pairs) whose upstream YouTube source, according to the dataset's conversion table, also appears in another split are excluded, so that no upstream source crosses splits; this leaves 706 training, 128 validation, and 136 test real videos. Each video contributes 101 consecutive frames at 25 fps (about four seconds) at the original spatial extent of the c23 release; the same frames are used for every arm, so the arms of one source are aligned frame by frame.

Shared detection and donor map. For every frame of every source video one face is detected with the insightface `buffalo_l` detector (the five-point landmarks and the detection box are stored

Table 4: Checkpoints. FF++ AUC: Deepfakes / Face2Face / FaceSwap / NeuralTextures, video level, our crops. DFB: DeepfakeBench release (Yan et al., 2023b); DF40: the DF40 release (Yan et al., 2024).

Checkpoint	Weights	Training data (publisher)	Inference	FF++ AUC (DF/F2F/FS/NT)
Xception (Chollet, 2017)	DFB	FF++ c23	DFB code, our crops	0.985/0.992/0.985/0.964
EfficientNet-B4 (Tan & Le, 2019)	DFB	FF++ c23	DFB code	0.950/0.966/0.973/0.941
CapsuleNet (Nguyen et al., 2019)	DFB	FF++ c23	DFB code	0.913/0.892/0.911/0.805
CORE (Ni et al., 2022)	DFB	FF++ c23	DFB code	0.986/0.991/0.979/0.968
F3Net (Qian et al., 2020)	DFB	FF++ c23	DFB code	0.984/0.984/0.985/0.971
FFD (Dang et al., 2020)	DFB	FF++ c23	DFB code	0.974/0.971/0.977/0.939
SPSL (Liu et al., 2021)	DFB	FF++ c23	DFB code	0.971/0.969/0.978/0.898
SRM (Luo et al., 2021)	DFB	FF++ c23	DFB code	0.987/0.986/0.975/0.968
UCF (Yan et al., 2023a)	DFB	FF++ c23	DFB code (cls head)	0.987/0.989/0.989/0.958
RECCE (Cao et al., 2022)	DFB	FF++ c23	DFB code	0.987/0.982/0.969/0.973
Meso4, MesoInception (Afchar et al., 2018); CNN-Aug (Wang et al., 2020)	DFB	FF++ c23	DFB code	saturated
FSFM ViT-B (Wang et al., 2025)	authors (HF)	FF++ c23 fine-tune	authors' code	0.992/0.985/0.986/0.916
Effort CLIP-L/14 (Yan et al., 2025)	authors	FF++ c23 fine-tune	authors' DFB fork	0.994/0.949/0.984/0.843
Xception, EffNet-B4, F3Net (DF40-FS)	DF40	DF40 FS split (has SimSwap, InSwapper)	DFB code	.81–.87/.63–.71/.91–.99/.59–.64
CLIP-B/16 (DF40-FS)	DF40	DF40 FS split	DF40 clip detector (HF CLIP vision + linear head)	0.975/0.734/0.984/0.584
CLIP-L/14 (DF40-FS)	DF40	DF40 FS split	DF40 clip detector	0.985/0.841/0.989/0.608
Xception (DF40-all)	DF40	pooled DF40 training split	DFB code	0.853/0.714/0.906/0.728
CLIP-B/16 (DF40-all)	DF40	pooled DF40 training split	DF40 clip detector	0.926/0.751/0.958/0.701
FTCN (Zheng et al., 2021)	authors	FF++ c23	authors' code	full-frame
SBI (c23, raw) (Shiohara & Yamasaki, 2022)	authors	FF++ reals, self-blending	authors' code	full-frame
LipForensics (Haliassos et al., 2021)	authors	FF++ c23	authors' code	full-frame
RealForensics (FF++, all-but-FS) (Haliassos et al., 2022)	authors	FF++ c23	authors' code	full-frame
CADDM (EfficientNet-B3) (Dong et al., 2023)	authors	FF++ c23	authors' code	full-frame

per frame), and the same detection drives all four arms. The donor map $d(\cdot)$ is derived once from a random permutation of the 136 test videos (seed 303): each video's donor is its successor in the permuted order, self-mapping is excluded, and the same map is used for both generators, so each video serves once as a target and once as a donor; the training and validation videos use a map of their own split built the same way. The donor identity of F is taken from the first frame of the donor video with a detected face, and the self-identity of S from the first such frame of the same video, so S and F differ only in the identity vector handed to the generator.

SimSwap arms. SimSwap-224 (Chen et al., 2020) aligns the detected face to its 224 template through a similarity transform M, runs its generator on the aligned crop with the identity embedding of the reference face (its own ArcFace), and pastes the output back with a soft mask: the white face template is eroded by a 40×40 kernel and blurred with a 41×41 Gaussian, and the result is composited into the original frame through M^{-1}. N replaces the generator output by the aligned crop itself and runs the identical paste-back, so it contains exactly two warps, one feathered paste, and no synthesis; S and F run the generator with the self and donor identity respectively. All arms are written as lossless PNG frames (C0). No watermark is added and no JPEG intermediate exists anywhere in the pipeline.

InSwapper arms. InSwapper-128 (InsightFace contributors, 2023) uses the same `buffalo_l` detection, its own 128 alignment template, and the insightface paste-back (a white mask eroded and feathered with kernel sizes set by the face size). N, S, and F share this alignment and paste-back, N again skipping the generator. R is the same original frame as in the SimSwap set, so the two generator sets share their R arm and differ in the other three.

Encodings. C0 is the lossless PNG sequence. C23 re-encodes every arm's PNG sequence with the same ffmpeg command (libx264, CRF 23, yuv420p, 25 fps) and decodes it back to frames named as the source frames, so the four arms of a source pass through one identical lossy step; the AUCs under C23 in Table 1 and Table 8 compare arms that have all been compressed once.

Verification. Two checks are applied to every frame of every arm and stored with the arm metadata. First, the `buffalo_l` recogniser (w600k_r50) embeds the 112-pixel aligned faces of R, S, and F on the shared landmarks and we record the cosine similarity of S and F to R and of F to the donor; for SimSwap this recogniser is independent of the generator's own ArcFace, for InSwapper it is the embedding the generator conditions on, so the InSwapper check is a consistency check rather than an independent one. Second, the mean absolute difference between N and R over the full frame measures the footprint of the pipeline in grey levels. Table 5 summarises both checks over the 136 test videos. The self-swap keeps the identity (mean cosine 0.76 for SimSwap and 0.82 for InSwapper; two videos per generator fall below 0.4), the identity swap removes it (0.19 and 0.06; three SimSwap videos stay above 0.4) and moves towards the donor (0.56 and 0.89, the InSwapper value being high because the generator optimises exactly this embedding), and the pipeline footprint is 0.10 and 0.16 grey levels on average over the full frame with a 95th percentile below 0.6. The spatial support of the difference is fixed by construction rather than measured: the paste-back copies every pixel outside the soft mask unchanged, so N differs from R only within the mask's support, by the round trip of the aligned crop through the alignment; the full-frame mean measures the magnitude of that difference, not its location. Every frame of every test video has a detected face. Expression, gaze, and mouth motion are not verified; the released generators preserve them to the extent their designs allow, and any difference between S and F on these attributes is part of what F/S measures.

Table 5: Arm verification over the 136 test videos (per-video means of the per-frame values; cosine similarities of `buffalo_l` embeddings on the shared landmarks; the footprint is the mean absolute difference over the full frame on the 0–255 scale).

Generator	Quantity	mean	5th pct.	95th pct.	min	max
SimSwap	cos(S, R): self-swap vs original	0.76	0.63	0.85	0.29	0.86
SimSwap	cos(F, R): identity swap vs original	0.19	0.05	0.33	0.00	0.58
SimSwap	cos(F, donor): identity swap vs donor	0.56	0.43	0.68	0.08	0.71
SimSwap	mean —N − R— over the full frame (grey levels)	0.10	0.02	0.36	0.01	0.58
SimSwap	detector confidence of the shared detection	0.85	0.77	0.91	0.73	0.92
SimSwap	frames with a detected face	13736 of 13736 (136 videos × 101 frames)				
InSwapper	cos(S, R): self-swap vs original	0.82	0.71	0.91	0.18	0.93
InSwapper	cos(F, R): identity swap vs original	0.06	-0.04	0.18	-0.09	0.27
InSwapper	cos(F, donor): identity swap vs donor	0.89	0.84	0.93	0.75	0.95
InSwapper	mean —N − R— over the full frame (grey levels)	0.16	0.04	0.52	0.02	0.87
InSwapper	detector confidence of the shared detection	0.85	0.77	0.91	0.73	0.92
InSwapper	frames with a detected face	13736 of 13736 (136 videos × 101 frames)				

C METRICS, UNCERTAINTY, AND THRESHOLDS

Pairwise AUC. Every reported contrast X/Y is the area under the ROC curve of the video scores of arm X (treated as positive) against arm Y (negative) over the same 136 sources, computed with the rank statistic; ties count one half. The six pairwise contrasts of the four arms are defined on the same video scores but are not additive: F/R is not a function of N/R, S/N, and F/S, which is why Table 6 lists all six.

Paired win rate. W_{id} is the fraction of sources v whose identity swap scores above their own self-swap, $W_{\mathrm{id}} = \frac{1}{136}\sum_v \big([\, s(\mathsf{F}_v) > s(\mathsf{S}_v)\,] + \frac{1}{2}[\, s(\mathsf{F}_v) = s(\mathsf{S}_v)\,]\big)$, with the half weight for ties inside the sum. Because each source contributes exactly one comparison, W_{id} is the within-source counterpart of F/S: F/S compares every identity swap with every self-swap across sources, W_{id} only with the self-swap of the same source. A model with F/S close to 0.5 and W_{id} close to 0.5 ranks the identity swap above its own self-swap as often as below.

Bootstrap. Intervals are 95% percentile intervals from 2,000 resamples of the sources with replacement, using the number of sources stated for each analysis (136 for the face arms unless noted), and all arms of a source are drawn together, so the resampling unit is the source and the pairing between arms is preserved. The intervals therefore reflect the source-to-source variation of a fixed set of checkpoints and a fixed donor map; they do not model the variation that a different donor map or a re-run of a stochastic generator would introduce. Every interval is pointwise, a 95% interval for one contrast of one checkpoint, not a simultaneous guarantee over a table; the prespecified intervention contrasts are analysed on their own, and the subgroup and implementation-grid analyses are exploratory and

labelled as such. For the training interventions we report the dispersion across three seeds (sample standard deviation, $n = 3$) and source-level intervals, either for one seed or for the three-seed mean with one set of source indices shared by the fixed models (captions state which); no interval resamples training seeds, and the seed dispersion and the source intervals are different quantities.

Thresholds. Judged-fake rates use the fixed threshold 0.5 on each model's released probability (or 0 on the logit for the models we trained), without tuning on our data. For the full-frame pipeline that releases a threshold (FTCN, 0.0026 on its probability) we report the rate at 0.5 in the tables; at the released threshold FTCN judges 36% of the originals, 77% (SimSwap) and 96% (InSwapper) of the pipeline-only videos, and 99–100% of the swaps fake at C0, so the released threshold moves every rate up but leaves the ordering of the arms unchanged. The DF40-all checkpoints judge 68–71% of the originals fake at 0.5, so for them only the AUCs are interpretable; Table 6 still lists their rates.

Two labels. A_{syn} is the mean of the four contrasts that separate a synthesised arm from a non-synthesised one, $(\mathsf{S}/\mathsf{R} + \mathsf{S}/\mathsf{N} + \mathsf{F}/\mathsf{R} + \mathsf{F}/\mathsf{N})/4$, and A_{id} is the mean of the three that separate the identity swap from everything else, $(\mathsf{F}/\mathsf{R} + \mathsf{F}/\mathsf{N} + \mathsf{F}/\mathsf{S})/3$. They are reported for the trained models only, where the label contract is ours; for released checkpoints we report the pairwise contrasts and leave the aggregation to the reader.

D Full four-arm tables

Table 6 reports all six pairwise contrasts, W_{id}, judged-fake rates, and paired bootstrap intervals for the 22 frame-level checkpoints, including the three saturated DeepfakeBench checkpoints reported separately from the primary comparison; Tables 7 and 8 give the full-frame pipelines under C0 and C23. Values are recomputed from the stored per-video prediction files. Three regularities are visible in the full tables that the main text summarises. The pipeline contrast N/R is larger on InSwapper than on SimSwap for all 26 primary checkpoints, in line with the larger footprint of the InSwapper paste-back in Table 5. The re-rendering contrast S/N is highest for the DF40 checkpoints (0.79–0.99), which saw both generators in training, and for the video pipelines that read mouth motion (LipForensics and RealForensics, 0.85–0.97); it is lowest for the ten FaceForensics++ CNN and frequency checkpoints (0.55–0.74). The identity contrast F/S never exceeds 0.67 for the 21 checkpoints outside the five named in Section 4.3, and the lower end of its interval never exceeds 0.62 for them. Under C23 (Table 8) the drop of N/R is largest for FTCN and the two SBI models (0.15–0.23 on SimSwap, 0.11–0.20 on InSwapper), while LipForensics, RealForensics, and CADDM change by less than 0.05 in either direction (LipForensics down, CADDM up); so one common compression pass removes part of the pipeline response of the pipelines that had one, without removing their response to synthesis.

E Source-baseline centring

For each source v and model we subtract $b_v = (s(\mathsf{R}_v) + s(\mathsf{N}_v))/2$ from $s(\mathsf{S}_v)$ and $s(\mathsf{F}_v)$ and recompute F/S. The paired difference $s(\mathsf{F}_v) - s(\mathsf{S}_v)$ is unchanged, so W_{id} is identical by construction; only the cross-source ordering can change. The diagnostic asks whether the low F/S values of Section 4.3 are an artefact of source-dependent offsets: if a model scored every arm of an easy source low and every arm of a hard source high, identity swaps of easy sources would rank below self-swaps of hard sources and F/S would understate the within-source separation that W_{id} measures. Table 9 shows that this additive correction changes F/S by less than 0.04 for the ten FaceForensics++ frame-level models, FSFM, the three DF40 face-swap CNN and frequency models, LipForensics, and RealForensics, and lowers it for FTCN, SBI, CADDM, Effort, the DF40 CLIP checkpoints, and the DF40-all Xception (by up to 0.22; largest for the DF40-all CLIP-B/16 and the DF40-FS CLIP-L/14). The correction produces no large or consistent improvement in F/S (the largest increase is 0.033, FSFM on InSwapper); it does not rule out source-dependent scale or non-additive effects. For the CLIP checkpoints one reading of the decrease is that part of their cross-source F/S comes from sources whose four arms all score high, which the baseline removes; the diagnostic does not identify this uniquely, and their W_{id} (0.76–0.85) is the offset-free reading.

Table 6: Frame-level checkpoints, C0, video-level AUC (136 sources). Judged fake: fraction of R/N/S/F videos with mean probability above 0.5. Brackets: 95% paired bootstrap.

Model	Gen.	N/R	S/R	S/N	F/R	F/N	F/S [CI]	W_{id}	judged fake R/N/S/F
Xception	SimSwap	0.59	0.75	0.67	0.81	0.74	0.59 [0.54,0.63]	0.67	22/32/58/70%
Xception	InSwapper	0.69	0.86	0.73	0.87	0.75	0.51 [0.46,0.57]	0.49	22/51/76/83%
EfficientNet-B4	SimSwap	0.54	0.62	0.58	0.69	0.66	0.58 [0.54,0.62]	0.59	58/65/74/82%
EfficientNet-B4	InSwapper	0.69	0.81	0.66	0.84	0.71	0.54 [0.49,0.59]	0.51	58/82/95/95%
CapsuleNet	SimSwap	0.58	0.62	0.55	0.69	0.62	0.57 [0.52,0.61]	0.57	47/61/68/76%
CapsuleNet	InSwapper	0.67	0.75	0.60	0.78	0.61	0.50 [0.46,0.55]	0.49	47/71/81/86%
CORE	SimSwap	0.60	0.70	0.62	0.77	0.69	0.58 [0.53,0.62]	0.58	29/45/60/70%
CORE	InSwapper	0.72	0.87	0.72	0.87	0.73	0.50 [0.46,0.55]	0.51	29/66/89/86%
F3Net	SimSwap	0.59	0.64	0.56	0.72	0.65	0.58 [0.53,0.62]	0.60	22/32/43/52%
F3Net	InSwapper	0.70	0.84	0.71	0.84	0.69	0.47 [0.42,0.51]	0.46	22/46/76/76%
FFD	SimSwap	0.55	0.64	0.58	0.71	0.67	0.59 [0.54,0.64]	0.65	74/79/84/90%
FFD	InSwapper	0.59	0.74	0.67	0.75	0.68	0.51 [0.47,0.56]	0.52	74/90/94/91%
SPSL	SimSwap	0.51	0.61	0.61	0.70	0.69	0.61 [0.57,0.66]	0.68	68/70/88/91%
SPSL	InSwapper	0.56	0.61	0.55	0.68	0.63	0.59 [0.54,0.63]	0.57	68/78/87/93%
SRM	SimSwap	0.59	0.66	0.58	0.73	0.66	0.59 [0.55,0.63]	0.62	40/49/60/71%
SRM	InSwapper	0.68	0.79	0.65	0.83	0.69	0.53 [0.48,0.58]	0.50	40/65/79/85%
UCF	SimSwap	0.58	0.71	0.63	0.79	0.73	0.61 [0.57,0.65]	0.69	42/57/75/88%
UCF	InSwapper	0.68	0.84	0.72	0.87	0.75	0.52 [0.47,0.57]	0.57	42/71/90/94%
RECCE	SimSwap	0.61	0.70	0.60	0.76	0.68	0.59 [0.54,0.64]	0.64	31/47/59/69%
RECCE	InSwapper	0.73	0.89	0.74	0.89	0.74	0.49 [0.44,0.53]	0.53	31/64/89/89%
FSFM ViT-B	SimSwap	0.68	0.78	0.63	0.84	0.71	0.59 [0.55,0.64]	0.62	8/23/40/51%
FSFM ViT-B	InSwapper	0.81	0.92	0.72	0.94	0.75	0.52 [0.46,0.57]	0.53	8/43/74/74%
Effort CLIP-L/14	SimSwap	0.62	0.93	0.87	0.98	0.96	0.77 [0.73,0.82]	0.85	26/41/93/98%
Effort CLIP-L/14	InSwapper	0.69	0.91	0.80	0.98	0.93	0.79 [0.74,0.84]	0.81	26/56/90/97%
Xception (DF40-FS)	SimSwap	0.53	0.99	0.99	0.99	0.99	0.55 [0.51,0.59]	0.59	11/15/97/99%
Xception (DF40-FS)	InSwapper	0.59	0.95	0.91	0.97	0.94	0.56 [0.51,0.61]	0.54	11/21/88/91%
EffNet-B4 (DF40-FS)	SimSwap	0.54	0.95	0.94	0.95	0.94	0.51 [0.47,0.56]	0.54	21/25/96/97%
EffNet-B4 (DF40-FS)	InSwapper	0.60	0.91	0.85	0.90	0.86	0.53 [0.48,0.57]	0.47	21/32/88/88%
F3Net (DF40-FS)	SimSwap	0.54	0.99	0.98	0.99	0.99	0.57 [0.53,0.61]	0.61	7/10/97/99%
F3Net (DF40-FS)	InSwapper	0.61	0.95	0.91	0.97	0.94	0.58 [0.53,0.63]	0.65	7/15/81/90%
CLIP-B/16 (DF40-FS)	SimSwap	0.58	0.91	0.88	0.97	0.95	0.72 [0.67,0.77]	0.79	19/25/82/95%
CLIP-B/16 (DF40-FS)	InSwapper	0.65	0.87	0.79	0.95	0.91	0.73 [0.69,0.78]	0.79	19/32/80/90%
CLIP-L/14 (DF40-FS)	SimSwap	0.58	0.93	0.90	0.99	0.98	0.76 [0.71,0.81]	0.81	36/46/96/100%
CLIP-L/14 (DF40-FS)	InSwapper	0.64	0.88	0.81	0.98	0.95	0.74 [0.69,0.79]	0.76	36/53/88/100%
Xception (DF40-all)	SimSwap	0.54	0.99	0.98	0.99	0.99	0.53 [0.49,0.56]	0.57	71/75/100/100%
Xception (DF40-all)	InSwapper	0.64	0.94	0.88	0.97	0.94	0.63 [0.59,0.67]	0.73	71/85/100/100%
CLIP-B/16 (DF40-all)	SimSwap	0.60	0.97	0.94	0.99	0.99	0.71 [0.67,0.75]	0.80	68/81/100/100%
CLIP-B/16 (DF40-all)	InSwapper	0.71	0.92	0.79	0.98	0.92	0.72 [0.68,0.77]	0.78	68/94/99/100%
Meso4 (saturated)	SimSwap	0.51	0.51	0.50	0.52	0.51	0.51 [0.50,0.53]	0.51	100/100/100/100%
Meso4 (saturated)	InSwapper	0.52	0.53	0.51	0.52	0.49	0.49 [0.47,0.51]	0.49	100/100/100/100%
MesoInception (saturated)	SimSwap	0.63	0.64	0.51	0.70	0.58	0.56 [0.53,0.59]	0.64	88/96/97/97%
MesoInception (saturated)	InSwapper	0.74	0.77	0.53	0.78	0.53	0.50 [0.46,0.54]	0.49	88/100/100/99%
CNN-Aug (saturated)	SimSwap	0.56	0.60	0.55	0.64	0.59	0.54 [0.51,0.57]	0.57	100/100/100/100%
CNN-Aug (saturated)	InSwapper	0.61	0.62	0.52	0.62	0.52	0.51 [0.47,0.55]	0.49	100/100/100/100%

Table 7: Full-frame pipelines, C0 (lossless arms), video-level AUC: the contrasts of Table 1 together with S/R, F/R, and F/N.

