DOCTOR TRIGGER: A FRAMEWORK FOR RELEASE-BOUND FACE VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Face swapping and generative editing enable realistic facial manipulation, while online redistribution complicates publisher oversight of circulating face images. Existing approaches establish provenance, detect manipulation, recover embedded watermarks, verify image integrity, or disrupt unauthorized face manipulation. These capabilities do not necessarily establish consistency with a specific publisher-enrolled release, limiting the evidence available for verification and human review. To address this gap, we introduce DOCTOR TRIGGER, a framework for verifying circulating faces against private publisher-held release records. Before publication, the publisher embeds a weak keyed Fourier-phase signal in each face while preserving visual fidelity and stores its spectral reference. Afterward, verification compares the circulating face with that reference and uses key projection to measure retention of the release-specific signal. Therefore, relation and key evidence jointly help publishers identify suspicious copies for review, potentially reducing the circulation of manipulated images. Experimental results across FF++, DFDC, and Celeb-DF show that spectral agreement separates tested face swaps and generative manipulations from enrolled releases, while the keyed signal distinguishes published versions from their unreleased originals under the evaluated key-preserving benign transformations. DOCTOR TRIGGER complements provenance metadata with release-specific face verification and record-linked evidence for further assessment.

1 INTRODUCTION

Recent face and head swapping systems produce realistic identity changes (Wang et al., 2025a; Kang et al., 2025), while compression and resizing also alter images as they circulate through online sharing and reposting. For publishers such as news organizations and public institutions, the question is: *does a publicly distributed face still correspond to the specific version that was released?* The publisher can retain release-specific evidence before distribution for later verification.

To support trust in circulating images, C2PA binds signed provenance assertions to content and supports reconnection through soft bindings (Coalition for Content Provenance and Authenticity, 2026). Complementing provenance, recent detectors learn visual-language cues, discrepancies, and large-scale forgery representations (Guo et al., 2025; Yang et al., 2025; Jung et al., 2026). Reference-based detection further compares regional identity and texture with genuine references (Ming et al., 2026). However, these capabilities may not establish whether a circulating face remains consistent with a specific publisher-enrolled release.

To address this need, DOCTOR TRIGGER uses private records for **release-bound face verification**: assessing circulating faces against specific publisher-enrolled releases. First, the publisher applies a weak keyed phase displacement before publication and stores a spectral reference. A publisher-operated verification service, such as a newsroom verification desk, then uses the record to assess face consistency and release-signal retention in user-submitted copies. Finally, the resulting status guides routine handling or further assessment, as illustrated in Fig. 1. The name reflects this workflow: *Doctor* evokes assessment against a retained history followed by triage; *Trigger* denotes the weak phase signal added before publication while preserving visual fidelity.

Conceptually, the reference assesses face consistency; the key helps identify the published version. Although original-phase matching, released-image similarity, and ArcFace features (Deng et al., 2019) separate many clean face re-renderings, these comparisons may not distinguish the published version from its visually similar original under thresholds calibrated for benign processing. Key

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.

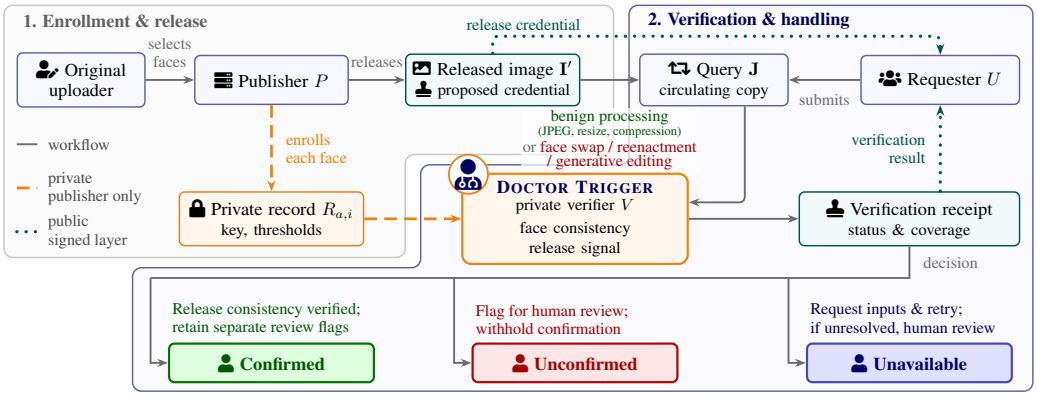

Figure 1: **Two-stage publisher-side verification interface.** The original uploader provides enrollment content; a requester, who may be a different person or platform, submits a circulating query and receives the verification result. The release-to-query arrow traces processing during circulation. Private records remain with the publisher (orange); credentials and status form the public layer (teal). The lower row proposes handling for each status; confirmation can coexist with a review flag. C2PA integration and public signing are proposed; retention tests use HMAC-authenticated credentials.

projection adds evidence of the signal introduced during enrollment, helping identify the enrolled release and providing **release binding**.

From a complementary perspective, robust watermarking preserves embedded messages through processing (Sander et al., 2025; Alam et al., 2025; Bui et al., 2025), while semi-fragile watermarking aims to tolerate benign transformations while detecting manipulation (Neekhara et al., 2024). However, message survival alone does not establish face consistency. Beyond message recovery, face-specific methods provide further evidence: RecoverMark supports localization and recovery (An et al., 2026). Recently, LIDMark has combined landmarks and source identifiers for detection, localization, and tracing (Wu et al., 2026). Our private record supports separate assessments of face consistency and release-signal retention without requiring access to the original image at verification.

Building on these complementary roles, DOCTOR TRIGGER supports a **human-in-the-loop workflow** through its verification statuses: *unconfirmed* directs reviewers to record discrepancies, while *unavailable* requests missing information or human assistance. These outcomes *trigger* the human *doctor* to assess the case. We contribute:

- We formulate **release-bound face verification** using a private per-release record, complementing public provenance metadata and query-only forensic detection.
- We introduce a weak keyed phase displacement with complementary evidence: spectral agreement assesses face consistency to distinguish evaluated swaps and generative manipulations, while signed first-order projection measures release-signal retention, separating correct from wrong keys and releases from unreleased originals under evaluated benign processing.
- We evaluate release verification across datasets, isolating the contribution of key evidence through private-reference controls and assessing practical cost through deployment measurements. To further guide deployment, we analyze processing and attack risks, motivating channel validation and stronger binding between release signals and facial content.

2 PROBLEM SETUP

Enrollment and query. The publisher verifies a circulating face against a specific release enrolled before publication. Let $\mathbf{I} \in \mathbb{R}^{N_h \times N_w \times 3}$ denote the original RGB face crop, with height N_h and width N_w in pixels. Enrollment adds a weak keyed signal to produce the released crop $\mathbf{I}'$ and retains a private record R. For asset a and face slot i, this record is $R_{a,i}$; we omit the indices when considering one face. At verification, a request supplies the circulating query $\mathbf{J}$ and an identifier resolving its record. Our experiments verify individual crops with known enrollment geometry; the proposed multi-face interface verifies each associated face slot separately.

Processing channel. Once released, a face may undergo delivery processing. A benign channel $\mathcal{P}$ specifies recompression, resizing, or another such operation; $\mathcal{C}$ contains allowed, calibrated

channels. Let $\mathcal{R}$ contain the evaluated face re-rendering operators M, such as swapping and diffusion resynthesis (Chen et al., 2020; Rombach et al., 2022). We compare two processing histories:

$$\mathcal{H}_c:\ \mathbf{J}=\mathcal{P}(\mathbf{I}'),\qquad \mathcal{H}_d:\ \mathbf{J}=\mathcal{P}(M(\mathbf{I}')),\quad \mathcal{P}\in\mathcal{C},\ M\in\mathcal{R}. \tag{1}$$

Here $\mathcal{H}_c$ represents delivery processing alone, while $\mathcal{H}_d$ includes face re-rendering before delivery; neither specifies the editor's intent. A supported channel must therefore preserve useful separation between these histories. Benign queries calibrate thresholds, while manipulation tests assess detection performance. The evaluated verifier receives the channel and crop geometry as inputs rather than inferring an unknown processing history.

Verification and human assessment. Verification assesses face consistency and release-signal retention to support publisher decisions. In the proposed workflow, routine cases proceed automatically, while record discrepancies or incomplete verification prompt human assessment. Reviewers use the available scores and record information to examine discrepancies, request missing inputs, and decide whether further verification is needed before acting on a circulating copy. The decision policy formalizes this handoff through verification statuses and a separate review flag.

3 The Doctor Trigger Framework

Doctor Trigger prepares each face for verification before publication and assesses circulating copies against the retained release record. We first describe how the publisher creates the released image and its record, then introduce the verification scores and the policy that maps them to confirmed, unconfirmed, or unavailable outcomes.

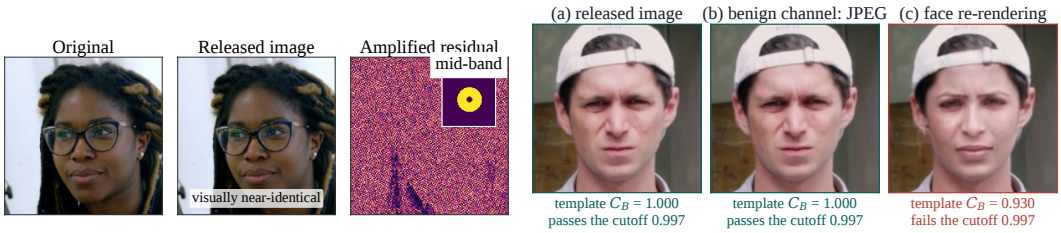

Figure 2: **Release evidence with preserved visual fidelity.** Left: a weak keyed phase displacement adds a release-specific signal while preserving facial appearance. Right: template agreement C_B illustrates why retaining a spectral reference enables face comparison after JPEG processing and reveals changes caused by re-rendering.

Fig. 2 links spectral agreement to face consistency and signed key projection to release-signal retention. Fig. 3 shows enrollment, from per-face key generation to encoding and visual-fidelity assessment.

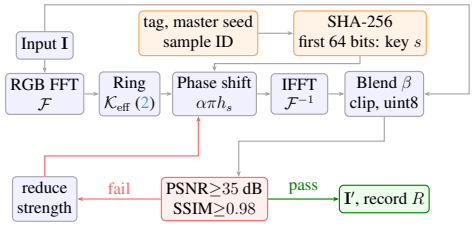

Figure 3: **Enrollment for later verification.** A per-face key determines the signal embedded in the release. The fidelity gate checks the encoded image and reduces signal strength when needed, preserving facial appearance while preparing the release and private record for verification.

