

Online Resource 1: Extended Experimental Results, Statistical Tests, Ablations, Failure Analysis, Robustness Results, and Leakage-Control Procedure

Samarth Kukreja^{1*}

^{1*}Independent Researcher, Faridabad, Haryana, India.

Corresponding author(s). E-mail(s): samarthkukreja2003@gmail.com;

Journal: Signal, Image and Video Processing

Article: Cross-Dataset Generalization in Deepfake Detection: A Controlled Study of Spatial, Frequency, and Video-Level Baselines

1 Extended Results

This supplement preserves the complete result tables, statistical tests, manipulation-level breakdowns, ablations, and robustness analyses condensed from the main manuscript. The headline metrics retain the original evaluations; Tables S3–S7 use a separate regenerated-score evaluation at the same training seed. Dataset abbreviations are FaceForensics++ (FF++), Celeb-DF, and DeepFake Detection (DFD). Metric abbreviations are area under the receiver operating characteristic curve (AUC), confidence interval (CI), equal error rate (EER), and test-time augmentation (TTA).

Training of the five red–green–blue (RGB) backbones across the four main experiments is complete, with two additional probes reported below: a frequency-only ResNet-18 baseline for Exp1/Exp2 and a no-retraining score-level temporal aggregation baseline. Unless stated, all metrics are *video-level* (face-crop scores aggregated per video). Accuracy and F1 use a fixed threshold of 0.5; AUC is threshold-independent and EER is derived from the error-rate curves. Reported 95% bootstrap intervals apply to AUC.

1.1 Backbone comparison

Table S1 reports video-level AUC for every backbone and experiment, and Table S2 gives the corresponding accuracy, F1 and EER. These values are interpreted as single-seed point estimates; the paired tests below assess held-out-video uncertainty but not training-seed variability. ConvNeXt-Tiny and ResNet-50 yield the two highest observed AUCs in Exp1, Exp2 and Exp4. ConvNeXt-Tiny is highest by point estimate, but its gap over ResNet-50 is not significant under the reported Holm-adjusted procedures in Exp1/Exp2. ViT-Base does not lead under this protocol, and Xception has the lowest in-distribution point estimate. These observations do not establish a general convolutional advantage or a stable architectural ranking. Because the five backbones share a common baseline pipeline, architecture is the deliberately varied factor, but each architecture is not independently optimized. The EfficientNet-B4 ablation further shows that removing progressive unfreezing can materially improve that model.

1.2 Statistical significance of backbone differences

Because all results are single-seed, we test whether the point-estimate gaps in Table S1 are statistically supported before drawing architectural

Table S1 Video-level AUC (95% bootstrap CI) per backbone across the four experiments. The best *point estimate* per column is in **bold**; because several CIs overlap, the bolded value is not always statistically separated from the next model — see the paired significance tests (Tables S3–S4). Exp3 shows weak or near-chance transfer, so its column ranking should be interpreted cautiously.

Backbone	Exp1: FF++→FF++	Exp2: FF++→Celeb-DF	Exp3: Celeb-DF→FF++	Exp4: Mixed→DFD
EfficientNet-B4	0.900 (0.85–0.94)	0.713 (0.63–0.80)	0.506 (0.47–0.54)	0.788 (0.75–0.82)
Xception	0.736 (0.65–0.81)	0.573 (0.47–0.67)	0.543 (0.51–0.57)	0.735 (0.70–0.77)
ResNet-50	0.924 (0.88–0.97)	0.799 (0.72–0.87)	0.589 (0.56–0.62)	0.812 (0.78–0.84)
ConvNeXt-Tiny	0.957 (0.92–0.99)	0.806 (0.72–0.88)	0.576 (0.54–0.61)	0.915 (0.89–0.93)
ViT-Base	0.867 (0.80–0.93)	0.668 (0.57–0.76)	0.588 (0.56–0.62)	0.713 (0.67–0.75)

Table S2 Video-level Accuracy, F1 and EER per backbone, grouped by experiment (lower EER is better).

