# FAITHFUL IS NOT FORENSIC: RIFT FOR AUDITING DEEPFAKE DETECTORS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A deepfake detector can be accurate and faithfully explained while relying on evidence that does not account for the manipulation that changed an image's authenticity. We introduce RIFT, Robust Intervention-based Forensic Testing, a post hoc audit for frozen detectors that uses only scalar scores. Given a matched pristine and manipulated pair, RIFT applies the same intervention to a candidate region in both images. Manipulation Reliance ($M$) measures how much that region accounts for the matched authenticity contrast; Nuisance Instability ($Q$) is a complementary fixed-family stability diagnostic; and FSS is a conservative joint summary. Across six admissible detectors, the matched effect exceeds score-matched same-authenticity placebos by $0.198$ to $0.666$, with every paired interval above zero. The audit also withholds interpretation where raw metrics would mislead: a detector that passes the in-domain competence gate and shows a positive raw GT-over-control margin fails the placebo test ($-0.035$, $95\%$ CI $[-0.050, -0.019]$), so no forensic claim is made for it. A shared-cue sanity check and geometry-matched spatial controls confirm that the matched construction behaves as defined, while operator changes, reserved nuisances, a failed prespecified stress test, and controlled FF++ residual transfer bound the admissible interpretation. RIFT requires no retraining, gradients, internal features, or architecture-specific hooks.

## 1 INTRODUCTION

Deepfake detection has progressed from handcrafted artifact analysis to learned spatial, frequency, reconstruction, identity, and foundation model representations (Qian et al., 2020; Cao et al., 2022; Shiohara & Yamasaki, 2022; Dong et al., 2022; Yan et al., 2023a; Sun et al., 2024; Cui et al., 2025). Strong benchmark accuracy, however, does not reveal which evidence supports a detector's decision and does not guarantee robustness under manipulation or domain shift (Rössler et al., 2019; Li et al., 2020; Zi et al., 2020; Yan et al., 2024; Yermakov et al., 2026b).

Post hoc explanation methods identify features or regions that influence model output (Selvaraju et al., 2020; Sundararajan et al., 2017; Petsiuk et al., 2018; Ribeiro et al., 2016; Fong & Vedaldi, 2017), while faithfulness tests ask whether those explanations track the behavior of the model (Adebayo et al., 2018; Hooker et al., 2019; Yeh et al., 2019; Fong et al., 2019; Chen et al., 2024a). These are necessary questions, but they do not by themselves determine whether influential evidence is tied to the authenticity change represented by a matched pristine and manipulated pair. A feature may be genuinely predictive because it reflects compression, acquisition, background, identity distribution, or another shortcut (Geirhos et al., 2020; Bai et al., 2025; Yermakov et al., 2026a; A et al., 2026; Wang et al., 2026). A faithful explanation can therefore faithfully expose a dependency that is weakly related to the manipulation of forensic interest.

We call the missing audit target **forensic specificity**: whether a candidate region accounts for the matched authenticity contrast. We separately characterize whether that measured dependence is stable when authenticity is preserved. This is an operational audit notion rather than a causal guarantee about the true generative mechanism of a deepfake.

We introduce RIFT, Robust Intervention-based Forensic Testing. Given a matched pristine image $x^R$, manipulated image $x^F$, and candidate region $S$, RIFT applies the same regional intervention to both images. Manipulation Reliance ($M$) measures how much neutralizing $S$ reduces the detector's



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.

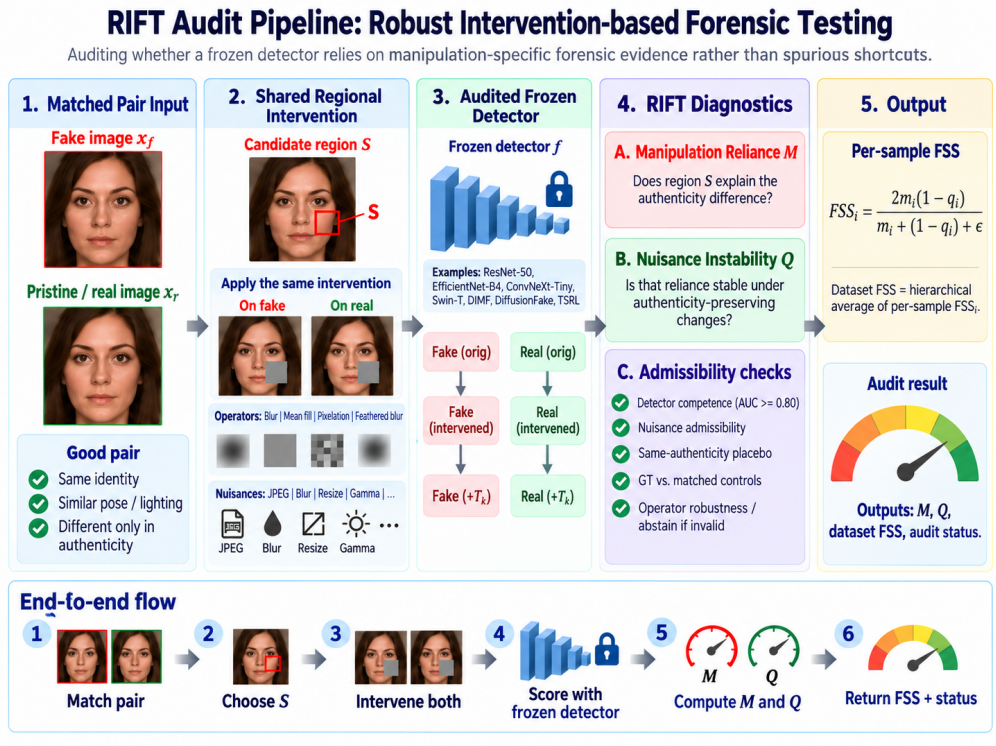



Figure 1: **RIFT audit pipeline.** The same regional intervention is applied to a matched pristine and manipulated pair and scored by a frozen detector. $M$ measures how much the region accounts for the matched authenticity contrast. $Q$ records the change in that reliance under a fixed nuisance family. Competence, placebo, spatial control, and operator checks determine which interpretations are admissible.

matched authenticity contrast. Nuisance Instability ($Q$) then records how that reliance changes under a fixed family of JPEG, blur, resize, and gamma transformations. FSS combines reliance and stability at the sample level when a single summary is useful. We treat $M$ as the primary manipulation specific diagnostic, $Q$ as a complementary stability diagnostic, and FSS as a summary rather than an optimized replacement for $M$.

The audit separates evidence discovery from evidence evaluation. The candidate region can be a known manipulation mask, an attribution region, or the output of a separate search procedure. RIFT evaluates the supplied region while the detector remains frozen. It uses only scalar inference scores and therefore does not require gradients, features, attention maps, or detector-specific instrumentation.

Our evaluation is designed around the assumptions required for interpretation. A shared predictive cue provides a controlled sanity check of the matched-pair construction. We then audit CNN, Transformer, and specialized forensic detectors using exact manipulation support and geometry-matched spatial controls. Same-authenticity placebos test whether $M$ exceeds generic intervention sensitivity. Reserved nuisances and severity sweeps examine the stability diagnostic while detector competence is monitored. Multiple intervention operators test boundary artifacts, failed audits are retained rather than removed, and a controlled residual transfer study examines the method under target-domain context changes.

The validity checks change conclusions rather than decorate them. TSRL (Lei et al., 2026) passes the in-domain competence gate (AUC 0.830) and has a positive raw GT-over-control FSS margin (0.094, 95% CI $[0.084, 0.104]$), so accuracy and raw spatial separation alone would suggest manipulation-grounded evidence. Its matched effect, however, falls below the same-authenticity placebo, and none of its nuisance conditions pass the competence gate. RIFT therefore withholds a forensic interpretation for TSRL, while the same checks support a manipulation-specific interpretation of $M$ for the other six audited detectors.

**Contributions.** Our contributions are threefold.

1. **Forensic specificity as a matched-pair audit target.** We formalize a distinction between evidence that influences a detector and evidence that accounts for a matched authenticity change.
2. **RIFT: a score-only audit with an explicit stability diagnostic.** RIFT uses Manipulation Reliance ($M$) as the primary matched-pair diagnostic, Nuisance Instability ($Q$) as a complementary stability diagnostic, and FSS as a conservative sample-level summary.
3. **Validity-driven evaluation across frozen detectors.** Controlled pair checks, geometry-matched controls, placebos, nuisance and operator analyses, failed audits, and competence-conditioned transfer characterize when a manipulation-specific interpretation is supported and when it must be withheld, including a detector whose accuracy and raw spatial margin pass but whose placebo check fails.

## 2 Related Work

### 2.1 Generalizable Deepfake Detection

Deepfake detectors use spatial artifacts, frequency cues, reconstruction inconsistencies, blending traces, and identity information (Qian et al., 2020; Cao et al., 2022; Shiohara & Yamasaki, 2022; Dong et al., 2022). UCF targets common forgery features across manipulation domains (Yan et al., 2023a), while more recent methods use diffusion supervision and pretrained vision representations (Sun et al., 2024; Cui et al., 2025). DeepfakeBench and DF40 show that strong in-domain performance does not ensure cross-domain robustness (Yan et al., 2023b; 2024). These methods improve predictive models. RIFT instead audits a frozen detector after training.

### 2.2 Explainability, Faithfulness, and Comparative Explanations

Post hoc methods identify influential features or regions (Ribeiro et al., 2016; Sundararajan et al., 2017; Selvaraju et al., 2020; Petsiuk et al., 2018; Fong & Vedaldi, 2017), and faithfulness tests evaluate whether those explanations reflect model behavior (Adebayo et al., 2018; Hooker et al., 2019; Yeh et al., 2019; Fong et al., 2019). Necessity and sufficiency provide intervention-based views of whether selected evidence is required for, or sufficient to support, a prediction (Chen et al., 2024a). Deepfake-specific methods increasingly add spatial, textual, and multimodal explanations (Chen et al., 2024b; Guo et al., 2025; Yu et al., 2025; Sun et al., 2025; Xu et al., 2026; Jiang et al., 2026).

Contrastive and counterfactual explanation methods instead explain a prediction relative to an alternative decision, missing feature, or reference input (Dhurandhar et al., 2018; Jacovi et al., 2021). Integrated Gradients is also defined relative to a chosen baseline (Sundararajan et al., 2017). RIFT is related to this comparative view, but the audit object is different: the same intervention is applied to a matched pristine and manipulated pair, and the change in their detector score contrast is measured under independent competence and validity checks.

### 2.3 Shortcut Learning and Forensic Evidence

Shortcut learning occurs when a model exploits predictive features that are not the intended concept (Geirhos et al., 2020). In deepfake detection, generalization failures have been linked to spurious correlations, compositing artifacts, and shortcut sensitive internal mechanisms (Bai et al., 2025; Yermakov et al., 2026a; A et al., 2026; Wang et al., 2026). Forensic Rewiring modifies internal attention behavior to suppress shortcut circuits (A et al., 2026). RIFT is complementary: it neither retrains nor edits the detector. It tests a proposed evidence region externally through matched input interventions and scalar scores.

## 3 Method: Robust Intervention-based Forensic Testing

RIFT first asks whether a candidate region accounts for the detector score difference created by a matched authenticity change. Manipulation Reliance ($M$) operationalizes this primary question. Nuisance Instability ($Q$) then records how the measured reliance changes under a fixed family of

transformations that preserve authenticity. FSS provides a conservative scalar summary when one value is useful.