3.1 Enrolling Faces for Release

Enrollment prepares a released image $\mathbf{I}'$ and private record R for later verification: the reference captures the enrolled face, while a weak keyed signal identifies its published version. The master seed and sample identifier derive a per-face key s that determines the signal pattern.

Selecting where to place the signal. We use luminance to select one frequency ring and key pattern shared across RGB channels. Define $\mathbf{Y}_I = 0.299\mathbf{I}_R + 0.587\mathbf{I}_G + 0.114\mathbf{I}_B$ and write its Fourier coefficients as $F_I(k) = A_I(k)e^{i\phi_I(k)}$. Here k indexes a two-dimensional Fourier bin, A_I and ϕ_I are its magnitude and phase, and i in $e^{i\phi_I}$ is the imaginary unit. For horizontal and vertical frequencies f_x, f_y in cycles per pixel, define $\kappa(k) = \sqrt{f_x^2 + f_y^2}/\sqrt{0.5^2 + 0.5^2}$. A key s seeds standard-normal draws $u_s(k)$ and the antisymmetric sign pattern $h_s(k) = \text{sign}(u_s(k) - u_s(-k))$. The support is

$$\mathcal{K}_{\text{eff}} = \{k : \kappa_{\min} \le \kappa(k) \le \kappa_{\max},\ A_I(k) > A_{\min},\ h_s(k) \neq 0\}, \tag{2}$$

The ring spans radii $\kappa_{\min}$ to $\kappa_{\max}$ and retains coefficients above the amplitude floor $A_{\min} = 10^{-6} \max_k A_I(k)$ for phase comparison. The key pattern pairs each bin with its conjugate frequency $-k$, taken modulo the grid dimensions; sign returns -1, 0, or 1, assigning opposite signs to each active pair. Self-conjugate bins have $h_s = 0$ and fall outside the active support. This luminance-derived support and pattern are shared across RGB channels.

Preserving visual fidelity. We control the signal with phase strength α, in units of π radians, and blend weight $\beta \in [0, 1]$; $c \in \{R, G, B\}$ indexes color channels. For each channel, enrollment multiplies the active Fourier coefficients by $e^{i\alpha\pi h_s(k)}$, leaving their unblended magnitudes unchanged. The inverse transform is real because $h_s(-k) = -h_s(k)$. Writing $\widetilde{F}$ for the three edited channel spectra and $\mathcal{F}^{-1}$ for the inverse two-dimensional discrete Fourier transform, we blend the edited image with the original, $\widetilde{\mathbf{I}}' = (1-\beta)\mathbf{I} + \beta\mathcal{F}^{-1}(\widetilde{F})$, where $\widetilde{\mathbf{I}}'$ denotes the unquantized blended image. We clip to $[0, 255]$ and truncate to uint8. This signal-dependent quantization shapes the encoded release $\mathbf{I}'$; before it, the phase displacement is

$$\delta_s(k) = \arg\{(1-\beta) + \beta e^{i\alpha\pi h_s(k)}\}. \tag{3}$$

Here arg returns the principal complex phase in radians. The shared RGB multiplier also displaces luminance, with active magnitudes scaled by $\sqrt{1 - 2\beta(1-\beta)(1-\cos\alpha\pi)}$, a second-order change for small α. To enforce visual fidelity, enrollment reduces strength until the encoded release meets PSNR ≥ 35 dB and SSIM ≥ 0.98 against the original.

Retaining evidence for verification. For later comparison, the template $\mathcal{T}_I$ stores pre-quantization phase; subtracting the displacement recovers original phase in exact arithmetic. Using $\mathrm{wrap}(x) = (x + \pi) \bmod 2\pi - \pi$ to map angles to $[-\pi, \pi)$, we obtain, for $k \in \mathcal{K}_{\text{eff}}$,

$$\mathcal{T}_I(k) = \mathrm{wrap}(\phi_I(k) + \delta_s(k)), \qquad \widehat{\phi}_I(k) = \mathrm{wrap}(\mathcal{T}_I(k) - \delta_s(k)). \tag{4}$$

The record also stores original phase and amplitude, key, strengths, support, and crop geometry. Explicit original phase avoids reconstruction quantization error. Nine encoded-release projections—one full-ring and eight bandwise—normalize verification scores without reloading release pixels.

Before quantization, every active bin receives $\delta_s(k) = \pm\delta_{\text{eff}}$. With $\beta = 0.15$, $\alpha = 0.05$ on FF++/Celeb-DF gives $\delta_{\text{eff}} \simeq 0.0235$ rad; $\alpha = 0.015$ on DFDC gives 0.0071 rad. Section 3.2 uses this reference and signed displacement to assess face consistency and release-signal retention.

3.2 Verifying Faces against Release Records

Direct verification combines template agreement C for face consistency with normalized key projection $G_s^{(\text{release})}$ for release-signal retention. For eligible queries, confirmation requires both scores to meet their calibrated thresholds. Original-reference agreement H provides a comparison for analysis. Fig. 4 connects these scores to verification and review.

Spectral agreement. Query luminance $\mathbf{Y}_J$, computed with the RGB weights of $\mathbf{Y}_I$, supplies Fourier phase ϕ_J. We compare it with the stored original phase $\widehat{\phi}_I$ and release template $\mathcal{T}_I$. Let $r_k(\mathbf{J}) = \mathrm{wrap}(\phi_J(k) - \widehat{\phi}_I(k))$ be the query residual. Weights $w_k = A_I(k)/\max_j A_I(j)$ emphasize frequencies that were strong in the original image; k and j index frequency bins. We compute

$$H(\mathbf{J}; R) = \frac{\sum_{\mathcal{K}_{\text{eff}}} w_k \cos r_k(\mathbf{J})}{\sum_{\mathcal{K}_{\text{eff}}} w_k}, \qquad C(\mathbf{J}; R) = \frac{\sum_{\mathcal{K}_{\text{eff}}} w_k \cos(\phi_J(k) - \mathcal{T}_I(k))}{\sum_{\mathcal{K}_{\text{eff}}} w_k}. \tag{5}$$

Larger H indicates agreement with the original face, while larger C indicates agreement with the release template. Their complements $1 - H$ and $1 - C$ are the original-reference error and template mismatch. Since H favors the original image and neither score tests a candidate key, high agreement alone does not identify the published version.

To exclude frequencies degraded by processing, a retained sub-band B restricts the sums to $\mathcal{K}_{\text{eff}} \cap B$, giving H_B and C_B. Relation-only evaluation uses $H_B \geq \tau_H$ (original reference) or $C_B \geq \tau_C$ (release template), with its cutoff fitted to the first percentile of channel-processed benign training scores. Sub-band detection uses $\kappa \leq 0.15$; joint comparisons use full-ring C.

Signed key projection. To resolve the version ambiguity, we project the residual onto its enrolled displacement δ_s. This signed response captures small displacements at first order:

$$G_s(\mathbf{J}; R) = \frac{\sum_{k \in \mathcal{K}_{\text{eff}}} w_k \delta_s(k) \sin r_k(\mathbf{J})}{\sum_{k \in \mathcal{K}_{\text{eff}}} w_k \delta_s(k)^2 + \epsilon_0}. \tag{6}$$

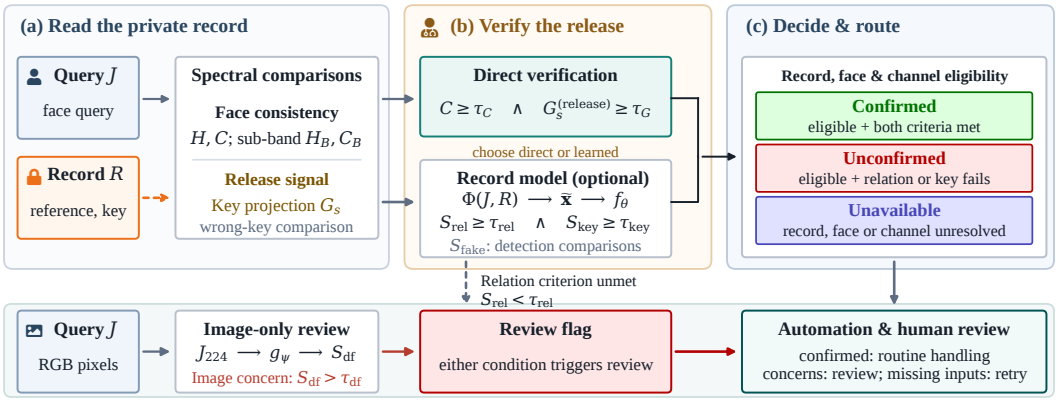

Figure 4: **From release evidence to verification and review.** Direct thresholds and the record model offer two paths for assessing face consistency and release evidence. Eligibility and both verification criteria determine the status, guiding automated handling or human assistance. Optional image-based review adds a separate flag.

Here $\epsilon_0 = 10^{-12}$ stabilizes the denominator. An ideal unquantized release has $r_k = \delta_s(k)$ and $G_s \simeq 1$, whereas the original has zero residual. Contributions from independent wrong keys cancel in expectation. The projection thus supplies release-version evidence that complements spectral agreement. App. C.4 characterizes the empirical wrong-key distribution.

Normalized key evidence. Direct verification measures retention relative to the enrolled release using $G_s^{(\text{release})} = G_s(\mathbf{J}) / \max(|G_s(\mathbf{I}')|, 10^{-6})$. It pairs this normalized projection with spectral agreement, calibrating a threshold for each criterion under the processing channel. Reference controls use the template projection as the denominator, denoted $G_s^{(\text{template})}$.

3.3 Making Verification and Review Decisions

The record model integrates spectral statistics into verification scores, while an image-only detector assesses manipulation from query pixels without access to the release record. The policy determines release consistency through direct thresholds or learned scores and uses a separate review flag to guide further assessment.

Key-contrast features. The record model also uses the contrast between the enrolled key and $m = 8$ wrong projection keys, evaluated on the same query and record. Writing G_j^- for wrong-key response j, we compute

$$(\overline{G}^-, G^-_{\max}, G_\Delta, G_z) = \left(\frac{1}{m} \sum_j G_j^-, \max_j G_j^-, G_s - \max_j G_j^-, \frac{G_s - \overline{G}^-}{\text{std}_j\, G_j^- + 10^{-8}} \right). \tag{7}$$

These features use the stored release response for each full-ring or bandwise comparison to normalize both enrolled- and wrong-key projections. The bar and max denote mean and maximum, and std_j uses divisor m. The margin G_Δ compares the enrolled key with the strongest sampled wrong key; G_z scales separation by their spread. Changing the projection key leaves H and C unchanged.