Experiment	Backbone	Acc	F1	EER
Exp1: FF++→FF++	EfficientNet-B4	0.799	0.741	0.194
	Xception	0.671	0.680	0.322
	ResNet-50	0.859	0.835	0.174
	ConvNeXt-Tiny	0.886	0.866	0.087
	ViT-Base	0.765	0.759	0.202
Exp2: FF++→Celeb-DF	EfficientNet-B4	0.659	0.633	0.324
	Xception	0.535	0.552	0.442
	ResNet-50	0.744	0.680	0.227
	ConvNeXt-Tiny	0.767	0.727	0.240
	ViT-Base	0.659	0.569	0.372
Exp3: Celeb-DF→FF++	EfficientNet-B4	0.498	0.287	0.490
	Xception	0.484	0.232	0.455
	ResNet-50	0.516	0.152	0.445
	ConvNeXt-Tiny	0.523	0.094	0.449
	ViT-Base	0.528	0.167	0.444
Exp4: Mixed→DFD	EfficientNet-B4	0.688	0.598	0.271
	Xception	0.614	0.686	0.331
	ResNet-50	0.694	0.582	0.243
	ConvNeXt-Tiny	0.710	0.589	0.159
	ViT-Base	0.650	0.602	0.338

conclusions. For Exp1 and Exp2 and each pair of backbones we compute, on the common set of test videos, (i) a *paired bootstrap* on the AUC difference (2000 resamples over videos, with a 95% CI and an empirical two-sided bootstrap-tail value) and (ii) an *exact McNemar test* on the binary decisions at the fixed 0.5 threshold. The pairwise outcomes for the two experiments with clear signal are reported in Table S3 (in-distribution Exp1) and Table S4 (cross-dataset FF++→Celeb-DF). **Scope of these tests:** the paired bootstrap resamples *test videos*, whereas exact McNemar tests use discordant binary decisions; they capture test-set sampling uncertainty *at a single training seed only*; they do *not* account

for training/initialisation (seed) variance, which would require retraining and is out of scope here. To reduce false discoveries across the ten pairwise backbone comparisons in each reported experiment, raw bootstrap and McNemar p-values are accompanied by Holm-Bonferroni adjusted q-values, corrected separately for each experiment and test family. The tests are computed on a regenerated, seeded five-view TTA per-video score set whose individual AUCs differ from the original Table S1 evaluations by at most 0.003210 AUC. Stochastic brightness TTA is a possible, but unconfirmed, explanation. Tables S3–S7 use this regenerated-score evaluation, not the original headline evaluation; pairwise Δ AUC values

therefore need not equal differences between the headline entries.

The original generating script was not recovered. A reconstructed numerical procedure reproduces the stored bootstrap intervals and tail values: align and sort by video identifier, reset NumPy’s `default_rng(42)` for each model pair, and draw 2,000 paired video samples with replacement. The CI uses the 2.5th and 97.5th percentiles of Δ_b , the resampled AUC difference. The reported p_{boot} is twice the smaller fraction of strictly negative and strictly positive differences, capped at one; exact zeros count in neither tail. Stored zeros, displayed as < 0.001 , mean no draws entered the relevant strict tail, not zero underlying probability. Numerical reproduction does not establish calibration of this tail calculation as a hypothesis test; bootstrap-based significance labels are conditional on this convention.

Under this convention, only the larger gaps survive correction. In the in-distribution setting, ResNet-50 and ConvNeXt-Tiny are each stronger than Xception on the AUC bootstrap and McNemar tests after Holm adjustment; EfficientNet-B4 remains separated from Xception on the AUC bootstrap, while its McNemar result is only nominal. ViT-Base is also stronger than Xception on AUC ($\Delta\text{AUC} = 0.131$, $q_{\text{boot}} = 0.007$), but not on the fixed-threshold McNemar test ($q_{\text{McN}} = 0.304$). ConvNeXt-Tiny’s gap over EfficientNet-B4 is nominal before correction ($p_{\text{boot}} = 0.020$, $p_{\text{McN}} = 0.007$) but borderline after correction ($q_{\text{boot}} = 0.100$, $q_{\text{McN}} = 0.0504$), so it should not be read as a decisive adjusted result. The comparison that matters most, ConvNeXt-Tiny versus ResNet-50, is not significant after adjustment in Exp1 ($q_{\text{boot}} = 0.147$, $q_{\text{McN}} = 0.688$) and is not statistically significant in Exp2 ($\Delta\text{AUC} = 0.007$, $q_{\text{boot}} = 0.875$). I report ConvNeXt-Tiny as the strongest observed point estimate, not as a uniquely superior architecture. Pairwise tests for Exp3 and Exp4 are not reported. Full $\Delta\text{AUC}/\text{CI}/p/q$ values are listed in the significance tables.