### 3.1 Problem Formulation

Let $f$ be a pretrained detector that remains frozen throughout the audit and let $g_f(x)$ denote its scalar fake score. Probabilities are converted to log odds before auditing. Let $(x_i^R, x_i^F)$ be a tightly matched pristine and manipulated pair and $S$ a candidate spatial region. Because detector score scales differ, intervention effects are normalized by a robust detector specific scale estimated once on a calibration set:

$$s_f = P_{95}(g_f(\mathcal{D}_{\text{cal}})) - P_5(g_f(\mathcal{D}_{\text{cal}}))\,. \tag{1}$$

The scale is fixed for the detector and is not adapted to a region, nuisance, or target domain. Because this normalization is detector specific, absolute $M$ and FSS values are not used to rank detector quality across architectures. The primary comparisons are within detector.

### 3.2 Manipulation Reliance

Let $O_S(x)$ neutralize only region $S$. The same support and operator are applied to both images. Define the positive clean authenticity gap and the gap after intervention:

$$D_i = [g_f(x_i^F) - g_f(x_i^R)]_+, \tag{2}$$

$$D_i^{-S} = [g_f(O_S(x_i^F)) - g_f(O_S(x_i^R))]_+. \tag{3}$$

The per pair Manipulation Reliance is

$$m_i(S) = \text{clip}_{[0,1]}\left(\frac{[D_i - D_i^{-S}]_+}{s_f + \eta}\right). \tag{4}$$

A high $m_i$ means that neutralizing $S$ removes a substantial portion of the matched authenticity contrast. A low value means that the region contributes little to that contrast even if it influences the detector on one image. Pairs with a nonpositive clean detector ordering are retained: the positive part sets $D_i = 0$, so they contribute $m_i = 0$ rather than being discarded.

### 3.3 Nuisance Instability

Let $\mathcal{N} = \{T_1, \ldots, T_K\}$ be a fixed family of transformations that preserve the authenticity label. For each $T_n$, both members of the matched pair are transformed and Eq. 4 is recomputed with the same $S$ and $O_S$, giving $m_{i,n}(S)$. We define

$$q_{i,n}(S) = \frac{|m_{i,n}(S) - m_i(S)|}{m_{i,n}(S) + m_i(S) + \epsilon_q}, \tag{5}$$

clipped to $[0, 1]$, and

$$q_i(S) = \frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}} q_{i,n}(S). \tag{6}$$

The primary family is fixed before evaluation so that difficult transformations cannot be removed after results are observed. Accordingly, $q_i$ is first a raw fixed family stress summary. A low value indicates little change in reliance across that family, while a high value indicates substantial change. A positive forensic stability interpretation is made only for nuisance conditions under which detector competence is preserved. Gate-failing conditions remain visible in the raw summary but are not used as positive construct-validity evidence for $Q$. Table 9 reports, for each detector, which of the four primary conditions pass the gate.

### 3.4 Forensic Specificity Score

When a single summary is useful, we combine manipulation reliance and stability at the sample level:

$$\text{FSS}_i(S) = \frac{2m_i(S)[1 - q_i(S)]}{m_i(S) + [1 - q_i(S)] + \epsilon}. \tag{7}$$

FSS is computed before any video or dataset level averaging. It is not reconstructed from aggregate $M$ and $Q$. Because $Q$ is a fixed family stress summary, FSS is also interpreted together with the validity status of the corresponding nuisance conditions. We do not claim that FSS should outperform $M$ on every GT versus control ranking task.

### 3.5 Audit Validity

Metric computation is separated from interpretation. First, the detector must satisfy in domain competence with AUC at least $0.80$. Second, a nuisance condition supports a stability interpretation only when clean AUC is at least $0.80$, nuisance AUC is at least $0.75$, and the AUC decrease is at most $0.10$. Conditions that fail this gate remain in the fixed family stress summary for transparency but are not treated as evidence that manipulation reliance itself is unstable. Third, matched intervention effects are compared with score matched real versus real and fake versus fake placebos. Fourth, the GT region is compared with nonoverlapping geometry matched spatial controls whenever the strict construction is feasible. Finally, the audit is repeated with alternative neutralization operators. A failed requirement does not erase the raw measurements; it withholds the corresponding interpretation. Exact computation, aggregation, and the end to end algorithm appear in Appendix A.

## 4 Experiments

### 4.1 Evaluation Protocol

**Datasets and forensic ground truth.** The primary audit uses FaceForensics++ (FF++) c23 (Rössler et al., 2019) with Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The forensic manifest is constructed from the FF++ test split and contains $4{,}581$ matched frame pairs from $280$ pristine and manipulated video pairs. Exact manipulation support defines the GT region and is used only for auditing. For the four standard backbones, detector training uses the FF++ training split, score calibration uses the validation split, and the forensic audit uses the test manifest. Controlled residual transfer additionally uses Celeb-DF-v2 (Li et al., 2020), DFD, WildDeepfake (Zi et al., 2020), and DiffSwap (Zhao et al., 2023). Natural OOD AUC is measured separately on the original target datasets.

**Audited detectors.** The common panel contains ResNet-50 (He et al., 2016), EfficientNet-B4 (Tan & Le, 2019), ConvNeXt-Tiny (Liu et al., 2022), and Swin-T (Liu et al., 2021), trained under the same FF++ c23 binary detection protocol and frozen before auditing. The shared protocol is not tuned separately to maximize each architecture's benchmark AUC. We additionally audit DIMF (Kumar & Wang, 2026), DiffusionFake in its EfficientNet-B4 configuration (Sun et al., 2024), and the frozen TSRL Student (Lei et al., 2026). RIFT does not retrain, fine tune, or instrument these models; it accesses only their scalar scores. Detector specific preprocessing and checkpoint constraints are reported in Appendix B.3.

**Regions, operators, and nuisances.** The GT support is compared with three nonoverlapping controls obtained from exact translations of the same mask whenever the strict construction is feasible: an appearance matched background region, a maximally displaced region, and a deterministic random valid region. The primary neutralization is local blur with kernel size $31$ and $\sigma = 8$ at $224 \times 224$. Mean fill, pixelation, and smooth boundary feathered blur test operator robustness. The fixed primary nuisance family is JPEG quality $60$, blur $\sigma = 1.5$, resize scale $0.75$, and gamma $0.8$. Brightness $0.9$, contrast $0.85$, and saturation $0.8$ are reserved for construct validity and never enter the primary $Q$ or FSS. Full constructions and severity grids appear in Appendix B.

**Model influence baseline.** We compare RIFT with a perturbation based Faith score formed from necessity and sufficiency on the manipulated image, following prior intervention based faithfulness work (Fong & Vedaldi, 2017; Fong et al., 2019; Chen et al., 2024a). Faith asks whether a region strongly and compactly supports the detector's fake prediction but does not consult the matched pristine image. Its definition appears in Appendix A.2.

### 4.2 Evaluation Design

**Controlled shared cue sanity check.** A ConvNeXt-Tiny detector is trained with a checkerboard cue correlated with the fake label while retaining clean supervision. During auditing, the identical cue is inserted into both members of each matched pair. The cue can therefore remain influential while being absent from the pairwise authenticity difference by construction. This experiment verifies that the matched intervention behaves as defined and quantifies what the pristine counterpart adds beyond single image Faith.

**Cross-detector audit and GT-over-control ranking.** We then remove the engineered cue and audit ordinary FF++ examples. The GT support is compared with geometry matched translated controls. Because the pristine and manipulated images are tightly paired and the manipulation is localized by the GT mask, this comparison is a strong implementation sanity check rather than sufficient evidence by itself. We therefore evaluate it jointly with same authenticity placebos, operator changes, and nuisance validity tests. At the video group level we also measure how often GT is ranked above the strongest control under Faith and FSS.

**Construct validity and robustness.** Manipulation Reliance is tested against score matched real versus real and fake versus fake placebos under the same intervention. Nuisance Instability is tested against the reserved nuisance family and through severity sweeps while detector competence is monitored. We repeat the audit under four operators, compare FSS with $M$, $1 - Q$, and $M(1 - Q)$, and evaluate the fixed competence gates over prespecified threshold grids. A prespecified nonforensic proxy stress test is retained even though its stronger hypothesis is not achieved.

**Controlled residual transfer.** For each target dataset, genuine images form the pristine side of a controlled pair and receive a fixed FF++ manipulation residual only within the corresponding support. The residual construction, nuisance settings, detector checkpoints, and score scales remain fixed across target domains. This is a controlled FF++ contrast in a new image domain, not a native manipulation from the target dataset. Results are interpreted only when controlled Pair AUC is at least $0.80$. Natural OOD AUC is measured independently and is not used to tune the construction.

**Statistics.** Faith, $M$, $Q$, and FSS are computed at the lowest sample level before aggregation. Reported values follow frame/sample $\rightarrow$ video group $\rightarrow$ manipulation family $\rightarrow$ equal weight macro. Confidence intervals use $10{,}000$ paired video group bootstrap replicates stratified by manipulation family. The same resampled groups are used for paired region comparisons, so frames from the same video are not treated as independent observations.

## 5 Results

Unless stated otherwise, intervals are paired $95\%$ video group bootstrap intervals with $10{,}000$ replicates.

### 5.1 Controlled Sanity Check: Influence Need Not Explain the Matched Contrast

We first verify the defining distinction in a setting where it is known by construction. The detector has learned a predictive checkerboard cue, but at audit time the identical cue is placed in both members of the matched pair. The cue can therefore remain influential while contributing nothing to the authenticity change represented by the pair.

From shortcut strength $0.50$ onward, Faith ranks the cue above the GT region. At strength $1.00$, Faith is $0.223$ for the cue and $0.074$ for GT. In contrast, $M$ is $0.017$ versus $0.588$, and FSS is $0.024$ versus $0.447$. The GT minus cue FSS margin is $0.413$ with $95\%$ CI $[0.393, 0.432]$. This is not a failure of faithfulness: Faith correctly reports that the cue influences the detector. The experiment isolates the additional information supplied by the matched pristine counterpart. Influence on one image does not determine whether a region accounts for the authenticity difference between the pair.

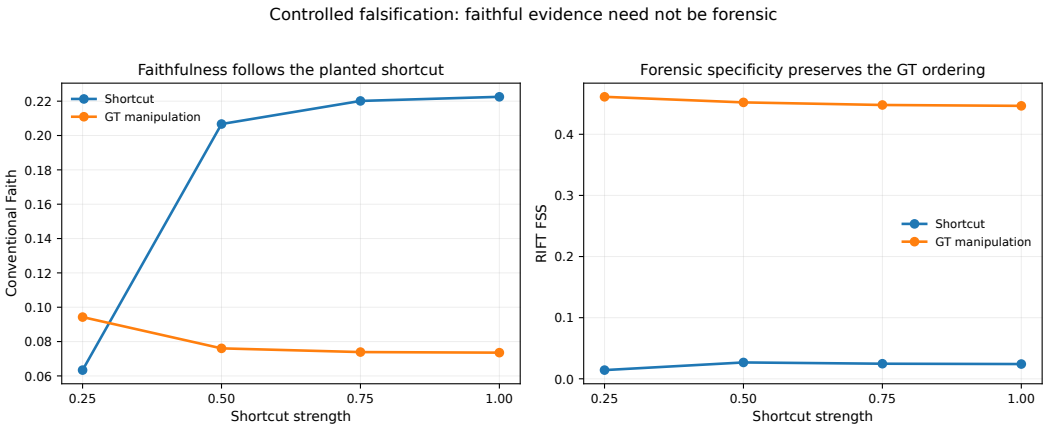



Figure 2: **Controlled matched pair sanity check.** As shortcut strength increases, the planted cue becomes increasingly influential under the single image Faith baseline. Because the same cue is present in both members of the pair, its contribution to the matched authenticity contrast remains small under $M$. Full values are reported in Appendix C.1.