Record model. The record model combines spectral agreement and key evidence across frequencies to learn their contribution to verification. Its input summarizes the query–record comparison through full-ring statistics $\mathbf{f}_0 \in \mathbb{R}^{11}$ and eight radial-band feature vectors $\mathbf{f}_\ell \in \mathbb{R}^8$, $\ell = 1, \ldots, 8$. The fixed feature map

$$\mathbf{x} = \Phi(\mathbf{J}, R) = [\mathbf{f}_0; \mathbf{f}_1; \cdots; \mathbf{f}_8] \in \mathbb{R}^{75} \tag{8}$$

concatenates agreement, phase-residual, normalized-projection, and wrong-key statistics; semicolons denote concatenation. App. D.1 specifies the features. Training-set means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{\sigma}$ place them on comparable scales before prediction:

$$\tilde{\mathbf{x}} = (\mathbf{x} - \boldsymbol{\mu}) \oslash \boldsymbol{\sigma}, \qquad (S_{\text{rel}}, S_{\text{key}}, S_{\text{fake}}) = f_\theta(\tilde{\mathbf{x}}). \tag{9}$$

Here $\oslash$ denotes elementwise division, with near-zero scales replaced by one. The MLP f_θ has parameters θ and sigmoid outputs: S_{rel} scores face consistency, S_{key} scores correct-key release

association, and S_{fake} scores re-rendering. The first two guide confirmation; the third supports detection comparisons.

Training objectives. Training keeps enrollment and Φ fixed. Benign, swapped, and wrong-key queries have binary targets $(y_{\text{rel}}, y_{\text{key}}, y_{\text{fake}})$ of $(1,1,0)$, $(0,0,1)$, and $(1,0,0)$, respectively. This supervision makes S_{key} a score for the correct-key release class, while G_s measures the retained signal. We train the three heads jointly:

$$\mathcal{L}(\theta) = \text{BCE}(S_{\text{fake}}, y_{\text{fake}}) + \lambda_{\text{rel}}\text{BCE}(S_{\text{rel}}, y_{\text{rel}}) + \lambda_{\text{key}}\text{BCE}(S_{\text{key}}, y_{\text{key}}) + \lambda_{\text{agr}}\mathbb{E}_{y_{\text{rel}}=1}\|z_{\text{rel}} - z_{\text{key}}\|_2^2. \quad (10)$$

Here BCE denotes binary cross-entropy, and the λ coefficients weight the objectives. The agreement term penalizes the distance between auxiliary projections $z_{\text{rel}}, z_{\text{key}} \in \mathbb{R}^8$ on $y_{\text{rel}} = 1$ queries; all three classification heads use the shared hidden features. App. D.1 details the architecture and evaluates how the agreement penalty affects verification by comparing training with and without this term.

Independent image-based review. Image-based review adds manipulation cues from the query pixels to the record-based assessment. A detector g_ψ, trained separately by cross-entropy, maps the resized RGB query $\mathbf{J}_{224} \in [0,1]^{224\times224\times3}$ to a softmax score $S_{\text{df}} = g_\psi(\mathbf{J}_{224})$. A vision-language model such as M2F2-Det (Guo et al., 2025) could provide diagnostic reports for reviewers to interpret suspected manipulation and guide further assessment, subject to latency and computational budgets.

Decision policy. Eligibility determines whether verification can proceed. We set $\mathsf{eligible}(J) = 1$ when the record is resolved, face association and geometry are valid, and the channel is supported; the selected path then compares scores with its thresholds. Otherwise, $\mathsf{eligible}(J) = 0$ yields unavailable regardless of scores, indicating missing prerequisites rather than a failed verification criterion:

$$\mathsf{status}(J) = \begin{cases} \text{confirmed}, & \mathsf{eligible}(J)=1,\ C \geq \tau_C \ \wedge\ G_s^{(\text{release})} \geq \tau_G \\ \text{unconfirmed}, & \mathsf{eligible}(J)=1,\ C < \tau_C \ \vee\ G_s^{(\text{release})} < \tau_G \end{cases} \quad \text{(direct, Sec. 3.2)}$$

$$\mathsf{status}(J) = \begin{cases} \text{confirmed}, & \mathsf{eligible}(J)=1,\ S_{\text{rel}} \geq \tau_{\text{rel}} \ \wedge\ S_{\text{key}} \geq \tau_{\text{key}} \\ \text{unconfirmed}, & \mathsf{eligible}(J)=1,\ S_{\text{rel}} < \tau_{\text{rel}} \ \vee\ S_{\text{key}} < \tau_{\text{key}} \\ \text{unavailable}, & \mathsf{eligible}(J)=0. \end{cases} \quad \text{(learned, Sec. 3.3)} \quad (11)$$

Calibration sets the tolerance for normal variation using benign training queries. Direct thresholds τ_C, τ_G share a lower-tail quantile selected to keep joint calibration rejection at most 1%; using C_B requires calibration on that sub-band. Learned thresholds $\tau_{\text{rel}}, \tau_{\text{key}}$ use the first percentiles of target-dataset benign training scores with model weights fixed. These per-head cutoffs can yield joint rejection above 1%. App. H.2 compares both paths at a matched joint calibration budget.

For human assessment of eligible queries, image-only review flags $S_{\text{df}} > \tau_{\text{df}}$; with the record model, $\widehat{y}_{\text{review}} = \mathbf{1}[S_{\text{rel}} < \tau_{\text{rel}}] \vee \mathbf{1}[S_{\text{df}} > \tau_{\text{df}}]$ combines relation discrepancies and image-based concerns. Here τ_{df} is the calibrated image-only cutoff, $\mathbf{1}$ an indicator, and $\wedge/\vee$ denote AND/OR. Together, statuses and flags guide routine handling, review of discrepancies, and requests for the records or processing information needed to complete verification and inform publisher decisions.

4 Experiments

Datasets and enrollment. We enroll 7,366 real faces from FF++, DFDC, and Celeb-DF (Rössler et al., 2019; Dolhansky et al., 2020; Li et al., 2020), using native-resolution crops located with SCRFD (Guo et al., 2022) and cross-checked with YuNet or a Haar-cascade face detector. Each face receives a release and private record; verification reads that record without reloading the original.

Queries and splits. To isolate manipulation effects, each released image is paired with a SimSwap (Chen et al., 2020) face swap; both are verified against the same record, yielding 14,732 queries. Replacement face identities come from manipulated images in the same dataset and split.

Metrics and training settings. Face-swap detection AUC compares swaps with paired releases, excluding wrong-key queries; Table 1 also covers StarGAN, UniFace, InSwapper, and SD inpainting. Key-discrimination AUC compares correct and wrong keys, while FPR measures benign rejection. Pass/rejection rates describe the specified criterion, not necessarily joint confirmation. Transfer tests freeze FF++- or DFDC-trained weights; App. D gives preparation, calibration, and training details.

4.1 Enrollment: creating the release

We evaluate release distortion and sensitivity to face re-rendering on the same assets. The release must therefore preserve visual quality while providing evidence that tolerates benign processing and remains sensitive to changes in face content.

Table 1: **Frequency-restricted spectral agreement** (AUC ↑). Reading only $\kappa \leq 0.15$ (C_B) improves separation under benign processing, particularly blur, compared with the full ring (C). The FF++ comparison includes the band-diagnostic cohort and is exploratory (App. G.2).

	Benign processing (FF++)				Clean SimSwap			Other face manipulations (FF++)			
Direct score	JPEG	Blur	WebP	Pixel.	FF++	DFDC	CDF	StarGAN	UniFace	InSwap	SD-inp
Template $1 - C_B$	**0.999**	**0.936**	**0.995**	**1.000**	**1.000**	**0.992**	**0.991**	**1.000**	**1.000**	**1.000**	**1.000**
$1 - C$ (full ring)	0.994	0.597	0.979	0.996	**1.000**	0.978	0.984	**1.000**	**1.000**	**1.000**	**1.000**

Visual quality. Fig. 5 shows the fidelity of phase editing, blending, and uint8 encoding. On the same face, our release achieves 51.2 dB PSNR, compared with 29.9 for StegaStamp and 41.4 for KAD-Net. The private record makes the weak signal measurable against a known reference, supporting high-fidelity publication at the cost of retaining per-release evidence. Accordingly, the comparison reflects different payload and reference budgets. App. E reports fidelity across datasets and resolutions.

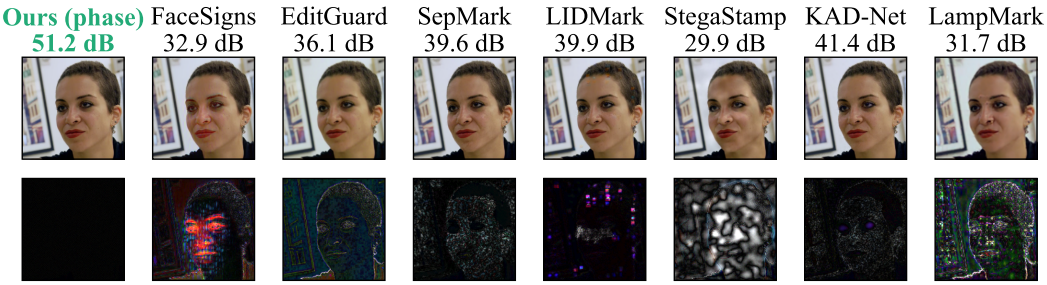

Figure 5: **Visual fidelity of released images.** Phase editing, blending, and uint8 encoding yield small release distortion. Top: the same FF++ face under each marking configuration, with PSNR relative to the original. Bottom: 10×-amplified residuals; payload and reference budgets differ across methods.

4.2 Face and Release Verification

Complementary relation and key evidence. Fig. 6 shows distinct responses to face re-rendering and projection-key changes. Evaluated re-renderings increase original-reference error $1 - H$, while key projection can remain high, particularly on Celeb-DF and DFDC. Changing only the projection key instead suppresses the key response and leaves H unchanged. The two scores thus provide separate evidence of face consistency and release-signal retention for the joint decision. Dataset-specific calibration accounts for variation in their distributions, as examined in App. H.3.