Model	Gen.	N/R	S/R	S/N	F/R	F/N	F/S [CI]	W_{id}	judged fake R/N/S/F
FTCN	SimSwap	0.80	0.99	0.91	0.99	0.93	0.58 [0.56,0.61]	0.74	1/16/84/91%
FTCN	InSwapper	0.95	1.00	0.86	1.00	0.89	0.61 [0.58,0.64]	0.76	1/65/95/96%
SBI (c23)	SimSwap	0.89	1.00	0.96	1.00	0.98	0.64 [0.60,0.69]	0.73	1/26/93/99%
SBI (c23)	InSwapper	0.97	0.99	0.79	0.99	0.79	0.54 [0.49,0.58]	0.58	1/63/89/85%
SBI (raw)	SimSwap	0.88	0.99	0.93	1.00	0.98	0.67 [0.62,0.71]	0.76	1/28/92/97%
SBI (raw)	InSwapper	0.96	0.98	0.74	0.99	0.74	0.53 [0.49,0.58]	0.60	1/63/86/84%
LipForensics	SimSwap	0.56	0.88	0.85	0.91	0.89	0.55 [0.51,0.59]	0.62	0/0/15/19%
LipForensics	InSwapper	0.62	0.97	0.94	0.98	0.97	0.62 [0.58,0.66]	0.64	0/1/46/65%
RealForensics	SimSwap	0.55	0.94	0.93	0.96	0.95	0.57 [0.53,0.61]	0.65	7/9/77/84%
RealForensics	InSwapper	0.61	0.98	0.97	0.99	0.99	0.62 [0.58,0.66]	0.66	7/11/93/97%
RealForensics (all-but-FS)	SimSwap	0.53	0.92	0.91	0.94	0.93	0.55 [0.52,0.59]	0.57	0/1/35/43%
RealForensics (all-but-FS)	InSwapper	0.57	0.98	0.97	1.00	0.99	0.61 [0.57,0.65]	0.74	0/1/76/86%
CADDM	SimSwap	0.49	0.95	0.96	0.99	0.99	0.72 [0.67,0.77]	0.80	1/1/59/82%
CADDM	InSwapper	0.53	0.80	0.78	0.92	0.91	0.70 [0.65,0.75]	0.75	1/2/22/51%

Table 8: Full-frame pipelines, C23 (all four arms re-encoded with identical libx264 settings), video-level AUC.

Model	Gen.	N/R	S/R	S/N	F/R	F/N	F/S [CI]	W_{id}	judged fake R/N/S/F
FTCN	SimSwap	0.65	0.93	0.86	0.95	0.88	0.55 [0.52,0.57]	0.65	0/8/46/53%
FTCN	InSwapper	0.84	0.99	0.84	0.99	0.88	0.59 [0.57,0.63]	0.64	0/24/78/88%
SBI (c23)	SimSwap	0.67	0.88	0.76	0.93	0.85	0.64 [0.59,0.69]	0.74	1/5/22/42%
SBI (c23)	InSwapper	0.79	0.89	0.65	0.88	0.66	0.52 [0.47,0.57]	0.53	1/11/27/30%
SBI (raw)	SimSwap	0.65	0.84	0.72	0.91	0.84	0.66 [0.61,0.71]	0.75	1/4/17/39%
SBI (raw)	InSwapper	0.76	0.85	0.63	0.86	0.66	0.53 [0.48,0.57]	0.53	1/7/20/23%
LipForensics	SimSwap	0.53	0.86	0.84	0.88	0.86	0.53 [0.50,0.57]	0.55	0/0/10/14%
LipForensics	InSwapper	0.57	0.91	0.88	0.94	0.92	0.58 [0.54,0.63]	0.59	0/1/26/32%
RealForensics	SimSwap	0.56	0.92	0.89	0.93	0.91	0.55 [0.52,0.59]	0.57	6/12/71/80%
RealForensics	InSwapper	0.63	0.97	0.94	0.98	0.96	0.59 [0.55,0.64]	0.65	6/17/89/93%
RealForensics (all-but-FS)	SimSwap	0.54	0.88	0.86	0.89	0.87	0.52 [0.49,0.56]	0.54	1/1/24/26%
RealForensics (all-but-FS)	InSwapper	0.59	0.96	0.93	0.98	0.96	0.58 [0.54,0.61]	0.67	1/2/57/66%
CADDM	SimSwap	0.53	0.71	0.68	0.83	0.82	0.68 [0.63,0.73]	0.78	1/1/5/18%
CADDM	InSwapper	0.57	0.62	0.56	0.76	0.71	0.65 [0.60,0.70]	0.62	1/1/2/10%

Table 9: F/S before and after source-baseline centring (C0), three decimals; change = centred minus raw.

	SimSwap			InSwapper		
Model	raw	centred	change	raw	centred	change
Xception	0.585	0.582	-0.004	0.515	0.533	+0.018
EfficientNet-B4	0.581	0.572	-0.009	0.542	0.537	-0.005
CapsuleNet	0.566	0.593	+0.027	0.502	0.519	+0.018
CORE	0.579	0.579	-0.001	0.504	0.514	+0.010
F3Net	0.576	0.579	+0.004	0.468	0.484	+0.016
FFD	0.595	0.562	-0.033	0.515	0.512	-0.003
SPSL	0.611	0.593	-0.017	0.586	0.587	+0.002
SRM	0.588	0.593	+0.004	0.530	0.539	+0.009
UCF	0.608	0.597	-0.012	0.523	0.528	+0.005
RECCE	0.589	0.576	-0.013	0.486	0.509	+0.023
FSFM ViT-B	0.591	0.611	+0.020	0.521	0.554	+0.033
Effort CLIP-L/14	0.772	0.668	-0.104	0.792	0.693	-0.099
Xception (DF40-FS)	0.551	0.547	-0.004	0.558	0.569	+0.011
EffNet-B4 (DF40-FS)	0.514	0.515	+0.001	0.525	0.516	-0.009
F3Net (DF40-FS)	0.571	0.551	-0.019	0.580	0.601	+0.020
CLIP-B/16 (DF40-FS)	0.719	0.654	-0.066	0.735	0.666	-0.069
CLIP-L/14 (DF40-FS)	0.761	0.573	-0.188	0.741	0.616	-0.125
Xception (DF40-all)	0.535	0.501	-0.034	0.630	0.508	-0.122
CLIP-B/16 (DF40-all)	0.711	0.505	-0.206	0.724	0.509	-0.215
FTCN	0.580	0.574	-0.006	0.606	0.515	-0.091
SBI (c23)	0.642	0.555	-0.086	0.538	0.512	-0.026
SBI (raw)	0.667	0.593	-0.074	0.532	0.528	-0.005
LipForensics	0.551	0.576	+0.025	0.619	0.610	-0.009
RealForensics	0.569	0.596	+0.026	0.619	0.616	-0.003
RealForensics (all-but-FS)	0.550	0.572	+0.022	0.608	0.617	+0.009
CADDM	0.720	0.690	-0.029	0.700	0.683	-0.017

F EXPLORATORY STRATIFICATION BY DONOR SIMILARITY, AND DONOR DEPENDENCE

Stratification by donor similarity (exploratory). The identity swap of a source adopts the donor to a varying degree: the cosine similarity between the swapped face and the donor (Table 5) ranges over the sources from 0.08 to 0.71 for SimSwap (median 0.56) and is concentrated at 0.90 for InSwapper, which optimises this embedding. Table 10 computes F/S and W_{id} separately for the sources at or above and below the median donor similarity, for every primary checkpoint. The stratification is descriptive and was added after the main results were known. The similarity was never randomised; it depends on the target–donor pair, pose, and face quality as well as on the swap, and the two halves differ in their self-swaps as much as in their identity swaps, so F/S in a half is a property of both arms and the split identifies no mechanism. What the split shows is only that the higher-similarity half does not consistently yield higher F/S. On SimSwap the difference (high minus low) is negative for 20 of 26 checkpoints on the unrounded values (18 at the two-decimal precision of the table, where three pairs tie; mean -0.03); on InSwapper, where the split spans a narrow range of similarities, it is negative for 16 of 26 (one tie at table precision; mean -0.01). For the five checkpoints with F/S above

0.70, a subgroup selected on the same results, the SimSwap difference is -0.14 to -0.00 (CADDM 0.65 against 0.79, Effort 0.74 against 0.80) and the InSwapper difference is -0.08 to -0.01 on the unrounded values. The checkpoints share the sources and the split, and several share training data or backbones, so the sign pattern across them is one observation, not 26 independent ones, and the individual differences are within the interval width of a 68-source half. The split does not show a consistent rise of F/S with donor similarity; because the similarity and the composition of the halves are not controlled, this neither confirms nor rules out an identity-related mechanism, and the same caution applies to the last sentence of Section 4.3.

Table 10: F/S and W_{id} within the half of the sources with cosine similarity to the donor at or above the median (high) and below it (low), C0, 68 sources per half.

	SimSwap				InSwapper			
Model	F/S high	F/S low	W_{id} high	W_{id} low	F/S high	F/S low	W_{id} high	W_{id} low
Xception	0.58	0.59	0.66	0.68	0.47	0.55	0.44	0.53
EfficientNet-B4	0.59	0.57	0.57	0.60	0.53	0.55	0.46	0.56
CapsuleNet	0.54	0.60	0.51	0.62	0.49	0.51	0.52	0.46
CORE	0.56	0.59	0.53	0.63	0.48	0.53	0.50	0.51
F3Net	0.55	0.60	0.54	0.66	0.43	0.50	0.38	0.54
FFD	0.56	0.62	0.60	0.69	0.50	0.54	0.50	0.54
SPSL	0.61	0.61	0.65	0.71	0.57	0.59	0.57	0.56
SRM	0.58	0.59	0.60	0.63	0.52	0.53	0.53	0.47
UCF	0.58	0.63	0.66	0.72	0.51	0.54	0.56	0.59
RECCE	0.56	0.61	0.59	0.69	0.44	0.53	0.47	0.59
FSFM ViT-B	0.58	0.60	0.60	0.65	0.53	0.52	0.51	0.54
Effort CLIP-L/14	0.74	0.80	0.78	0.91	0.75	0.83	0.78	0.84
Xception (DF40-FS)	0.56	0.55	0.60	0.59	0.56	0.54	0.57	0.51
EffNet-B4 (DF40-FS)	0.51	0.51	0.49	0.59	0.53	0.51	0.50	0.44
F3Net (DF40-FS)	0.60	0.54	0.62	0.60	0.62	0.54	0.71	0.60
CLIP-B/16 (DF40-FS)	0.69	0.75	0.76	0.81	0.73	0.74	0.75	0.82
CLIP-L/14 (DF40-FS)	0.76	0.76	0.82	0.79	0.72	0.76	0.76	0.76
Xception (DF40-all)	0.56	0.51	0.60	0.54	0.60	0.66	0.75	0.71
CLIP-B/16 (DF40-all)	0.68	0.75	0.76	0.84	0.71	0.74	0.74	0.82
FTCN	0.55	0.61	0.69	0.78	0.61	0.60	0.75	0.78
SBI (c23)	0.65	0.64	0.70	0.76	0.54	0.53	0.54	0.62
SBI (raw)	0.66	0.68	0.75	0.77	0.54	0.52	0.59	0.60
LipForensics	0.53	0.57	0.59	0.66	0.65	0.59	0.66	0.62
RealForensics	0.55	0.59	0.66	0.63	0.65	0.60	0.72	0.60
RealForensics (all-but-FS)	0.52	0.57	0.56	0.59	0.61	0.61	0.72	0.76
CADDM	0.65	0.79	0.74	0.87	0.69	0.72	0.72	0.78

Donor dependence. Every source is the target of its own four arms and the donor of one other source's identity swap, so the 136 rows are not independent units and the paired bootstrap of Section 3, which resamples rows, does not carry the dependence through donors. As a sensitivity check we fix a subset in which no source is both a target row and the donor of another row (a maximal independent set of the donor assignment, chosen greedily in video-id order, 58 sources for each generator) and recompute the contrasts on it. Table 11 lists F/S and N/R on the full set and on the subset for every primary checkpoint. On the subset N/R moves by at most 0.04, S/N by at most 0.06, and F/S by at most 0.04 on SimSwap and 0.09 on InSwapper (CADDM, 0.70 to 0.61; the four other checkpoints above 0.70 stay above it); the number of checkpoints with N/R above 0.5 is unchanged. The fixed-support label gain of Section 5 (three-seed mean of $T_{\mathrm{id}} - T_{\mathrm{syn}}$ in F/S) is +0.42 on the full set and +0.44 [0.40,0.48] on the subset for SimSwap, and +0.02 against +0.03 [-0.05,0.11] for InSwapper. The subset removes the donor-side dependence within the resampled rows but not the fact that one donor assignment was used; the second assignment below addresses that.

Second donor assignment. To check whether the identity contrast and the fixed-support result depend on the particular pairing of targets and donors, we regenerated the identity-swap arm of the 136 test sources with a second permutation (seed 404; self-mapping excluded as before) for both generators, reusing R, N, and S, which do not depend on the donor, and the trained models, whose training arms are unchanged. Table 12 lists F/S and F/R under the two assignments for the 19 frame-level checkpoints (the full-frame pipelines were not re-run). F/S moves by at most 0.03 on SimSwap and 0.02 on InSwapper, within the width of the source intervals for every checkpoint, and F/R by at most 0.02 at the table's precision. The fixed-support label gain (three-seed mean of $T_{\mathrm{id}} - T_{\mathrm{syn}}$ in F/S, computed from unrounded values) is +0.42 under the first assignment and

Table 11: Donor-endpoint-disjoint subset (58 sources per generator, no source both a target row and the donor of another row): F/S and N/R on the full set (136) and on the subset, C0; brackets: 95% paired bootstrap over the subset's sources.

	SimSwap				InSwapper			
Model	F/S full	F/S subset	N/R full	N/R subset	F/S full	F/S subset	N/R full	N/R subset
Xception	0.59	0.59 [0.52,0.66]	0.59	0.61	0.51	0.55 [0.47,0.63]	0.69	0.71
EfficientNet-B4	0.58	0.57 [0.51,0.63]	0.54	0.55	0.54	0.52 [0.46,0.59]	0.69	0.71
CapsuleNet	0.57	0.54 [0.49,0.60]	0.58	0.58	0.50	0.46 [0.40,0.53]	0.67	0.67
CORE	0.58	0.55 [0.48,0.62]	0.60	0.61	0.50	0.54 [0.47,0.62]	0.72	0.73
F3Net	0.58	0.56 [0.49,0.64]	0.59	0.60	0.47	0.48 [0.40,0.56]	0.70	0.71
FFD	0.59	0.60 [0.52,0.68]	0.55	0.57	0.51	0.53 [0.46,0.60]	0.59	0.62
SPSL	0.61	0.58 [0.53,0.66]	0.51	0.51	0.59	0.56 [0.49,0.63]	0.56	0.56
SRM	0.59	0.59 [0.54,0.65]	0.59	0.59	0.53	0.53 [0.46,0.60]	0.68	0.68
UCF	0.61	0.59 [0.52,0.66]	0.58	0.59	0.52	0.54 [0.46,0.62]	0.68	0.69
RECCE	0.59	0.58 [0.51,0.65]	0.61	0.62	0.49	0.52 [0.45,0.60]	0.73	0.74
FSFM ViT-B	0.59	0.61 [0.55,0.68]	0.68	0.68	0.52	0.53 [0.44,0.61]	0.81	0.81
Effort CLIP-L/14	0.77	0.76 [0.68,0.83]	0.62	0.61	0.79	0.81 [0.74,0.88]	0.69	0.70
Xception (DF40-FS)	0.55	0.55 [0.48,0.62]	0.53	0.52	0.56	0.56 [0.49,0.64]	0.59	0.58
EffNet-B4 (DF40-FS)	0.51	0.51 [0.44,0.58]	0.54	0.54	0.53	0.48 [0.42,0.54]	0.60	0.59
F3Net (DF40-FS)	0.57	0.57 [0.51,0.64]	0.54	0.55	0.58	0.58 [0.50,0.66]	0.61	0.63
CLIP-B/16 (DF40-FS)	0.72	0.74 [0.67,0.81]	0.58	0.58	0.73	0.75 [0.67,0.83]	0.65	0.66
CLIP-L/14 (DF40-FS)	0.76	0.77 [0.69,0.84]	0.58	0.58	0.74	0.79 [0.74,0.85]	0.64	0.64
Xception (DF40-all)	0.53	0.52 [0.47,0.57]	0.54	0.54	0.63	0.60 [0.53,0.66]	0.64	0.64
CLIP-B/16 (DF40-all)	0.71	0.70 [0.63,0.77]	0.60	0.59	0.72	0.73 [0.65,0.80]	0.71	0.72
FTCN	0.58	0.59 [0.56,0.63]	0.80	0.76	0.61	0.59 [0.56,0.64]	0.95	0.91
SBI (c23)	0.64	0.63 [0.57,0.71]	0.89	0.88	0.54	0.56 [0.49,0.63]	0.97	0.97
SBI (raw)	0.67	0.66 [0.60,0.73]	0.88	0.86	0.53	0.55 [0.48,0.62]	0.96	0.95
LipForensics	0.55	0.55 [0.48,0.62]	0.56	0.56	0.62	0.60 [0.54,0.67]	0.62	0.61
RealForensics	0.57	0.61 [0.55,0.67]	0.55	0.54	0.62	0.62 [0.55,0.69]	0.61	0.60
RealForensics (all-but-FS)	0.55	0.57 [0.51,0.63]	0.53	0.52	0.61	0.59 [0.55,0.64]	0.57	0.55
CADDM	0.72	0.71 [0.64,0.78]	0.49	0.49	0.70	0.61 [0.54,0.69]	0.53	0.53

+0.37 under the second on SimSwap (T_{syn} 0.56 to T_{id} 0.94; the joint identity head reaches 0.95), and +0.02 against +0.00 on InSwapper (0.64 to 0.64). The large gain on the training generator is therefore present under both assignments, and under the second assignment no confirmed gain on the unseen generator is observed either. The first assignment remains the pre-specified analysis and the second is a sensitivity check: the lower F/S of T_{id} under it (0.94 against 0.97) is the sensitivity observed between the two assignments tested, not an estimate of the variation over donor assignments in general.

Table 12: Second donor assignment (seed 404): F/S and F/R of the 19 frame-level checkpoints under the first (1) and the second (2) assignment of donors to the 136 test sources, C0; brackets: 95% paired bootstrap over sources for F/S under the second assignment.