Table 1: **Cross detector RIFT audit on FF++ Forensic GT.** For detectors supporting the strict spatial protocol, $\Delta\text{FSS} = \text{FSS}_{GT} - \max(\text{FSS}_{BG}, \text{FSS}_{disp}, \text{FSS}_{rand})$. $Q$ and FSS are raw fixed-family summaries. For the five full-control detectors, one or two primary nuisance conditions fail the competence gate (Table 9), so their $Q_{GT}$ is a stress summary rather than stability evidence. Absolute $M$ and FSS values are not used to rank detector quality across architectures.

| Frozen detector | ID AUC ↑ | $M_{GT}$ | $Q_{GT}$ | $\text{FSS}_{GT}$ | ΔFSS [95% CI] |
|---|---|---|---|---|---|
| ResNet-50 | 0.897 | 0.315 | 0.394 | 0.367 | 0.340 [0.319,0.360] |
| EfficientNet-B4 | 0.869 | 0.299 | 0.401 | 0.347 | 0.300 [0.279,0.321] |
| ConvNeXt-Tiny | 0.947 | 0.487 | 0.379 | 0.492 | 0.452 [0.429,0.475] |
| Swin-T | 0.940 | 0.431 | 0.410 | 0.449 | 0.394 [0.373,0.414] |
| DIMF | 0.963 | 0.517 | 0.331 | 0.536 | 0.471 [0.447,0.494] |
| DiffusionFake* | 0.999 | 0.720 | 0.084 | 0.796 | — |
| TSRL† | 0.830 | 0.083 | 0.458 | 0.109 | 0.094 [0.084,0.104] |

*DiffusionFake uses detector native preprocessing that does not preserve the identical translated mask construction; only GT quantities are reported. †TSRL passes the initial ID competence threshold but fails independent placebo and nuisance checks, so its raw values are diagnostic only.

### 5.2 Matched Manipulation Dependence Across Frozen Detectors

We next audit ordinary FF++ examples using the known manipulation support and strict spatial controls.

The main evidence in Table 1 is the within detector GT versus control comparison, not the absolute ordering of detectors by $M$ or FSS. Across this panel, ID AUC, $M_{GT}$, and $\text{FSS}_{GT}$ have the same detector ordering. We therefore do not interpret larger absolute audit values as evidence that one architecture is intrinsically more forensic than another. All five detectors that support the strict control construction have positive GT-minus-control FSS intervals, with margins from $0.300$ to $0.471$. These margins do not depend on the gate-failing conditions that enter $Q$: $M$ alone gives the same GT-versus-control separation (backbone-macro AUROC $0.9717$ for $M$ versus $0.9707$ for FSS; Table 11). Full control values are reported in Appendix C.2.

At the video group level, Faith ranks GT above the strongest spatial control on $64.7\%$ of groups on average, compared with $99.1\%$ for FSS. Detector specific net gains range from $+0.252$ to $+0.411$, and every paired interval excludes zero. This result mainly quantifies the benefit of evaluating the matched authenticity contrast rather than only single image influence; it is not used alone as proof of

forensic grounding. TSRL moves in the opposite direction, consistent with its independent validity failures. Full per-detector ranking results are reported in Appendix C.3.

### 5.3 Manipulation Reliance Validity and Nuisance Stability Evidence

For $M$, the matched pristine and manipulated intervention effect exceeds the stronger score matched same authenticity placebo for ResNet-50, EfficientNet-B4, ConvNeXt-Tiny, Swin-T, DIMF, and DiffusionFake by $0.211$, $0.198$, $0.357$, $0.310$, $0.383$, and $0.666$, respectively. Every paired interval excludes zero. TSRL reverses the expected relation: its matched effect is $0.184$ versus a strongest placebo of $0.219$, giving $-0.035[-0.050, -0.019]$. This independent failure prevents a manipulation specific interpretation for TSRL despite its positive raw spatial margin.

For $Q$, reserved brightness, contrast, and saturation transformations provide an independent stability check. After conditioning on $M$ and manipulation family, the partial Spearman interval excludes zero for EfficientNet-B4 ($0.216[0.084, 0.338]$), ConvNeXt-Tiny ($0.140[0.020, 0.257]$), DIMF ($0.230[0.101, 0.349]$), and DiffusionFake ($0.287[0.179, 0.387]$). ResNet-50 and Swin-T remain inconclusive. We therefore treat this as detector dependent evidence that $Q$ contains stability information beyond $M$, not as universal validation of an independent second discriminator.

Severity sweeps provide complementary evidence. DiffusionFake remains competence admissible at all tested JPEG levels: $Q_{GT}$ increases from $0.058$ to $0.282$ as quality decreases from $95$ to $40$, while AUC remains $0.987$ at the strongest level. Gate failing conditions for other detectors are retained in the plots but are not used as positive construct validity evidence. Complete placebo, reserved nuisance, and severity results appear in Appendices C.4 and C.5.

### 5.4 Robustness and Failed Audits

The GT versus control ordering is stable across intervention choices. Blur, mean fill, pixelation, and feathered blur yield positive paired intervals for all $20/20$ detector and operator combinations across the five valid full control detectors. The feathered operator removes the sharp binary boundary while preserving the same conclusion, reducing concern that the ordering is a seam artifact. Full values appear in Appendix C.6.

The scalarization is not responsible for the separation. Across the four standard backbones, GT versus control AUROC is $0.9717$ for $M$ and $0.9707$ for FSS, with pairwise ranking $0.9929$ and $0.9938$, respectively. The corresponding $1 - Q$ values are substantially weaker. We therefore treat $M$ as the primary manipulation specific discriminator, $Q$ as a stability diagnostic, and FSS as a conservative summary rather than a universally better metric. Full comparisons are in Appendix C.7.

Failed tests are retained. TSRL passes the initial ID competence gate but fails the placebo and nuisance checks. A prespecified nonforensic proxy also fails the required $M$ noninferiority condition, so the present experiments do not establish a case in which $M$ is fooled and $Q$ alone repairs the ordering. These negative results bound the claims rather than being redefined after observation. Details appear in Appendices C.9 and C.8.

### 5.5 Controlled FF++ Residual Transfer Across Target Domains

This experiment evaluates transfer of a fixed FF++ authenticity contrast rather than native manipulations from the target datasets. Genuine target domain images receive a deterministic FF++ residual inside the corresponding support, with the construction fixed across detectors and domains. WildDeepfake is the only target domain for which multiple detectors satisfy the controlled Pair AUC criterion of $0.80$, so Table 2 reports only those competence admissible cases. The complete grid is in Appendix C.10.

All four competence admissible detectors have lower FSS on the controlled WildDeepfake construction than on FF++. DIMF retains natural WildDeepfake AUC $0.936$ and controlled Pair AUC $0.835$, while its FSS decreases from $0.536$ to $0.256$. DiffusionFake reaches Pair AUC $0.913$ while its FSS decreases from $0.796$ to $0.360$. These results neither evaluate native WildDeepfake manipulations nor imply that FSS predicts natural OOD accuracy. They show only that, under a fixed transferred FF++ contrast, a detector can remain above the prespecified competence threshold while the mea-

Table 2: **Competence admissible controlled residual transfer on WildDeepfake.** Pair AUC measures competence on the constructed FF++ residual contrast. Natural OOD AUC is measured separately on the original dataset and is not used to construct or tune the audit.

| Detector | Pair AUC | ID FSS | OOD $M_{GT}$ | OOD $Q_{GT}$ | OOD $\text{FSS}_{GT}$ | FSS drop | Natural AUC |
|---|---|---|---|---|---|---|---|
| ConvNeXt-Tiny | 0.852 | 0.492 | 0.243 | 0.261 | 0.342 | 0.150 | — |
| Swin-T | 0.856 | 0.449 | 0.200 | 0.359 | 0.286 | 0.164 | — |
| DIMF | 0.835 | 0.536 | 0.181 | 0.411 | 0.256 | 0.280 | 0.936 |
| DiffusionFake | 0.913 | 0.796 | 0.251 | 0.209 | 0.360 | 0.436 | 0.798 |

sured matched dependence and its fixed family stability summary decrease. We therefore treat this experiment as a controlled transfer stress test.

## 6 Discussion

A faithful explanation should reveal any feature that genuinely influences a detector, including a shortcut. RIFT asks an additional matched pair question: does the candidate region account for the score contrast between a manipulated image and its pristine counterpart? The shared cue experiment is a sanity check of this distinction, while the ordinary sample analysis measures how often access to the matched counterpart changes GT versus control ranking.

The experiments support $M$ as the principal measure of matched manipulation dependence. $Q$ addresses the narrower question of stability under a fixed nuisance family. Its reserved nuisance associations are positive for several detectors but inconclusive for others, and the proxy experiment does not show a case in which $Q$ alone repairs an $M$ matched failure. We therefore treat $Q$ as a diagnostic and FSS as a conservative summary, not as a universally better replacement for $M$.

Absolute $M$ and FSS values can covary with detector competence and are not used to rank architectures. Interpretation is conditional on competence and the placebo, spatial, nuisance, and operator checks. TSRL illustrates why this matters: it passes the ID AUC threshold but fails independent construct checks. The target domain experiment is likewise a controlled stress test, not evidence that FSS predicts natural OOD accuracy.

## 7 Limitations and Future Work

RIFT requires matched or controlled pristine and manipulated pairs for its strongest interpretation. The GT versus translated control comparison is also a deliberately strong sanity check when the pair differs mainly inside the known manipulation support, so it is not treated as sufficient evidence by itself. Absolute $M$ values can covary with detector competence, and $Q$ only characterizes the fixed nuisance family under the stated competence gates. For five of the six admissible detectors, at least one primary nuisance condition fails the competence gate, so stability conclusions for those detectors rest on the admissible conditions and the reserved-nuisance analysis. Input interventions alter image statistics; placebos, geometry matched controls, and alternative operators reduce but do not eliminate this concern. The present audit is spatial and can miss manipulation aligned shortcuts that are stable under all tested nuisances. Future work should extend the intervention unit to temporal, frequency, and distributed evidence and develop defensible counterfactuals for unpaired suspect media.

## 8 Conclusion

We introduced RIFT, a post hoc audit of whether evidence used by a frozen deepfake detector accounts for a matched authenticity change. Manipulation Reliance provides the primary matched pair diagnostic, while Nuisance Instability records how that reliance changes under a fixed nuisance family. Controlled sanity checks, matched controls, placebos, operator tests, reserved nuisances, and failed audits define when the resulting interpretation is supported and when it should be withheld. These measurements complement predictive accuracy and faithful explanation rather than replacing either one.

## Reproducibility Statement

The main protocol is specified in Sections 3 and 4. The appendix records the complete algorithm, stabilizers, aggregation, split usage, detector constraints, control construction, nuisance settings, validation studies, and full result tables. The supplementary code stores fixed configurations, checkpoint identifiers, per sample outputs, and bootstrap settings required to regenerate the reported results. Anonymized code is available at https://github.com/aiai-9/RIFT

## Ethics Statement

This work concerns defensive auditing of deepfake detectors and does not generate impersonation content. Region level diagnostics can reveal detector sensitivities and therefore have potential dual use. We restrict the study to defensive evaluation and do not optimize or recommend evasion strategies. Qualitative examples are used only to illustrate the audit protocol, and no claims are made about the identity, intent, or authenticity of depicted individuals beyond the benchmark annotations.

## Statement on AI Use

Generative AI tools were used to assist with manuscript organization, language editing, readability, and critical feedback on the presentation of the methodology and experiments. All experimental design decisions, implementations, reported numerical results, mathematical statements, citations, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the content of the submission.

## References

Nishitha A, Rachit Verma, and Bhaskarjyoti Das. Forensic rewiring: Circuit-level shortcut suppression for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8560–8569, 2026.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Ningning Bai, Xiaofeng Wang, Ruidong Han, Jianpeng Hou, Qin Wang, and Shanmin Pang. Towards generalizable face forgery detection via mitigating spurious correlation. *Neural Networks*, 182: 106909, 2025. doi: 10.1016/j.neunet.2024.106909.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4122, 2022.