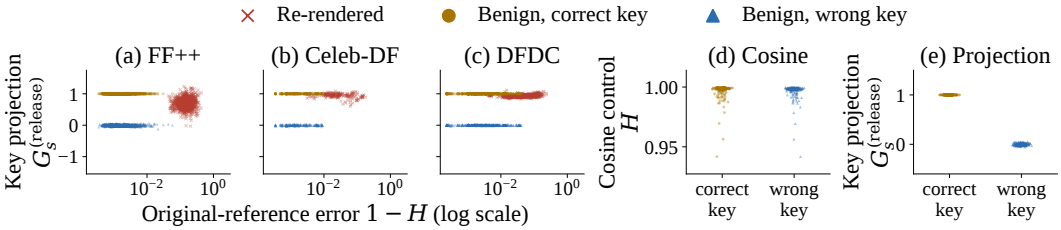

Figure 6: **Complementary evidence for face consistency and release-signal retention.** Left: re-rendering changes the face spectrum and raises original-reference error $1 - H$, while the keyed signal can persist. Right: a wrong projection key disrupts alignment with the release signal (e) without changing the face comparison (d). Joint verification therefore combines spectral agreement with release-normalized key projection to assess both aspects of the enrolled release.

Benign channels. Original-image amplitude weights emphasize strong enrolled coefficients, while C_B concentrates verification on the lower frequencies preserved by processing. Table 1 shows how this design improves face-swap separation. A fixed-band evaluation on additional source videos further supports its effectiveness under blur and reposting. Channel selection nevertheless requires reliable benign acceptance and manipulation rejection at the deployed threshold; App. G.2 examines these operating points and the retention of release-key evidence.

Comparison with peer methods. Each watermarking peer embeds its own mark into the same aligned source crops and undergoes SimSwap with the same replacement identities. Table 2 compares their verification signals. The methods differ in what each verifier reads and what its verdict establishes:

- **Manifest.** C2PA verifies signed assertions about an asset's provenance (Coalition for Content Provenance and Authenticity, 2026), but does not itself compare a circulating face with a publisher's private enrollment reference.
- **Embedded marks.** Robust message recovery (Tancik et al., 2020; Sander et al., 2025; Lu et al., 2025) can succeed after a face swap, leaving little evidence of replacement: WAM and VINE-R yield AUC 0.50. Fragile marks and decoder mismatch (Neekhara et al., 2024; Wu et al., 2023; He et al., 2026), or landmark displacement (Wu et al., 2026), are more directly sensitive to manipulation, but their effectiveness varies across datasets in Table 2. A recovered mark alone therefore cannot establish consistency with the enrolled face.
- **Image detectors.** Recent detectors learn manipulation cues through discrepancy learning and vision-language representations (Yang et al., 2025; Guo et al., 2025; Jung et al., 2026). Our ResNet-50 control reaches AUC 0.68 on Celeb-DF and 0.76 on DFDC with fixed FF++ weights, illustrating the importance of target-dataset evaluation. In our framework, image-only scores guide review alongside verification against the private release record.
- **Private face record.** Our verifier compares the query with the enrolled release and reads the release-specific keyed displacement under calibrated benign channels. The record-model row reports S_{fake}; release confirmation additionally requires both relation and key criteria to be satisfied.

Table 2: **Face-swap detection within and across datasets** (AUC ↑). Both record-based paths separate swaps across datasets. Learned scores reach near-perfect within-dataset AUC, while direct comparison transfers better to Celeb-DF. †: fixed public weights; —: no dataset-specific training.

			Per-dataset evaluation			From FF++		From DFDC	
Method	Detection score	Bits	FF++	CDF	DFDC	→CDF	→DFDC	→CDF	→FF++
StegaStamp	Robust bit error	100	0.91	0.57	0.52	0.59	0.54	0.56	0.80
FaceSigns†	Fragile bit error	128	**1.00**	0.70	0.70	—	—	—	—
SepMark†	Decoder mismatch	128	0.69	0.50	0.50	—	—	—	—
WAM†	Bit error	32	0.50	0.50	0.50	—	—	—	—
KAD-Net†	Fragile bit error	128	0.54	0.50	0.50	—	—	—	—
VINE-R†	Bit error	100	0.50	0.50	0.50	—	—	—	—
LIDMark	Landmark distance (px)	152	**1.00**	0.74	0.84	0.90	0.74	0.92	0.98
Ours	Image-only S_{df}	none	0.99	0.98	**1.00**	0.68	0.76	0.93	0.95
	Direct spectral $1 - C_B$	none	**1.00**	0.99	0.99	**0.99**	0.99	**0.99**	**1.00**
	Record model S_{fake}	none	**1.00**	**1.00**	**1.00**	0.98	**1.00**	0.97	**1.00**

Effect of key projection. With thresholds calibrated on each dataset's benign training releases, Table 3 shows that references admit most unreleased originals, whereas G_s rejects them while accepting JPEG releases. The key therefore helps distinguish the published version from its original.

Joint decision policy across datasets. At a matched 1% joint calibration budget, direct verification rejects more transferred swaps, and both paths reject all tested wrong keys. Under JPEG and reposting, the record model accepts more benign releases but also more swaps. Publishers can therefore start with direct verification and use the record model when reducing benign rejection is valuable and target-channel swap rejection remains adequate. App. H.2 quantifies this tradeoff.

Table 3: **Release evidence from the key.** Direct reference tests detect swaps but admit never-released originals under JPEG calibration. Key projection G_s distinguishes releases from originals even after JPEG processing.

	Clean swap AUC			Pass rate (%)	
Direct test	FF++	DFDC	CDF	Original pass ↓	Release pass ↑
Original H	1.000	0.977	0.978	100	99.25
Template C	1.000	0.978	0.984	100	99.25
SSIM	1.000	1.000	1.000	100	99.5
ArcFace	1.000	1.000	1.000	99.25	99.0
Key G_s	–	–	–	0	99.5

Verification cost and storage. Relation-only C_B takes 5.9 ms in memory; NPZ loading, full-ring features, and MLP take 149.9 ms. Separately, a 2.3 KiB sub-band payload preserves detection/key AUCs to three decimals on 150 FF++ assets, excluding keys and metadata. App. J specifies the timed paths and codec configurations.

4.3 Verification after Image Processing

To examine how delivery affects verification, we process 200 FF++ releases and their SimSwap, UniFace reenactment, and SD inpainting counterparts through each channel. Thresholds follow Sec. 3.2, using all 1,144 benign training releases. Providing the enrolled records isolates spectral robustness, while credential retention measures whether the record link survives.

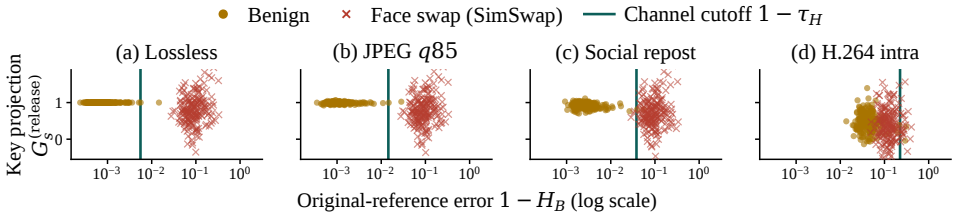

Figure 7: **Direct spectral evidence across channels.** Vertical lines mark $1 - \tau_H$ for channel-calibrated agreement H_B; release-normalized key projection is shown separately. Table 4 reports channel-calibrated pass/reject rates.

Fig. 7 and Table 4 show strong manipulation separation after lossless read-back, metadata stripping, screenshots, JPEG $q85$, and one repost. However, repeated reposting, intra-frame H.264, and cropping weaken swap rejection despite high benign pass rates; reenactment and SD inpainting remain more separable. A benign channel $\mathcal{P}$ can therefore erase evidence distinguishing the histories in Eq. (1). The supported set $\mathcal{C}$ must reflect verification performance, not merely benign processing: recalibration cannot restore lost separation. Clip averaging partly recovers swap rejection, as detailed in App. G.2; more stable features and geometry-aware comparison may further extend channel coverage.

Table 4: **Credential retention and direct spectral agreement across delivery channels.** Relation tests use H_B for stills, masked templates for crops, and mean C_B for 120 SimSwap clips. Channel cutoffs target 1% benign rejection. Credential retention requires an HMAC-authenticated asset ID.

		Channel-calibrated relation tests target benign rejection: 1%			
Channel	Credential retained	Face swap reject ↑	Reenactment reject ↑	SD inpaint. reject ↑	Benign pass ↑
Lossless re-encode	Retained	100%	100%	100%	99%
Credential stripping	Lost	100%	100%	100%	99%
Simulated screenshot (1.25×)	Lost	100%	100%	100%	99%
JPEG $q85$, credential kept	Retained	100%	100%	100%	99%
Social repost (0.75×, $q70$)	Lost	98.5%	100%	100%	99%
Repost of a repost	Lost	63%	100%	100%	99%
H.264 intra, per image	Lost	11%	99.5%	97.5%	99%
H.264 intra, clip mean (n=120)	Lost	65%	–	–	98.3%
Center crop 90%	Retained	6%	99%	81.5%	99%

Takeaway. Because processing affects relation and key evidence differently, channel selection requires evaluating both signals together. Joint tests in App. H.2 show that reposting increases swap acceptance for both paths. Publishers should therefore validate joint decisions on intended delivery channels and provide a trusted record lookup when credentials are lost.

5 Conclusion

Doctor Trigger enables publishers to assess circulating faces against specific enrolled releases. A private spectral reference and weak keyed signal provide complementary evidence of face consistency and release-signal retention while preserving visual fidelity. Experiments across three datasets support direct verification without a trained model under calibrated processing conditions, making enrollment a basis for subsequent verification.

More broadly, Doctor Trigger connects verification to publisher decision-making: confirmed queries support routine handling, unconfirmed queries identify discrepancies for review, and unavailable queries prompt additional information or human assistance. These outcomes distinguish record discrepancies from missing prerequisites to guide follow-up. Linking evidence to release records, the framework complements provenance credentials and image-only detection.

Ethics Statement

Publishers such as news organizations and public institutions could retain enrollment records, while publisher-operated verification services assess copies submitted by users or reviewers. Deployment requires enrollment consent, controlled access to private records, and retention and deletion policies. Verification results could support review of circulating images, but public reports should clearly communicate coverage and uncertainty: consistency with a released version does not establish the truth of the depicted event. Private spectral records contain sensitive face-derived information and require protection, as examined in App. K.4.