	SimSwap				InSwapper			
Model	F/S (1)	F/S (2)	F/R (1)	F/R (2)	F/S (1)	F/S (2)	F/R (1)	F/R (2)
Xception	0.59	0.57 [0.53,0.62]	0.81	0.81	0.51	0.51 [0.46,0.56]	0.87	0.87
EfficientNet-B4	0.58	0.58 [0.54,0.62]	0.69	0.69	0.54	0.53 [0.48,0.57]	0.84	0.83
CapsuleNet	0.57	0.57 [0.52,0.61]	0.69	0.69	0.50	0.52 [0.47,0.56]	0.78	0.78
CORE	0.58	0.57 [0.53,0.62]	0.77	0.77	0.50	0.50 [0.45,0.54]	0.87	0.88
F3Net	0.58	0.59 [0.55,0.62]	0.72	0.72	0.47	0.46 [0.41,0.50]	0.84	0.82
FFD	0.59	0.60 [0.55,0.65]	0.71	0.71	0.51	0.52 [0.47,0.57]	0.75	0.76
SPSL	0.61	0.60 [0.55,0.65]	0.70	0.69	0.59	0.60 [0.55,0.64]	0.68	0.69
SRM	0.59	0.57 [0.53,0.61]	0.73	0.72	0.53	0.51 [0.46,0.56]	0.83	0.81
UCF	0.61	0.60 [0.55,0.65]	0.79	0.79	0.52	0.52 [0.47,0.58]	0.87	0.87
RECCE	0.59	0.57 [0.53,0.62]	0.76	0.76	0.49	0.47 [0.43,0.52]	0.89	0.87
FSFM ViT-B	0.59	0.59 [0.55,0.63]	0.84	0.83	0.52	0.52 [0.47,0.57]	0.94	0.93
Effort CLIP-L/14	0.77	0.75 [0.71,0.80]	0.98	0.97	0.79	0.77 [0.73,0.82]	0.98	0.97
Xception (DF40-FS)	0.55	0.56 [0.52,0.60]	0.99	0.99	0.56	0.57 [0.52,0.62]	0.97	0.96
EffNet-B4 (DF40-FS)	0.51	0.51 [0.46,0.56]	0.95	0.96	0.53	0.52 [0.46,0.57]	0.90	0.92
F3Net (DF40-FS)	0.57	0.59 [0.54,0.63]	0.99	0.99	0.58	0.58 [0.54,0.63]	0.97	0.97
CLIP-B/16 (DF40-FS)	0.72	0.72 [0.67,0.76]	0.97	0.97	0.73	0.73 [0.68,0.77]	0.95	0.96
CLIP-L/14 (DF40-FS)	0.76	0.79 [0.74,0.83]	0.99	0.99	0.74	0.75 [0.70,0.80]	0.98	0.98
Xception (DF40-all)	0.53	0.55 [0.51,0.59]	0.99	0.99	0.63	0.61 [0.57,0.65]	0.97	0.96
CLIP-B/16 (DF40-all)	0.71	0.73 [0.69,0.77]	0.99	0.99	0.72	0.70 [0.66,0.74]	0.98	0.97

G A THIRD FACE-SWAP GENERATOR: GHOST

As a check that the two-generator picture of Section 4 is not tied to SimSwap and InSwapper, we built the four arms of the same 136 sources with a third released generator, GHOST (Groshev et al., 2022) (the released U-Net variant `G_unet_2blocks`: backbone `unet`, two blocks per AAD residual block, its own ArcFace identity encoder, 224-pixel arcface-template alignment, and a landmark-mask paste-back), which is absent from the FaceForensics++ and DF40 training data of every checkpoint; held out refers to these documented forgery-generation methods used for detector training. Its generator follows the FaceShifter design; FaceShifter outputs form a separate subset of the FaceForensics++ release, outside the four manipulations of the FaceForensics++ training sets (the DeepfakeBench configurations list it as a separate dataset), and neither method is among the 40 of DF40. The detection and the donor map are shared with the other two generators; the generator's own alignment, identity embedding, generator (crops resized to 256 for the network and back to 224, as released), and paste-back (106-point landmark convex hull, eroded and blurred with the released schedule chosen from the landmark offset between the swapped and the target crop) are used for N (aligned original crop passed through the same paste-back, offset 0), S (source: the same person's reference crop), and F (source: the donor's reference crop). N, S, and F thus use the same paste-back algorithm, but their realised masks are not held fixed across arms, so the GHOST S/N and F/S contrasts include changes in the blending mask. All 136 sources and all 13,736 frames were built; the same-stack identity check gives mean cosine 0.72 between S and R, 0.18 between F and R, and 0.53 between F and the donor, with F closer to the donor than to the original for 135 of 136 sources. Scoring, metrics, and the 32-frame sampling follow Section 4. The original arm is shared with the other two generators; re-scoring it in this run reproduces the per-video scores of the main comparison exactly for six of the seven full-frame pipelines, and within 0.002 (frame-level checkpoints) and 0.009 (FTCN) in probability, differences consistent with floating-point effects of batch composition; the GHOST contrasts use the R scores of this run.

Table 13 places the GHOST contrasts next to the SimSwap and InSwapper values. On the 19 primary frame-level checkpoints, N/R exceeds 0.5 for all 19 (0.54–0.72; SimSwap 0.51–0.68, InSwapper 0.56–0.81), F/S is 0.46–0.71 (SimSwap 0.51–0.77, InSwapper 0.47–0.79) and lies below F/R for 18 of the 19 (the exception is CapsuleNet, whose F/R on GHOST is 0.46), and the four CLIP-based checkpoints (Effort and the three DF40 CLIP models) remain the highest at 0.68–0.71, as on the other two generators. The seven DF40 checkpoints separate the GHOST self-swaps from its pipeline-only outputs at S/N 0.87–0.99 despite the absence of GHOST from their documented detector-training methods, while their F/S stays at 0.52–0.71, below their F/R of 0.97–0.99. Among the seven full-frame pipelines (Table 13, lower block; C0), N/R is 0.52–0.98 (SimSwap 0.49–0.89, InSwapper 0.53–0.97); the two SBI models judge 76% and 81% of the GHOST pipeline-only videos fake against 1% of the originals (SimSwap 26% and 28%, InSwapper 63% for both), and 11% and 6% after the common C23 re-encoding. F/S is 0.49–0.60 against F/R 0.99–1.00: CADDM, which reaches 0.72 and 0.70 on SimSwap and InSwapper, falls to 0.59, and LipForensics and both RealForensics models separate the self-swaps from the pipeline-only arm at S/N 1.00 while their F/S stays at 0.52–0.60.

H POST-PROCESSING FLIPS AND FOREIGN REAL SETS

Procedure. Figure 3 plots the decision flips summarised in Section 4.4 and Table 14 lists them. For the frame-level checkpoints the input is the FaceForensics++ c23 test set under the crop rule of Appendix A: 136 real videos and 136 videos of each manipulation, eight uniformly sampled frames per video (the four-arm analyses and Table 18 use 32 frames per source). Blur is a Gaussian filter with standard deviation 2 or 4 pixels applied to the 299-pixel crop before the model's own resize; JPEG uses quality 30 on the crop with the OpenCV encoder. A frame counts as flipped when its decision at threshold $\tau = 0.5$ changes between the clean and the processed crop, and the denominator is the set of frames whose clean decision was correct: for blur the table reports $p^{\mathrm{R}}_{\mathrm{flip}} = \#\{s(R) < \tau,\, s(T(R)) \geq \tau\}/\#\{s(R) < \tau\}$ over real frames, and for JPEG the analogous fraction of NeuralTextures and Deepfakes fake frames with $s(F) \geq \tau$ that receive $s(T(F)) < \tau$; these are not the fractions of all real or all fake frames. For the full-frame pipelines the processing is applied to every frame of the excerpt before the authors' preprocessing and the rates are video-level judged-fake rates before and after processing, on our SimSwap arms (R for blur, F for JPEG).

Table 13: Three face-swap generators on the same 136 sources (C0; video-level AUC over sources, mean per-frame probability over 32 frames for the frame-level checkpoints, the pipelines' own aggregates below the rule). S = SimSwap, I = InSwapper, G = GHOST. The GHOST columns F/R and judged-N give the conventional contrast and the share of pipeline-only videos judged fake at the fixed threshold. The three low-discrimination DeepfakeBench checkpoints are listed after the primary ones.

	N/R			S/N			F/S			F/R	judged N
Model	S	I	G	S	I	G	S	I	G	G	G
Xception	0.59	0.69	0.67	0.67	0.73	0.75	0.59	0.51	0.46	0.84	0.49
EfficientNet-B4	0.54	0.69	0.59	0.58	0.66	0.73	0.58	0.54	0.52	0.81	0.72
CapsuleNet	0.58	0.67	0.58	0.55	0.60	0.40	0.57	0.50	0.48	0.46	0.62
CORE	0.60	0.72	0.65	0.62	0.72	0.64	0.58	0.50	0.49	0.76	0.56
F3Net	0.59	0.70	0.64	0.56	0.71	0.67	0.58	0.47	0.47	0.77	0.39
FFD	0.55	0.59	0.61	0.58	0.67	0.56	0.59	0.51	0.47	0.63	0.85
SPSL	0.51	0.56	0.61	0.61	0.55	0.63	0.61	0.59	0.50	0.73	0.79
SRM	0.59	0.68	0.64	0.58	0.65	0.66	0.59	0.53	0.51	0.78	0.57
UCF	0.58	0.68	0.65	0.63	0.72	0.64	0.61	0.52	0.53	0.78	0.69
RECCE	0.61	0.73	0.66	0.60	0.74	0.61	0.59	0.49	0.48	0.74	0.57
FSFM ViT-B	0.68	0.81	0.68	0.63	0.72	0.76	0.59	0.52	0.54	0.91	0.23
Effort CLIP-L/14	0.62	0.69	0.72	0.87	0.80	0.87	0.77	0.79	0.71	0.99	0.57
Xception (DF40-FS)	0.53	0.59	0.54	0.99	0.91	0.99	0.55	0.56	0.55	0.99	0.15
EffNet-B4 (DF40-FS)	0.54	0.60	0.56	0.94	0.85	0.96	0.51	0.53	0.53	0.98	0.26
F3Net (DF40-FS)	0.54	0.61	0.57	0.98	0.91	0.98	0.57	0.58	0.53	0.99	0.11
CLIP-B/16 (DF40-FS)	0.58	0.65	0.60	0.88	0.79	0.91	0.72	0.73	0.70	0.97	0.28
CLIP-L/14 (DF40-FS)	0.58	0.64	0.60	0.90	0.81	0.94	0.76	0.74	0.71	0.99	0.47
Xception (DF40-all)	0.54	0.64	0.59	0.98	0.88	0.97	0.53	0.63	0.52	0.98	0.79
CLIP-B/16 (DF40-all)	0.60	0.71	0.58	0.94	0.79	0.87	0.71	0.72	0.68	0.97	0.82
Meso4	0.51	0.52	0.51	0.50	0.51	0.45	0.51	0.49	0.52	0.48	1.00
MesoInception	0.63	0.74	0.63	0.51	0.53	0.49	0.56	0.50	0.52	0.64	0.96
CNN-Aug	0.56	0.61	0.56	0.55	0.52	0.41	0.54	0.51	0.51	0.48	1.00
FTCN	0.80	0.95	0.70	0.91	0.86	0.99	0.58	0.61	0.55	1.00	0.09
SBI (c23)	0.89	0.97	0.98	0.96	0.79	0.88	0.64	0.54	0.49	1.00	0.76
SBI (raw)	0.88	0.96	0.98	0.93	0.74	0.87	0.67	0.53	0.50	1.00	0.81
LipForensics	0.56	0.62	0.55	0.85	0.94	1.00	0.55	0.62	0.52	1.00	0.00
RealForensics (FF++)	0.55	0.61	0.54	0.93	0.97	1.00	0.57	0.62	0.57	1.00	0.07
RealForensics (all but FS)	0.53	0.57	0.52	0.91	0.97	1.00	0.55	0.61	0.60	1.00	0.01
CADDM (B3)	0.49	0.53	0.58	0.96	0.78	0.97	0.72	0.70	0.59	0.99	0.04

Reading. Blur removes high-frequency content and drives real-to-fake flips: at $\sigma = 4$ the FaceForensics++ CNN and frequency checkpoints flip nearly every real frame, and the three CLIP models, whose input resolution is 224 and whose features are coarser, flip 61–77%. Under JPEG 30 the FaceForensics++ CNN and frequency models flip 15–78% of the NeuralTextures frames to real (FSFM 96%, Effort 48%), while the DF40 checkpoints, trained on generator outputs that were themselves compressed, flip 0–5%. The same asymmetry appears at video level: FTCN and SBI, the two pipelines with the largest pipeline response in Table 1, also move the most under blur (81% and 90% of real videos judged fake at $\sigma = 2$), and RealForensics moves from 7% to 64%; LipForensics, which reads mouth motion, moves least (0% to 10%). Under the synthesis contract of Section 3 the blur flips are false alarms and the JPEG flips are misses; the direction and size of the flips are consistent with the pipeline response being a response to the high-frequency statistics of the input.

Foreign real sets. 587 Celeb-DF real videos (Li et al., 2020) and 300 YouTube real videos from the same collection are scored with our FaceForensics++-trained Xception (seed 303), an EfficientNet-B4 trained the same way, and an SBI model trained on the FaceForensics++ reals with the authors' recipe; video-level judged-fake rates at threshold 0 on the logit are 0.0% / 3.6% / 11.7% (Xception), 3.7% / 27.8% / 27.3% (EfficientNet-B4), and 3.7% / 93.0% / 89.7% (SBI) for FaceForensics++ / Celeb-DF / YouTube reals. The SBI model, whose positives are blended real faces and whose negatives are FaceForensics++ reals, judges nine out of ten foreign reals fake without any manipulation being present; the two supervised models judge 3.6–27.8% of them fake. These rates are the foreign-domain counterpart of the pipeline response: a change of source distribution, which alters content, subjects, and capture conditions together, moves the decision without any manipulation being present.

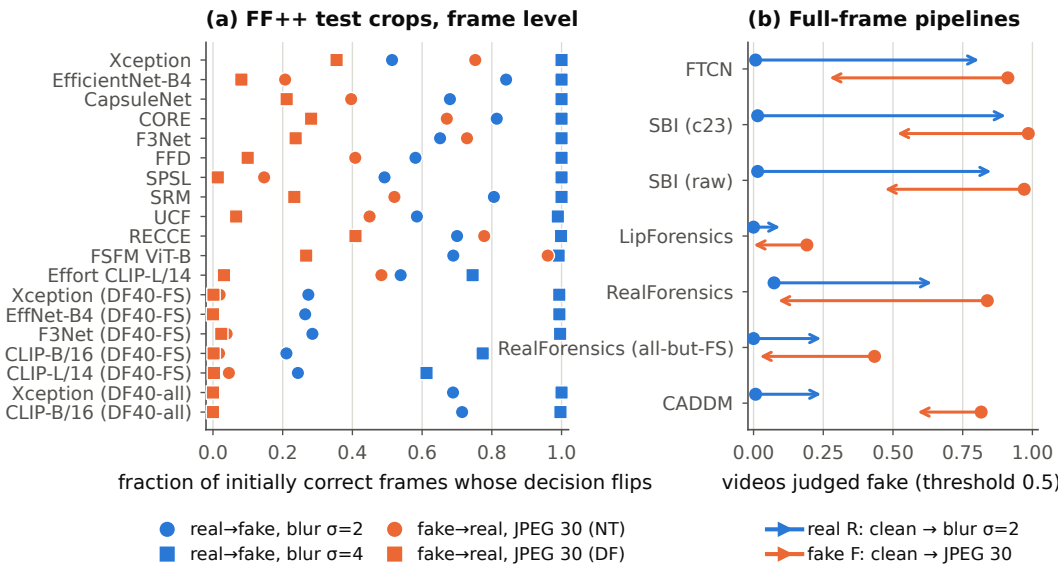

Figure 3: **Non-generative processing flips decisions.** (a) Fraction of initially correctly classified FaceForensics++ test frames whose decision flips: real to fake after Gaussian blur ($\sigma = 2$ circles, $\sigma = 4$ squares) and fake to real after JPEG quality 30 (NeuralTextures circles, Deepfakes squares). (b) Full-frame pipelines at video level: arrows run from clean to processed; blue is the fraction of real videos judged fake, orange the fraction of identity-swap videos judged fake.

Table 14: Decision flips on FaceForensics++ test crops (frame level) and on full-frame pipelines (video level): real→fake under Gaussian blur (σ 2 and 4), fake→real under JPEG quality 30 for NeuralTextures (NT) and Deepfakes (DF); denominators as defined in the procedure above.

Model	blur σ=2	blur σ=4	JPEG 30, NT	JPEG 30, DF
Xception (frames)	0.51	1.00	0.75	0.35
EfficientNet-B4 (frames)	0.84	1.00	0.21	0.08
CapsuleNet (frames)	0.68	1.00	0.40	0.21
CORE (frames)	0.81	1.00	0.67	0.28
F3Net (frames)	0.65	1.00	0.73	0.24
FFD (frames)	0.58	1.00	0.41	0.10
SPSL (frames)	0.49	1.00	0.15	0.01
SRM (frames)	0.81	1.00	0.52	0.23
UCF (frames)	0.59	0.99	0.45	0.07
RECCE (frames)	0.70	1.00	0.78	0.41
FSFM ViT-B (frames)	0.69	0.99	0.96	0.27
Effort CLIP-L/14 (frames)	0.54	0.74	0.48	0.03
Xception (DF40-FS) (frames)	0.27	0.99	0.02	0.00
EffNet-B4 (DF40-FS) (frames)	0.26	0.99	0.00	0.00
F3Net (DF40-FS) (frames)	0.29	1.00	0.04	0.02
CLIP-B/16 (DF40-FS) (frames)	0.21	0.77	0.02	0.00
CLIP-L/14 (DF40-FS) (frames)	0.24	0.61	0.05	0.00
Xception (DF40-all) (frames)	0.69	1.00	0.00	0.00
CLIP-B/16 (DF40-all) (frames)	0.71	1.00	0.00	0.00
FTCN (videos judged fake, clean→proc.)	0.01→0.81	0.01→0.92	F: 0.91→0.27	
SBI (c23) (videos judged fake, clean→proc.)	0.01→0.90	0.01→0.68	F: 0.99→0.51	
SBI (raw) (videos judged fake, clean→proc.)	0.01→0.85	0.01→0.68	F: 0.97→0.47	
LipForensics (videos judged fake, clean→proc.)	0.00→0.10	0.00→0.35	F: 0.19→0.00	
RealForensics (videos judged fake, clean→proc.)	0.07→0.64	0.07→0.85	F: 0.84→0.09	
RealForensics (all-but-FS) (videos judged fake, clean→proc.)	0.00→0.24	0.00→0.71	F: 0.43→0.02	
CADDM (videos judged fake, clean→proc.)	0.01→0.24	0.01→0.68	F: 0.82→0.59	

H.1 Sharpness as a cue: within-video variation, the pipeline arm, and region-selective blur

The flips above show that non-generative processing moves decisions; this section asks, with inference on the released checkpoints only, which sharpness-related quantity the scores follow. It is an

exploratory analysis added after the main results. We measure sharpness proxies on the common 299-pixel C0 crops before checkpoint-specific resizing and normalisation: q is the natural logarithm of one plus the variance of the 3×3 Laplacian of the grey crop inside the central square of 60% side length, and c is q minus the same quantity on the ring outside the central square of 80% side length; the square and the ring are fixed geometric proxies for the face interior and the surrounding context, not landmark-defined regions or verified complements of the paste-back mask. The pixel change of an arm is the RMS difference, over all pixels and channels of the crop, to the same-index N frame. Every checkpoint scores the same 32 frames per arm as in Section 4 under its released recipe; the fake logit is $\ell = \log(\tilde{p}/(1-\tilde{p}))$ with the per-frame probability $\tilde{p}$ clipped to $[10^{-6}, 1-10^{-6}]$, and video AUROCs recomputed from the mean per-frame probabilities differ from the reported values by at most 0.0005 over the 168 four-arm contrasts of the 21 re-scored checkpoints. Correlations are computed within each source over its 32 frames and summarised over the 136 sources with a source-level bootstrap.

Natural variation within a video. Across sources, the mean within-video Spearman correlation between the fake logit and q on R is negative for all 19 primary frame-level checkpoints, ranging from -0.30 to -0.13, with every source-bootstrap interval below zero (Table 17); on the self-swaps the mean correlation is negative for 19 of 19 (SimSwap) and 17 of 19 (InSwapper) checkpoints.

What the arms do to sharpness. The pipeline arm reduces the sharpness proxy in the face interior: the per-source median change of q from R to N is negative for 99% of the sources, with a median over sources of -0.40 (SimSwap) and -0.62 (InSwapper), and c changes by -0.25 and -0.23, so the interior loses more of the proxy than the ring. This relative reduction in inner-region sharpness is consistent with the local resolution-mismatch motivation of Li & Lyu (2019). The generators add a smaller change (median change of q from N to S -0.11 and -0.03, to F -0.18 and -0.04); of the frames whose original crop lies in the lowest tercile of q (pooled over the sources of a generator), 24% (SimSwap) and 40% (InSwapper) have a self-swap with a higher q than the pipeline-only frame. On SimSwap, the self-swap frames the generator sharpened receive a smaller mean score increase over N than the frames it blurred for all 19 checkpoints (Table 17; for F3Net, SPSL, and EfficientNet-B4 the sharpened frames are scored like N or below it). On InSwapper the two subsets show no uniform ordering across checkpoints (the sharpened subset is higher for 11 of 19) and the sharpened subset has a positive mean S$-$N change for all 19; a decrease in measured q is therefore not required for a positive re-rendering response.

Region-selective blur. For six checkpoints fixed before this analysis (Xception, SPSL, SRM, FSFM, and the CLIP-B/16 checkpoints trained on DF40-FS and on DF40-all) we blur the whole crop, only the central square, or only the outer ring (masks feathered with a Gaussian of standard deviation 8 pixels), at $\sigma \in \{1, 2, 4\}$, and apply JPEG at quality 90, 60, and 30, on the four arms and on the FaceForensics++ real and NeuralTextures crops, re-measuring q and c on the processed crops. At both reported blur strengths, inner-region blur produces the largest mean fake-score increase for all six checkpoints (Table 18): at $\sigma = 2$ on real frames, Xception moves by +1.5, +7.5, and +0.5 logits (whole, inner, outer), SPSL by -0.0, +5.3, and +0.4, FSFM by +2.2, +3.4, and -0.0, and the DF40-FS CLIP-B/16 by +2.1, +10.0, and -0.4; outer-ring blur stays below whole-crop blur for five of the six (all but SPSL). This spatial selectivity is consistent with sensitivity to inner–outer sharpness differences, the relation the pipeline arm alters, but it does not isolate that relation from the spatial support and transition boundaries of the blur. JPEG behaves differently: in our measurements it raises the Laplacian proxy (median change of q +0.6 at quality 30 on real crops) while leaving the mean gradient magnitude almost unchanged, so q does not track physical sharpness under JPEG, and the tested Gaussian/JPEG grid yielded no eligible q-matched pairs under the matching rule (same frame, changes of q agreeing within 0.05); it therefore does not support a matched-q explanation of the JPEG-induced flips. The direction of the JPEG response differs across these checkpoints and their training groups: JPEG 30 flips 75%, 52%, 96%, and 15% of the NeuralTextures frames to real for Xception, SRM, FSFM, and SPSL and 0–2% for the two DF40 CLIP checkpoints (Table 14), whose fake scores on real frames rise under it instead (26% and 89% of initially real-judged frames flip at quality 30, Table 18).