1.3 Cross-dataset generalization

Every backbone degrades from the in-distribution setting (Exp1) to the cross-dataset settings, so

the in-domain numbers overstate deployment performance. The $\text{FF}++ \rightarrow \text{Celeb-DF}$ (Exp2) and $\text{Mixed} \rightarrow \text{DFD}$ (Exp4) transfers retain useful signal (AUC 0.57–0.92), but **the reverse Celeb-DF $\rightarrow \text{FF}++$ transfer (Exp3) collapses to near chance (~ 0.51 – 0.59) for every backbone**. Within this protocol, the asymmetry is consistent with source-domain scale and diversity contributing to the cross-dataset result: models trained on the larger, more varied $\text{FF}++$ transfer better than models trained on Celeb-DF and evaluated on $\text{FF}++$ ’s manipulation families. However, the present experiments do not isolate dataset scale and diversity from other differences between the datasets, so this interpretation should not be treated as a causal finding. The vector summary in the main paper visualizes the same AUC pattern.

1.4 Per-manipulation analysis (Exp1)

We report the manipulation-family breakdown for the in-distribution $\text{FF}++$ test set in Exp1. For each backbone and each family we report the video-level AUC of that family’s fakes against the shared pool of real videos (Table S5); a lower value marks a manipulation that the trained model finds harder to detect. NeuralTextures has the lowest observed family AUC for four of the five backbones (0.66–0.87), while Xception’s lowest value occurs for Face2Face (0.60). Under the fixed seed and protocol, the family-difficulty ordering is broadly similar across backbones. ConvNeXt-Tiny has the highest observed point estimate for every family (0.87–0.99), but these family-level rankings are descriptive and were not tested for stability across training seeds.

1.5 Error analysis (Confusion Matrices)

Table S6 shows the Exp1 confusion counts at a threshold of 0.5 (80 real and 69 fake videos). ConvNeXt-Tiny, EfficientNet-B4, and ResNet-50 have the three lowest false-positive counts, with $\text{FP} = 3, 4$, and 5 , respectively. EfficientNet-B4 misses 26 fake videos. ViT-Base has more false positives ($\text{FP} = 21$), while Xception has the largest number, with 32 of the 80 real videos classified as fake. Xception has 17 false negatives.

Table S3 Pairwise backbone comparison from regenerated scores on Exp1: FF++→FF++ (149 videos). $\Delta\text{AUC} = \text{AUC}(\text{A}) - \text{AUC}(\text{B})$ with paired-bootstrap 95% CI; p_{boot} and p_{McN} are raw empirical bootstrap-tail values and exact McNemar p-values, while q_{boot} and q_{McN} are Holm-Bonferroni adjusted within each experiment/test family. Bold marks Holm-adjusted $q < 0.05$.

A	B	ΔAUC (95% CI)	p_{boot}	q_{boot}	p_{McN}	q_{McN}
EfficientNet-B4	Xception	+0.163 [+0.088, +0.249]	< 0.001	< 0.001	0.014	0.075
EfficientNet-B4	ResNet-50	-0.022 [-0.068, +0.018]	0.262	0.444	0.078	0.304
EfficientNet-B4	ConvNeXt-Tiny	-0.059 [-0.110, -0.010]	0.020	0.100	0.007	0.0504
EfficientNet-B4	ViT-Base	+0.032 [-0.018, +0.087]	0.222	0.444	0.511	0.688
Xception	ResNet-50	-0.186 [-0.273, -0.111]	< 0.001	< 0.001	< 0.001	< 0.001
Xception	ConvNeXt-Tiny	-0.222 [-0.311, -0.151]	< 0.001	< 0.001	< 0.001	< 0.001
Xception	ViT-Base	-0.131 [-0.223, -0.046]	0.001	0.007	0.076	0.304
ResNet-50	ConvNeXt-Tiny	-0.037 [-0.076, -0.000]	0.049	0.147	0.344	0.688
ResNet-50	ViT-Base	+0.055 [+0.005, +0.108]	0.031	0.124	0.013	0.075
ConvNeXt-Tiny	ViT-Base	+0.091 [+0.035, +0.154]	0.001	0.007	0.001	0.011

Table S4 Pairwise backbone comparison from regenerated scores on Exp2: FF++→Celeb-DF (129 videos). $\Delta\text{AUC} = \text{AUC}(\text{A}) - \text{AUC}(\text{B})$ with paired-bootstrap 95% CI; p_{boot} and p_{McN} are raw empirical bootstrap-tail values and exact McNemar p-values, while q_{boot} and q_{McN} are Holm-Bonferroni adjusted within each experiment/test family. Bold marks Holm-adjusted $q < 0.05$.