AI use statement

Generative AI tools assisted with editing and polishing the manuscript, drafting and revising portions of the text, identifying related work, checking bibliographic information, and developing mathematical proofs and their written presentation. AI assistance also supported code revision, execution of supplementary experiments, analysis of results, and consistency checks across the method description, reported measurements, and released code. These activities informed revisions to the interpretation and scope of empirical claims. The authors retain responsibility for the correctness, originality, and final presentation of all text, proofs, results, and code.

References

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

Inzamamul Alam, Md Tanvir Islam, Simon S. Woo, and Khan Muhammad. SpecGuard: Spectral projection-based advanced invisible watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17984–17993, 2025.

Haonan An, Xiaohui Ye, Guang Hua, Yihang Tao, Hangcheng Cao, Xiangyu Yu, and Yuguang Fang. RecoverMark: Robust watermarking for localization and recovery of manipulated faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8587–8597, 2026.

Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. Proactive image manipulation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15365–15374, 2022. doi: 10.1109/CVPR52688.2022.01495.

Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hyemin Boo, Eunsang Lee, and Jiyoung Lee. Referee: Reference-aware audiovisual deepfake detection. In *Interspeech 2026*, pp. 2434–2438, 2026. doi: 10.21437/Interspeech.2026-1246.

Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18629–18639, 2025.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2003–2011, 2020.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Coalition for Content Provenance and Authenticity. C2PA technical specification, version 2.4, 2026.

Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. ID-Reveal: Identity-aware DeepFake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15108–15117, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jessica Fridrich. Methods for tamper detection in digital images. In *ACM Workshop on Multimedia and Security*, pp. 19–23, 1999.

Enis Golaszewski, Neal Krawetz, Alan T. Sherman, Edward Zieglar, Sai K. Matukumalli, Roberto Yus, Carson L. Kegley, Michael Barthel, William Bowman, Bharg Barot, and Kaur Kullman. Verifying provenance of digital media: Security analysis of C2PA and its implementation. Cryptology ePrint Archive, Paper 2026/804, 2026.

Mathias Graf, Marco Willi, Melanie Mathys, Michael Aerni, Christian Schwarzer, Martin Melchior, and Michael H. Graber. DeepSignature: Digitally signed, content-encoding watermarks for robust and transparent image authentication. arXiv preprint arXiv:2604.23016, 2026.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations (ICLR)*, 2022.

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*, pp. 105–116, 2025.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Sijia He, Yunfeng Diao, Yongming Li, Chen Sun, Liejun Wang, and Zhiqing Guo. Kad-net: Kolmogorov-arnold and differential-aware networks for robust and sensitive proactive deepfake forensics. *Knowledge-Based Systems*, 331:114692, 2026. doi: 10.1016/j.knosys.2025.114692.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2137–2146, 2018.

Inho Jung, Hyeongjun Choi, Binh M. Le, Hohyun Na, and Simon S. Woo. A rich knowledge space for scalable deepfake detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Taewoong Kang, Sohyun Jeong, Hyojin Jang, and Jaegul Choo. Zero-shot head swapping in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10805–10814, 2025.

Zhimao Lai, Yun Zhang, Dong Li, Zhuangxi Yao, Chuntao Wang, and Chuan Qin. A new robust training-free proactive deepfake detection scheme using watermarking and identity-aware hashing. *Expert Systems with Applications*, 308:131126, 2026. doi: 10.1016/j.eswa.2026.131126.

Shulin Lan, Kanlin Liu, Yazhou Zhao, Chen Yang, Yingchao Wang, Xingshan Yao, and Liehuang Zhu. Facial features matter: A dynamic watermark based proactive deepfake detection approach. arXiv preprint arXiv:2411.14798, 2024.

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)*, 2020.

Ching-Yung Lin and Shih-Fu Chang. Semi-fragile watermarking for authenticating jpeg visual content. In *SPIE Security and Watermarking of Multimedia Contents II*, 2000.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Liyue Ming, Peisong He, Beibei Li, Haoliang Li, and Xinghao Jiang. Exploring transferable inconsistencies with regional guidance for reference-based deepfake detection. *Pattern Recognition*, 179:113795, 2026. doi: 10.1016/j.patcog.2026.113795.

Paarth Neekhara, Shehzeen Hussain, Xinqiao Zhang, Ke Huang, Julian McAuley, and Farinaz Koushanfar. FaceSigns: Semi-fragile watermarks for media authentication. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11):1–21, 2024. doi: 10.1145/3640466.

J. J. K. Ó Ruanaidh, W. J. Dowling, and F. M. Boland. Phase watermarking of digital images. In *Proceedings of the 3rd IEEE International Conference on Image Processing (ICIP)*, pp. 239–242, 1996. doi: 10.1109/ICIP.1996.560428.

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)*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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)*, 2019.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations (ICLR)*, 2025.

John C. Simmons and Joseph M. Winograd. Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography. arXiv:2405.12336, 2024.

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 (NeurIPS)*, 2024.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Runqi Wang, Yang Chen, Sijie Xu, Tianyao He, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. DynamicFace: High-quality and consistent face swapping for image and video using composable 3D facial priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13438–13447, 2025a.

Tianyi Wang, Mengxiao Huang, Harry Cheng, Xiao Zhang, and Zhiqi Shen. Lampmark: Proactive deepfake detection via training-free landmark perceptual watermarks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

Tianyi Wang, Harry Cheng, Ming-Hui Liu, and Mohan Kankanhalli. Fractalforensics: Proactive deepfake detection and localization via fractal watermarks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b. doi: 10.1145/3746027.3754544.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Junjiang Wu, Liejun Wang, and Zhiqing Guo. All in one: Unifying deepfake detection, tampering localization, and source tracing with a robust landmark-identity watermark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14106–14115, 2026.

Xiaoshuai Wu, Xin Liao, and Bo Ou. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *European Conference on Computer Vision (ECCV)*, 2022.

Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4662–4670, 2022. doi: 10.1609/aaai.v36i4.20391.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23850–23859, 2025.

Yuankun Yang, Chenyue Liang, Hongyu He, Xiaoyu Cao, and Neil Zhenqiang Gong. Faceguard: Proactive deepfake detection. arXiv preprint arXiv:2109.05673, 2021.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Lei Zhang, Zhiqing Guo, Dan Ma, Wenzhong Yang, and Gaobo Yang. Whether, which, and whose: Solving the triple challenge of deepfake proactive forensics in multi-face scenarios. arXiv preprint arXiv:2604.26342, 2026.

Yuan Zhao, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, and Xin Yu. Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4602–4611, 2023.

Haiyang Zheng, Nan Pu, Wenjing Li, Teng Long, Nicu Sebe, and Zhun Zhong. Open-world deepfake attribution via confidence-aware asymmetric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13378–13386, 2026.

Tong Zhou, Ruyi Ding, Gaowen Liu, Charles Fleming, Ramana Rao Kompella, Yunsi Fei, Xiaolin Xu, and Shaolei Ren. MetaSeal: Defending against image attribution forgery through content-dependent cryptographic watermarks. *Transactions on Machine Learning Research*, 2026.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14800–14809, 2021.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Fei-Fei Li. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*, 2018.

Appendix Overview

Background and Analysis
§ A Notation . . . 14
§ B Related Work . . . 17
§ C Analysis of Relation and Key Evidence . . . 18
Experimental Evidence
§ D Experimental Setup and Reproducibility . . . 20
§ E Release Fidelity and Visual Examples . . . 22
§ F Private References and Release-Key Evidence . . . 24
§ G Channel Robustness and Spectral Verification . . . 26
§ H Operating Points and Calibration . . . 28
§ I Sensitivity of the Release Mechanism . . . 30
Deployment and Failure Cases
§ J Deployment Interface and Cost . . . 31
§ K Verification Limits and Record Privacy . . . 33

A Notation

Symbols are defined where they enter the method or analysis and collected here for reference. Bold symbols denote images or feature vectors; calligraphic symbols denote sets, operators, hypotheses, or the phase template. Subscripts I and J identify original and query quantities. All angles are in radians.

Images and the record	
N_h, N_w	crop height and width in pixels; RGB supplies the third dimension
$\mathbf{I}$	aligned original crop, before publication (Sec. 2)
$\mathbf{I}'$	released crop, the original after the phase displacement
$\mathbf{J}$	query: the aligned crop received at verification
s	secret key of one enrolled face
$R, R_i, R_{a,i}$	R is the record for a single crop; R_i indexes a face and $R_{a,i}$ additionally indexes its asset a. The private record contains phase and amplitude arrays, support, key, strengths, and release projections (Sec. 3.1)
$\mathcal{T}_I,\ \mathcal{T}_I(k)$	stored phase template: luminance phase after the ideal edit, before release quantization
$A_I(k),\ \phi_I(k)$	original luminance amplitude and phase; $F_I(k) = A_I(k)e^{i\phi_I(k)}$
$\widehat{\phi}_I,\ \widehat{\phi}_I(k)$	original phase obtained from the record, stored explicitly in the evaluated implementation and recoverable in exact arithmetic by Eq. (4)
$\phi_J,\ \phi_J(k)$	query luminance phase, as an array and at one bin
$\mathbf{Y}_I, \mathbf{Y}_J$	original/query luminance, $0.299R + 0.587G + 0.114B$ for the respective RGB pixels
$F_I, \widetilde{F}$	original luminance spectrum and the three phase-edited RGB spectra, respectively
$\widetilde{\mathbf{I}}'$	blended release before clipping and uint8 truncation
c, I_c	color-channel index $c \in \{R, G, B\}$ and the corresponding original channel
$\mathcal{F}, \mathcal{F}^{-1}$	two-dimensional discrete Fourier transform and its inverse
P, U, V	publisher, requesting user, and publisher-side verifier in Fig. 1
The ring and the displacement	
$k, -k$	a Fourier-bin index and its conjugate index, modulo the grid dimensions
f_x, f_y	horizontal and vertical frequency in cycles per pixel
B	selected frequency sub-band; direct comparison uses $B = \{k : \kappa(k) \leq 0.15\}$
$\mathcal{K}_{\text{eff}}$	active bins, Eq. (2)