Xuexin Chen, Ruichu Cai, Zhengting Huang, Yuxuan Zhu, Julien Horwood, Zhifeng Hao, Zijian Li, and José Miguel Hernández-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6486–6502. PMLR, 2024a.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024b.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025. doi: 10.1109/CVPR52734.2025.01789.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives, 2018. URL https://arxiv.org/abs/1802.07623.

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. 9468–9478, 2022.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2950–2958, 2019. doi: 10.1109/ICCV.2019.00304.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3429–3437, 2017.

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. doi: 10.1038/s42256-020-00257-z.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–116, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability, 2021. URL https://arxiv.org/abs/2103.01378.

Ning Jiang, Dingheng Zeng, Yanhong Liu, Haiyang Yi, Shijie Yu, Minghe Weng, Haifeng Shen, and Ying Li. Explainable deepfake detection with rl enhanced self-blended images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Sahil Kumar and Honggang Wang. Dual-identity mamba fusion for source-aware deepfake detection. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026.

Zhanhe Lei, Zhongyuan Wang, Jikang Cheng, Baojin Huang, Yuhong Yang, Zhen Han, Chao Liang, and Dengpan Ye. Tutor-student reinforcement learning: A dynamic curriculum for robust deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41436–41445, 2026.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision (ECCV)*, pp. 86–103, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019. doi: 10.1109/ICCV.2019.00009.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.

Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3218.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19576–19586, 2025.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Yihui Wang, Yonghui Yang, Jilong Liu, Fengbin Zhu, Le Wu, and Tat-Seng Chua. Suppressing forgery-specific shortcuts for generalizable deepfake detection. *arXiv preprint arXiv:2606.01843*, 2026.

Wenbo Xu, Wei Lu, Xiangyang Luo, and Jiantao Zhou. Mare: Multimodal alignment and reinforcement for explainable deepfake detection via vision-language models. *arXiv preprint arXiv:2601.20433*, 2026.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22412–22423, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4534–4565, 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 *Advances in Neural Information Processing Systems*, volume 37, pp. 29387–29434, 2024.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10965–10976, 2019.

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

Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 773–783, 2026b. doi: 10.1109/WACV61042.2026.00082.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72925–72943. PMLR, 2025.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8568–8577, 2023.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020.

# A Additional RIFT Method Details

This section records the implementation details omitted from Section 3 for space. The definitions are identical to those used for all reported results.

## A.1 Numerical Stabilization and Score Normalization

The detector-specific normalization scale is

$$s_f = P_{95}(g_f(\mathcal{D}_{\text{cal}})) - P_5(g_f(\mathcal{D}_{\text{cal}}))\,, \tag{8}$$

estimated once from the validation calibration set. The implementation uses

$$\eta = 10^{-8}, \qquad \epsilon = 10^{-8}, \qquad \epsilon_q = 10^{-6}, \tag{9}$$

for score normalization, harmonic aggregation, and relative-instability stabilization, respectively. These values are fixed globally and are not tuned per detector, evidence region, nuisance, or domain.

## A.2 Perturbation-Based Faithfulness Baseline

The main experiments compare RIFT with a perturbation-based Faith baseline that measures model influence on the manipulated image without consulting its matched pristine counterpart.

Necessity measures the reduction in fake score after neutralizing $S$:

$$N_i(S) = \text{clip}_{[0,1]}\left(\frac{g_f(x_i^F) - g_f(O_S(x_i^F))}{s_f + \eta}\right). \tag{10}$$

Sufficiency measures how well the selected region preserves the original fake score when its complement $\bar{S}$ is neutralized:

$$U_i(S) = 1 - \text{clip}_{[0,1]}\left(\frac{\left|g_f(x_i^F) - g_f(O_{\bar{S}}(x_i^F))\right|}{s_f + \eta}\right). \tag{11}$$

The baseline is their harmonic mean:

$$\text{Faith}_i(S) = \frac{2N_i(S)U_i(S)}{N_i(S) + U_i(S) + \epsilon}. \tag{12}$$

Faith therefore asks whether $S$ strongly and compactly supports the detector's fake prediction. Because it is evaluated on the manipulated image alone, it does not test whether the same evidence is specific to the matched authenticity change. This distinction is isolated by the controlled shared cue check in Section 5.1.

## A.3 Per-Sample Computation and Hierarchical Aggregation

For every evidence region and every matched pair, RIFT first computes $m_i$, each $q_{i,n}$, their average $q_i$, and finally $\text{FSS}_i$. Nonlinear quantities are therefore formed before any video- or dataset-level averaging.

Let $\mathcal{I}_v$ denote the frames belonging to video group $v$, $\mathcal{V}_c$ the video groups for manipulation family $c$, and $\mathcal{C}$ the evaluated manipulation families. For a generic per sample quantity $z_i$, the reported macro value is obtained hierarchically as

$$z_v = \frac{1}{|\mathcal{I}_v|}\sum_{i\in\mathcal{I}_v} z_i, \tag{13}$$

$$z_c = \frac{1}{|\mathcal{V}_c|}\sum_{v\in\mathcal{V}_c} z_v, \tag{14}$$

$$z_{\text{macro}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} z_c. \tag{15}$$

This procedure is applied independently to $m_i$, $q_i$, $\mathrm{FSS}_i$, and the Faith baseline. In particular,

$$\mathrm{Avg}[\mathrm{FSS}_i] \neq \frac{2M(1-Q)}{M+(1-Q)+\epsilon} \tag{16}$$

in general, and the right-hand expression is not used to generate the reported dataset-level FSS.

### A.4 End-to-End Audit

**Algorithm 1:** RIFT audit of a candidate region $S$ under a frozen detector $f$

**Input:** frozen detector $f$; matched pairs $\{(x_i^R, x_i^F)\}_{i=1}^N$; candidate region $S$; neutralization operator $O_S$; fixed primary nuisance family $\mathcal{N}$; calibration set $\mathcal{D}_{\mathrm{cal}}$
**Output:** raw $(M, Q, \mathrm{FSS})$ for $S$ and an audit-validity status

$s_f \leftarrow P_{95}(g_f(\mathcal{D}_{\mathrm{cal}})) - P_5(g_f(\mathcal{D}_{\mathrm{cal}}))$
**foreach** *matched pair* $(x_i^R, x_i^F)$ **do**
  $D_i \leftarrow [g_f(x_i^F) - g_f(x_i^R)]_+$
  $D_i^{-S} \leftarrow [g_f(O_S(x_i^F)) - g_f(O_S(x_i^R))]_+$
  $m_i \leftarrow \mathrm{clip}_{[0,1]}\big([D_i - D_i^{-S}]_+/(s_f+\eta)\big)$
  **foreach** $T_n \in \mathcal{N}$ **do**
    recompute $m_{i,n}$ on $\big(T_n(x_i^R), T_n(x_i^F)\big)$ using the same $S$ and $O_S$
    $q_{i,n} \leftarrow |m_{i,n} - m_i|/(m_{i,n} + m_i + \epsilon_q)$
    $q_{i,n} \leftarrow \mathrm{clip}_{[0,1]}(q_{i,n})$
  **end**
  $q_i \leftarrow |\mathcal{N}|^{-1} \sum_{n\in\mathcal{N}} q_{i,n}$
  $\mathrm{FSS}_i \leftarrow 2m_i(1-q_i)/\big(m_i + (1-q_i) + \epsilon\big)$
**end**
$M \leftarrow$ hierarchical average of $\{m_i\}$
$Q \leftarrow$ hierarchical average of $\{q_i\}$
FSS $\leftarrow$ hierarchical average of $\{\mathrm{FSS}_i\}$
evaluate detector competence, nuisance competence, same authenticity placebo, spatial-control, and operator robustness conditions
**return** raw $(M, Q, \mathrm{FSS})$ with the corresponding audit-validity status

Algorithm 1 makes explicit that nuisance competence is not used to select a favorable subset of transformations before computing $Q$. The primary family is fixed in advance. Competence is evaluated separately and controls the strength of the interpretation attached to the resulting audit quantities.

### A.5 Admissibility Checks

The primary detector competence requirement is

$$\mathrm{AUC}_{\mathrm{ID}} \geq 0.80. \tag{17}$$

For a nuisance condition $T_n$, competence is preserved only when

$$\mathrm{AUC}_{\mathrm{clean}} \geq 0.80, \qquad \mathrm{AUC}_n \geq 0.75, \qquad \mathrm{AUC}_{\mathrm{clean}} - \mathrm{AUC}_n \leq 0.10. \tag{18}$$

The same authenticity placebo compares the matched fake versus pristine intervention effect against score-matched real–real and fake–fake pairs under the same operator. A valid manipulation-specific interpretation requires the matched effect to exceed the stronger placebo.

The strict spatial-control protocol compares the GT manipulation support with an appearance-matched exact translation, a maximally displaced exact translation, and a deterministic random valid translation whenever such translations exist. These controls preserve the GT mask's area, shape, perimeter, and boundary complexity; their construction is detailed in Appendix B.2.

Finally, operator robustness repeats the audit with local blur, mean fill, pixelation, and smooth-boundary feathered blur. The feathered operator specifically tests whether a GT versus control ordering can be explained by a sharp intervention boundary.

A failed gate does not erase the corresponding raw measurements. Such values may be retained for diagnostic analysis, as in the TSRL audit, but they are not treated as positive evidence of forensic specificity. Likewise, when a strict control is structurally unavailable under detector-native preprocessing, the unsupported comparison is reported as unavailable rather than inferred.

# B Experimental Protocol Details

This section provides the dataset, detector, intervention, control, and statistical details underlying the experiments in Section 4. All detector checkpoints are frozen before auditing, and no experiment modifies the detector using RIFT outputs.

## B.1 Datasets and Forensic Ground Truth

**FaceForensics++.** The primary forensic audit uses FF++ c23 across Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The forensic manifest is generated from the FF++ test split and contains $4{,}581$ matched frame pairs from $280$ pristine and manipulated video pairs. Each record contains the manipulation family, video-group identity, pristine frame, manipulated frame, and binary manipulation support. The manipulation mask is used only as an evaluation region or as the geometric template for spatial controls; it is never supplied to the frozen detector as an input feature.

For the four standard backbones, training uses the FF++ training split, score scale calibration uses the FF++ validation split, and the forensic audit uses the test manifest. The audit manifest is therefore not reused for detector training or score calibration. Pairs with a nonpositive clean detector ordering are retained by the metric definition: Eq. 2 sets $D_i = 0$ and the corresponding Manipulation Reliance is zero rather than removing the pair.

**OOD datasets.** Controlled forensic-specificity transfer is evaluated on Celeb-DF-v2, DFD, Wild-Deepfake, and DiffSwap. These target datasets are not used to retrain or adapt the audited detectors. Natural OOD AUC is computed independently on the original datasets and is kept separate from the controlled transferred-pair audit.

## B.2 Strict Spatial Control Construction

For each GT manipulation mask, the common protocol searches for valid non-overlapping translations of exactly the same mask. Three controls are constructed whenever the required translations exist.

The appearance-matched background control is the valid translation whose pristine-region RGB mean and standard deviation are closest to those of the GT region. The displaced control is the valid translation with the greatest spatial displacement from the GT location. The random control is a deterministic random valid translation.

Because these controls translate rather than reshape the GT mask, they preserve its area, shape, perimeter, and boundary complexity. Strict mode does not replace an unavailable translated control with arbitrary pixels. When a detector or preprocessing path prevents the common geometric construction, the unsupported GT-minus-control comparison is reported as unavailable.

## B.3 Audited Detector Protocols

**Standard backbones.** ResNet-50, EfficientNet-B4, ConvNeXt-Tiny, and Swin-T are trained under the same FF++ c23 binary detection protocol and audited from frozen best checkpoints. The shared training protocol is not tuned separately to maximize each architecture's AUC. The common RIFT intervention is applied before each architecture's detector specific normalization or internal resizing.