Feature-space displacement. For the six checkpoints we also record the representation the released classifier reads, the input of its final linear layer (2048 dimensions for Xception, SPSL, and SRM;

768 for FSFM and the two CLIP-B/16 checkpoints), on the same frames, together with that layer's weights; the released probability is recovered from the stored vector to within 8×10^{-4} for FSFM, whose released inference runs in bfloat16, and 3×10^{-7} for the others. All released detector parameters remain frozen; only the diagnostic linear probes described below are fitted. Displacements are Euclidean distances between the stored vectors of the same frame index. On SimSwap, face-interior blur ($\sigma = 4$) of a real frame produces a median displacement of 2.6–4.6 times the median displacement of the identity swap for the four FaceForensics++-trained checkpoints and comparable magnitudes for the two DF40 CLIP-B/16 checkpoints (1.1 and 1.0); on InSwapper the ratios are 1.9–3.4 and 1.1, 1.0 (Tables 15 and 16). On SimSwap, the pipeline arm displaces the vector by 0.1–0.4 and the change between consecutive sampled frames of the same real video by 0.3–0.6 of the swap displacement. Source-mean blur and swap displacements are positively aligned, with median cosine similarities of 0.70–0.79 (consecutive-frame changes: -0.16 to 0.19), and in the full feature space 99–100% of the face-blurred real frames lie closer to the F centroid than to the R centroid, against 68–98% of the F frames themselves (Figure 4).

Blur-derived directions and the released read-out. We estimate directions of the representation from real frames and their blurred versions only, without labels: the rows of a displacement matrix are the per-frame displacements under face-interior blur at $\sigma = 2$ and $\sigma = 4$ (both strengths stacked with equal weight, one row per frame, no centring, no normalisation), and U_k holds its top-k right singular vectors. With $P_k = U_k U_k^\top$, the released read-out $z(\mathbf{f}) = \mathbf{w}^\top \mathbf{f} + b$, where $\mathbf{w}$ and b are the differences between the two class rows of the released linear layer, is restricted to $z_{\text{only}} = \mathbf{w}^\top P_k \mathbf{f} + b$ or to $z_{\text{remove}} = \mathbf{w}^\top (I - P_k)\mathbf{f} + b$ with the original coefficients. Video scores are means of z over the 32 frames, and all released-read-out AUROCs in Tables 15 and 16 use the split-half protocol: U_k is estimated on half of the sources and the AUROC computed on the other half, averaged over the two assignments (the probe column uses source-level five-fold cross-validation); the full read-out is evaluated under the same protocol and therefore differs slightly from the mean-probability values of Table 6. Restricting each released read-out to the single leading blur-derived direction changes the held-out F/R AUROC by at most 0.03 across the six checkpoints and both generators. On SimSwap, the leading direction of natural within-video change (consecutive sampled frames of the same real video) retains it within 0.02 for 5 of the six (SPSL: 0.74 against 0.67), so this result is not specific to blur-derived directions; removing the top four blur-derived directions lowers the AUROC to 0.45–0.56 for the four FaceForensics++-trained checkpoints and to 0.79 and 0.91 for the DF40 CLIP-B/16 checkpoints (InSwapper: 0.33–0.54, and 0.79 and 0.88, Table 16), removing four random directions changes it by at most 0.01, and removing the top four natural-change directions also changes it, by amounts that differ across checkpoints (0.36–0.60; 0.90 and 0.66). On the FaceForensics++ real and NeuralTextures crops, the leading blur-derived direction retains the real-versus-NeuralTextures AUROC of Xception, SPSL, SRM, and FSFM (0.97, 0.88, 0.97, 0.91 against 0.96, 0.88, 0.97, 0.92), and removing the top four lowers it to 0.46, 0.33, 0.10, 0.45 (values below 0.5 mean that the remaining component of the weight vector ranks the fakes below the reals: the discrimination along the original direction falls or reverses); the two DF40 CLIP-B/16 checkpoints separate these crops weakly (0.59, 0.72) and change little. A linear probe refit on the complement of the top 32 blur-derived directions, with source-level five-fold cross-validation in which the directions, the standardisation, the 128-component PCA, and the logistic regression are all fitted on the training-fold sources, separates F from R with AUROC 0.87–0.99 (SimSwap; 0.86–0.99 InSwapper): F/R information remains linearly decodable outside the top 32 blur-derived directions. This does not establish that the residual information is independent of sharpness or unused by the released read-out.

On SimSwap, the remaining 12 primary checkpoints with a linear read-out (Table 15, lower block) show the same pattern with one exception. Their blur displacement ratios are 0.8–7.9; for 11 of the 12 the source-mean blur and swap displacements are positively aligned (median cosine 0.56–0.99), 89–100% of the blurred frames lie nearer the F centroid, and the single leading blur-derived direction retains the held-out F/R AUROC within 0.02; removing the top four blur-derived directions lowers the AUROC for all 12 (0.15–0.84 against 0.69–1.00) and four random directions change none of them by more than 0.01. The exception is CORE, whose blurred frames move far (7.9 times the swap displacement) but away from F (median cosine -0.32; no blurred frame nearer the F centroid), so that the leading blur-derived direction does not retain its read-out (0.49 against 0.77); on InSwapper its blurred frames do align with F (median cosine 0.77; 100% nearer the F centroid), but the leading blur-derived direction still does not retain the read-out (0.61 against 0.87, Table 16).

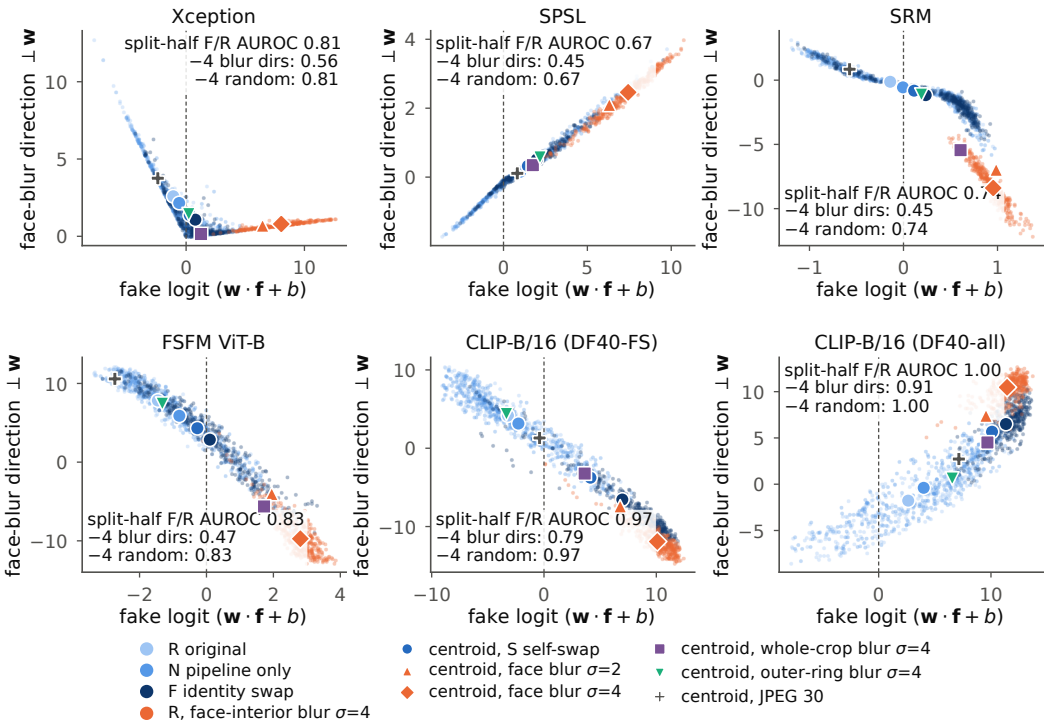

Figure 4: Feature space of the six prespecified checkpoints on the SimSwap four-arm crops. Each panel places frames on the checkpoint's own fake logit $z = \mathbf{w}^\top \mathbf{f} + b$ (horizontal; dashed line: logit 0, probability 0.5) and on the leading right singular vector of the blur-displacement matrix, then orthogonalised against $\mathbf{w}$ and unit-normalised (vertical; raw projections, not centred). Points: every third source and every fourth frame of R, N, F, and R with face-interior blur $\sigma = 4$; markers: centroids over all frames of these groups and of S, face-interior blur $\sigma = 2$, whole-crop blur $\sigma = 4$, outer-ring blur $\sigma = 4$, and JPEG 30. The nearest-centroid shares of Table 15 are computed in the full feature space, not from these coordinates. Corner text: split-half video-level F/R AUROC of the released read-out, after removing the top four blur-derived directions, and after removing four random directions.

What these results license. The 19 primary frame-level checkpoints show negative mean within-video score–q associations on real crops. The six prespecified checkpoints additionally show spatially selective responses to controlled blur. For the six, directions estimated from real frames and their blurred versions carry F/R discrimination close to that of the released read-out, removing them lowers or reverses the discrimination along the original weight vector, and the orthogonal complement still holds linearly decodable F/R information; this neither shows that the checkpoints learned only sharpness nor that face-interior blur displaces every representation farther than the swap. These findings support sharpness-related sensitivity of the released decisions under the tested crop and processing settings, not exclusive reliance on sharpness, and the generator step shows a response that a decrease in q does not account for. Replicating a video-level target across frames may let generation-correlated processing cues support classification even where manipulation evidence varies across frames; the fixed-support intervention of Section 5 shows that changing the supervisory target alters score ordering, but it does not establish that framewise label propagation caused the sharpness dependence observed here.

I Stepwise score contrasts of standard detectors

For an Xception trained on FaceForensics++ with the DeepfakeBench recipe (seed 303), the mean video logit on SimSwap C0 moves from -6.9 (R) to -4.7 (N) to -2.6 (S) to -1.2 (F); the adjacent

Table 15: Feature-space displacement and the released read-out on the SimSwap four-arm crops (32 frames per source). Displacement: median over frames of the Euclidean displacement of the penultimate feature at the same frame index, in units of the median F−R displacement, for N, face-interior blur $\sigma = 4$ of R, and the change between consecutive sampled frames of the same real video. Cosine: median over sources of the cosine between the per-source mean displacement and that of F. Near F: share of face-blurred R frames closer to the F centroid than to the R centroid in the full feature space. AUROC: video-level F/R AUROC of the released read-out (mean logit over the 32 frames) under the split-half protocol for every column of the released-read-out block, including the full read-out: full; restricted to the single leading blur-derived direction; after removing the top four blur-derived directions, the top four directions of natural frame-to-frame change, or four random directions. Probe: source-level five-fold linear probe refit on the complement of the top 32 blur-derived directions (directions, standardisation, PCA, and logistic regression fitted on the training folds). Rows below the rule: the remaining primary frame-level checkpoints (CapsuleNet has no linear read-out).

	displacement / swap			cosine with F−R		near F	F/R AUROC of the released read-out (split-half)					probe
Model	N	face $\sigma4$	natural	face $\sigma4$	N	face $\sigma4$	full	1 dir.	−4 face	−4 natural	−4 random	−32 face
Xception	0.29	2.58	0.60	+0.77	+0.98	1.00	0.81	0.81	0.56	0.36	0.81	0.92
SPSL	0.28	3.25	0.60	+0.71	+0.66	1.00	0.67	0.67	0.45	0.56	0.67	0.87
SRM	0.39	4.62	0.59	+0.70	+0.86	1.00	0.74	0.74	0.45	0.55	0.74	0.92
FSFM ViT-B	0.33	2.58	0.42	+0.76	+0.79	1.00	0.83	0.84	0.47	0.60	0.83	0.98
CLIP-B/16 (DF40-FS)	0.13	1.08	0.31	+0.79	+0.41	0.99	0.97	0.97	0.79	0.90	0.97	0.98
CLIP-B/16 (DF40-all)	0.18	0.99	0.44	+0.79	+0.56	0.99	1.00	1.00	0.91	0.66	1.00	0.99
EfficientNet-B4	0.23	2.58	0.59	+0.97	+0.89	1.00	0.69	0.69	0.43	0.55	0.69	0.76
CORE	0.37	7.93	0.56	-0.32	+0.88	0.00	0.77	0.49	0.65	0.42	0.77	0.90
F3Net	0.32	4.82	0.59	+0.59	+0.93	1.00	0.72	0.72	0.59	0.46	0.72	0.89
FFD	0.32	5.64	0.61	+0.68	+0.80	1.00	0.69	0.69	0.51	0.73	0.69	0.88
UCF	0.26	1.98	0.45	+0.96	+0.91	1.00	0.80	0.79	0.47	0.49	0.80	0.78
RECCE	0.36	2.41	0.65	+0.79	+0.96	1.00	0.76	0.77	0.54	0.45	0.76	0.88
Effort CLIP-L/14	0.18	0.85	0.33	+0.56	+0.31	0.97	0.98	0.98	0.62	0.86	0.98	0.98
Xception (DF40-FS)	0.08	1.18	0.18	+0.95	+0.64	1.00	0.99	0.99	0.84	0.93	0.99	1.00
EffNet-B4 (DF40-FS)	0.07	1.46	0.16	+0.98	+0.77	0.89	0.95	0.95	0.82	0.51	0.95	0.99
F3Net (DF40-FS)	0.08	1.14	0.15	+0.98	+0.67	1.00	0.99	0.99	0.75	0.96	0.99	1.00
CLIP-L/14 (DF40-FS)	0.14	0.85	0.31	+0.74	+0.39	0.96	0.99	0.99	0.74	0.91	0.99	0.99
Xception (DF40-all)	0.06	0.89	0.12	+0.99	+0.73	0.92	1.00	1.00	0.15	0.32	1.00	0.99

Table 16: As Table 15, for the InSwapper four-arm crops.

	displacement / swap			cosine with F−R		near F	F/R AUROC of the released read-out (split-half)					probe
Model	N	face $\sigma4$	natural	face $\sigma4$	N	face $\sigma4$	full	1 dir.	−4 face	−4 natural	−4 random	−32 face
Xception	0.55	2.25	0.52	+0.80	+0.99	1.00	0.87	0.87	0.54	0.49	0.87	0.93
SPSL	0.55	3.38	0.62	+0.62	+0.85	1.00	0.66	0.66	0.33	0.44	0.66	0.86
SRM	0.51	3.36	0.43	+0.83	+0.96	1.00	0.83	0.85	0.52	0.72	0.83	0.92
FSFM ViT-B	0.48	1.87	0.30	+0.83	+0.85	1.00	0.94	0.94	0.42	0.63	0.94	0.98
CLIP-B/16 (DF40-FS)	0.21	1.05	0.31	+0.72	+0.42	0.99	0.95	0.95	0.79	0.87	0.95	0.95
CLIP-B/16 (DF40-all)	0.33	0.97	0.43	+0.71	+0.63	0.99	0.98	0.98	0.88	0.61	0.98	0.99
EfficientNet-B4	0.47	1.77	0.40	+0.99	+0.98	1.00	0.84	0.84	0.43	0.73	0.84	0.81
CORE	0.56	6.01	0.42	+0.77	+0.95	1.00	0.87	0.61	0.80	0.51	0.87	0.94
F3Net	0.55	3.77	0.46	+0.69	+0.96	1.00	0.84	0.84	0.61	0.43	0.84	0.92
FFD	0.48	5.83	0.63	+0.68	+0.89	1.00	0.72	0.71	0.49	0.75	0.72	0.87
UCF	0.42	1.55	0.36	+0.98	+0.95	0.99	0.87	0.87	0.54	0.60	0.87	0.88
RECCE	0.57	1.83	0.49	+0.87	+0.97	1.00	0.88	0.89	0.60	0.49	0.88	0.93
Effort CLIP-L/14	0.27	0.78	0.31	+0.49	+0.34	0.98	0.97	0.97	0.64	0.86	0.97	0.99
Xception (DF40-FS)	0.20	1.50	0.23	+0.91	+0.86	1.00	0.97	0.97	0.74	0.94	0.97	0.96
EffNet-B4 (DF40-FS)	0.20	1.86	0.21	+0.98	+0.87	0.93	0.91	0.90	0.75	0.49	0.91	0.92
F3Net (DF40-FS)	0.19	1.29	0.17	+0.97	+0.82	1.00	0.97	0.97	0.73	0.94	0.97	0.95
CLIP-L/14 (DF40-FS)	0.22	0.82	0.31	+0.67	+0.42	0.97	0.98	0.99	0.73	0.86	0.98	0.97
Xception (DF40-all)	0.19	1.26	0.18	+0.98	+0.89	0.98	0.97	0.97	0.31	0.49	0.97	0.97

differences are $+2.25$ [2.03, 2.46] (N−R), $+2.03$ [1.72, 2.38] (S−N), and $+1.38$ [0.88, 1.85] (F−S) with source-level bootstrap intervals, and judged-fake rates rise from 0% to 9%, 21%, and 38%. On InSwapper the first step is larger (-6.9 to -3.0, N/R 0.85) and the last smaller (F/S 0.53 [0.48, 0.57]). Tables 19 and 20 repeat the contrasts and the AUCs for the other two seeds, for the same architecture trained with the corruption augmentation of the DeepfakeBench recipe, for an SBI-trained Xception, and for an Xception whose real training frames passed through a sham pipeline (Gaussian smoothing

Table 17: Sharpness and scores on the released frame-level checkpoints (inference only). Left: within-source Spearman correlation between the fake logit and the inner-square sharpness proxy q over the 32 frames of each source, mean over the 136 sources [95% source bootstrap], on the originals (shared by the generators) and on the self-swaps of each generator. Right: mean change of the fake logit from N to S, pooled over all frames whose self-swap has a higher q than the pipeline-only frame (sharper; 11% of the frames on SimSwap and 34% on InSwapper) and over the frames with $q(\mathsf{S}) \leq q(\mathsf{N})$ (blurrier).

	ρ(logit, q) within source			S−N logit, SimSwap		S−N logit, InSwapper	
Model	R	S SimSwap	S InSwapper	sharper	blurrier	sharper	blurrier
Xception	-0.27 [-0.31,-0.22]	-0.27 [-0.33,-0.23]	-0.20 [-0.25,-0.15]	+0.61	+0.93	+1.22	+1.09
EfficientNet-B4	-0.25 [-0.29,-0.20]	-0.30 [-0.36,-0.25]	-0.30 [-0.35,-0.25]	-0.06	+0.55	+0.95	+0.86
CapsuleNet	-0.26 [-0.30,-0.21]	-0.29 [-0.34,-0.24]	-0.20 [-0.25,-0.15]	-0.03	+0.06	+0.07	+0.08
CORE	-0.25 [-0.30,-0.21]	-0.28 [-0.33,-0.23]	-0.26 [-0.31,-0.21]	+0.14	+0.77	+1.69	+1.47
F3Net	-0.27 [-0.31,-0.22]	-0.26 [-0.31,-0.21]	-0.23 [-0.28,-0.18]	-0.41	+0.58	+1.87	+1.65
FFD	-0.18 [-0.24,-0.13]	-0.22 [-0.28,-0.17]	-0.20 [-0.26,-0.15]	-0.27	+0.40	+0.96	+0.84
SPSL	-0.18 [-0.23,-0.13]	-0.18 [-0.23,-0.13]	-0.15 [-0.21,-0.10]	-0.06	+0.52	+0.26	+0.09
SRM	-0.23 [-0.28,-0.19]	-0.29 [-0.33,-0.24]	-0.17 [-0.23,-0.12]	+0.06	+0.13	+0.22	+0.20
UCF	-0.27 [-0.32,-0.23]	-0.34 [-0.39,-0.29]	-0.33 [-0.38,-0.28]	+0.41	+1.62	+2.52	+2.14
RECCE	-0.30 [-0.34,-0.25]	-0.30 [-0.35,-0.26]	-0.25 [-0.29,-0.20]	-0.01	+0.94	+2.02	+1.84
FSFM ViT-B	-0.18 [-0.22,-0.14]	-0.27 [-0.31,-0.22]	-0.28 [-0.32,-0.23]	+0.41	+0.55	+0.86	+0.95
Effort CLIP-L/14	-0.25 [-0.30,-0.20]	-0.21 [-0.26,-0.16]	-0.10 [-0.16,-0.04]	+1.48	+1.90	+1.35	+1.43
Xception (DF40-FS)	-0.20 [-0.26,-0.15]	-0.18 [-0.23,-0.12]	-0.13 [-0.19,-0.08]	+8.69	+11.31	+6.45	+7.16
EffNet-B4 (DF40-FS)	-0.18 [-0.24,-0.13]	-0.12 [-0.17,-0.07]	-0.13 [-0.19,-0.08]	+3.67	+5.45	+2.89	+3.65
F3Net (DF40-FS)	-0.22 [-0.27,-0.17]	-0.15 [-0.20,-0.10]	-0.13 [-0.18,-0.07]	+8.81	+11.25	+7.23	+7.96
CLIP-B/16 (DF40-FS)	-0.16 [-0.21,-0.11]	-0.14 [-0.19,-0.08]	-0.08 [-0.14,-0.03]	+4.24	+6.69	+4.10	+4.72
CLIP-L/14 (DF40-FS)	-0.13 [-0.18,-0.09]	-0.05 [-0.09,+0.00]	+0.02 [-0.03,+0.07]	+5.13	+7.76	+5.19	+5.12
Xception (DF40-all)	-0.14 [-0.18,-0.09]	-0.16 [-0.21,-0.10]	+0.06 [+0.00,+0.10]	+7.48	+9.85	+5.01	+5.70
CLIP-B/16 (DF40-all)	-0.19 [-0.24,-0.15]	-0.18 [-0.23,-0.13]	-0.11 [-0.15,-0.06]	+4.07	+6.34	+3.24	+3.11

Table 18: Region-selective blur and JPEG on six checkpoints. Cells: mean change of the fake logit over all 32 frames of the 136 four-arm originals (R, C0; a different frame set from the 8-frame FaceForensics++ crops of Table 14, whose whole-crop rates it reproduces within 0.03) and, in parentheses, the pooled fraction of initially real-judged frames that flip to fake (the aggregation of Table 14), for Gaussian blur of the whole crop, of the inner square only, and of the outer ring only at $\sigma = 2$ and 4, and for JPEG quality 30.