κ	normalized radius of a Fourier bin; $\kappa_{\min}, \kappa_{\max}$ bound the ring and $A_{\min}$, a 10^{-6} fraction of the largest amplitude, is the amplitude floor, together defining $\mathcal{K}_{\text{eff}}$
$u_s(k)$	standard-normal field over the Fourier grid, drawn from a generator seeded with the key s
$h_s(k), h$	the key's sign pattern $\text{sign}(u_s(k) - u_s(-k))$: ± 1 on every active bin, the same in every channel, with $h_s(-k) = -h_s(k)$
$\Delta\phi_{s,k}$	pre-blend phase challenge $\alpha\pi h_s(k)$, before the release blend
$\delta_s(k), \delta_i, \delta$	signed keyed displacement; $i = k$ and bare δ abbreviate the binwise quantity in the analysis/figures. Its effective value is given by Eq. (3)
α, β	dimensionless phase strength (pre-blend angle $\alpha\pi$) and blend weight; α_i, β_i are the values for face i
δ_{eff}	displacement magnitude on every active bin, determined by α and β (Sec. 3.1)
w_k	original-image amplitude weight on a bin, Eq. (6)
$r_k(\mathbf{J}), r_i, r$	wrapped phase drift of the query against $\widehat{\phi}_I$ (Sec. 3.2)
Readings from the record	
$H(\mathbf{J}; R)$	amplitude-weighted agreement with the original phase, Eq. (5)
H_B, C_B	the original/template readings restricted to sub-band B
$1 - H, 1 - H_B$	original-reference error over the full ring or sub-band; larger values indicate poorer agreement
$C(\mathbf{J}; R)$	template agreement: the same reading against $\mathcal{T}_I$ instead
$1 - C, 1 - C_B$	template mismatch over the full ring or sub-band; distinct from original-reference error
$G_s(\mathbf{J}; R)$	key projection: signed first-order agreement with the key, Eq. (6)
s^-, m, j	independently sampled wrong key, number of wrong keys, and their index $j = 1, \ldots, m$; j is a dummy bin index in the amplitude maximum; i, j in Fig. 11 instead index queries and records
$G_j^-, \overline{G^-}, G_{\max}^-$	the same projection under m independent wrong keys, their mean and maximum, Eq. (7)
G_Δ, G_z	key margin above the strongest sampled wrong key, and standardized separation from the wrong-key mean, Eq. (7)
ϵ_0	projection denominator stabilizer, 10^{-12}; distinct from the attack budget ϵ
$G_s^{(\text{release})}$	G_s divided by the release's own stored projection (Sec. 3.2)
$G_s^{(\text{template})}$	G_s divided by the template's projection instead, used by the reference controls
Scores, labels, and the policy	
$\Phi, \mathbf{x}$	fixed query-and-record feature map and its 75-dimensional output, Eq. (8)
$\mathbf{f}_0, \mathbf{f}_\ell, \ell$	11 full-ring features, eight features from radial band ℓ, and band index $\ell = 1, \ldots, 8$
$[\cdot;\cdot], \oslash$	vector concatenation and elementwise division
$\widetilde{\mathbf{x}}, \boldsymbol{\mu}, \boldsymbol{\sigma}$	standardized model input, training feature means, and training standard deviations with near-zero scales replaced by one
f_θ, θ	learned record MLP and its parameters; maps $\widetilde{\mathbf{x}}$ to three sigmoid scores
$\mathbf{J}_{224}, g_\psi, \psi$	resized RGB query in $[0,1]^{224\times224\times3}$, image-only classifier, and its parameters; g_ψ outputs S_{df}
$S_{\text{rel}}, S_{\text{key}}, S_{\text{fake}}$	sigmoid record-model scores for face consistency, a benign release paired with its correct key, and re-rendering; larger values favor these training targets (Sec. 3.3)
S_{df}	image-only manipulation score, from $\mathbf{J}$ alone; larger is more suspicious
$y_{\text{rel}}, y_{\text{key}}, y_{\text{fake}}$	binary training labels assigned by query construction, with 1 indicating the named target (App. D.1)

$\lambda_{\text{rel}}, \lambda_{\text{key}}, \lambda_{\text{agr}}, \mathcal{L}$	weights of relation BCE, key BCE, and representation agreement, and the record-model training objective $\mathcal{L}(\theta)$, Eq. (10)
$z_{\text{rel}}, z_{\text{key}}$	eight-dimensional representations from linear maps of the hidden layer, pulled together for training rows with $y_{\text{rel}} = 1$, Eq. (10)
τ_H, τ_C	direct agreement cutoffs fitted separately for the selected reference, support, channel, and aggregation; relation-only tests use first-percentile benign-training calibration, while joint decisions use the budget in Sec. 3.3
τ_G	cutoff on release-normalized key projection in the direct joint policy
$\tau_{\text{rel}}, \tau_{\text{key}}, \tau_{\text{df}}$	record and review thresholds in the decision policy (Sec. 3.3)
τ	generic score cutoff; the theory applies it to H, while empirical tables specify the reading and calibration protocol
$\mathsf{eligible}(J), \mathsf{status}(J)$	query eligibility indicator and three-state verdict for its associated record and channel, Eq. (11)
$\widehat{y}_{\text{review}}$	binary review flag, separate from verification status; the learned-path form combines relation failure and image-only concern
Channels and hypotheses	
$\mathcal{C}$	the benign channels a deployment supports
$\mathcal{P}$	the processing channel taken by a query, $\mathcal{P} \in \mathcal{C}$ under $\mathcal{H}_c$
$\mathcal{R}, M$	the set of face re-renderings, and one re-rendering in that set
$\mathcal{H}_c, \mathcal{H}_d$	benign-channel and re-rendering-then-channel hypotheses, Eq. (1)
Local noise and concentration analysis	
i, q_i, a_i	active-bin index $i = k$, $q_i = \cos r_i$, and normalized weight $a_i = w_i / \sum_j w_j$; this i is not a face index
b, B_b, a_b, Q_b	block index, its bin set, total weight $\sum_{i \in B_b} a_i$, and weighted mean cosine within the block
n_w, n_{eff}	effective sample sizes for independent weighted bins and independent weighted blocks, respectively
μ_c, b_a, μ_d	nominal benign mean, nonnegative allowance for benign mean reduction, and upper bound on the re-rendered mean; b_a is a scalar allowance, not a block weight
t	positive score deviation in the concentration proof; in Fig. 12(a), the target benign rejection rate
ξ_i	residual channel noise at bin i in the local model $r_i = \delta_i + \xi_i$
ρ	attenuation factor $\sum_i w_i \delta_i^2 / (\sum_i w_i \delta_i^2 + \epsilon_0)$ in the projection expansion
Operators and mathematical conventions	
$\mathbb{R}, i$	real numbers and the imaginary unit in complex exponentials; elsewhere i is an explicitly scoped index
$\arg, \text{wrap}, \text{sign}$	principal complex argument, wrapping to $[-\pi, \pi)$, and sign in $\{-1, 0, 1\}$
$\mathbb{E}, \Pr, \text{std}_j$	expectation, probability, and standard deviation across wrong keys with divisor m
$\text{BCE}, \lVert \cdot \rVert_2, O(\cdot)$	binary cross-entropy, Euclidean norm, and asymptotic order of a remainder
$\mathbf{1}[\cdot], \wedge, \vee$	indicator (one when its condition holds), logical AND, and logical OR
$\cap, \cup, \in$	set intersection, union, and membership
Evaluation and figure conventions	
n	number of observations for the stated experiment; captions specify the counting unit
q, p	JPEG quality setting and augmentation application probability; q_i in the theory instead denotes a cosine reading
ϵ, L_∞	attack budget $(8/255)$ and maximum absolute pixel-change norm on images scaled to $[0, 1]$
$\uparrow, \downarrow, \pm$	larger/smaller is preferable; mean plus/minus standard deviation where specified by the caption

PSNR, SSIM	peak signal-to-noise ratio in decibels and structural similarity, relative to the original/reference image
ROC, AUC	receiver operating characteristic and area under that curve
TPR, FPR, FRR	true-positive rate, false-positive rate, and benign false-rejection rate
JPEG $q90$, blur $k7$	JPEG quality 90 and Gaussian blur with a 7×7 kernel; $k7$ is a channel label, not Fourier-bin index k
FF++, DFDC, CDF	FaceForensics++, Deepfake Detection Challenge, and Celeb-DF v2; CelebDF is also used as a dataset label
CRF, QP	constant-rate-factor setting and quantization parameter in video-codec conditions

B Related Work

Provenance credentials. C2PA Content Credentials bind signed assertions about an asset's issuer and history to its content (Coalition for Content Provenance and Authenticity, 2026). Their validation depends on certificate handling, timestamps, and validator behavior, as studied for C2PA v2.2 (Golaszewski et al., 2026). Content-carried signals can reconnect circulating copies with provenance information after metadata loss (Simmons & Winograd, 2024). Doctor Trigger complements this infrastructure with a private reference for face-level comparison against an enrolled release.

Reference-based face verification. Identity embeddings such as ArcFace assess whether two faces depict the same person (Deng et al., 2019). Reference-based forensics also models person-specific behavior: ID-Reveal learns temporal facial signatures (Cozzolino et al., 2021), while Referee uses a reference video as an audiovisual identity anchor (Boo et al., 2026). More recently, regional identity and texture comparisons address identity-preserving generative manipulations (Ming et al., 2026). These approaches establish an important reference-based alternative to query-only detection. Doctor Trigger instead associates the reference with a particular released asset and combines face consistency with evidence of its keyed release signal.

Content authentication. Fragile watermarking detects content changes, while semi-fragile designs tolerate specified processing such as JPEG compression (Fridrich, 1999; Lin & Chang, 2000). Recent methods couple authentication evidence to image content: MetaSeal uses content-dependent cryptographic watermarks for attribution (Zhou et al., 2026), and DeepSignature combines signed content-encoding watermarks with latent-space tamper verification (Graf et al., 2026). Thus, content-dependent verification is an established direction. Our formulation centers on publisher-held spectral records and separate face-relation and release-key evidence under calibrated processing; the reference controls assess what the key adds beyond image and identity similarity.

Image-only face analysis. Image-only detectors identify manipulation traces through spatial and frequency cues (Afchar et al., 2018; Qian et al., 2020; Luo et al., 2021), temporal consistency, and cross-modal evidence (Haliassos et al., 2021; Zhou & Lim, 2021). Generalization approaches include diffusion-guided disentanglement (Sun et al., 2024), multi-generator discrepancy learning (Yang et al., 2025), and multimodal pretraining (Jung et al., 2026). Related attribution methods identify source generators under closed- or open-world assumptions (Yang et al., 2022; Zheng et al., 2026). In a publisher workflow, these query-only signals can support review alongside the release-specific evidence supplied by a publisher-held record.