A	B	ΔAUC (95% CI)	p_{boot}	q_{boot}	p_{McN}	q_{McN}
EfficientNet-B4	Xception	+0.144 [+0.023, +0.266]	0.019	0.114	0.018	0.140
EfficientNet-B4	ResNet-50	-0.084 [-0.160, -0.001]	0.044	0.176	0.099	0.358
EfficientNet-B4	ConvNeXt-Tiny	-0.092 [-0.175, -0.008]	0.026	0.130	0.035	0.210
EfficientNet-B4	ViT-Base	+0.047 [-0.048, +0.145]	0.329	0.658	1.000	1.000
Xception	ResNet-50	-0.228 [-0.353, -0.099]	< 0.001	< 0.001	< 0.001	0.004
Xception	ConvNeXt-Tiny	-0.236 [-0.369, -0.102]	0.001	0.009	< 0.001	0.002
Xception	ViT-Base	-0.097 [-0.232, +0.036]	0.144	0.432	0.046	0.232
ResNet-50	ConvNeXt-Tiny	-0.007 [-0.076, +0.063]	0.875	0.875	0.690	1.000
ResNet-50	ViT-Base	+0.131 [+0.048, +0.216]	0.002	0.016	0.090	0.358
ConvNeXt-Tiny	ViT-Base	+0.138 [+0.052, +0.224]	0.003	0.021	0.020	0.140

Table S5 Per-manipulation detection from regenerated scores on Exp1 (FF++ in-distribution): video-level AUC of each manipulation family’s fakes against the shared real videos. Lower values mark manipulation types a backbone finds harder.

Backbone	Deepfakes	Face2Face	FaceShifter	FaceSwap	NeuralTextures
EfficientNet-B4	0.940	0.887	0.912	0.915	0.810
Xception	0.860	0.599	0.652	0.850	0.697
ResNet-50	0.953	0.940	0.851	0.974	0.819
ConvNeXt-Tiny	0.973	0.983	0.924	0.985	0.874
ViT-Base	0.934	0.862	0.861	0.930	0.656

Looking at these counts gives a somewhat different picture from AUC alone. The relevant error can depend on how the detector is used. In a moderation queue, missing a fake may be more concerning, whereas in forensic triage, a high number of false positives could mean that more real videos have to be checked manually.

1.6 Deeper Failure-Case Analysis

The saved video predictions allow the errors to be counted directly for Exp1 and Exp2 (Table S7). In Exp1, ConvNeXt-Tiny makes 17 errors out of 149 videos, compared with 21 for ResNet-50. Xception has a much higher error count, with 49 misclassified videos, most of them false positives.

Table S6 Video-level confusion counts from regenerated scores for Exp1 on FF++ (80 real and 69 fake videos; threshold 0.5). TP and TN denote correctly classified fake and real videos, respectively. FP denotes a real video classified as fake, and FN denotes a fake video classified as real.

Backbone	TP	TN	FP	FN
EfficientNet-B4	43	76	4	26
Xception	52	48	32	17
ResNet-50	53	75	5	16
ConvNeXt-Tiny	55	77	3	14
ViT-Base	55	59	21	14

Error counts increase for every backbone in Exp2. ConvNeXt-Tiny makes 30 errors out of 129 videos, while ResNet-50 makes 33. The two remain close in terms of error count, consistent with the paired test results. EfficientNet-B4 and ViT-Base both have error rates close to one-third. Xception again produces a relatively large number of false positives on real videos.

The manipulation breakdown shows where many Exp1 errors occur. NeuralTextures is particularly difficult for the three strongest backbones: ConvNeXt-Tiny misses 5/9 videos, ResNet-50 misses 6/9, and EfficientNet-B4 misses 7/9. Real videos show a different pattern. ConvNeXt-Tiny and ResNet-50 produce only 3/80 and 5/80 false positives, respectively, while ViT-Base and Xception produce 21/80 and 32/80.

1.7 Ablations (EfficientNet-B4)

Tables S8 and S9 ablate the training and inference choices on the EfficientNet-B4 baseline (video-level AUC). Label smoothing has negligible effect; TTA improves AUC in Exp1/Exp2 but slightly reduces it in Exp3/Exp4. Training augmentation helps, with the largest observed gain in Exp4 (0.79 vs. 0.62 without it). For video aggregation, mean and median are best and clearly beat max/top- k , justifying mean as the default.