	$\sigma = 2$			$\sigma = 4$			JPEG 30
Model	whole	inner	outer	whole	inner	outer	
Xception	+1.5 (0.54)	+7.5 (1.00)	+0.5 (0.16)	+2.3 (1.00)	+9.1 (1.00)	+1.3 (0.47)	-1.3 (0.00)
SPSL	-0.0 (0.50)	+5.3 (1.00)	+0.4 (0.51)	+0.8 (1.00)	+6.4 (1.00)	+1.2 (0.91)	-0.2 (0.27)
SRM	+0.7 (0.80)	+1.1 (1.00)	+0.1 (0.13)	+0.7 (1.00)	+1.1 (1.00)	+0.3 (0.40)	-0.4 (0.02)
FSFM ViT-B	+2.2 (0.69)	+3.4 (0.92)	-0.0 (0.01)	+3.2 (0.99)	+4.3 (1.00)	+0.1 (0.02)	-1.3 (0.00)
CLIP-B/16 (DF40-FS)	+2.1 (0.21)	+10.0 (0.90)	-0.4 (0.02)	+6.8 (0.76)	+13.3 (0.99)	-0.1 (0.04)	+2.8 (0.26)
CLIP-B/16 (DF40-all)	+3.3 (0.72)	+6.9 (0.99)	+1.4 (0.39)	+7.1 (1.00)	+8.8 (1.00)	+3.9 (0.81)	+4.5 (0.89)

or down-up resampling blended into the face region of half of the real frames, mild JPEG on both classes), on both generators. The pattern is stable across seeds (the pipeline step is $+2.2$ to $+2.9$ on SimSwap and the identity step $+1.4$ to $+1.5$), the augmented model has a smaller pipeline step ($+1.2$), the SBI-trained model has the largest ($+8.2$), in line with its training positives being blended reals, and the sham-pipeline model has none ($+0.06$, N/R 0.51) while keeping its identity step, which is the free-sampling precursor of the matched-negatives result in Section 5.3. These are paired score contrasts on the raw-logit scale of one model: the three differences sum to F−R by construction, a monotone rescaling of the score would change their proportions without changing any AUC, and they should not be read as the shares of three additive mechanisms.

J GENERATED-VIDEO CONTROL: CONSTRUCTION, DETECTORS, AND FULL RESULTS

Sources and released outputs. AIGVDBench (Ma et al., 2026) releases, for every source video of its test split, the outputs of several generators under three tasks: text-to-video from the caption of the source, image-to-video from its first frame, and video-to-video from the whole source. The released

Table 19: Stepwise raw-logit contrasts of the Xception variants we trained (C0, 136 sources): mean video logit per arm and the three adjacent differences; brackets: 95% paired bootstrap over sources.

Detector	mean logit R/N/S/F	N−R	S−N	F−S
SimSwap				
Xception (seed 303)	-6.9/-4.7/-2.6/-1.2	+2.25 [2.03,2.46]	+2.03 [1.72,2.38]	+1.38 [0.88,1.85]
Xception (seed 304)	-7.3/-5.0/-1.6/-0.2	+2.25 [1.99,2.53]	+3.46 [3.07,3.85]	+1.40 [0.84,1.94]
Xception (seed 305)	-7.6/-4.7/-1.8/-0.2	+2.85 [2.54,3.15]	+2.98 [2.50,3.46]	+1.52 [0.93,2.11]
SBI-trained Xception (our run)	-5.0/3.2/8.3/8.9	+8.24 [7.92,8.55]	+5.10 [4.76,5.44]	+0.54 [0.38,0.71]
Xception + augmentation	-7.3/-6.1/-3.1/-1.3	+1.17 [1.02,1.32]	+3.02 [2.63,3.40]	+1.78 [1.21,2.36]
Xception + sham-pipeline reals	-7.6/-7.5/-6.0/-4.4	+0.06 [-0.10,0.23]	+1.50 [1.18,1.81]	+1.60 [1.10,2.11]
InSwapper				
Xception (seed 303)	-6.9/-3.0/-1.4/-1.0	+3.92 [3.65,4.18]	+1.61 [1.27,1.95]	+0.35 [-0.17,0.80]
Xception (seed 304)	-7.3/-2.9/-0.6/-0.4	+4.37 [4.04,4.74]	+2.32 [1.89,2.78]	+0.21 [-0.41,0.86]
Xception (seed 305)	-7.6/-3.0/-0.5/0.2	+4.64 [4.29,4.99]	+2.49 [2.06,2.88]	+0.65 [0.06,1.24]
SBI-trained Xception (our run)	-5.0/5.3/7.3/7.7	+10.33 [9.86,10.79]	+1.98 [1.75,2.23]	+0.36 [0.15,0.56]
Xception + augmentation	-7.3/-5.1/-3.4/-2.6	+2.16 [1.95,2.36]	+1.73 [1.33,2.11]	+0.79 [0.16,1.41]
Xception + sham-pipeline reals	-7.6/-7.3/-6.2/-5.5	+0.29 [0.06,0.51]	+1.04 [0.71,1.36]	+0.73 [0.25,1.22]

Table 20: Pairwise AUCs and judged-fake rates of the same Xception variants on the same scores (C0, 136 sources; judged fake at logit 0).

Detector	N/R	S/R	F/R	F/S	judged fake R/N/S/F
SimSwap					
Xception (seed 303)	0.73	0.87	0.90	0.60	0/9/21/38%
Xception (seed 304)	0.70	0.90	0.92	0.60	1/7/32/47%
Xception (seed 305)	0.74	0.89	0.91	0.60	1/11/29/48%
SBI-trained Xception (our run)	0.99	1.00	1.00	0.62	4/89/99/99%
Xception + augmentation	0.60	0.82	0.89	0.64	1/5/17/34%
Xception + sham-pipeline reals	0.51	0.66	0.76	0.64	1/1/1/13%
InSwapper					
Xception (seed 303)	0.85	0.92	0.93	0.53	0/18/35/38%
Xception (seed 304)	0.84	0.92	0.92	0.52	1/22/44/48%
Xception (seed 305)	0.85	0.93	0.94	0.55	1/19/45/53%
SBI-trained Xception (our run)	0.99	1.00	1.00	0.60	4/96/99/99%
Xception + augmentation	0.69	0.79	0.84	0.57	1/8/18/22%
Xception + sham-pipeline reals	0.55	0.65	0.71	0.58	1/1/2/6%

files of one generator share one container and codec configuration, listed in Table 21 as read from the files with ffprobe. We draw 300 sources at random (seed 303) from the test split; the sources are 1920×1080 at 29.97 fps with 82–113 frames, and each generator's outputs for these sources are the G arms. All source videos have a released video-to-video or image-to-video output for every family used here, and text-to-video outputs exist for CogVideoX1.5 and LTX.

Table 21: Released output settings per generator family (read from the released files) and the frame counts of the released tasks. The format control N of a family reproduces the first five columns; the sources are 1920×1080, 29.97 fps, 82–113 frames.

Family	resolution	fps	frames (N)	x264 CRF	V2V frames	I2V frames	T2V frames
CogVideoX1.5-5B	720×480	16	49	25	49	–	33
LTX-Video	768×512	24	49	25	49	49	97
Stable Video Diffusion	512×320	8	16	23	–	16	–
Pyramid-Flow	640×384	24	89	25	–	89	–

Format control. For each family, N resamples the source to the released resolution with bicubic interpolation (the released aspect ratios are 3:2 for CogVideoX1.5 and LTX, 16:10 for SVD, and 5:3 for Pyramid-Flow against the 16:9 source, so the source is rescaled non-uniformly to the released width and height without cropping, as the released outputs are), picks the released number of frames at uniformly spaced positions (`linspace` over the source frames, rounded), and encodes with libx264 at the released preset and CRF (`preset medium`, CRF 25 for CogVideoX1.5, LTX, and Pyramid-Flow, CRF 23 for SVD), yuv420p, and the released frame rate. The preset and CRF are read from the x264 settings string that libx264 writes into the released files (SEI user data), and the codec, profile, pixel format, frame rate, frame count, and resolution were compared field by field with `ffprobe`; the group-of-pictures structure and the rate-control state of the released encoder are not recovered from the files, so the control fixes the encoder parameters it can read and does not claim

identical compression distortion. A fixed CRF does not fix the bitrate: the control files come out two to three times larger than the released files, an interaction of content and encoding that the control does not isolate, and we do not equalise the bitrate because that would change the CRF the released files fix. The control also changes the time axis: taking the released frame count at uniform positions and playing it at the released frame rate lengthens or shortens the clip (the 2.7–3.8 s sources become 3.1 s for CogVideoX1.5, 2.0 s for LTX and SVD, and 3.7 s for Pyramid-Flow), as the released outputs do, so N is a format-and-temporal-resampling control rather than a codec change alone. The content, motion, and reconstruction statistics of N and G differ by construction; N isolates what a real video acquires by passing through the released output format. The sensitivity of N/R to the kernel and the CRF is measured with the grid of Appendix J.1.

VAE round trip. V and N are built from the same uncompressed resampled frames N_0 of a source: N encodes N_0, and V encodes the output of the family's video VAE applied to N_0 (`AutoencoderKLCogVideoX` from `THUDM/CogVideoX1.5-5B` and `AutoencoderKLLTXVideo` from `Lightricks/LTX-Video`, in bfloat16, the mode of the encoder distribution rather than a sample, tiling and slicing enabled for memory), with the same encoder settings. The two arms therefore share one encoding pass and differ only by the VAE round trip before it; V is not a re-encoding of the N file. SVD's public pipeline includes a temporal-decoder VAE and Pyramid-Flow releases its own, but we have not verified that either reproduces the reconstruction behind these released videos, so no V arm is built for them. The VAE version may also differ from the one used to produce the CogVideoX1.5 and LTX videos; the arm measures the response to a same-family reconstruction trace, not to the exact trace in G. PSNR of the VAE output against N_0 is 33.0 dB for CogVideoX1.5 (range 20.9–41.8 over the 300 sources) and 29.3 dB for LTX (18.4–39.0), so the reconstruction differs measurably at the pixel level; we do not assume that it preserves every semantic detail, and V is read as a reconstruction-sensitivity control, not as a counterfactual that changes only a fingerprint.

Detectors and inference contracts. Each detector is run with its authors' released weights, and the main-text rows use the inference contract of the authors' released code; where we also ran a variant, the variant is listed separately (Table 22) rather than folded into the authors' name. DeCoF (Ma et al., 2024) is the detector released with the benchmark (a frozen CLIP-L/14 image encoder followed by a two-layer temporal transformer over frame tokens); the authors' rule takes eight frames spread over the first 32 frames of the clip, crops the centre square, passes a JPEG intermediate (quality 95 in our run; the authors' quality is not stated), resizes to 224, and applies ImageNet normalisation, and the score is the softmax probability of the fake class. Because the first 32 frames span 2 s of a 16 fps CogVideoX1.5 output but the whole 16-frame, 2 s SVD output, we also score every arm under three variants: eight frames spread uniformly over the whole clip with the JPEG step kept, the authors' frames without the JPEG step, and both changes together. WaveRep (Corvi et al., 2025) is the authors' DINOv2 ViT-B/14 checkpoint (registers, 504-pixel input, the four-generator weights) trained with wavelet-domain augmentation; the authors' demo code decodes every frame, centre-crops it to 504 (zero-padding a smaller side), normalises with ImageNet statistics, and scores the video by the mean per-frame logit, calling it synthetic when the sigmoid of that mean exceeds 0.5; the main-text rows follow this contract, run through the demo's own frame reader and transform. The toolkit variant (Cakiroglu et al., 2026), which scores 16 frames of a two-second window with the same weights and spatial processing, is reported alongside. D3 (Zheng et al., 2025) is training-free; the authors' released code extracts three seconds at 8 fps to JPEG frames with `ffmpeg`, starting at $\lfloor U(0, L-3) \rfloor$ seconds for a clip of length $L > 3$ s and at 0 otherwise, which is 0 for every arm of the main comparison (all shorter than 4 s; our extraction draws the start per video with a fixed seed and uses 0 whenever the clip is at most 3 s) and 1 s for 4% of the 4.04 s LTX text-to-video clips, takes the first 16 frames (a span of 1.875 s), crops 10% off both ends of the longer side, resizes to 224, normalises with ImageNet statistics on the BGR array their loader reads, embeds each frame with XCLIP ViT-B/16, and scores the video with $d_t = \|h_{t+1} - h_t\|_2$, $u_t = d_{t+1} - d_t$, and $s = \mathrm{SD}_t(u_t)$, the temporal standard deviation of the first differences of consecutive-embedding distances, which their evaluation treats as a score for the real class; we use $-s$ so that higher means more generated, and the head has no released threshold. The main-text rows implement this operator and extraction line by line from the released repository (`Zig-HS/D3`, commit `c798fbc`). The toolkit's D3 wrapper computes a different statistic, the standard deviation over time of the vector second difference of the embeddings averaged over feature dimensions, from eight frames spread uniformly over the clip with CLIP normalisation; because it is not the authors' operator we report it as a D3-inspired

variant. The toolkit decoder skips one source video with a corrupt tail, so its rows use 299 sources; the author-contract D3 and demo-contract WaveRep rows use all 300 (the demo reader decodes the corrupt file up to the damaged frame).

Full results. Tables 22 to 24 give every contrast with paired-bootstrap intervals for each detector, contract variant, and generator family, including the released image-to-video output of LTX, the VAE round trip against the original, and the exploratory text-to-video against video-to-video contrast (T2V changes content, camera, and duration as well as the generation task, so this contrast is not an identity axis and is listed only to show that DeCoF and WaveRep separate the two released tasks weakly, at 0.61–0.72). The judged-fake rates give the picture at each detector's threshold: DeCoF (0.5 on its probability) calls 1% of the originals fake, 9–17% of the format controls, 14–16% of the VAE round trips, and 11–19% of the released outputs of CogVideoX1.5, LTX, and Pyramid-Flow (59% for SVD), so under its threshold a resized and re-encoded real is about as likely to be flagged as a generated one; WaveRep (a positive mean logit, the demo's rule) calls 3% of the originals and 3–6% of the format controls fake against 86–98% of the video-to-video outputs and 56–99% of the image-to-video outputs; D3's head is an unthresholded scalar and its rates are not defined. **Contract variants.** For DeCoF, keeping the JPEG step but spreading the eight frames over the whole clip changes no contrast by more than 0.02 (N/R 0.83, 0.81, 0.91, 0.87 against 0.82, 0.80, 0.89, 0.86), whereas dropping the JPEG intermediate lowers N/R by 0.02–0.05 (0.79, 0.78, 0.84, 0.84) and raises G/N by at most 0.02; the two changes together give N/R 0.81, 0.79, 0.86, 0.84. The reading is the same under all four contracts. For WaveRep, the toolkit variant (16 frames of a two-second window instead of every frame) gives N/R 0.53, 0.59, 0.65, 0.52 against 0.52, 0.58, 0.64, 0.51 for the demo contract and G/N 0.93–1.00 against 0.94–1.00, so the frame budget moves the format response by at most 0.01 and leaves the reconstruction and generation responses in place. For D3 the contract decides the reading: the authors' operator and extraction give N/R 0.38, 0.34, 0.31, 0.40 and G/R 0.63, 0.49, 0.37, 0.52, the toolkit's D3-inspired statistic on eight frames spread over the clip gives N/R 0.50, 0.50, 0.51, 0.50 and G/R 0.81, 0.68, 0.52, 0.64. The two procedures differ in the operator, the number and positions of the sampled frames, the spatial preprocessing, the colour channels and normalisation, and the JPEG intermediate, so the comparison is between two complete procedures and does not attribute the reversal to any one of them; the control changes the playback speed, the temporal sampling, and the image processing together, and which of these drives D3's ranking is not isolated by this comparison (Appendix J.1 applies the control's operations one at a time). Both are valid measurements of stated procedures, but only the authors' contract is a reproduction of the released D3; the other is a different method variant. On the 299 sources both contracts scored, every contrast of both D3 rows and both WaveRep rows differs by at most 0.002 from the value on its full sample (`evidence/gai_metrics/common299.json`), so the difference in sample size does not enter the comparison.

J.1 Sensitivity of the format control to its implementation

The format control fixes the released resolution, frame rate, frame count, and encoder, but two of its steps admit alternatives: the resampling kernel and the rate setting. A pre-specified grid crosses two kernels (bicubic, the main-text choice, and area averaging) with two CRF values (the released value and CRF 18, which roughly doubles the bitrate) for the CogVideoX1.5 and LTX families; the SVD and Pyramid-Flow controls are not varied. A second set of variants, added after the grid results were inspected, applies the control's three operations one at a time (below). Every variant N' is scored by the same detectors and Tables 25 and 26 report its separation from the original (N'/R), from the main-text control (N'/N, which is 0.5 when the two implementations are indistinguishable to the detector), and from the VAE round trip and the released generation of the family (V/N', G/N').

The kernel matters little and the rate matters. Replacing bicubic by area averaging at the released CRF leaves every contrast of DeCoF and D3 within 0.02 of the main-text value (N'/N 0.50–0.51) and raises WaveRep's format response by 0.05–0.06 (0.52 to 0.57 and 0.58 to 0.63). Lowering the CRF to 18 moves two of the three detectors in opposite directions. DeCoF's format response falls from 0.80–0.82 to 0.70–0.74 and its separation of the generated videos from the control rises from 0.51–0.57 to 0.61–0.69 (the CRF-18 control is ranked below the released-CRF control, N'/N 0.37–0.39), so part of what DeCoF reads in the format control is the compression level of the released files; the response that remains at CRF 18 is not attributed to a single remaining factor, because CRF 18 is still a lossy encoding and the resampling and the encoding interact. WaveRep moves the other

Table 22: Generated-video control, part 1: format control and VAE round trip (video-level AUC with 95% paired bootstrap over sources; judged fake: fraction of R/N/V/G videos above each detector's threshold, V absent for SVD and Pyramid-Flow). Row labels: DeCoF is the authors' rule, "uniform" spreads its eight frames over the whole clip, "no JPEG" drops the JPEG intermediate; WaveRep (all frames) is the demo contract and (16 frames) the toolkit variant; D3 (authors') is the authors' operator and extraction and D3-inspired (toolkit) the toolkit statistic (Appendix J).