Watermarking and provenance. Classical and neural watermarking encode ownership or provenance signals before distribution (Cox et al., 1997; Zhu et al., 2018; Tancik et al., 2020; Fernandez et al., 2023). Robust schemes preserve recoverable evidence through editing (Wen et al., 2023; Lu et al., 2025; Sander et al., 2025; Bui et al., 2025). SpecGuard embeds messages through spectral projections in learned features (Alam et al., 2025). Artificial fingerprinting embeds identifying signals in generator training data for later attribution (Yu et al., 2021). Our spectral design relates to phase-domain watermarking (Ó Ruanaidh et al., 1996) and proactive release templates (Asnani et al., 2022). Tree-Ring (Wen et al., 2023) embeds a Fourier-ring pattern in diffusion latents during generation; by comparison, Doctor Trigger edits an existing image's phase and retains a private record for later comparison.

Proactive deepfake defenses. FaceSigns (Neekhara et al., 2024) and SepMark (Wu et al., 2023) use semi-fragile decoder or tracer outputs. Identity-WM (Zhao et al., 2023), LampMark (Wang et al., 2024), and LIDMark (Wu et al., 2026) connect embedded signals to identity or landmarks. FaceGuard detects manipulation through disagreement between extracted and expected watermarks (Yang et al., 2021). MaLP (Asnani et al., 2023) and FractalForensics (Wang et al., 2025b) localize manipulated regions, while multi-face proactive forensics identifies affected faces and identities (Zhang et al., 2026). FaceProtect derives dynamic watermarks from facial features (Lan et al., 2024). RecoverMark (An et al., 2026) embeds recoverable face information in surrounding regions; TFPDD (Lai et al., 2026) embeds an identity-aware hash without training. Within this family, DOCTOR TRIGGER reads face-relation agreement and release-key retention separately from a private spectral record. Its analytic readings require no fitted model weights, and a learned record model combines their features for joint decisions.

C ANALYSIS OF RELATION AND KEY EVIDENCE

Spectral agreement measures preservation of the enrolled phase relation, while signed projection measures alignment with the release signal. We analyze these complementary roles under explicit noise and dependence assumptions, then compare their scales with the observed wrong-key distribution. Thresholds in the experiments are calibrated empirically.

C.1 ORIGINAL-REFERENCE AGREEMENT CONCENTRATION

For $i = k \in \mathcal{K}_{\text{eff}}$, let $q_i = \cos r_i(\mathbf{J})$ and normalize the amplitude weights as $a_i = w_i / \sum_j w_j$, so that $H = \sum_i a_i q_i$. If the weighted coefficients were independent, the natural effective sample size would be

$$n_w = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} = \frac{1}{\sum_i a_i^2}. \tag{12}$$

To include exact dependence within conjugate pairs or larger groups, partition the active bins into blocks B_b. Let $a_b = \sum_{i \in B_b} a_i$ and $Q_b = a_b^{-1} \sum_{i \in B_b} a_i q_i$. Then $H = \sum_b a_b Q_b$, $Q_b \in [-1, 1]$, and $n_{\text{eff}} = 1/\sum_b a_b^2 \le n_w$.

Proposition 1 (Separation under independent spectral blocks)**.** *Let μ_c be a nominal benign-score mean, $b_a \ge 0$ an allowance for a reduction in that mean (e.g., alignment error), and μ_d an upper bound on the re-rendered mean. Let τ be the acceptance cutoff on H. Assume the block variables Q_b are independent under each hypothesis. If $\mathbb{E}[H \mid \mathcal{H}_c] \ge \mu_c - b_a$, $\mathbb{E}[H \mid \mathcal{H}_d] \le \mu_d$, and $\mu_d < \tau < \mu_c - b_a$, then*

$$\Pr(H < \tau \mid \mathcal{H}_c) \le \exp[-(\mu_c - b_a - \tau)^2 n_{\text{eff}}/2], \tag{13}$$

$$\Pr(H \ge \tau \mid \mathcal{H}_d) \le \exp[-(\tau - \mu_d)^2 n_{\text{eff}}/2]. \tag{14}$$

Proof. Each weighted block $a_b Q_b$ has range length $2a_b$. For any deviation $t > 0$, the one-sided bounded-sum inequality gives $\Pr(H - \mathbb{E}H \ge t) \le \exp[-t^2/(2\sum_b a_b^2)]$, with the same bound for the lower tail. Substituting the two mean-to-threshold gaps yields the result. □

This model explains how weighting and dependence change the available evidence. However, it does not establish independent blocks for real codec residuals or estimate n_{eff} for the evaluated images. We therefore calibrate thresholds empirically and do not use this bound to claim a numerical false-accept rate. Ring cardinality is a support size, not a measure of cryptographic key entropy.

C.2 SIGNED RELEASE-KEY PROJECTION

For an ideal unquantized relation-preserving query, write

$$r_i(\mathbf{J}) = \delta_i + \xi_i,$$

where all sums in this local model run over active bins. Here $i = k$ is a frequency index (not a face-slot index), $\delta_i = \delta_s(k)$ is the enrolled keyed displacement, $w_i = w_k$, and ξ_i is weakly dependent residual noise with approximately zero mean and uncorrelated with the enrolled displacement. This

local model describes the continuous phase response. Since pixel truncation adds signal-dependent changes, verification uses the encoded release's measured projection for normalization. Substituting into Eq. (6) gives

$$G_s(\mathbf{J}; R) = \frac{\sum_i w_i \delta_i \sin(\delta_i + \xi_i)}{\sum_i w_i \delta_i^2 + \epsilon_0}. \tag{15}$$

For small $|\delta_i + \xi_i|$,

$$G_s(\mathbf{J}; R) = \rho + \frac{\sum_i w_i \delta_i \xi_i}{\sum_i w_i \delta_i^2 + \epsilon_0} + O\left(\frac{\sum_i w_i |\delta_i| \, |\delta_i + \xi_i|^3}{\sum_i w_i \delta_i^2 + \epsilon_0} \right), \tag{16}$$

where

$$\rho = \frac{\sum_i w_i \delta_i^2}{\sum_i w_i \delta_i^2 + \epsilon_0} \approx 1. \tag{17}$$

Here ρ is the attenuation caused by the denominator stabilizer, and $O(\cdot)$ denotes a remainder bounded in order by its argument as the residuals approach zero. Thus the expected correct-key projection stays near its enrolled value under the local noise model; concentration additionally requires control of the weighted noise variance. For an independently sampled wrong key s^-, the signed displacement directions are uncorrelated, so their contributions cancel in expectation:

$$\mathbb{E}[G_{s^-}(\mathbf{J}; R)] \approx 0, \qquad \mathbb{E}[G_s(\mathbf{J}; R)] \approx \rho. \tag{18}$$

The paired correct-versus-wrong-key experiment tests exactly this distinction by holding the image fixed and changing only the projection key.

C.3 Why Ordinary Cosine Is Weakly Key-Sensitive

Using the same residual model, now assume channel noise is independent of the key direction. Comparing the displaced residual with its undisplaced counterpart gives

$$\cos(\xi_i + \delta_i) - \cos \xi_i = -\delta_i \sin \xi_i - \frac{1}{2} \delta_i^2 \cos \xi_i + O(|\delta_i|^3). \tag{19}$$

If the residual noise is approximately symmetric about zero, then $\mathbb{E}[\sin \xi_i] \approx 0$, so the first-order contribution cancels in expectation:

$$\mathbb{E}[\cos(\xi_i + \delta_i) - \cos \xi_i] = -\frac{1}{2} \delta_i^2 \, \mathbb{E}[\cos \xi_i] + O(|\delta_i|^3). \tag{20}$$

For small residual noise, $\mathbb{E}[\cos \xi_i] \approx 1$, leaving a key-dependent cosine change of order $\delta_i^2/2$. Ordinary cosine therefore loses the sign of the weak displacement near alignment, whereas the signed key projection multiplies by the known displacement direction before aggregation and retains first-order information.

At the deployed settings, $|\delta_{\text{eff}}|$ is approximately 0.0235 rad on FF++/Celeb-DF and 0.0071 rad on DFDC, giving second-order cosine terms $\delta_{\text{eff}}^2/2$ of 2.76×10^{-4} and 2.52×10^{-5}, respectively. This explains the weak local cosine response to the release signal. In the paired wrong-key control, all inputs to H and C stay fixed, so their invariance is exact rather than an approximation at small displacement.

C.4 Wrong-Key Null

Equation (7) uses $m{=}8$ sampled wrong keys as an empirical null rather than as a security bound. An independent key produces a signed displacement pattern uncorrelated with the enrolled one, so its projection averages around zero over the active support. The median active-bin counts are $19{,}824$, $32{,}461$, and $157{,}684$ on FF++, Celeb-DF, and DFDC, respectively.

The empirical wrong-key mean is within $2{\times}10^{-4}$ of zero on each dataset; median per-asset spreads are 0.017, 0.012, and 0.006, respectively. The enrolled lossless query has release-normalized projection 1 by construction. These sampled distributions do not resolve the far-tail false-accept probability.

D EXPERIMENTAL SETUP AND REPRODUCIBILITY

This section specifies enrollment, model training, query generation, and score conventions used by the supplementary experiments.

Enrollment cohort and crop geometry. From 24,527 source frames across FF++, DFDC, and CELEB-DF (Rössler et al., 2019; Dolhansky et al., 2020; Li et al., 2020), faces are located with SCRFD (Guo et al., 2022) and accepted only when YuNet or a Haar-cascade face detector also locates an overlapping face box (24,209 accepted). We center native-resolution crops on the detected face, using width and height ratios 0.224 and 0.334 relative to the source frame, and save lossless PNGs without padding or resizing. We partition CELEB-DF by source video to prevent video-level overlap between splits. Each of the 7,366 enrolled real faces receives its own released image and a private NPZ record storing the enrolled phase, original phase and amplitude, keyed displacement, active bins, and the release's own key projection. Record-based verification reads the query and this stored record; it does not reload the original image.

D.1 HYPERPARAMETERS AND TRAINING SETUP

Enrollment and key settings. Enrollment writes the keyed displacement at $\alpha = 0.05$, so the pre-blend shift is 0.05π rad, with blend weight $\beta = 0.15$ on the ring $\kappa \in [0.10, 0.45]$ of normalized frequency ($\alpha = 0.015$ for the DFDC release configuration). Every face passed at the first rung, so these are also the per-face (α_i, β_i) stored in the record, where i indexes the enrolled face. The key s is the first 64 bits of SHA-256 over a fixed tag, a master seed, and the sample identifier; it seeds NumPy's PCG64 generator, whose standard-normal draws $u_s(k)$ over the Fourier grid give $h = \mathrm{sign}(u_s(k) - u_s(-k))$. Bins with $k = -k$ get $h = 0$ and lie outside $\mathcal{K}_{\mathrm{eff}}$. The amplitude floor is $A_{\min} = 10^{-6} \max_k A_I(k)$ on the luminance spectrum of the aligned face; it removes at most 2% of a face's ring bins (median below 0.05%). The wrong-key comparison draws m=8 keys; feature extraction uses the full ring and 8 radial bands. The MLP bands use normalized Euclidean distance from the center of the shifted array, divided into eight intervals over the active support. For rectangular crops this differs from the cycles-per-pixel radius κ used for enrollment and direct sub-band selection.