**DIMF.** DIMF is evaluated through its frozen source-free inference path. RIFT does not use training-only privileged information and does not alter model parameters. DIMF receives the identical

intervention, control, nuisance, placebo, and aggregation protocol as the standard backbones; no RIFT setting is tuned for it.

**Split usage for external checkpoints.** The DIMF, DiffusionFake, and TSRL checkpoints were trained without the FF++ test videos used to build the forensic audit manifest.

**DiffusionFake.** DiffusionFake (Sun et al., 2024) is evaluated in its EfficientNet-B4 configuration using its frozen checkpoint and native inference preprocessing. Although DiffusionFake uses diffusion-guided supervision during training, RIFT interacts only with its deployed scalar score. Its native preprocessing does not preserve the exact coordinate construction required by the common translated-mask controls; we therefore report its GT-region $M$, $Q$, and FSS without inferring a strict GT-minus-control margin. The GT audit retains $4{,}545$ of $4{,}581$ candidate samples; 36 samples ($0.79\%$) are excluded under the prespecified geometry-compatibility rule before aggregation.

**TSRL.** TSRL is audited using the final frozen Student checkpoint from its tutor–student procedure. It is retained throughout the evaluation even when independent validity requirements fail, allowing the audit to demonstrate the difference between a raw metric value and an admissible forensic interpretation.

**ViT-B/16 sensitivity run.** ViT-B/16 participates only in the detector competence sensitivity analysis. It is not pooled with the primary positive detector panel when the locked in-domain competence requirement is not satisfied.

### B.4 Controlled Shared Cue Protocol

The controlled shared cue check uses a ConvNeXt-Tiny detector with a checkerboard cue correlated with the fake label while clean supervision is retained. The controlled detector follows a $256 \times 256$ image protocol.

The shortcut is evaluated at strengths

$$0.25, \quad 0.50, \quad 0.75, \quad 1.00.$$

At audit time, the identical cue is inserted into both members of the matched pristine–manipulated pair. This construction is deliberate: the shortcut can continue to influence the detector, but because it is shared by the two audit images it cannot itself constitute the authenticity-changing difference.

The cue region and the GT manipulation region are independently evaluated using the perturbation based Faith baseline and the RIFT quantities. No selection is made using the resulting $M$, $Q$, or FSS values. The full dose-response results are reported in Appendix C.1.

### B.5 Intervention Operators

The common audit operates at $224 \times 224$ before detector specific internal resizing. The primary intervention is strong local Gaussian blur with kernel size 31 and $\sigma = 8$.

Operator robustness repeats the same audit using:

- local Gaussian blur;
- local mean fill;
- local pixelation; and
- smooth-boundary feathered blur.

The region support is unchanged across operators. The feathered variant is included specifically to reduce the sharp boundary introduced by a binary regional intervention. Operator-specific results are reported in Appendix C.6.

### B.6 Primary, Reserved, and Severity Nuisances

The primary nuisance family that defines $Q$ is fixed before evaluation:

$$\mathcal{N} = \{\text{JPEG}(60), \text{Blur}(1.5), \text{Resize}(0.75), \text{Gamma}(0.8)\}. \tag{19}$$

The numerical arguments denote JPEG quality, Gaussian blur $\sigma$, resize scale, and gamma, respectively. These four conditions participate in the raw fixed family $Q$ and therefore in per sample FSS. The family is not reduced after observing competence results. Competence is evaluated separately for each nuisance condition and determines whether a stability interpretation is admissible.

A disjoint reserved family is used only for construct-validity analysis:

$$\mathcal{N}_{\text{hold}} = \{\text{Brightness}(0.9), \text{Contrast}(0.85), \text{Saturation}(0.8)\}. \tag{20}$$

Reserved nuisances never enter the primary $Q$ or FSS.

The severity analysis uses:

$$\text{JPEG quality} : 95,\ 80,\ 60,\ 40, \tag{21}$$

$$\text{Blur } \sigma : 0.5,\ 1.0,\ 1.5,\ 2.0, \tag{22}$$

$$\text{Resize scale} : 0.95,\ 0.85,\ 0.75,\ 0.60, \tag{23}$$

$$\text{Gamma} : 0.95,\ 0.90,\ 0.80,\ 0.65. \tag{24}$$

Each severity point is evaluated together with detector competence. A gate-failing condition may be displayed diagnostically but is not interpreted as evidence of nuisance instability. The complete severity results appear in Appendix C.5.

### B.7 Same-Authenticity Placebo Protocol

The placebo study tests whether Manipulation Reliance can be explained by generic intervention sensitivity rather than an authenticity-changing contrast. For each detector, the matched fake versus pristine intervention effect is compared with score-matched real–real and fake–fake pairs under the same evidence region and intervention operator.

The stronger of the real–real and fake–fake effects defines the placebo reference. A manipulation-specific interpretation requires the matched fake versus pristine effect to exceed this reference. Full placebo results are reported in Appendix C.4.

### B.8 Reserved Validation of Nuisance Instability

Development $Q$ is computed only from the fixed primary nuisance family. To test whether it contains stability information beyond those transformations, we independently compute the corresponding instability under brightness, contrast, and saturation.

The primary reserved analysis measures the association between development and reserved instability while conditioning on Manipulation Reliance and manipulation family. A complementary matched-$M$ analysis compares samples within the caliper

$$|M_i - M_j| \leq 0.05.$$

Neither analysis modifies the primary nuisance family or the reported FSS.

### B.9 Scalarization and Gate Sensitivity Protocols

To determine whether the empirical ordering is created by the harmonic scalarization itself, we compare the perturbation-based Faith baseline, $M$, $1 - Q$, the product $M(1 - Q)$, and FSS using GT versus control AUROC and pairwise ranking.

Audit-threshold sensitivity is evaluated without retraining or recalibrating the detectors. The detector competence threshold is re-evaluated at

$$0.75, \quad 0.80, \quad 0.85.$$

The nuisance gate is evaluated under three settings:

$$\text{loose} : (0.70,\ 0.15), \tag{25}$$

$$\text{primary} : (0.75,\ 0.10), \tag{26}$$

$$\text{strict} : (0.80,\ 0.05), \tag{27}$$

where each pair denotes the minimum nuisance AUC and maximum allowed AUC drop. The primary paper uses the middle setting together with the locked in-domain AUC threshold of $0.80$.

### B.10 Reserved Nonforensic Proxy Stress Test

The proxy experiment tests a stronger potential role for $Q$: whether nuisance instability can reject a nonforensic region after that region has first matched the GT region under $M$.

A high-frequency proxy is selected on a video disjoint calibration split using only coverage and Manipulation Reliance. Neither $Q$ nor FSS participates in the selection. The selected proxy is then frozen and evaluated on the reserved evaluation split. The prespecified Manipulation Reliance non-inferiority margin is $-0.05$. Because the hypothesis is defined before observing $Q$, failure of this non-inferiority condition is retained as a failed prespecified stress test rather than followed by post hoc proxy reselection.

### B.11 Controlled FF++ Residual Transfer Construction

The transfer experiment tests the audit under a controlled FF++ authenticity contrast in a new target domain. For each target dataset, a genuine image is used as the pristine member of the pair. A deterministic residual from a fixed FF++ Forensic-GT template is transferred only within that template's manipulation support to construct its paired counterpart.

The residual pool and strength, spatial support, intervention operator, primary nuisance family, detector checkpoint, and detector specific score scale are held fixed across target domains. These pairs are designed only for the controlled RIFT transfer audit and are not treated as native OOD deepfakes.

Transferred audit quantities are interpreted only when the detector achieves controlled Pair AUC of at least $0.80$. Natural OOD AUC is measured separately on the original target dataset and is never used to tune the transferred construction or the RIFT quantities.

### B.12 Aggregation and Statistical Inference

All nonlinear audit quantities are computed before dataset-level aggregation. For a generic per sample quantity $z_i$, values are averaged in the order

$$\text{frame/sample} \rightarrow \text{video group} \rightarrow \text{manipulation family} \rightarrow \text{equal-weight macro}.$$

Consequently, Faith and FSS are never reconstructed from group-level averages of their constituent terms.

Uncertainty is estimated using $10{,}000$ paired bootstrap replicates at the video-group level, stratified by manipulation family. When comparing GT and control regions, the same resampled video-group identities are used for both regions before computing their difference. This preserves the paired experimental design and avoids treating multiple frames from the same video group as independent observations.

The same aggregation and resampling rules are used throughout the primary audit, placebo analysis, recovery study, operator comparison, and reported GT-minus-control margins unless explicitly stated otherwise.

## C Extended Results and Audit Diagnostics

This appendix reports the complete supporting analyses behind Section 5. The studies follow the same evidential progression as the main text: controlled pair sanity checks, cross detector spatial controls, ordinary sample recovery, construct validity of $M$, diagnostic evidence for $Q$, intervention and scalarization robustness, admissibility sensitivity, failed prespecified stress tests, and controlled residual transfer.

### C.1 Controlled Shared Cue Dose Response

Table 3 gives the complete numerical trajectory behind Figure 2. The planted cue becomes more influential under the perturbation based Faith baseline as cue strength increases, while remaining nearly irrelevant to the matched authenticity contrast measured by $M$.

Table 3: **Controlled shared cue response under increasing cue strength.** The cue is predictive for the controlled detector but is shared across the matched fake–pristine pair at audit time. $\Delta\mathrm{FSS}$ is GT FSS minus shortcut FSS.

| Strength | $\mathrm{Faith}_S$ | $\mathrm{Faith}_{GT}$ | $M_S$ | $M_{GT}$ | $Q_S$ | $Q_{GT}$ | $\mathrm{FSS}_S$ | $\mathrm{FSS}_{GT}$ | $\Delta$FSS [95% CI] |
|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 0.063 | 0.094 | 0.009 | 0.600 | 0.545 | 0.547 | 0.014 | 0.461 | 0.426 [0.407,0.445] |
| 0.50 | 0.207 | 0.076 | 0.019 | 0.594 | 0.461 | 0.551 | 0.027 | 0.452 | 0.418 [0.399,0.438] |
| 0.75 | 0.220 | 0.074 | 0.018 | 0.590 | 0.434 | 0.554 | 0.025 | 0.448 | 0.414 [0.395,0.433] |
| 1.00 | 0.223 | 0.074 | 0.017 | 0.588 | 0.430 | 0.556 | 0.024 | 0.447 | 0.413 [0.393,0.432] |

At strength $0.25$, the true manipulation still has the larger Faith value. From strength $0.50$ onward, the Faith ordering reverses toward the planted cue. The $M$ and FSS ordering does not: GT FSS remains between $0.447$ and $0.461$, while cue FSS remains between $0.014$ and $0.027$. The sanity check therefore holds at every tested strength rather than at one selected setting.

### C.2 Full Cross Detector Spatial-Control Audit

The common full-control protocol compares the GT manipulation support with three non-overlapping regions constructed from exact translations of the same mask whenever the strict protocol is feasible. Table 4 reports every control separately.