Detector	Family	judged fake R/N/V/G	N/R	V/R	V/N
DeCoF	CogVideoX1.5	1/11/14/11	0.82 [0.80,0.84]	0.85 [0.83,0.88]	0.56 [0.55,0.57]
DeCoF	LTX-Video	1/9/16/14	0.80 [0.78,0.82]	0.86 [0.84,0.89]	0.61 [0.59,0.62]
DeCoF	SVD	1/17/–/59	0.89 [0.87,0.92]	–	–
DeCoF	Pyramid-Flow	1/15/–/19	0.86 [0.84,0.88]	–	–
DeCoF, uniform, no JPEG	CogVideoX1.5	1/9/12/10	0.81 [0.79,0.83]	0.85 [0.82,0.87]	0.56 [0.55,0.57]
DeCoF, uniform, no JPEG	LTX-Video	1/8/14/9	0.79 [0.78,0.82]	0.87 [0.85,0.89]	0.62 [0.60,0.64]
DeCoF, uniform, no JPEG	SVD	1/14/–/45	0.86 [0.84,0.88]	–	–
DeCoF, uniform, no JPEG	Pyramid-Flow	1/10/–/16	0.84 [0.82,0.86]	–	–
DeCoF, uniform, JPEG kept	CogVideoX1.5	1/10/14/10	0.83 [0.81,0.85]	0.87 [0.85,0.89]	0.56 [0.55,0.57]
DeCoF, uniform, JPEG kept	LTX-Video	1/9/16/11	0.81 [0.80,0.84]	0.88 [0.86,0.90]	0.61 [0.60,0.63]
DeCoF, uniform, JPEG kept	SVD	1/17/–/59	0.91 [0.89,0.93]	–	–
DeCoF, uniform, JPEG kept	Pyramid-Flow	1/13/–/18	0.87 [0.84,0.89]	–	–
DeCoF, authors' frames, no JPEG	CogVideoX1.5	1/9/12/10	0.79 [0.77,0.82]	0.83 [0.81,0.86]	0.56 [0.55,0.57]
DeCoF, authors' frames, no JPEG	LTX-Video	1/9/14/12	0.78 [0.76,0.80]	0.85 [0.83,0.88]	0.61 [0.60,0.63]
DeCoF, authors' frames, no JPEG	SVD	1/14/–/45	0.84 [0.82,0.86]	–	–
DeCoF, authors' frames, no JPEG	Pyramid-Flow	1/12/–/17	0.84 [0.81,0.86]	–	–
WaveRep (all frames)	CogVideoX1.5	3/4/51/86	0.52 [0.48,0.55]	0.93 [0.92,0.95]	0.94 [0.92,0.95]
WaveRep (all frames)	LTX-Video	3/6/95/98	0.58 [0.55,0.61]	0.99 [0.99,1.00]	0.99 [0.98,0.99]
WaveRep (all frames)	SVD	3/4/–/99	0.64 [0.60,0.68]	–	–
WaveRep (all frames)	Pyramid-Flow	3/3/–/56	0.51 [0.47,0.55]	–	–
WaveRep (16 frames)	CogVideoX1.5	3/4/53/85	0.53 [0.49,0.56]	0.93 [0.91,0.95]	0.93 [0.92,0.95]
WaveRep (16 frames)	LTX-Video	3/6/95/98	0.59 [0.56,0.62]	0.99 [0.99,1.00]	0.99 [0.98,0.99]
WaveRep (16 frames)	SVD	3/4/–/99	0.65 [0.61,0.69]	–	–
WaveRep (16 frames)	Pyramid-Flow	3/3/–/58	0.52 [0.49,0.56]	–	–
D3 (authors')	CogVideoX1.5	–	0.38 [0.35,0.41]	0.38 [0.35,0.41]	0.50 [0.48,0.51]
D3 (authors')	LTX-Video	–	0.34 [0.31,0.37]	0.31 [0.28,0.34]	0.48 [0.46,0.50]
D3 (authors')	SVD	–	0.31 [0.28,0.34]	–	–
D3 (authors')	Pyramid-Flow	–	0.40 [0.37,0.42]	–	–
D3-inspired (toolkit)	CogVideoX1.5	–	0.50 [0.48,0.51]	0.48 [0.47,0.49]	0.48 [0.47,0.49]
D3-inspired (toolkit)	LTX-Video	–	0.50 [0.49,0.51]	0.45 [0.43,0.46]	0.44 [0.43,0.46]
D3-inspired (toolkit)	SVD	–	0.51 [0.49,0.52]	–	–
D3-inspired (toolkit)	Pyramid-Flow	–	0.50 [0.49,0.51]	–	–
AIGVDet	CogVideoX1.5	5/33/24/46	0.48 [0.44,0.53]	0.41 [0.36,0.45]	0.45 [0.44,0.46]
AIGVDet	LTX-Video	5/21/15/23	0.42 [0.38,0.46]	0.37 [0.33,0.40]	0.47 [0.45,0.48]
AIGVDet	SVD	5/2/–/0	0.61 [0.56,0.66]	–	–
AIGVDet	Pyramid-Flow	5/7/–/6	0.34 [0.29,0.38]	–	–
L3DE	CogVideoX1.5	8/10/18/24	0.55 [0.53,0.58]	0.68 [0.66,0.71]	0.63 [0.61,0.65]
L3DE	LTX-Video	8/12/57/59	0.57 [0.54,0.59]	0.86 [0.84,0.89]	0.83 [0.80,0.85]
L3DE	SVD	8/14/–/87	0.63 [0.60,0.66]	–	–
L3DE	Pyramid-Flow	8/17/–/71	0.62 [0.59,0.64]	–	–
Ivy-xDetector (authors' contract)	CogVideoX1.5	83/67/70/98	0.37 [0.35,0.39]	0.38 [0.35,0.40]	0.51 [0.49,0.52]
Ivy-xDetector (authors' contract)	LTX-Video	83/71/–/90	0.39 [0.36,0.41]	–	–
Ivy-xDetector (authors' contract)	SVD	83/46/–/79	0.25 [0.22,0.28]	–	–
Ivy-xDetector (authors' contract)	Pyramid-Flow	83/56/–/72	0.29 [0.27,0.31]	–	–
Ivy-xDetector, toolkit contract (16 frames)	CogVideoX1.5	12/7/8/15	0.35 [0.33,0.37]	0.36 [0.34,0.39]	0.51 [0.50,0.53]
Ivy-xDetector, toolkit contract (16 frames)	LTX-Video	12/7/10/9	0.33 [0.31,0.36]	0.42 [0.39,0.45]	0.58 [0.57,0.60]
Ivy-xDetector, toolkit contract (16 frames)	SVD	12/2/–/10	0.21 [0.18,0.23]	–	–
Ivy-xDetector, toolkit contract (16 frames)	Pyramid-Flow	12/3/–/3	0.26 [0.23,0.28]	–	–
Skyra (authors' contract)	CogVideoX1.5	6/97/99/99	0.96 [0.94,0.97]	0.96 [0.95,0.98]	0.51 [0.50,0.52]
Skyra (authors' contract)	LTX-Video	6/94/92/96	0.94 [0.92,0.96]	0.93 [0.91,0.95]	0.49 [0.48,0.51]
Skyra (authors' contract)	SVD	6/0/–/1	0.47 [0.46,0.48]	–	–
Skyra (authors' contract)	Pyramid-Flow	6/0/–/1	0.47 [0.46,0.48]	–	–

way: its format response rises from 0.52–0.58 to 0.69–0.75 at CRF 18 and to 0.73–0.80 with area averaging at CRF 18, while its separation of the VAE round trip and of the generated videos from the control stays at 0.83–0.99. A plausible reading, which the grid does not test, is that a detector trained on high-frequency residuals sees more of the resampling trace when compression removes less of it. D3 under the authors' contract does not move (N′/N 0.50–0.51; N'/R stays at 0.34–0.39, the reversed ranking of the main text), so its ranking is insensitive to the kernel and CRF variants tested here (the temporal step is examined below; other cues are not excluded); the toolkit variant does not move either.

Two consequences follow for the main-text reading. The qualitative picture is stable across the grid: at every setting DeCoF's format response is large and its separation of generation from the control

Table 23: Generated-video control, part 2: released generation against the original, the format control, and the VAE round trip (G: video-to-video output; image-to-video for SVD and Pyramid-Flow). Row labels as in Table 22.

Detector	Family	G/R	G/N	G/V
DeCoF	CogVideoX1.5	0.81 [0.79,0.84]	0.51 [0.47,0.54]	0.45 [0.41,0.48]
DeCoF	LTX-Video	0.84 [0.82,0.86]	0.57 [0.56,0.59]	0.47 [0.45,0.48]
DeCoF	SVD	0.97 [0.96,0.98]	0.77 [0.74,0.79]	–
DeCoF	Pyramid-Flow	0.88 [0.86,0.90]	0.54 [0.53,0.56]	–
DeCoF, uniform, no JPEG	CogVideoX1.5	0.82 [0.79,0.85]	0.53 [0.50,0.57]	0.48 [0.45,0.51]
DeCoF, uniform, no JPEG	LTX-Video	0.85 [0.82,0.87]	0.59 [0.57,0.60]	0.47 [0.45,0.49]
DeCoF, uniform, no JPEG	SVD	0.97 [0.95,0.98]	0.78 [0.76,0.81]	–
DeCoF, uniform, no JPEG	Pyramid-Flow	0.86 [0.84,0.89]	0.56 [0.54,0.58]	–
DeCoF, uniform, JPEG kept	CogVideoX1.5	0.83 [0.80,0.86]	0.51 [0.48,0.55]	0.46 [0.42,0.49]
DeCoF, uniform, JPEG kept	LTX-Video	0.85 [0.83,0.87]	0.57 [0.55,0.59]	0.46 [0.44,0.47]
DeCoF, uniform, JPEG kept	SVD	0.98 [0.97,0.99]	0.77 [0.74,0.80]	–
DeCoF, uniform, JPEG kept	Pyramid-Flow	0.88 [0.86,0.90]	0.54 [0.52,0.56]	–
DeCoF, authors' frames, no JPEG	CogVideoX1.5	0.80 [0.78,0.83]	0.53 [0.49,0.56]	0.47 [0.44,0.50]
DeCoF, authors' frames, no JPEG	LTX-Video	0.84 [0.81,0.86]	0.59 [0.58,0.61]	0.48 [0.46,0.50]
DeCoF, authors' frames, no JPEG	SVD	0.96 [0.95,0.97]	0.78 [0.76,0.81]	–
DeCoF, authors' frames, no JPEG	Pyramid-Flow	0.86 [0.84,0.88]	0.55 [0.53,0.56]	–
WaveRep (all frames)	CogVideoX1.5	0.98 [0.97,0.99]	0.98 [0.98,0.99]	0.76 [0.73,0.79]
WaveRep (all frames)	LTX-Video	1.00 [0.99,1.00]	0.99 [0.99,1.00]	0.61 [0.58,0.63]
WaveRep (all frames)	SVD	1.00 [1.00,1.00]	1.00 [1.00,1.00]	–
WaveRep (all frames)	Pyramid-Flow	0.93 [0.92,0.95]	0.94 [0.93,0.95]	–
WaveRep (16 frames)	CogVideoX1.5	0.98 [0.97,0.99]	0.98 [0.97,0.99]	0.75 [0.72,0.78]
WaveRep (16 frames)	LTX-Video	1.00 [0.99,1.00]	0.99 [0.99,1.00]	0.61 [0.58,0.64]
WaveRep (16 frames)	SVD	1.00 [1.00,1.00]	1.00 [1.00,1.00]	–
WaveRep (16 frames)	Pyramid-Flow	0.93 [0.91,0.95]	0.93 [0.92,0.95]	–
D3 (authors')	CogVideoX1.5	0.63 [0.59,0.67]	0.73 [0.70,0.76]	0.73 [0.70,0.76]
D3 (authors')	LTX-Video	0.49 [0.46,0.53]	0.66 [0.62,0.69]	0.68 [0.65,0.71]
D3 (authors')	SVD	0.37 [0.33,0.40]	0.59 [0.55,0.63]	–
D3 (authors')	Pyramid-Flow	0.52 [0.49,0.56]	0.63 [0.59,0.66]	–
D3-inspired (toolkit)	CogVideoX1.5	0.81 [0.78,0.84]	0.80 [0.78,0.83]	0.81 [0.79,0.84]
D3-inspired (toolkit)	LTX-Video	0.68 [0.65,0.72]	0.68 [0.65,0.71]	0.72 [0.69,0.75]
D3-inspired (toolkit)	SVD	0.52 [0.48,0.56]	0.51 [0.47,0.55]	–
D3-inspired (toolkit)	Pyramid-Flow	0.64 [0.60,0.67]	0.63 [0.60,0.67]	–
AIGVDet	CogVideoX1.5	0.60 [0.55,0.65]	0.59 [0.57,0.62]	0.64 [0.61,0.67]
AIGVDet	LTX-Video	0.44 [0.40,0.48]	0.52 [0.50,0.53]	0.55 [0.53,0.57]
AIGVDet	SVD	0.39 [0.34,0.44]	0.34 [0.32,0.36]	–
AIGVDet	Pyramid-Flow	0.30 [0.26,0.34]	0.45 [0.43,0.47]	–
L3DE	CogVideoX1.5	0.72 [0.69,0.75]	0.67 [0.63,0.70]	0.55 [0.51,0.58]
L3DE	LTX-Video	0.87 [0.85,0.90]	0.84 [0.81,0.87]	0.52 [0.48,0.55]
L3DE	SVD	0.96 [0.94,0.97]	0.94 [0.92,0.96]	–
L3DE	Pyramid-Flow	0.93 [0.91,0.94]	0.87 [0.85,0.90]	–
Ivy-xDetector (authors' contract)	CogVideoX1.5	0.67 [0.63,0.71]	0.80 [0.77,0.83]	0.80 [0.77,0.83]
Ivy-xDetector (authors' contract)	LTX-Video	0.55 [0.52,0.58]	0.67 [0.64,0.70]	–
Ivy-xDetector (authors' contract)	SVD	0.46 [0.43,0.50]	0.72 [0.69,0.75]	–
Ivy-xDetector (authors' contract)	Pyramid-Flow	0.40 [0.36,0.43]	0.62 [0.59,0.65]	–
Ivy-xDetector, toolkit contract (16 frames)	CogVideoX1.5	0.52 [0.49,0.55]	0.65 [0.62,0.68]	0.64 [0.61,0.67]
Ivy-xDetector, toolkit contract (16 frames)	LTX-Video	0.44 [0.41,0.47]	0.60 [0.58,0.63]	0.52 [0.49,0.55]
Ivy-xDetector, toolkit contract (16 frames)	SVD	0.46 [0.42,0.50]	0.75 [0.72,0.78]	–
Ivy-xDetector, toolkit contract (16 frames)	Pyramid-Flow	0.27 [0.24,0.30]	0.54 [0.52,0.57]	–
Skyra (authors' contract)	CogVideoX1.5	0.97 [0.95,0.98]	0.51 [0.50,0.52]	0.50 [0.49,0.51]
Skyra (authors' contract)	LTX-Video	0.95 [0.94,0.97]	0.51 [0.50,0.53]	0.52 [0.50,0.54]
Skyra (authors' contract)	SVD	0.48 [0.46,0.49]	0.51 [0.50,0.51]	–
Skyra (authors' contract)	Pyramid-Flow	0.47 [0.46,0.49]	0.50 [0.50,0.51]	–

is smaller than its separation from the originals; WaveRep's response to the VAE round trip and to generation exceeds its response to the format control; D3's format ranking stays reversed and its generation response is larger against the control than against the originals. The absolute size of a detector's format response, however, is a property of the implementation of N as much as of the detector (for WaveRep it spans 0.52–0.80 across the grid), so a format control should be reported with its encoder settings, and the released settings are the natural default because they are what a generated file actually carries. The V arm is not rebuilt for each variant, so V/N' compares one reconstruction arm with differently built controls rather than isolating the VAE effect at every setting.

One step at a time. The control applies three operations, and three further variants apply them singly, each finished with the released encoder settings so that every variant is a decodable file: $\mathsf{N}'_{\mathrm{enc}}$ keeps the source resolution, frame count, and frame rate and only re-encodes (libx264, released preset and CRF); $\mathsf{N}'_{\mathrm{sp}}$ rescales to the released resolution (bicubic) and re-encodes; $\mathsf{N}'_{\mathrm{tm}}$ takes the released frame count at uniform positions, plays it at the released frame rate, and re-encodes. The step rows

Table 24: Generated-video control, part 3: the released image-to-video output of LTX against the control, and the exploratory text-to-video contrasts (CogVideoX1.5, LTX; families and detectors without these arms omitted). Row labels as in Table 22.

Detector	Family	G_{I2V}/N	T2V/R	T2V/V2V
DeCoF	CogVideoX1.5	–	0.88 [0.86,0.91]	0.64 [0.60,0.68]
DeCoF	LTX-Video	0.57 [0.55,0.60]	0.89 [0.86,0.91]	0.61 [0.57,0.65]
DeCoF, uniform, no JPEG	CogVideoX1.5	–	0.89 [0.87,0.91]	0.64 [0.60,0.67]
DeCoF, uniform, no JPEG	LTX-Video	0.59 [0.57,0.61]	0.90 [0.88,0.93]	0.64 [0.60,0.68]
DeCoF, uniform, JPEG kept	CogVideoX1.5	–	0.90 [0.87,0.92]	0.64 [0.60,0.67]
DeCoF, uniform, JPEG kept	LTX-Video	0.57 [0.55,0.59]	0.90 [0.87,0.92]	0.61 [0.57,0.65]
DeCoF, authors' frames, no JPEG	CogVideoX1.5	–	0.88 [0.85,0.90]	0.64 [0.61,0.68]
DeCoF, authors' frames, no JPEG	LTX-Video	0.60 [0.58,0.62]	0.90 [0.88,0.92]	0.64 [0.60,0.68]
WaveRep (all frames)	CogVideoX1.5	–	0.99 [0.99,1.00]	0.65 [0.61,0.68]
WaveRep (all frames)	LTX-Video	1.00 [1.00,1.00]	1.00 [1.00,1.00]	0.72 [0.68,0.75]
WaveRep (16 frames)	CogVideoX1.5	–	0.99 [0.99,1.00]	0.64 [0.60,0.68]
WaveRep (16 frames)	LTX-Video	1.00 [1.00,1.00]	1.00 [1.00,1.00]	0.74 [0.71,0.78]
D3 (authors')	CogVideoX1.5	–	0.46 [0.42,0.50]	0.34 [0.31,0.38]
D3 (authors')	LTX-Video	0.59 [0.55,0.62]	0.51 [0.47,0.55]	0.51 [0.47,0.55]
D3-inspired (toolkit)	CogVideoX1.5	–	0.79 [0.76,0.82]	0.49 [0.45,0.53]
D3-inspired (toolkit)	LTX-Video	0.57 [0.54,0.61]	0.63 [0.59,0.67]	0.47 [0.43,0.51]
AIGVDet	CogVideoX1.5	–	0.73 [0.69,0.78]	0.61 [0.58,0.65]
AIGVDet	LTX-Video	0.50 [0.48,0.53]	0.49 [0.45,0.54]	0.55 [0.51,0.59]
L3DE	CogVideoX1.5	–	0.70 [0.67,0.74]	0.50 [0.45,0.54]
L3DE	LTX-Video	0.88 [0.86,0.90]	0.85 [0.82,0.88]	0.50 [0.46,0.53]
Ivy-xDetector, toolkit contract (16 frames)	CogVideoX1.5	–	0.56 [0.51,0.59]	0.53 [0.49,0.57]
Ivy-xDetector, toolkit contract (16 frames)	LTX-Video	0.59 [0.56,0.62]	0.47 [0.43,0.51]	0.53 [0.49,0.57]
Skyra (authors' contract)	CogVideoX1.5	–	0.96 [0.94,0.97]	0.49 [0.48,0.50]
Skyra (authors' contract)	LTX-Video	–	0.95 [0.93,0.96]	0.49 [0.48,0.51]

of Tables 25 and 26 give the result. At the tested encoder settings, spatial rescaling followed by encoding reproduces most of DeCoF's full-control ranking: N'/R is 0.83 and 0.79 with rescaling alone (CogVideoX1.5, LTX) against 0.52 with re-encoding alone and 0.51 with temporal resampling alone; the rescaled variant has a profile close to the full control (N'/N 0.50–0.53) while the other two rank well below it (0.18–0.22), and G/N' returns to the level of G/R (0.80–0.84 against 0.81–0.84) when the control lacks the rescaling. Temporal resampling followed by encoding reproduces D3's principal reversal: N'/R is 0.39 and 0.36 with temporal resampling alone against 0.50 with re-encoding alone and 0.48 with rescaling alone, and the increase of G/N' over G/R is largest for the temporally resampled variant (0.73 and 0.64 against 0.63 and 0.49; 0.50–0.65 for the other two). WaveRep's small format response appears with the rescaling step (0.52 and 0.62; re-encoding alone 0.47 and temporal resampling alone 0.47–0.49, small reversed rankings), and no single step moves its separation of the VAE round trip or of the generated videos from the control (0.94–1.00). Re-encoding alone at the released settings produces only small ranking effects for all three detectors (N'/R 0.47–0.52). These are conditional effects of the implemented transformations at these encoder settings, spatial for DeCoF and WaveRep and temporal for D3; they establish neither unique internal cues nor the necessity of a step, and they do not exclude interactions with the encoding and with each detector's own preprocessing (the CRF grid above shows one such interaction for DeCoF).