The most consequential finding is on *progressive unfreezing*. Removing it (i.e. fine-tuning all stages from the start) **improves EfficientNet-B4 in every experiment** — Exp1 0.900 $\rightarrow$ 0.927, Exp2 0.713 $\rightarrow$ 0.812, Exp3 0.506 $\rightarrow$ 0.546, Exp4 0.788 $\rightarrow$ 0.813 — an observed configuration-level improvement that does not isolate freezing from the accompanying optimizer and learning-rate changes. I identify the **no-progressive-unfreeze recipe as the recommended EfficientNet-B4 configuration**, and report it explicitly in

the ablation. I deliberately retain the common recipe in the headline backbone comparison (Table S1) because only EfficientNet-B4 was ablated; re-running the better recipe for the other four backbones would require additional training, which is outside the scope of this no-retraining study and is flagged in future work. The cross-configuration ranking changes: EfficientNet-B4 overtakes ResNet-50 in Exp1/Exp4 and both ResNet-50 and ConvNeXt-Tiny in Exp2, although it remains below ConvNeXt-Tiny in Exp1. These single-seed comparisons do not establish architectural superiority.

1.8 Frequency-only and temporal-score probes

Two additional low-cost baselines test whether simple alternatives to RGB mean aggregation reduce the transfer problem. The first is a frequency-only ResNet-18 trained on DCT/log-magnitude face crops for Exp1 and Exp2. The second is a score-level temporal baseline that trains a lightweight logistic-regression classifier over per-video score statistics, using existing crop-score CSVs without retraining the image models. Exp2 fits on Exp1 scores and tests on scores from separately trained Exp2 detectors of the same architecture; it is exploratory cross-checkpoint transfer, not a fixed-detector aggregation comparison.

The frequency result (Table S10) is mixed rather than a clear improvement. In Exp1, its AUC is well below that of the strongest RGB backbones in this study (0.653 versus 0.957 for ConvNeXt-Tiny and 0.924 for ResNet-50). In Exp2, it reaches an AUC of 0.752, higher than EfficientNet-B4 and ViT-Base but lower than ConvNeXt-Tiny and ResNet-50. The low accuracy at the fixed 0.5 threshold indicates that the AUC in Exp2 does not translate into similarly

Table S7 Failure-case summary from regenerated scores for Exp1 and Exp2 at the fixed 0.5 video-level threshold. FP and FN denote false positives and false negatives, respectively.

Experiment	Backbone	FP	FN	Total errors	Error rate
Exp1: FF++→FF++	EfficientNet-B4	4	26	30/149	0.201
	Xception	32	17	49/149	0.329
	ResNet-50	5	16	21/149	0.141
	ConvNeXt-Tiny	3	14	17/149	0.114
	ViT-Base	21	14	35/149	0.235
Exp2: FF++→Celeb-DF	EfficientNet-B4	28	15	43/129	0.333
	Xception	44	17	61/129	0.473
	ResNet-50	15	18	33/129	0.256
	ConvNeXt-Tiny	16	14	30/129	0.233
	ViT-Base	19	25	44/129	0.341

Table S8 Training ablations (EfficientNet-B4, video-level AUC). “−X” removes component X from the full recipe; experiment abbreviations follow Table S1.

Setting	Exp1	Exp2	Exp3	Exp4
Full (baseline)	0.900	0.713	0.506	0.788
− label smoothing	0.898	0.708	0.505	0.790
− progressive unfreeze	0.927	0.812	0.546	0.813
− train augmentation	0.888	0.704	0.518	0.619
− TTA	0.890	0.690	0.508	0.794

Table S9 Video-score aggregation ablation (EfficientNet-B4, video-level AUC); experiment abbreviations follow Table S1.

Setting	Exp1	Exp2	Exp3	Exp4
Mean (baseline)	0.900	0.713	0.506	0.788
Median	0.899	0.700	0.506	0.789
Max	0.861	0.642	0.492	0.759
Top- k	0.883	0.677	0.497	0.775

Table S10 Lightweight frequency-only ResNet-18 baseline. Inputs are discrete cosine transform (DCT)/log-magnitude face crops, not RGB images. Values are video-level metrics with AUC bootstrap CI.

Experiment	AUC (95% CI)	Acc	F1	EER
Exp1: FF++→FF++	0.653 (0.561–0.733)	0.617	0.544	0.409
Exp2: FF++→Celeb-DF	0.752 (0.657–0.835)	0.519	0.622	0.259

strong classification at that threshold. Spectral input appears to carry some transfer signal, but this simple ResNet-18 probe does not match the stronger RGB backbones.