Table 4: **Full GT versus control forensic-specificity audit.** The final margin subtracts the strongest of matched-background, maximally displaced, and deterministic-random controls from GT FSS.

| Detector | $\mathrm{FSS}_{GT}$ | $\mathrm{FSS}_{BG}$ | $\mathrm{FSS}_{disp}$ | $\mathrm{FSS}_{rand}$ | $\Delta$FSS | 95% CI |
|---|---|---|---|---|---|---|
| ResNet-50 | 0.3672 | 0.0276 | 0.0141 | 0.0123 | 0.3396 | [0.3185,0.3605] |
| EfficientNet-B4 | 0.3474 | 0.0474 | 0.0243 | 0.0287 | 0.3000 | [0.2793,0.3207] |
| ConvNeXt-Tiny | 0.4916 | 0.0391 | 0.0072 | 0.0153 | 0.4525 | [0.4291,0.4754] |
| Swin-T | 0.4493 | 0.0557 | 0.0137 | 0.0231 | 0.3935 | [0.3727,0.4145] |
| DIMF | 0.5359 | 0.0653 | 0.0351 | 0.0383 | 0.4706 | [0.4471,0.4939] |
| TSRL$^\dagger$ | 0.1086 | 0.0145 | 0.0042 | 0.0085 | 0.0941 | [0.0844,0.1045] |

$^\dagger$TSRL's raw spatial margin is shown for completeness but is not interpreted as a successful forensic audit because independent placebo and nuisance-validity requirements fail.

For each valid detector, matched background is the strongest spatial control, yet remains far below the true manipulation region. The displaced and random controls are smaller still. Because these controls preserve the GT geometry, their separation from GT cannot be attributed only to region area, mask shape, perimeter, or intervention size.

### C.3 Faith Versus FSS: GT-over-Control Ranking

The controlled shared cue check isolates the matched pair distinction under known construction. Table 5 reports how often the same matched comparison changes GT versus control ranking on ordinary video groups.

Across the five valid detectors, the average ranking rate increases from $64.7\%$ under Faith to $99.1\%$ under FSS. TSRL is the only detector for which the FSS ranking rate decreases, with net change $-0.054[-0.093, -0.014]$; this behavior agrees with its independent construct failures rather than being removed from the analysis.

Table 5: **Full sample-level recovery analysis.** Faith Rank and FSS Rank are the fractions of video groups in which GT is ranked above the strongest spatial control. Recovery is the fraction of Faith failures corrected by FSS; regression is the converse.

| Detector | Faith Rank | FSS Rank | Faith Fail | Recovery | Regression | Net Gain [95% CI] |
|---|---|---|---|---|---|---|
| ResNet-50 | 0.708 | 0.993 | 0.292 | 0.980 | 0.000 | +0.285 [0.234,0.335] |
| EfficientNet-B4 | 0.575 | 0.986 | 0.425 | 0.963 | 0.005 | +0.411 [0.353,0.468] |
| ConvNeXt-Tiny | 0.740 | 0.993 | 0.260 | 0.983 | 0.005 | +0.252 [0.202,0.303] |
| Swin-T | 0.621 | 0.996 | 0.379 | 0.988 | 0.000 | +0.375 [0.321,0.429] |
| DIMF | 0.589 | 0.989 | 0.411 | 0.977 | 0.005 | +0.401 [0.347,0.455] |
| TSRL† | 0.953 | 0.899 | 0.047 | 0.722 | 0.090 | -0.054 [-0.093,-0.014] |

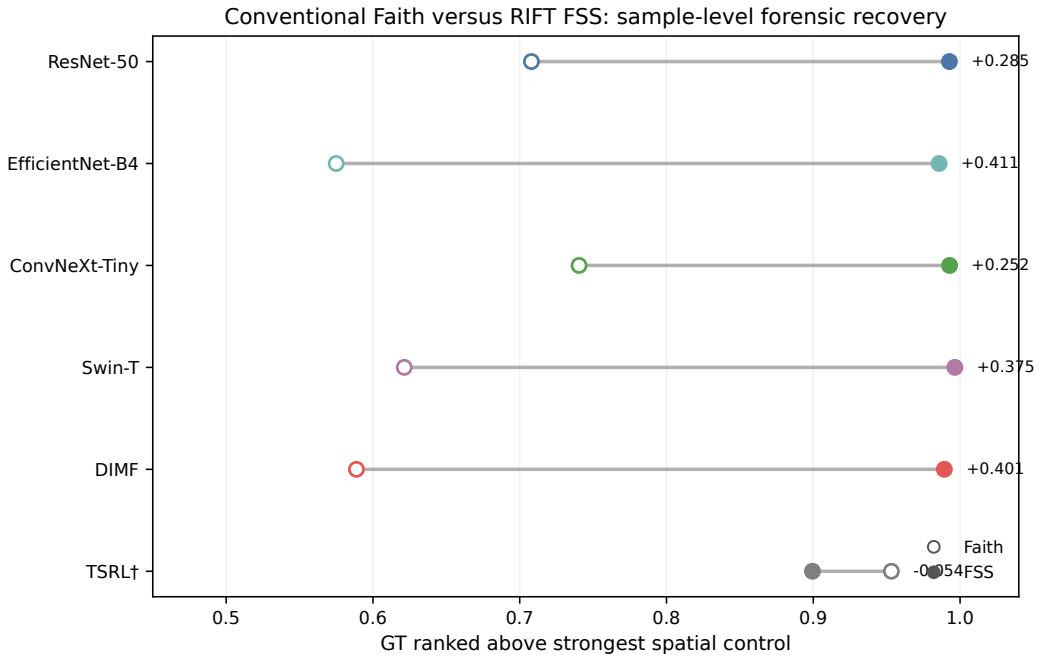



Figure 3: **GT-over-control ranking under Faith and FSS.** Open markers show the Faith GT-over-control ranking rate and filled markers show the corresponding FSS ranking rate. The gain reflects access to the matched pristine counterpart, which Faith does not use. All five valid full-control detectors improve; TSRL moves in the opposite direction, consistent with its independent audit-validity failures.

### C.4 Same Authenticity Placebo Tests

Manipulation Reliance should not become large merely because an intervention changes a detector score. We therefore compare the matched fake–pristine effect with score-matched real–real and fake–fake pairs under the same intervention.

Table 6: **Same-authenticity placebo test.** The final column subtracts the stronger real–real or fake–fake placebo from the matched fake–pristine effect.

| Detector | Matched F–R | R–R | F–F | Matched − strongest placebo [95% CI] |
|---|---|---|---|---|
| ResNet-50 | 0.3337 | 0.1223 | 0.1203 | 0.2114 [0.1902,0.2317] |
| EfficientNet-B4 | 0.3145 | 0.1148 | 0.1170 | 0.1975 [0.1747,0.2195] |
| ConvNeXt-Tiny | 0.4990 | 0.1417 | 0.1009 | 0.3573 [0.3280,0.3873] |
| Swin-T | 0.4403 | 0.1307 | 0.1023 | 0.3096 [0.2837,0.3352] |
| DIMF | 0.5297 | 0.1471 | 0.1169 | 0.3826 [0.3519,0.4131] |
| DiffusionFake | 0.7327 | 0.0668 | 0.0555 | 0.6660 [0.6490,0.6825] |
| TSRL | 0.1842 | 0.2188 | 0.2093 | -0.0346 [-0.0495,-0.0194] |

The six admissible evaluations produce positive matched-versus-placebo margins with intervals entirely above zero. TSRL exhibits the opposite ordering and is therefore retained as a failed construct-validity case.

### C.5 Reserved Nuisance Validation

#### C.5.1 Reserved Nuisance Prediction

Primary $Q$ is computed from the predefined JPEG, blur, resize, and gamma family, while brightness, contrast, and saturation are reserved for this reserved analysis. We test whether the primary measurement contains stability information that transfers beyond the nuisance family used in the primary audit.

Table 7: **Reserved nuisance construct validity.** The partial statistic controls for Manipulation Reliance and manipulation family. Matched-$M$ pairs use the caliper $|M_i - M_j| \leq 0.05$.

| Detector | $Q_{primary}$ vs. $Q_{reserve}$ | Partial $\rho$ given $M$ [95% CI] | Matched-$M$ pairs | Concordance [95% CI] |
|---|---|---|---|---|
| ResNet-50 | 0.333 | 0.118 [-0.016,0.240] | 128 | 0.445 [0.351,0.600] |
| EfficientNet-B4 | 0.360 | 0.216 [0.084,0.338] | 126 | 0.595 [0.455,0.696] |
| ConvNeXt-Tiny | 0.380 | 0.140 [0.020,0.257] | 129 | 0.504 [0.417,0.673] |
| Swin-T | 0.277 | 0.109 [-0.009,0.227] | 126 | 0.516 [0.396,0.648] |
| DIMF | 0.552 | 0.230 [0.101,0.349] | 127 | 0.551 [0.442,0.692] |
| DiffusionFake | 0.349 | 0.287 [0.179,0.387] | 131 | 0.542 [0.444,0.688] |
| TSRL[†] | 0.470 | 0.527 [0.431,0.608] | 128 | 0.664 [0.552,0.780] |

[†]TSRL is not counted as positive construct-validity evidence because its nuisance competence requirement fails.

EfficientNet-B4, ConvNeXt-Tiny, DIMF, and DiffusionFake retain positive partial associations with intervals excluding zero after conditioning on $M$ and manipulation family. ResNet-50 and Swin-T remain inconclusive. The matched-$M$ concordance estimates are weaker and are therefore not used as the principal evidence for an independent $Q$ effect.

#### C.5.2 Nuisance Severity Under Competence Gating

A second analysis tests whether $Q$ responds in the expected direction as authenticity-preserving perturbations become stronger, while separately checking whether the detector remains competent.

The severity grids are JPEG qualities $95, 80, 60, 40$; blur $\sigma = 0.5, 1.0, 1.5, 2.0$; resize scales $0.95, 0.85, 0.75, 0.60$; and gamma $0.95, 0.90, 0.80, 0.65$. Admissibility is evaluated per condition rather than inferred from the number of passing levels. A severity condition is interpreted only when

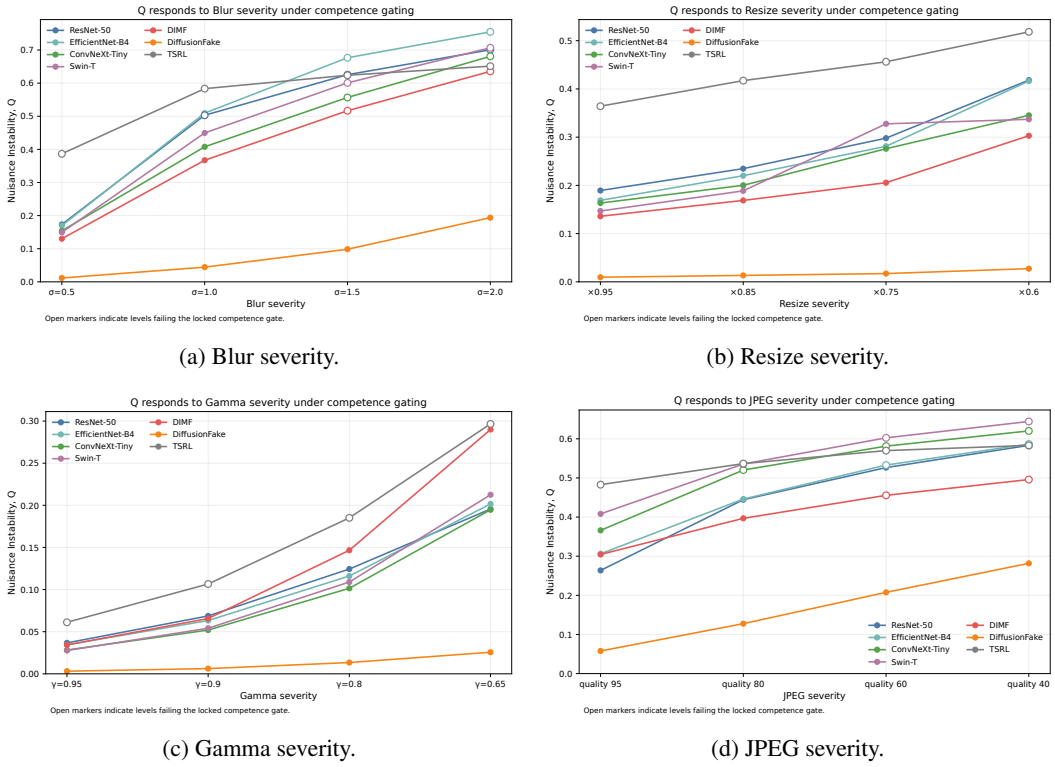
Table 8: **Competence-admissible nuisance severity levels.** Each entry gives the number of admissible levels out of four; parentheses report the severity–$Q$ Spearman correlation when defined.