J.2 Additional detectors

Beyond the three detectors of Section 6 we scored the same arms with AIGVDet (Bai et al., 2024), a two-stream ResNet-50 that reads RGB frames and an optical-flow visualisation (the authors' released weights; flow from torchvision RAFT, 16 frames). Table 27 reports it. Its released weights give no consistent ranking on this benchmark: the released video-to-video and image-to-video outputs are ranked against the originals with AUC 0.30–0.60 (0.73 for the CogVideoX1.5 text-to-video output) and against the format control with 0.34–0.59, and the format control itself with 0.34–0.61 against the original. Values below 0.5 are reversed rankings, not absent responses: on Pyramid-Flow the detector ranks the originals above both the generated videos and the format control. Its thresholded rates (5% of originals, 21–33% of the CogVideoX1.5 and LTX format controls, and 23–46% of their generated videos judged fake) show that it does push some resized and re-encoded reals over its threshold while ranking them inconsistently. We report it because a detector whose direction and transfer are unstable on a benchmark is also information about the benchmark, and because its judged-fake rates illustrate why AUC and thresholded rates are reported separately throughout. Three further released detectors were run: L3DE through the toolkit wrapper, and two vision-language detectors

Table 25: N-implementation grid and single-step variants, part 1: DeCoF under the authors' rule and under the uniform-frame, no-JPEG variant (CogVideoX1.5 and LTX; video-level AUC over 300 sources with 95% paired bootstrap). N′ is the variant: the first four rows of a block vary the kernel and the CRF of the full control, the three step rows apply one operation followed by the released encoding (*encode only*, *rescale*, *time resample*); N′/N compares the variant with the main-text control; the last column is the fraction of variant videos judged fake at the detector's threshold (0.5 on DeCoF's probability). G is the released video-to-video output.

Variant	N′/R	N′/N	V/N′	G/N′	judged fake
DeCoF (authors' rule), CogVideoX1.5					
bicubic, released CRF	0.82 [0.80,0.84]	–	0.56 [0.55,0.57]	0.51 [0.47,0.54]	11%
area, released CRF	0.82 [0.80,0.85]	0.51 [0.51,0.52]	0.55 [0.54,0.56]	0.50 [0.46,0.53]	11%
bicubic, CRF 18	0.73 [0.70,0.75]	0.38 [0.36,0.39]	0.68 [0.66,0.69]	0.63 [0.59,0.66]	6%
area, CRF 18	0.74 [0.71,0.76]	0.39 [0.37,0.40]	0.67 [0.65,0.68]	0.61 [0.58,0.65]	5%
step: encode only	0.52 [0.52,0.53]	0.20 [0.17,0.22]	0.84 [0.82,0.86]	0.80 [0.77,0.83]	1%
step: rescale	0.83 [0.81,0.85]	0.53 [0.52,0.54]	0.53 [0.51,0.54]	0.48 [0.44,0.51]	12%
step: time resample	0.51 [0.50,0.52]	0.18 [0.16,0.20]	0.85 [0.83,0.88]	0.81 [0.79,0.84]	1%
DeCoF (authors' rule), LTX					
bicubic, released CRF	0.80 [0.78,0.82]	–	0.61 [0.59,0.62]	0.57 [0.56,0.59]	9%
area, released CRF	0.80 [0.78,0.83]	0.51 [0.50,0.51]	0.60 [0.58,0.61]	0.56 [0.55,0.58]	9%
bicubic, CRF 18	0.70 [0.68,0.72]	0.37 [0.35,0.38]	0.73 [0.71,0.75]	0.69 [0.67,0.71]	5%
area, CRF 18	0.70 [0.68,0.72]	0.38 [0.36,0.39]	0.72 [0.70,0.74]	0.69 [0.67,0.71]	5%
step: encode only	0.52 [0.52,0.53]	0.22 [0.20,0.24]	0.85 [0.83,0.87]	0.82 [0.80,0.85]	1%
step: rescale	0.79 [0.77,0.82]	0.50 [0.48,0.51]	0.61 [0.59,0.63]	0.58 [0.56,0.59]	10%
step: time resample	0.51 [0.50,0.52]	0.20 [0.18,0.23]	0.86 [0.84,0.88]	0.84 [0.81,0.86]	1%
DeCoF, uniform frames, no JPEG, CogVideoX1.5					
bicubic, released CRF	0.81 [0.79,0.83]	–	0.56 [0.55,0.57]	0.53 [0.50,0.57]	9%
area, released CRF	0.81 [0.80,0.84]	0.51 [0.51,0.52]	0.54 [0.54,0.56]	0.52 [0.49,0.56]	9%
bicubic, CRF 18	0.69 [0.68,0.71]	0.35 [0.34,0.37]	0.70 [0.68,0.72]	0.67 [0.64,0.70]	4%
area, CRF 18	0.71 [0.69,0.72]	0.37 [0.35,0.38]	0.69 [0.67,0.70]	0.66 [0.63,0.69]	4%
DeCoF, uniform frames, no JPEG, LTX					
bicubic, released CRF	0.79 [0.78,0.82]	–	0.62 [0.60,0.64]	0.59 [0.57,0.60]	8%
area, released CRF	0.81 [0.79,0.83]	0.51 [0.51,0.52]	0.61 [0.59,0.62]	0.57 [0.56,0.59]	9%
bicubic, CRF 18	0.68 [0.67,0.71]	0.36 [0.35,0.38]	0.74 [0.72,0.76]	0.71 [0.69,0.74]	4%
area, CRF 18	0.69 [0.68,0.72]	0.37 [0.36,0.38]	0.73 [0.71,0.76]	0.70 [0.68,0.73]	3%

under their authors' contracts. L3DE (Chang et al., 2025), a 3D-CNN over DINOv2-G appearance, RAFT flow, and UniDepth-v2 depth (24 frames), responds to everything and most to generation: N/R 0.55/0.57/0.63/0.62, V/N 0.63/0.83, G/R 0.72/0.87/0.96/0.93, G/N 0.67/0.84/0.94/0.87, with 8% of originals, 10–17% of format controls, 24–59% of video-to-video outputs, and 67–87% of image-to-video outputs judged fake. Two vision-language detectors answer with a free-text analysis and a real/fake verdict; each was checked on its own benchmark (Table 28). Ivy-xDetector (Jiang et al., 2026) (Qwen2.5-VL-3B) was run under two contracts. The toolkit contract passes 16 uniformly sampled frames as images with the prompt of the model card; the authors' contract samples the video at 1 fps with at most six frames, passed as a video input, and uses the system prompt of the released training configuration. Its score is the softmax probability of the "fake" token against the "real" token at the position of the final `<conclusion>` answer, in $[0, 1]$ with higher meaning more generated; on every scored clip of both contracts the parsed conclusion and the 0.5 threshold on this probability give the same decision, so the judged-fake rates are verdict rates. Both contracts separate the Ivy-Fake test videos (AUC 1.00 and 1.00 on 16+16, Table 28); the authors' contract reaches only 0.69 on the ViF-Bench subset, so the check establishes that the execution is not degenerate, not that it reproduces the published benchmark. On the arms both contracts score the format control more real than the originals: the toolkit contract judges 12% of the originals fake against 2–7% of the controls (N/R 0.35/0.33/0.21/0.26, G/R 0.52/0.44/0.46/0.27, G/N 0.65/0.60/0.75/0.54), and the authors' contract judges 83% of the originals fake (248 of 299) against 46–71% of the controls and 72–98% of the released outputs (N/R 0.37/0.39/0.25/0.29, G/R 0.67/0.55/0.46/0.40, G/N 0.80/0.67/0.72/0.62; Table 27). Under the authors' contract the video reader does not decode the source with a corrupt tail, so every row uses the 299 sources with all of its arms scored; the LTX VAE arm was not evaluated under this contract. Both contracts rank the format controls below the originals while their absolute decision rates differ sharply; the control changes the resolution, the aspect ratio, the time axis, and the compression together, and each contract rescales what the model sees, so these are benchmark- and contract-dependent operating points and identify no cause. Where a control is scored more real than the originals, the D3 and Ivy rows share one operational pattern under which a fixed G separates from

Table 26: N-implementation grid and single-step variants, part 2: WaveRep (demo contract and toolkit variant) and D3 (authors' operator and toolkit variant); 300 sources for the demo and authors' contracts, 299 for the toolkit variants. Columns as in Table 25; judged fake at 0 on WaveRep's mean logit, undefined for D3, which has no released threshold.

Variant	N'/R	N'/N	V/N'	G/N'	judged fake
WaveRep (demo contract, all frames), CogVideoX1.5					
bicubic, released CRF	0.52 [0.48,0.55]	–	0.94 [0.92,0.95]	0.98 [0.98,0.99]	4%
area, released CRF	0.57 [0.54,0.61]	0.56 [0.55,0.58]	0.92 [0.90,0.93]	0.98 [0.97,0.99]	4%
bicubic, CRF 18	0.69 [0.66,0.71]	0.69 [0.67,0.71]	0.86 [0.84,0.88]	0.96 [0.95,0.97]	8%
area, CRF 18	0.73 [0.70,0.76]	0.73 [0.71,0.76]	0.83 [0.81,0.85]	0.95 [0.94,0.96]	11%
step: encode only	0.47 [0.46,0.48]	0.45 [0.42,0.48]	0.94 [0.93,0.96]	0.99 [0.98,0.99]	2%
step: rescale	0.52 [0.48,0.55]	0.50 [0.49,0.51]	0.94 [0.92,0.95]	0.98 [0.98,0.99]	4%
step: time resample	0.49 [0.48,0.49]	0.47 [0.43,0.50]	0.94 [0.92,0.95]	0.98 [0.98,0.99]	2%
WaveRep (demo contract, all frames), LTX					
bicubic, released CRF	0.58 [0.55,0.61]	–	0.99 [0.98,0.99]	0.99 [0.99,1.00]	6%
area, released CRF	0.63 [0.60,0.66]	0.56 [0.55,0.57]	0.98 [0.98,0.99]	0.99 [0.99,1.00]	9%
bicubic, CRF 18	0.75 [0.73,0.78]	0.69 [0.68,0.71]	0.97 [0.96,0.98]	0.99 [0.98,0.99]	14%
area, CRF 18	0.80 [0.78,0.83]	0.75 [0.73,0.77]	0.96 [0.95,0.97]	0.98 [0.97,0.99]	19%
step: encode only	0.47 [0.46,0.48]	0.39 [0.36,0.42]	1.00 [0.99,1.00]	1.00 [1.00,1.00]	2%
step: rescale	0.62 [0.59,0.65]	0.55 [0.54,0.56]	0.99 [0.98,0.99]	0.99 [0.99,1.00]	7%
step: time resample	0.47 [0.46,0.48]	0.39 [0.36,0.42]	1.00 [0.99,1.00]	1.00 [1.00,1.00]	2%
WaveRep, toolkit variant (16 frames), CogVideoX1.5					
bicubic, released CRF	0.53 [0.49,0.56]	–	0.93 [0.92,0.95]	0.98 [0.97,0.99]	4%
area, released CRF	0.58 [0.55,0.61]	0.56 [0.56,0.58]	0.91 [0.90,0.93]	0.98 [0.97,0.98]	5%
bicubic, CRF 18	0.70 [0.67,0.73]	0.69 [0.67,0.71]	0.86 [0.84,0.88]	0.96 [0.94,0.97]	9%
area, CRF 18	0.74 [0.71,0.77]	0.73 [0.72,0.76]	0.83 [0.80,0.85]	0.95 [0.93,0.96]	12%
WaveRep, toolkit variant (16 frames), LTX					
bicubic, released CRF	0.59 [0.56,0.62]	–	0.99 [0.98,0.99]	0.99 [0.99,1.00]	6%
area, released CRF	0.64 [0.61,0.67]	0.56 [0.55,0.57]	0.98 [0.98,0.99]	0.99 [0.99,1.00]	8%
bicubic, CRF 18	0.76 [0.73,0.79]	0.69 [0.67,0.71]	0.97 [0.96,0.98]	0.99 [0.98,0.99]	13%
area, CRF 18	0.80 [0.77,0.83]	0.74 [0.72,0.77]	0.96 [0.95,0.97]	0.98 [0.97,0.99]	18%
D3 (authors' operator), CogVideoX1.5					
bicubic, released CRF	0.38 [0.35,0.41]	–	0.50 [0.48,0.51]	0.73 [0.70,0.76]	–
area, released CRF	0.38 [0.35,0.41]	0.50 [0.48,0.52]	0.50 [0.48,0.51]	0.73 [0.70,0.76]	–
bicubic, CRF 18	0.39 [0.36,0.41]	0.51 [0.49,0.52]	0.49 [0.47,0.51]	0.73 [0.70,0.76]	–
area, CRF 18	0.39 [0.35,0.41]	0.51 [0.49,0.52]	0.49 [0.47,0.51]	0.73 [0.70,0.76]	–
step: encode only	0.50 [0.48,0.51]	0.61 [0.59,0.65]	0.38 [0.35,0.41]	0.63 [0.59,0.67]	–
step: rescale	0.48 [0.46,0.51]	0.61 [0.58,0.64]	0.39 [0.36,0.42]	0.65 [0.61,0.68]	–
step: time resample	0.39 [0.37,0.42]	0.51 [0.50,0.53]	0.48 [0.46,0.50]	0.73 [0.69,0.76]	–
D3 (authors' operator), LTX					
bicubic, released CRF	0.34 [0.31,0.37]	–	0.48 [0.46,0.50]	0.66 [0.62,0.69]	–
area, released CRF	0.34 [0.31,0.36]	0.50 [0.49,0.51]	0.48 [0.46,0.50]	0.66 [0.62,0.69]	–
bicubic, CRF 18	0.34 [0.31,0.36]	0.50 [0.49,0.51]	0.48 [0.46,0.50]	0.66 [0.62,0.69]	–
area, CRF 18	0.34 [0.32,0.37]	0.51 [0.49,0.52]	0.47 [0.45,0.49]	0.65 [0.62,0.68]	–
step: encode only	0.50 [0.48,0.52]	0.66 [0.63,0.69]	0.32 [0.29,0.35]	0.50 [0.46,0.53]	–
step: rescale	0.48 [0.46,0.50]	0.65 [0.62,0.67]	0.33 [0.30,0.36]	0.51 [0.48,0.54]	–
step: time resample	0.36 [0.33,0.38]	0.52 [0.51,0.54]	0.46 [0.43,0.48]	0.64 [0.60,0.67]	–
D3-inspired toolkit variant, CogVideoX1.5					
bicubic, released CRF	0.50 [0.48,0.51]	–	0.48 [0.47,0.49]	0.80 [0.78,0.83]	–
area, released CRF	0.50 [0.48,0.51]	0.50 [0.49,0.50]	0.49 [0.48,0.49]	0.80 [0.78,0.83]	–
bicubic, CRF 18	0.50 [0.49,0.52]	0.50 [0.49,0.51]	0.48 [0.47,0.49]	0.80 [0.78,0.83]	–
area, CRF 18	0.50 [0.49,0.52]	0.50 [0.49,0.51]	0.48 [0.47,0.49]	0.80 [0.78,0.83]	–
D3-inspired toolkit variant, LTX					
bicubic, released CRF	0.50 [0.49,0.51]	–	0.44 [0.43,0.46]	0.68 [0.65,0.71]	–
area, released CRF	0.50 [0.49,0.52]	0.50 [0.49,0.51]	0.44 [0.42,0.45]	0.68 [0.65,0.71]	–
bicubic, CRF 18	0.50 [0.49,0.51]	0.50 [0.49,0.50]	0.45 [0.43,0.46]	0.68 [0.65,0.71]	–
area, CRF 18	0.50 [0.49,0.51]	0.50 [0.49,0.50]	0.45 [0.43,0.46]	0.68 [0.65,0.71]	–

the control even where G/R is low; they need not share a mechanism. Skyra (Li et al., 2026) (the released `Skyra-RL` checkpoint on Qwen2.5-VL-7B) was run under two input serialisations, both with the authors' 16 frames at short side 256, their pixel budget, greedy decoding, and an 8,192-token limit: the toolkit serialisation places the frames before the prompt without timestamps, whereas the authors' inference engine interleaves each frame with a line "[T=t s]". On the 16+16 ViF-Bench videos the non-author serialisation yields degenerate verdicts (real for all 32, AUC 0.50) and the author-aligned configuration recovers 15 of the 16 generated and 15 of the 16 real decisions (AUC 0.94), which invalidates the non-transfer reading that the degenerate run had suggested; what differs between the two runs is the complete input configuration, so the recovery is not attributed to the

timestamp line alone. The toolkit's single-token score is not used for Skyra because it does not consistently target the complete final answer ("Fake" spans several tokens and the extraction can land on a token of the analysis); its score is the deterministically parsed final verdict (1 fake, 0 real; all 3,900 answers parsed), AUROC is computed from these binary verdicts with half credit for ties, so each AUC equals $\frac{1}{2} + \frac{1}{2}(p_A - p_B)$ for the judged-fake rates p of the two arms, and the non-author serialisation appears only in Table 28. For CogVideoX1.5 and LTX it judges 94–97% of the format controls, 92–99% of the VAE round trips, and 96–99% of the released outputs fake against 6% of the originals, so the released outputs are separated from the originals (G/R 0.95–0.97) but not from the format control (G/N 0.51; N/R 0.94–0.96): non-generative processing is sufficient to trigger the fake verdict on these source-matched controls, and the binary verdict adds little discrimination between G and N; this does not identify a shared internal cue. For SVD and Pyramid-Flow it answers real for controls and released outputs alike (0–1% judged fake; N/R 0.47, G/R 0.47–0.48), so its detection sensitivity on these two families is low, and the absence of false alarms on their controls is not evidence of robustness.

Table 27: Additional detectors on the generated-video arms (video-level AUC; brackets: 95% paired bootstrap; G is the released video-to-video output, image-to-video for SVD and Pyramid-Flow). n: sources with every arm of the row scored (the toolkit decoder and the Ivy-xDetector video reader skip one source with a corrupt tail; Skyra's reader decodes it). Judged fake: fraction of R/N/V/G videos above 0.5 on each detector's probability, which for Ivy-xDetector coincides with its parsed verdict; Skyra is scored by its parsed verdict, so its AUCs are binary-verdict AUROCs with half credit for ties. The LTX VAE arm was not evaluated under the Ivy-xDetector authors' contract; the non-author Skyra serialisation is reported only in Table 28.

Detector	Family	n	N/R	V/N	G/R	G/N	judged fake
AIGVDet	CogVideoX1.5	299	0.48 [0.44,0.53]	0.45 [0.44,0.46]	0.60 [0.55,0.65]	0.59 [0.57,0.62]	5/33/24/46%
AIGVDet	LTX	299	0.42 [0.38,0.46]	0.47 [0.45,0.48]	0.44 [0.40,0.48]	0.52 [0.50,0.53]	5/21/15/23%
AIGVDet	SVD	299	0.61 [0.56,0.66]	–	0.39 [0.34,0.44]	0.34 [0.32,0.36]	5/2/–/0%
AIGVDet	Pyramid-Flow	299	0.34 [0.29,0.38]	–	0.30 [0.26,0.34]	0.45 [0.43,0.47]	5/7/–/6%
L3DE	CogVideoX1.5	299	0.55 [0.53,0.58]	0.63 [0.61,0.65]	0.72 [0.69,0.75]	0.67 [0.63,0.70]	8/10/18/24%
L3DE	LTX	299	0.57 [0.54,0.59]	0.83 [0.80,0.85]	0.87 [0.85,0.90]	0.84 [0.81,0.87]	8/12/57/59%
L3DE	SVD	299	0.63 [0.60,0.66]	–	0.96 [0.94,0.97]	0.94 [0.92,0.96]	8/14/–/87%
L3DE	Pyramid-Flow	299	0.62 [0.59,0.64]	–	0.93 [0.91,0.94]	0.87 [0.85,0.90]	8/17/–/71%
Ivy-xDetector, authors' contract	CogVideoX1.5	299	0.37 [0.35,0.39]	0.51 [0.49,0.52]	0.67 [0.63,0.71]	0.80 [0.77,0.83]	83/67/70/98%
Ivy-xDetector, authors' contract	LTX	299	0.39 [0.36,0.41]	–	0.55 [0.52,0.58]	0.67 [0.64,0.70]	83/71/–/90%
Ivy-xDetector, authors' contract	SVD	299	0.25 [0.22,0.28]	–	0.46 [0.43,0.50]	0.72 [0.69,0.75]	83/46/–/79%
Ivy-xDetector, authors' contract	Pyramid-Flow	299	0.29 [0.27,0.31]	–	0.40 [0.36,0.43]	0.62 [0.59,0.65]	83/56/–/72%
Ivy-xDetector, toolkit contract	CogVideoX1.5	299	0.35 [0.33,0.37]	0.51 [0.50,0.53]	0.52 [0.49,0.55]	0.65 [0.62,0.68]	12/7/8/15%
Ivy-xDetector, toolkit contract	LTX	299	0.33 [0.31,0.36]	0.58 [0.57,0.60]	0.44 [0.41,0.47]	0.60 [0.58,0.63]	12/7/10/9%
Ivy-xDetector, toolkit contract	SVD	299	0.21 [0.18,0.23]	–	0.46 [0.42,0.50]	0.75 [0.72,0.78]	12/2/–/10%
Ivy-xDetector, toolkit contract	Pyramid-Flow	299	0.26 [0.23,0.28]	–	0.27 [0.24,0.30]	0.54 [0.52,0.57]	12/3/–/3%
Skyra, authors' contract	CogVideoX1.5	300	0.96 [0.94,0.97]	0.51 [0.50,0.52]	0.97 [0.95,0.98]	0.51 [0.50,0.52]	6/97/99/99%
Skyra, authors' contract	LTX	300	0.94 [0.92,0.96]	0.49 [0.48,0.51]	0.95 [0.94,0.97]	0.51 [0.50,0.53]	6/94/92/96%
Skyra, authors' contract	SVD	300	0.47 [0.46,0.48]	–	0.48 [0.46,0.49]	0.51 [0.50,0.51]	6/0/–/1%
Skyra, authors' contract	Pyramid-Flow	300	0.47 [0.46,0.48]	–	0.47 [0.46,0.49]	0.50 [0.50,0.51]	6/0/–/1%

Table 28: Own-benchmark check of the two vision-language detectors under the two contracts: 16 real and 16 generated videos from each model's own evaluation set (ViF-Bench for Skyra, one generated video per generator; the Ivy-Fake test split for Ivy-xDetector, two per generator with Kinetics-400 and MSR-VTT reals), scored on the same harness as the arms. AUC on the soft score (the verdict for Skyra), balanced accuracy of the verdicts, the fraction of real and of generated videos answered "fake", and the number of clips without a parsable verdict.