The temporal-score probe (Table S11) is also mostly negative. It gives a small Exp2 AUC gain for EfficientNet-B4 (0.714 → 0.725), but it reduces AUC for the other four backbones in

Exp2 and for all five backbones in the Exp1 split. The internal Exp1 split shows no AUC improvement, but Exp2 confounds dataset shift with a detector-checkpoint change. These results do not rule out mean aggregation as a limitation. A fixed-detector test would require source-calibration and target-test scores from the same checkpoint.

Table S11 Score-level temporal aggregation probe. “Mean” is the matched video mean score; “Temporal” is logistic regression over per-video score statistics. Exp1 fits on 104 videos and evaluates on 45 (an internal 70/30 split); Exp2 fits on Exp1 FF++ scores and tests on separately trained Exp2 detector scores from Celeb-DF (exploratory cross-checkpoint transfer).

Experiment	Backbone	Mean AUC	Temporal AUC	Δ AUC
Exp1: FF++ $\rightarrow$ FF++	EfficientNet-B4	0.823	0.704	−0.119
	Xception	0.754	0.625	−0.129
	ResNet-50	0.873	0.847	−0.026
	ConvNeXt-Tiny	0.968	0.915	−0.054
	ViT-Base	0.841	0.794	−0.048
Exp2: FF++ $\rightarrow$ Celeb-DF	EfficientNet-B4	0.714	0.725	+0.011
	Xception	0.570	0.557	−0.013
	ResNet-50	0.798	0.786	−0.012
	ConvNeXt-Tiny	0.806	0.747	−0.059
	ViT-Base	0.667	0.642	−0.026

Table S12 Robustness of EfficientNet-B4: clean video-level AUC and minimum observed AUC across tested severities of each corruption (drop in parentheses).

Experiment	Clean	JPEG	Blur	Noise	Downscale
Exp1: FF++ $\rightarrow$ FF++	0.890	0.734 (−0.156)	0.740 (−0.150)	0.476 (−0.414)	0.780 (−0.109)
Exp2: FF++ $\rightarrow$ Celeb-DF	0.690	0.640 (−0.050)	0.560 (−0.130)	0.465 (−0.225)	0.588 (−0.101)
Exp3: Celeb-DF $\rightarrow$ FF++	0.508	0.502 (−0.006)	0.495 (−0.013)	0.459 (−0.049)	0.491 (−0.017)
Exp4: Mixed $\rightarrow$ DFD	0.793	0.692 (−0.101)	0.589 (−0.205)	0.489 (−0.304)	0.648 (−0.145)

1.9 Robustness to image degradations

Table S12 reports the minimum observed EfficientNet-B4 video-level AUC across the tested severities of each corruption. To keep the comparison clean, both the clean baseline and the corrupted evaluations use *single-view* (no-TTA) inference; each clean baseline comes from its corresponding robustness evaluation, not the separate no-TTA ablation or the TTA headline evaluation. **Additive noise is by far the most damaging corruption in every experiment** (maximum observed AUC drops of 0.41/0.23/0.30 on Exp1/Exp2/Exp4), collapsing the detector toward chance. The *milder* corruptions, however, do *not* order consistently across experiments: JPEG compression is the mildest on the two cross-dataset transfers (Exp2 drop 0.05, Exp4 drop 0.10), but on the in-distribution Exp1 *downscaling* is the mildest (drop 0.11) and JPEG is in fact the second-most damaging after noise (drop 0.16). Even the DFD detector (clean AUC 0.79) falls to ~ 0.49 under heavy noise, showing a measured robustness failure instead of a broad

resilience claim. (Exp3 is omitted from interpretation because its clean baseline is already near chance.)

2 Leakage-Control Procedure

Algorithm 1: Leakage-Controlled Experiment Construction

Input: datasets (FF++, Celeb-DF, DFD), official FF++ source-video splits
Output: train/validation/test manifest per experiment

For Exp1, assign each FF++ real video to its source-video split;
foreach Exp1 FF++ fake video generated from source-video identifiers T and S **do**
 if T and S belong to the same split **then**
 retain the video in that split;
 else
 discard the cross-partition video;
 end if

For Exp4, exclude DeepFakeDetection/DFD videos from training and validation;
Set training and validation data from the source dataset or datasets and the test data from the target dataset;
Downsample fake videos to the real-video count within each split;
Keep all crops from a source video in the same split;
Assert zero train–test source-video overlap for Exp1;