| Detector | JPEG | Blur | Resize | Gamma |
|---|---|---|---|---|
| ResNet-50 | 3/4 (1.00) | 1/4 | 4/4 (1.00) | 4/4 (1.00) |
| EfficientNet-B4 | 2/4 | 2/4 | 4/4 (1.00) | 4/4 (1.00) |
| ConvNeXt-Tiny | 1/4 | 2/4 | 4/4 (1.00) | 4/4 (1.00) |
| Swin-T | 1/4 | 2/4 | 4/4 (1.00) | 4/4 (1.00) |
| DIMF | 2/4 | 2/4 | 4/4 (1.00) | 4/4 (1.00) |
| DiffusionFake | 4/4 (1.00) | 4/4 (1.00) | 4/4 (1.00) | 4/4 (1.00) |
| TSRL$^{\dagger}$ | 0/4 | 0/4 | 0/4 | 0/4 |

Table 9: **Admissibility of the four primary nuisance settings.** A check indicates that the exact condition used in the fixed primary family passes the locked competence gate. These statuses affect interpretation, not membership in the raw fixed family $Q$.

| Detector | JPEG 60 | Blur 1.5 | Resize 0.75 | Gamma 0.8 |
|---|---|---|---|---|
| ResNet-50 | ✓ | ✗ | ✓ | ✓ |
| EfficientNet-B4 | ✗ | ✗ | ✓ | ✓ |
| ConvNeXt-Tiny | ✗ | ✗ | ✓ | ✓ |
| Swin-T | ✗ | ✗ | ✓ | ✓ |
| DIMF | ✗ | ✗ | ✓ | ✓ |
| DiffusionFake | ✓ | ✓ | ✓ | ✓ |
| TSRL | ✗ | ✗ | ✗ | ✗ |

the predefined competence requirements remain satisfied; individual gate failures are shown by open markers in Figure 4.




(a) Blur severity.

(b) Resize severity.

(c) Gamma severity.

(d) JPEG severity.

Figure 4: **Nuisance-severity response.** Open markers denote levels that fail the locked competence gate and are reported diagnostically rather than used as construct-validity evidence.

DiffusionFake provides the cleanest fully admissible trajectory. JPEG $Q_{GT}$ rises from $0.058$ to $0.282$ as quality decreases from $95$ to $40$ while AUC remains $0.987$ at the strongest setting. Blur $Q$ rises from $0.012$ to $0.194$ while the strongest tested condition retains AUC $0.997$. Resize and gamma are monotonic across all admissible levels.

### C.6 Intervention Operator and Boundary Robustness

A perturbation-based audit can be confounded if its result depends on a particular neutralization operator or on the artificial boundary introduced by the intervention. We therefore repeat the full-control audit with local blur, local mean fill, local pixelation, and a smooth-boundary feathered blur.

Table 10: **GT-versus-strongest-control ΔFSS across intervention operators.** Feathered blur removes the sharp binary intervention boundary. Every displayed interval excludes zero.

| Detector | Local blur | Local mean fill | Local pixelate | Feathered blur |
|---|---|---|---|---|
| ResNet-50 | 0.340 [0.319,0.360] | 0.372 [0.350,0.393] | 0.330 [0.309,0.350] | 0.326 [0.304,0.347] |
| EfficientNet-B4 | 0.300 [0.279,0.321] | 0.329 [0.307,0.350] | 0.324 [0.302,0.345] | 0.288 [0.266,0.309] |
| ConvNeXt-Tiny | 0.452 [0.429,0.475] | 0.474 [0.451,0.497] | 0.452 [0.429,0.474] | 0.445 [0.421,0.469] |
| Swin-T | 0.394 [0.373,0.414] | 0.431 [0.410,0.451] | 0.409 [0.388,0.430] | 0.386 [0.365,0.408] |
| DIMF | 0.471 [0.447,0.494] | 0.481 [0.458,0.505] | 0.447 [0.424,0.471] | 0.463 [0.440,0.487] |

DiffusionFake does not support the identical strict translated-control construction. Under its detector-native GT-only protocol, $\mathrm{FSS}_{GT}$ is $0.796$, $0.797$, $0.741$, and $0.783$ under blur, mean fill, pixelation, and feathered blur, respectively.

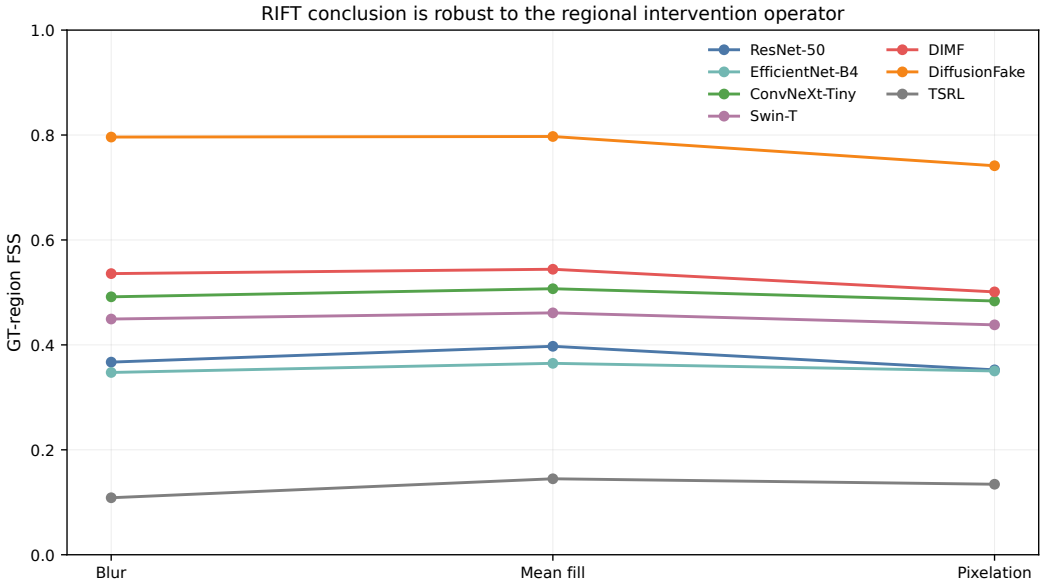



Figure 5: **Intervention robustness.** Absolute GT-region FSS varies moderately across operators while the forensic ordering remains stable for the valid full-control detectors.

All $20/20$ valid detector and operator combinations retain positive GT versus control intervals. The feathered results remain close to the hard blur results, providing a direct control against an explanation based on a sharp intervention seam.

### C.7 Metric Comparison and Scalarization

FSS is intended as a conservative joint summary of manipulation reliance and stability, not as an optimized classifier of GT regions. We therefore compare it directly with its components and with the product $M(1-Q)$.

For the standard backbones and DIMF, $M$ and FSS are nearly tied. This result narrows the interpretation of the method: $M$ is the primary matched manipulation diagnostic, $Q$ records fixed family stability, and FSS summarizes their conjunction. No claim is made that FSS universally improves GT versus control discrimination.

### C.8 Audit Gate Sensitivity

The primary audit uses an ID competence threshold of $\mathrm{AUC} \geq 0.80$, nuisance AUC of at least $0.75$, and a maximum nuisance AUC drop of $0.10$. We reclassify the saved validation outputs over a prespecified threshold grid without retraining or recalibrating any detector.

The valid detector panel remains ID-admissible throughout the tested range. Tightening the nuisance gate reduces the number of interpretable severity conditions but does not reverse the principal valid/invalid pattern: DiffusionFake remains admissible at all 19 tested settings, whereas TSRL remains inadmissible at all 19.

### C.9 Reserved Nonforensic Proxy Stress Test

To test whether $Q$ can reject a nonforensic region after controlling for high $M$, a high-frequency proxy is selected on a video-disjoint calibration split using only coverage and $M$. Neither $Q$ nor FSS participates in selection. The resulting proxy is then frozen and evaluated on the evaluation split.

The intended universal hypothesis is not supported. DiffusionFake exhibits the largest competence-admissible nuisance gap, $\Delta Q = +0.235[+0.224, +0.246]$, but its proxy already fails the prespecified

Table 11: **GT versus control discrimination by audit quantity.** Backbone macro is the equal-weight average over ResNet-50, EfficientNet-B4, ConvNeXt-Tiny, and Swin-T.

| Evaluation | Metric | AUROC ↑ | Pairwise Ranking ↑ |
|---|---|---|---|
| Backbone macro | Faith | 0.7095 | 0.6711 |
| | $M$ | 0.9717 | 0.9929 |
| | $1-Q$ | 0.6683 | 0.5641 |
| | $M(1-Q)$ | 0.9726 | 0.9920 |
| | FSS | 0.9707 | 0.9938 |
| DIMF | Faith | 0.6755 | 0.5960 |
| | $M$ | 0.9820 | 0.9893 |
| | $1-Q$ | 0.7781 | 0.7046 |
| | $M(1-Q)$ | 0.9827 | 0.9893 |
| | FSS | 0.9816 | 0.9893 |
| TSRL$^{\dagger}$ | Faith | 0.9493 | 0.9532 |
| | $M$ | 0.8991 | 0.9175 |
| | $1-Q$ | 0.4874 | 0.3055 |
| | $M(1-Q)$ | 0.9070 | 0.9356 |
| | FSS | 0.8974 | 0.9067 |

Table 12: **Sensitivity of the detector-level competence gate.**

| Detector | ID AUC | $\geq 0.75$ | $\geq 0.80$ | $\geq 0.85$ |
|---|---|---|---|---|
| ResNet-50 | 0.897 | ✓ | ✓ | ✓ |
| EfficientNet-B4 | 0.869 | ✓ | ✓ | ✓ |
| ConvNeXt-Tiny | 0.947 | ✓ | ✓ | ✓ |
| Swin-T | 0.940 | ✓ | ✓ | ✓ |
| DIMF | 0.963 | ✓ | ✓ | ✓ |
| DiffusionFake | 0.999 | ✓ | ✓ | ✓ |
| TSRL | 0.830 | ✓ | ✓ | ✗ |
| ViT-B/16 | 0.787 | ✓ | ✗ | ✗ |

Table 13: **Nuisance-gate sensitivity.** Entries give the number of competence-admissible nuisance severity levels out of 19. Column headers show (minimum nuisance AUC, maximum AUC drop).

| Detector | Loose $(0.70, 0.15)$ | Primary $(0.75, 0.10)$ | Strict $(0.80, 0.05)$ |
|---|---|---|---|
| ResNet-50 | 18/19 | 15/19 | 11/19 |
| EfficientNet-B4 | 18/19 | 15/19 | 11/19 |
| ConvNeXt-Tiny | 17/19 | 14/19 | 11/19 |
| Swin-T | 18/19 | 14/19 | 11/19 |
| DIMF | 18/19 | 15/19 | 10/19 |
| DiffusionFake | 19/19 | 19/19 | 19/19 |
| TSRL | 0/19 | 0/19 | 0/19 |

Table 14: **Reserved nonforensic proxy stress test.** $\Delta M = M_P - M_{GT}$, $\Delta Q = Q_P - Q_{GT}$, and $\Delta\text{FSS} = \text{FSS}_{GT} - \text{FSS}_P$. M-NI denotes non-inferiority under the prespecified $-0.05$ margin.