Contract	Benchmark	n	AUC	bal. acc.	real→fake	gen.→fake	no verdict
Skyra, authors' contract (timestamped frames)	ViF-Bench	16+16	0.94	0.94	6%	94%	0
Skyra, authors' contract (timestamped frames)	Ivy-Fake	16+16	0.75	0.75	0%	50%	0
Skyra, toolkit prompt (no timestamps)	ViF-Bench	16+16	0.50	0.50	0%	0%	0
Skyra, toolkit prompt (no timestamps)	Ivy-Fake	16+16	0.62	0.62	0%	25%	0
Ivy-xDetector, authors' contract (1 fps video, $\leq$6 frames)	ViF-Bench	16+16	0.69	0.66	31%	62%	0
Ivy-xDetector, authors' contract (1 fps video, $\leq$6 frames)	Ivy-Fake	16+16	1.00	0.94	12%	100%	0
Ivy-xDetector, toolkit contract (16 frames)	ViF-Bench	16+16	0.63	0.53	0%	6%	0
Ivy-xDetector, toolkit contract (16 frames)	Ivy-Fake	16+16	1.00	0.88	0%	75%	0

K TRAINING DETAILS AND PER-SEED VALUES

Data. The interventions use the SimSwap four arms of the 706 training and 128 validation real videos of the split in Appendix B, with 16 uniformly spaced frames per video and the crop rule of Appendix A at 299 pixels; evaluation uses the 32-frame test crops of both generators (and of GHOST in Appendix K.1). The four arms of a frame are loaded together and receive one horizontal-flip draw, so an augmentation never distinguishes arms.

Model and optimisation. All recipes fine-tune an ImageNet-initialised Xception-299 (the DeepfakeBench backbone) with AdamW, peak learning rate 10^{-4}, weight decay 10^{-4}, a one-cycle schedule with 10% warm-up over five epochs, binary cross-entropy with logits, and seeds 303, 304, 305 that fix the initialisation of the head, the data order, and the flip draws. The negatives recipes (U_{base}, U_{matched}, U_{both}) sample frames with a weighted sampler that balances negatives and positives, batch 64 frames, and evaluate the epoch with the best validation video AUC on the recipe's own task. The fixed-support recipes (T_{syn}, T_{id}, T_{joint}) sample 16 source frames per step and use all four arms of each (64 crops per step), with identical minibatches, initialisation, and flip draws for a given seed across the three recipes, and evaluate the last epoch without selection; the epoch with the best validation AUC on each recipe's own axis differs from the last epoch by at most 0.006 in F/S for T_{id} and the identity head of T_{joint} and by at most 0.025 for T_{syn}, whose validation axis is the synthesis axis. T_{joint} has two linear heads on the shared backbone and minimises $\frac{1}{2}(\mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{id}})$; each head is evaluated on its own axis.

Training exposure (from the run logs). One epoch of a negatives recipe is n draws with replacement from a sampler whose weights are $1/n_+$ for the F frames and $1/n_-$ for the negative frames, so every batch of 64 holds positives and negatives in equal expectation, where n is the number of training frames of the recipe: 22,592 for U_{base} (706 R and 706 F videos $\times$ 16 frames), 22,592 for U_{matched} (706 N and 706 F), and 33,888 for U_{both} (706 R, 706 N, 706 F); with the last incomplete batch dropped this is 353, 353, and 529 optimizer steps per epoch and 1,765, 1,765, and 2,645 over the five epochs, the one-cycle schedule spanning each recipe's own total. U_{base} and U_{matched} therefore differ only in which arm supplies the negatives; U_{both} also takes about 1.5 times as many updates per epoch, so its comparison with the other two is not exposure-matched. Checkpoints are selected by the unrounded validation video AUROC on the recipe's own task, ties resolved by the earliest epoch; it reaches exactly 1.0 in the first epoch for every seed of U_{base} and U_{matched} and for seeds 303 and 305 of U_{both} (seed 304: 0.99997 at epoch 1 and 1.0 at epoch 2), so the selected epochs are 1/1/1, 1/1/1, and 1/2/1 (seeds 303/304/305), that is, after 353 updates for the first two recipes and 529 or 1,058 for the third. One epoch of a fixed-support recipe visits each of the 11,296 training source frames (706 sources $\times$ 16 frames) once, in a seeded shuffled order shared by T_{syn}, T_{id}, and T_{joint}, 16 source frames per step with all four arms of each and equal weight, so 706 steps per epoch and 3,530 in total; the last epoch is evaluated, and the epochs with the best validation AUC on each recipe's own axis are 1/1/1 (T_{syn}), 5/3/5 (T_{id}), and 4/5/4 (T_{joint}). The earlier free-sampling U_{edit} drew 45,184 frames per epoch (706 steps, 3,530 in total) and selected epochs 3/5/5.

Sensitivity to the selected epoch. Because the negatives recipes select epoch 1 (seed 304 of U_{both}: epoch 2), we retrained the three recipes with identical arguments and seeds, saved every epoch, and scored all 45 checkpoints with the evaluation code and test manifests of Table 2 (Table 29). The retrained runs select the same epochs (1/1/1, 1/1/1, and 1/2/1) and reproduce the reported selected-epoch contrasts exactly (every retrained selected-epoch AUC equals the reported value). Over the five epochs the gap widens: N/R of U_{base} rises from 0.71 to 0.77 on SimSwap and from 0.91 to 0.94 on InSwapper, while that of U_{matched} falls from 0.43 to 0.30 and from 0.55 to 0.37 (last epoch 0.77/0.94 and 0.30/0.37); S/R stays at 1.00 on SimSwap for every recipe, and on InSwapper the cross-generator loss of U_{matched} noted in Section 5.3 grows with training (S/R 0.79 at epoch 1, 0.55–0.62 afterwards). The three-seed mean F/S stays within 0.64–0.70 across the evaluated recipes, epochs, and generators. Pipeline-matched negatives yield lower N/R than original negatives at every evaluated epoch on both generators, so the pipeline-response reduction is not specific to the selected checkpoint.

Targets. With arms ordered (R, N, S, F), T_{syn} uses the targets $(0, 0, 1, 1)$, T_{id} uses $(0, 0, 0, 1)$, and T_{joint} trains both. The negatives recipes use the positives F and the negatives R (U_{base}), N (U_{matched}), or both (U_{both}), which is a target vector of $(0, \cdot, \cdot, 1)$ with the unused arms dropped from the sampler

Table 29: Negatives recipes at every epoch: retrained with identical arguments and seeds and every epoch saved; cells are seed mean ± sample standard deviation over seeds 303/304/305 of the video-level AUC on the 136 test sources of each generator, and FF++ is the video AUC on the four FaceForensics++ manipulations averaged over manipulations. Table 2 reports epoch 1 (seed 304 of U_{both}: epoch 2).

		SimSwap (training generator)			InSwapper (unseen)			
Recipe	Epoch	N/R	S/R	F/S	N/R	S/R	F/S	FF++
U_{base}	1	0.71±0.04	1.00±0.00	0.65±0.02	0.91±0.04	0.98±0.01	0.66±0.02	0.53±0.03
	2	0.76±0.04	1.00±0.00	0.65±0.01	0.95±0.02	0.99±0.01	0.67±0.02	0.54±0.03
	3	0.77±0.02	1.00±0.00	0.64±0.02	0.95±0.01	0.99±0.00	0.67±0.02	0.53±0.01
	4	0.78±0.02	1.00±0.00	0.64±0.03	0.95±0.02	0.99±0.01	0.66±0.03	0.54±0.01
	5	0.77±0.05	1.00±0.00	0.64±0.03	0.94±0.04	0.98±0.02	0.66±0.03	0.54±0.02
U_{matched}	1	0.43±0.01	1.00±0.00	0.66±0.01	0.55±0.02	0.79±0.02	0.67±0.01	0.47±0.02
	2	0.36±0.03	1.00±0.00	0.66±0.04	0.43±0.03	0.62±0.10	0.67±0.02	0.45±0.03
	3	0.34±0.05	1.00±0.00	0.66±0.03	0.43±0.13	0.62±0.21	0.68±0.02	0.44±0.02
	4	0.30±0.04	1.00±0.00	0.67±0.02	0.35±0.10	0.55±0.16	0.70±0.01	0.44±0.02
	5	0.30±0.04	1.00±0.00	0.67±0.02	0.37±0.10	0.59±0.13	0.70±0.01	0.44±0.02
U_{both}	1	0.53±0.05	1.00±0.00	0.65±0.02	0.67±0.12	0.83±0.11	0.66±0.04	0.51±0.02
	2	0.49±0.04	1.00±0.00	0.64±0.04	0.62±0.09	0.79±0.07	0.64±0.01	0.49±0.02
	3	0.50±0.00	1.00±0.00	0.65±0.03	0.65±0.04	0.82±0.04	0.67±0.00	0.51±0.01
	4	0.47±0.02	1.00±0.00	0.66±0.02	0.61±0.09	0.77±0.11	0.67±0.01	0.50±0.02
	5	0.49±0.00	1.00±0.00	0.66±0.02	0.64±0.05	0.82±0.06	0.68±0.01	0.50±0.02

rather than labelled. An earlier free-sampling recipe U_{edit} (negatives R, N, S; positives F) reached F/S 0.968 ± 0.007 on SimSwap and 0.830 ± 0.016 on InSwapper with S/R 0.27 on InSwapper; the InSwapper value did not reappear under the fixed-support protocol and is reported here only for completeness, with the difference between the two protocols (frame sampling, the presence of all four arms in every step, and last-epoch versus selected-epoch evaluation) left unresolved.

Per-seed values. Table 30 lists per-seed F/S and N/R for the fixed-support recipes. The seed dispersion of F/S on the training generator is below 0.01 for T_{syn}, T_{id}, and the identity head of T_{joint}; on the unseen generator it is 0.01–0.04, and the per-seed paired gain of T_{id} over T_{syn} on InSwapper is 0.023 ± 0.019 (Section 5.2). On InSwapper the joint identity head's pooled A_{id} exceeds that of T_{id} (0.74 against 0.62) because F/R rises from 0.56 to 0.79 and F/N from 0.61 to 0.74: swaps are ranked higher above unprocessed and pipeline-only videos, while their ranking above self-swaps barely moves.

K.1 Mixed-generator training and the held-out GHOST

The fixed-support recipes of Section 5.1 were retrained on the four arms of both training generators: the SimSwap and InSwapper training and validation arms of the same 706 and 128 real videos, 16 frames per video and generator, pooled into one set in which every minibatch group holds the four arms of one frame from one generator. Initialisation, optimiser, schedule, the five epochs, the seeds, and the last-epoch rule are unchanged, so a mixed run takes 7,060 updates against 3,530 for a SimSwap-only run; an update-matched control trains the SimSwap-only T_{syn} and T_{id} for ten epochs (7,060 updates). All models and the checkpoints of Table 2 were scored on the test four arms of SimSwap, InSwapper, and GHOST (Appendix G); the re-scored checkpoints reproduce their SimSwap and InSwapper values exactly (Table 31). On GHOST, which no training run saw, the SimSwap-only identity target raises F/S from 0.56 to 0.69 (gain $+0.14$ $[+0.09, +0.18]$) and W_{id} from 0.60 to 0.76, whereas the same checkpoints show no confirmed gain on InSwapper ($+0.02$ $[-0.02, +0.07]$; ten epochs: $+0.06$ $[+0.01, +0.10]$). Training on both generators lifts the InSwapper F/S of T_{id} to 0.98 (0.62 for T_{syn}), but its gain on GHOST, $+0.18$ $[+0.13, +0.23]$, shows no confirmed difference from that of the update-matched control, $+0.17$ $[+0.12, +0.22]$ (difference $+0.01$ $[-0.04, +0.06]$, one set of source indices for both). Mixed training raises the GHOST F/S of T_{id} (0.77 against 0.72 for the control) and of T_{syn} (0.59 against 0.55) alike, and the joint model's identity head reaches 0.76. Because the GHOST S and F use its adaptive paste-back, the GHOST F/S includes changes in the blending mask. Within this training procedure the label gain therefore transfers partially and unevenly across the tested unseen generators; on GHOST, mixed training shows no confirmed increase in this gain over the update-matched SimSwap-only control. These additional evaluations and training runs

Table 30: Fixed-support recipes, per-seed values (last epoch): F/S, N/R, S/N, and F/N per head and generator.

		F/S SimSwap			F/S InSwapper		
Recipe	Head	s303	s304	s305	s303	s304	s305
T_{syn}	syn	0.545	0.552	0.548	0.640	0.658	0.668
T_{id}	id	0.974	0.978	0.967	0.670	0.660	0.706
T_{joint}	syn	0.502	0.453	0.427	0.654	0.685	0.684
T_{joint}	id	0.977	0.978	0.976	0.654	0.725	0.701
N/R			SimSwap			InSwapper	
T_{syn}	syn	0.414	0.453	0.479	0.496	0.553	0.650
T_{id}	id	0.440	0.421	0.427	0.419	0.391	0.421
T_{joint}	syn	0.503	0.462	0.477	0.684	0.652	0.637
T_{joint}	id	0.470	0.417	0.463	0.578	0.518	0.586
S/N			SimSwap			InSwapper	
T_{syn}	syn	1.000	1.000	1.000	0.688	0.673	0.753
T_{id}	id	0.640	0.602	0.694	0.438	0.398	0.400
T_{joint}	syn	1.000	1.000	1.000	0.735	0.668	0.725
T_{joint}	id	0.741	0.770	0.676	0.586	0.494	0.617
F/N			SimSwap			InSwapper	
T_{syn}	syn	1.000	1.000	1.000	0.809	0.813	0.869
T_{id}	id	0.996	0.994	0.998	0.620	0.578	0.639
T_{joint}	syn	1.000	1.000	1.000	0.829	0.828	0.853
T_{joint}	id	0.995	0.996	0.999	0.713	0.721	0.776

were added after the other results; as for the exploratory additions listed in Section 7, their intervals do not account for this selection.

Table 31: Identity target trained on one or two generators and tested on the held-out GHOST (three seeds, last epoch; mean±s.d. over seeds). Bottom: label gain on F/S with source-level paired-bootstrap intervals; gains are computed from unrounded AUCs. The GHOST S and F use its adaptive paste-back (Appendix G).

	F/S SimSwap	F/S InSwapper	F/S GHOST	W_{id} GHOST	F/R GHOST
SimSwap training arms, 5 epochs (the checkpoints of Table 2, re-scored)					
T_{syn}, syn head	0.55±0.00	0.66±0.01	0.56±0.01	0.60±0.01	0.91±0.08
T_{id}, id head	0.97±0.01	0.68±0.02	0.69±0.01	0.76±0.02	0.77±0.03
T_{joint}, id head	0.98±0.00	0.69±0.04	0.63±0.00	0.69±0.01	0.87±0.04
SimSwap training arms, 10 epochs (update-matched to the mixed runs)					
T_{syn}, syn head	0.54±0.01	0.65±0.01	0.55±0.01	0.56±0.00	0.93±0.06
T_{id}, id head	0.97±0.00	0.71±0.04	0.72±0.00	0.77±0.01	0.80±0.03
SimSwap and InSwapper training arms, 5 epochs					
T_{syn}, syn head	0.52±0.02	0.62±0.02	0.59±0.00	0.62±0.01	1.00±0.00
T_{id}, id head	0.96±0.00	0.98±0.00	0.77±0.02	0.84±0.01	0.92±0.02
T_{joint}, id head	0.97±0.00	0.98±0.00	0.76±0.01	0.85±0.01	0.99±0.01
Label gain on F/S, T_{id} minus T_{syn} (three-seed mean, 95% source-level paired bootstrap)					
SimSwap, 5 epochs	+0.42 [.40,.45]	+0.02 [−.02,.07]	+0.14 [.09,.18]		
SimSwap, 10 epochs	+0.43 [.40,.46]	+0.06 [.01,.10]	+0.17 [.12,.22]		
SimSwap+InSwapper	+0.45 [.42,.48]	+0.36 [.32,.39]	+0.18 [.13,.23]		

L Compute, software, and reproducibility

All generation, scoring, and training ran on one cluster partition with NVIDIA H200 GPUs (one GPU per job); Table 32 lists representative wall-clock costs. Building the two face-swap four-arm sets for the 136 test videos took 37 s (SimSwap) and 20 s (InSwapper) per video on one GPU, dominated by per-frame detection and the two generator passes; the training and validation arms (834 videos)

were built the same way. One training run of a fixed-support recipe takes 3.5–6 minutes and one negatives recipe 2–3 minutes, so the 18 training runs behind Table 2 amount to about one GPU hour and the 15 runs of Appendix K.1 to about 1.7; the epoch sweep of Table 29 (nine retraining runs and the scoring of 45 checkpoints) took about 40 minutes on one eight-GPU node. Scoring the 29 checkpoints on the four arms of both generators and both encodings is dominated by the full-frame pipelines, whose authors' preprocessing (detection, landmarks, alignment) runs once per pipeline and arm. For the generated-video control, building the format control and the VAE round trip for all families costs 8 s per 1080p source, and scoring one arm of 300 videos costs 20–30 s for DeCoF and D3 and about four minutes for the two-stream AIGVDet. Software: PyTorch 2.14.0 with CUDA 12.6 (torchvision 0.29.0), diffusers 0.39.0 for the VAEs, transformers 4.57.6 (5.17.0 in the face environment), ffmpeg 7.0.2 (static build, via imageio-ffmpeg) with libx264 for every encoding, PyAV 17.1.0 and OpenCV 4.11 for decoding, insightface 2.0 with onnxruntime 1.23.2 (`buffalo_l`, `inswapper_128.onnx`) for detection, recognition, and InSwapper, the authors' SimSwap repository, DeepfakeBench and the DF40 codebase for the frame-level checkpoints, and the VidAudit toolkit for the WaveRep, D3, AIGVDet, and L3DE wrappers. We do not redistribute the source or the derived videos; users obtain the source datasets from their distributors and comply with the applicable terms, and the released scripts rebuild the arms. The construction scripts, the donor map, the per-frame detections, the per-video and per-frame predictions of every checkpoint, the stored penultimate features with the released classifier parameters and source-split records of Appendix H.1, and the evaluation code will be released after the review period; the per-video predictions allow every AUC, interval, and video-level judged-fake rate to be recomputed without the videos, the frame-level flip rates need the per-frame predictions, and the construction checks and timings come from separate records.

Data use and ethics. The face arms are derived from the YouTube-sourced videos of FaceForensics++, accessed under the distributor's terms, with three released face-swap generators; we recruited no participants and collected no footage, and dataset access is not treated as evidence of individual consent to every downstream use. Every donor is another subject of the same dataset, so no new identity is introduced, and no complete swapped video is distributed (Figure 1 shows a few labelled frames). The generated-video arms are derived from the AIGVDBench test split under its licence. The protocol is a measurement tool for detectors and adds no capability to forgery generation; the dual-use risk of failure diagnostics is discussed in the ethics statement.

Table 32: Representative wall-clock costs on one H200 GPU (means over jobs; ranges over videos or seeds where given).

Step	Cost
SimSwap four arms, one 101-frame video (detection, alignment, generator twice, paste-back, PNG)	37 s (range 13–131)
InSwapper four arms, one 101-frame video	20 s (range 8–59)
Training, negatives recipe (Xception-299, five epochs, 706 videos $\times$ 16 frames), one seed	111–164 s
Training, fixed-support recipe (four arms per source), one seed	205–375 s
Format control N and VAE round trip V for one 1080p source (all four families)	8 s
DeCoF, one arm of 300 videos (eight frames each)	29 s
DeCoF uniform-frame contract, one arm of 300 videos	19 s
AIGVDet (RGB and optical-flow streams), one arm of 300 videos	254 s