| Detector | $\Delta M$ [95% CI] | M-NI | $\Delta Q$ [95% CI] | Q-caught | $\Delta$FSS [95% CI] | FSS-restored | All nuisance gates |
|---|---|---|---|---|---|---|---|
| DIMF | -0.412 [-0.443,-0.383] | No | +0.184 [+0.168,+0.199] | Yes | +0.395 [+0.369,+0.421] | Yes | No |
| DiffusionFake | -0.648 [-0.666,-0.630] | No | +0.235 [+0.224,+0.246] | Yes | +0.675 [+0.657,+0.692] | Yes | Yes |
| TSRL | -0.060 [-0.071,-0.049] | No | +0.007 [-0.011,+0.026] | No | +0.075 [+0.064,+0.086] | Yes | No |

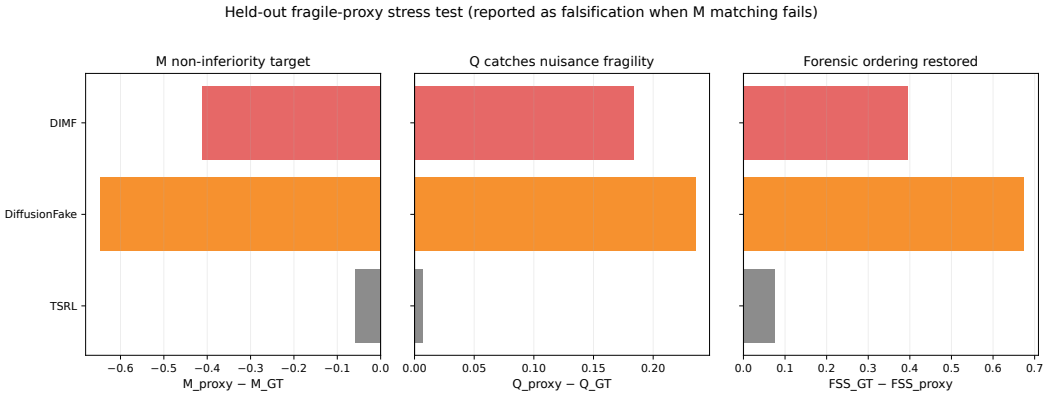



Figure 6: **Reserved nonforensic proxy stress test.** The prespecified universal double-dissociation criterion is not satisfied because the proxy fails $M$ non-inferiority for every tested detector. The experiment is retained as a failed prespecified stress test rather than redefined after observing the result.

$M$ non-inferiority requirement. The experiment therefore cannot establish that $Q$ repairs an $M$-matched failure.

### C.10 Controlled Residual Transfer Results

The controlled pair construction is specified in Appendix B.11. This section reports the complete target domain grid and separates controlled Pair AUC from natural OOD AUC.

#### C.10.1 Full Per-Domain Controlled Transfer Grid

The full grid explains why the main paper restricts its controlled transfer interpretation to WildDeepfake. It is the only tested target domain for which multiple detectors satisfy the locked controlled Pair AUC threshold: ConvNeXt-Tiny (0.852), Swin-T (0.856), DIMF (0.835), and DiffusionFake (0.913). All four nevertheless show lower FSS than in-domain. For the remaining detector–domain combinations, insufficient controlled-pair competence prevents a strong controlled-transfer interpretation regardless of the reported raw $M$, $Q$, or FSS values.

### C.11 Qualitative Audit Examples

Table 15: **Per-domain controlled residual transfer results.** Pair AUC measures detector competence on the controlled transferred authenticity pair. Parentheses give 95% video/sample-group bootstrap CIs for the GT manipulation region. Bold Pair AUC values satisfy the locked 0.80 controlled-pair competence criterion and therefore support the strongest controlled-transfer interpretation. Natural OOD AUC is measured independently on the original dataset; — indicates that it is not reported for that detector.

| Detector | Domain | Pair AUC | $M_{GT}$ | $Q_{GT}$ | $\text{FSS}_{GT}$ | FSS drop from ID | Natural OOD AUC |
|---|---|---|---|---|---|---|---|
| ResNet-50 | Celeb-DF-v2 | 0.579 | 0.070 (0.062,0.078) | 0.384 (0.366,0.401) | 0.099 (0.089,0.109) | 0.269 | — |
| ResNet-50 | DFD | 0.530 | 0.026 (0.017,0.037) | 0.428 (0.393,0.461) | 0.040 (0.029,0.053) | 0.328 | — |
| ResNet-50 | WildDeepfake | 0.686 | 0.084 (0.080,0.087) | 0.365 (0.357,0.374) | 0.133 (0.128,0.138) | 0.234 | — |
| ResNet-50 | DiffSwap | 0.572 | 0.041 (0.038,0.043) | 0.380 (0.371,0.388) | 0.067 (0.063,0.071) | 0.300 | — |
| EfficientNet-B4 | Celeb-DF-v2 | 0.546 | 0.050 (0.044,0.056) | 0.367 (0.349,0.384) | 0.077 (0.070,0.085) | 0.270 | — |
| EfficientNet-B4 | DFD | 0.513 | 0.019 (0.013,0.026) | 0.361 (0.329,0.392) | 0.030 (0.022,0.038) | 0.317 | — |
| EfficientNet-B4 | WildDeepfake | 0.688 | 0.066 (0.063,0.069) | 0.331 (0.323,0.340) | 0.112 (0.107,0.116) | 0.236 | — |
| EfficientNet-B4 | DiffSwap | 0.590 | 0.053 (0.051,0.055) | 0.345 (0.337,0.353) | 0.090 (0.086,0.093) | 0.258 | — |
| ConvNeXt-Tiny | Celeb-DF-v2 | 0.708 | 0.190 (0.175,0.205) | 0.448 (0.432,0.464) | 0.225 (0.210,0.241) | 0.266 | — |
| ConvNeXt-Tiny | DFD | 0.569 | 0.051 (0.040,0.063) | 0.456 (0.423,0.489) | 0.074 (0.060,0.090) | 0.417 | — |
| ConvNeXt-Tiny | WildDeepfake | **0.852** | 0.243 (0.237,0.248) | 0.261 (0.254,0.267) | 0.342 (0.335,0.349) | 0.150 | — |
| ConvNeXt-Tiny | DiffSwap | 0.741 | 0.163 (0.157,0.168) | 0.354 (0.348,0.360) | 0.229 (0.222,0.235) | 0.263 | — |
| Swin-T | Celeb-DF-v2 | 0.647 | 0.104 (0.095,0.114) | 0.446 (0.428,0.462) | 0.137 (0.126,0.149) | 0.312 | — |
| Swin-T | DFD | 0.546 | 0.028 (0.021,0.036) | 0.420 (0.387,0.453) | 0.045 (0.035,0.056) | 0.404 | — |
| Swin-T | WildDeepfake | **0.856** | 0.200 (0.195,0.205) | 0.359 (0.352,0.366) | 0.286 (0.279,0.292) | 0.164 | — |
| Swin-T | DiffSwap | 0.658 | 0.102 (0.098,0.106) | 0.431 (0.423,0.440) | 0.147 (0.141,0.153) | 0.302 | — |
| DIMF | Celeb-DF-v2 | 0.665 | 0.163 (0.150,0.176) | 0.445 (0.428,0.462) | 0.204 (0.190,0.218) | 0.332 | 0.962 |
| DIMF | DFD | 0.565 | 0.056 (0.045,0.068) | 0.444 (0.409,0.477) | 0.082 (0.068,0.097) | 0.454 | 0.799 |
| DIMF | WildDeepfake | **0.835** | 0.181 (0.176,0.186) | 0.411 (0.403,0.420) | 0.256 (0.249,0.262) | 0.280 | 0.936 |
| DIMF | DiffSwap | 0.778 | 0.074 (0.070,0.077) | 0.405 (0.396,0.414) | 0.117 (0.112,0.122) | 0.419 | 0.999 |
| DiffusionFake | Celeb-DF-v2 | 0.665 | 0.069 (0.061,0.078) | 0.256 (0.239,0.273) | 0.107 (0.095,0.119) | 0.689 | 0.577 |
| DiffusionFake | DFD | 0.534 | 0.014 (0.009,0.020) | 0.293 (0.258,0.328) | 0.025 (0.016,0.034) | 0.772 | 0.475 |
| DiffusionFake | WildDeepfake | **0.913** | 0.251 (0.244,0.257) | 0.209 (0.202,0.215) | 0.360 (0.352,0.368) | 0.436 | 0.798 |
| DiffusionFake | DiffSwap | 0.608 | 0.071 (0.067,0.075) | 0.275 (0.266,0.283) | 0.111 (0.105,0.117) | 0.685 | 0.804 |
| TSRL† | Celeb-DF-v2 | 0.425 | 0.019 (0.016,0.023) | 0.337 (0.318,0.357) | 0.027 (0.023,0.031) | 0.082 | 0.527 |
| TSRL† | DFD | 0.462 | 0.021 (0.014,0.028) | 0.444 (0.404,0.483) | 0.028 (0.021,0.037) | 0.080 | 0.433 |
| TSRL† | WildDeepfake | 0.555 | 0.033 (0.029,0.038) | 0.498 (0.486,0.511) | 0.038 (0.034,0.041) | 0.071 | 0.552 |
| TSRL† | DiffSwap | 0.505 | 0.028 (0.025,0.031) | 0.482 (0.471,0.493) | 0.034 (0.031,0.036) | 0.075 | 0.606 |

†TSRL is already invalidated by the primary audit-validity tests and is included here only as a diagnostic failure. Pair AUC below 0.80 denotes a competence-limited controlled transfer condition; corresponding RIFT values are descriptive and are not used as primary controlled-transfer evidence.

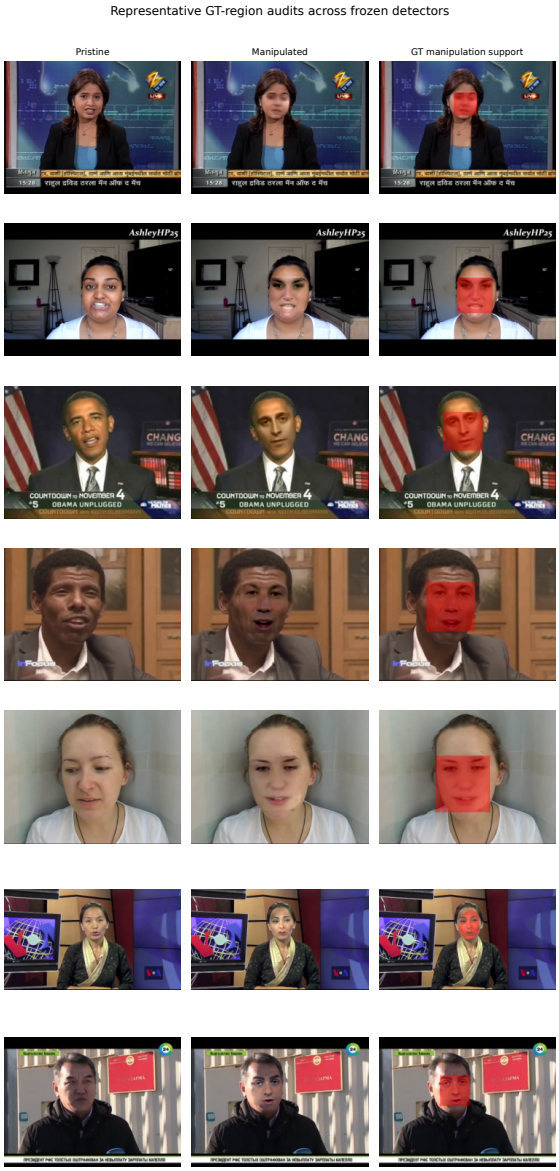



Figure 7: **Representative qualitative audit inputs.** Examples show the matched pristine and manipulated images together with the GT manipulation support used by the audit. They are illustrative only; all forensic conclusions are based on grouped quantitative analyses.