Record model and features. The feature map Φ in Eq. (8) supplies the record model with 75 inputs: 11 global readings and 8 readings in each of the 8 radial bands ($11 + 8 \times 8 = 75$). All key projections in this feature list use the release normalization of Sec. 3.2; bare r and δ denote the binwise residual r_k and displacement $\delta_s(k)$. The full-ring block $\mathbf{f}_0$ contains H, C, the amplitude-weighted mean and unweighted standard deviation of the wrapped residual r, the weighted mean of $\cos r$, G_s, $\overline{G^-}$, $G^-_{\max}$, G_Δ, G_z, and the weighted mean of $\delta \sin r$. Each block $\mathbf{f}_\ell$ has H, C, G_s, a band key-distance column, the weighted residual mean and unweighted standard deviation, and the weighted means of $\cos r$ and $\delta \sin r$. Because the mean of $\cos r$ equals H and the band key-distance column equals that band's G_s, only 58 of the 75 inputs are distinct. Each input is standardized with its mean and standard deviation over the training split of the training dataset (benign, swap, and wrong-key rows), giving $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ in Eq. (9). Standard deviations below 10^{-6} are replaced by one, and missing feature values are imputed with the corresponding training mean. These statistics and the trained model are frozen during test scoring. The default $\Phi(\mathbf{J}, R)$ uses the enrolled key; wrong-key training rows substitute the candidate projection direction while keeping the query, original phase, and template fixed.

Training targets. The record model learns from benign, re-rendered, and wrong-key queries, with the targets in Eq. (10) assigned according to query construction. This supervision maps spectral features to relation, key, and manipulation scores for the decision policy. The outputs serve as decision scores rather than calibrated posterior probabilities.

Architecture and optimization. The MLP has two 64-unit hidden layers with ReLU activations, three scalar output heads, and two eight-dimensional linear projections for the agreement loss. Sigmoid converts each output logit to its reported score; training evaluates BCE directly on logits for numerical stability. The record model is trained for 1,000 AdamW steps (learning rate 10^{-3}, weight decay 10^{-4}) on batches of 1,024 rows drawn with replacement, with $\lambda_{\mathrm{rel}} = 1$, $\lambda_{\mathrm{key}} = 0.3$, $\lambda_{\mathrm{agr}} = 0.1$, and three random seeds, using the objective in Eq. (10).

Agreement-penalty ablation. To assess the auxiliary penalty, we train with $\lambda_{\text{agr}} = 0$ and 0.1 using matched seeds and FF++ training data. Face-swap AUCs are $0.99972/0.99970$ on FF++, $0.99618/0.99664$ on DFDC, and $0.97385/0.97535$ on Celeb-DF. Removing the penalty changes joint swap acceptance by $0/-0.105/+0.159$ percentage points under target-dataset first-percentile calibration. The penalty has a small, setting-dependent effect. The projections are auxiliary: equal outputs, including zero, minimize their distance independently of the classification scores.

Image-only control and calibration. The image-only control uses ResNet-50 at 224^2. Within-dataset evaluations train on the corresponding dataset; transfer evaluations retain FF++-trained weights. Training uses the main training splits with the following augmentation: JPEG quality $q = 35$–80 with application probability $p = 0.4$, Gaussian blur with kernel width 3–7 pixels at $p = 0.2$, Gaussian noise, and HSV jitter, all applied at native resolution before resizing. The joint-policy thresholds τ_{rel} and τ_{key} use the first percentiles of clean benign training scores, as detailed in App. H.2. Channel experiments separately calibrate the direct relation and review cutoffs on transformed benign training queries. The threshold-transfer diagnostic in App. H.3 fits quantiles to test scores separately from the main evaluation thresholds.

D.2 Evaluation Protocol

Re-rendering engines. Every engine operates on the released image and is scored against that image's own record, and all use released weights without retraining. SimSwap (Chen et al., 2020) supplies the main re-rendered query; the four re-rendering columns add StarGAN-v2 (Choi et al., 2020) (authors' CelebA-HQ EMA checkpoint, 100k steps), UniFace (Xu et al., 2022) (released 500k-step swap checkpoint), InsightFace `inswapper_128` (released ONNX), and Stable Diffusion v1.5 (Rombach et al., 2022) inpainting over the detected face box (25 steps, guidance 7.5, fixed seed, no identity donor, so the face is re-rendered by the diffusion prior alone). The inpainting mask uses a face box detected on the released image with InsightFace's `buffalo_l` detector. StarGAN-v2 and SD inpainting also cover all 636 DFDC and 209 Celeb-DF test releases with unchanged generation settings; Table 12 reports their direct verification results. The feature table names G_Δ `G_margin`. For Table 4, UniFace reenactment uses the separate released 1M-step checkpoint: the enrolled release supplies appearance, while the existing paired attack-source image drives pose and expression. Inputs are resized to 256×256. The same 200 enrollments, delivery channels, and benign-calibrated cutoffs are used for SimSwap, reenactment, and SD inpainting; only SimSwap has clip-level evaluation.

Peer training settings. FaceSigns, SepMark, KAD-Net, WAM, and VINE-R use one fixed public encoder–decoder checkpoint per method across all datasets in Table 2. Their first three columns identify test datasets, not training sources. StegaStamp, LIDMark, and our learned controls use dataset-specific weights; transfer fixes source-trained weights for target-dataset evaluation, including the watermark encoder and decoder. Direct template comparison has no trained weights, so its transfer columns repeat the corresponding target-dataset evaluation. Record-model AUCs average three seeds and use only benign and swapped queries; wrong-key rows belong to the separate key-discrimination tests.

Peer-score conventions. Peer scores have different units: message bit-error rates for FaceSigns, StegaStamp, and KAD-Net; decoder disagreement for SepMark; landmark distance in pixels for LIDMark; and cosine mismatch for our direct test. We therefore compare score order through ROC-AUC, which is invariant to strictly increasing rescaling. Reported signals are oriented toward larger values for suspicious queries, and their means were checked for each method–engine pair. KAD-Net uses its fragile decoder's bit-error rate, whereas SepMark uses its dual-decoder disagreement.

Under StarGAN-v2, LIDMark's mean decoded landmark distance reaches 2.4×10^{14} pixels, compared with 13.1 on benign queries. Its AUC of 1.000 reflects separation driven by extreme decoder outputs, not accurate or numerically stable landmark recovery.

Credential retention. C2PA integration is proposed. The credential-retention experiment uses HMAC-SHA256-authenticated JSON in PNG text or JPEG comment fields to isolate metadata survival; it does not implement a C2PA manifest or validator. When this metadata is stripped, comparison remains possible only if another trusted route resolves the corresponding record.

E RELEASE FIDELITY AND VISUAL EXAMPLES

Table 6 compares release distortion on a shared external pool at three delivery resolutions. Methods use their benchmark configurations with different payload and reference budgets. The following examples show release edits, benign-channel queries, and re-renderings.

Across all 7,366 enrollments, DOCTOR TRIGGER records an average PSNR of 51.20 dB (51.22/51.15/51.21 on FF++/DFDC/CELEB-DF, respectively), with a minimum of 49.3 dB and SSIM rounding to 1.00 at two decimal places. These values are measured from the released PNG after writing to and reading back from disk.

Table 6: **Release fidelity across datasets and resolutions.** PSNR ↑ / SSIM ↑ on 40 held-out faces per dataset, resized to 128/256/512 pixels. All methods use the same crops and are measured against the corresponding original; payload sizes differ across methods.

	PSNR (dB) ↑ / SSIM ↑								
	FF++			CELEB-DF			DFDC		
Method	128	256	512	128	256	512	128	256	512
StegaStamp	29.2/0.98	28.8/0.98	28.8/0.98	29.2/0.98	28.9/0.97	28.7/0.97	29.9/0.98	29.4/0.98	29.5/0.98
FaceSigns	36.2/1.00	34.1/1.00	35.2/1.00	35.3/1.00	34.0/1.00	34.7/1.00	34.6/0.99	33.3/0.99	33.9/0.99
SepMark	41.4/1.00	39.7/1.00	40.2/1.00	41.1/1.00	39.7/1.00	40.1/1.00	41.1/1.00	39.4/1.00	39.9/1.00
EditGuard	41.5/1.00	38.4/1.00	37.0/1.00	38.9/1.00	38.1/1.00	35.5/1.00	40.2/1.00	37.4/1.00	36.3/1.00
LampMark	35.0/1.00	32.6/0.99	33.7/1.00	35.4/1.00	34.5/1.00	34.9/1.00	36.4/1.00	33.5/0.99	34.6/0.99
KAD-Net	43.7/1.00	41.1/1.00	41.9/1.00	43.3/1.00	40.9/1.00	41.5/1.00	43.4/1.00	40.8/1.00	41.6/1.00
LIDMark	42.6/1.00	41.3/1.00	41.8/1.00	36.6/1.00	35.2/1.00	35.7/1.00	42.7/1.00	41.2/1.00	41.8/1.00
Ours	51.2/1.00	51.2/1.00	51.4/1.00	51.3/1.00	51.6/1.00	51.2/1.00	51.1/1.00	51.2/1.00	51.2/1.00

Visual comparisons. Fig. 8 makes the weak release signal visible through amplified residuals. Fig. 9 then shows how spectral agreement responds to processing and face replacement across datasets.



Figure 8: **Release residuals.** Differences from the original images are scaled separately for visibility. The displayed edits have maximum magnitude $1/255$ and form a spatially diffuse pattern.

Key pattern and residual. Fig. 10 makes the key–template relation visible on one enrolled asset. The key is a seed that expands into a sign pattern over the ring (a); (α, β) only set its size, producing the keyed displacement δ_s (b), and the template is the face's own phase carrying that displacement (c). A benign query's residual against the original phase looks like noise (d), but its *signs* echo the key's pattern: 82% of active bins agree with the correct key's signs ($G_s = 1.08$, e) against 49%, chance, for a wrong key ($G_s = 0.01$, f). Here G_s is the unnormalized projection of Eq. (6); panel (f)

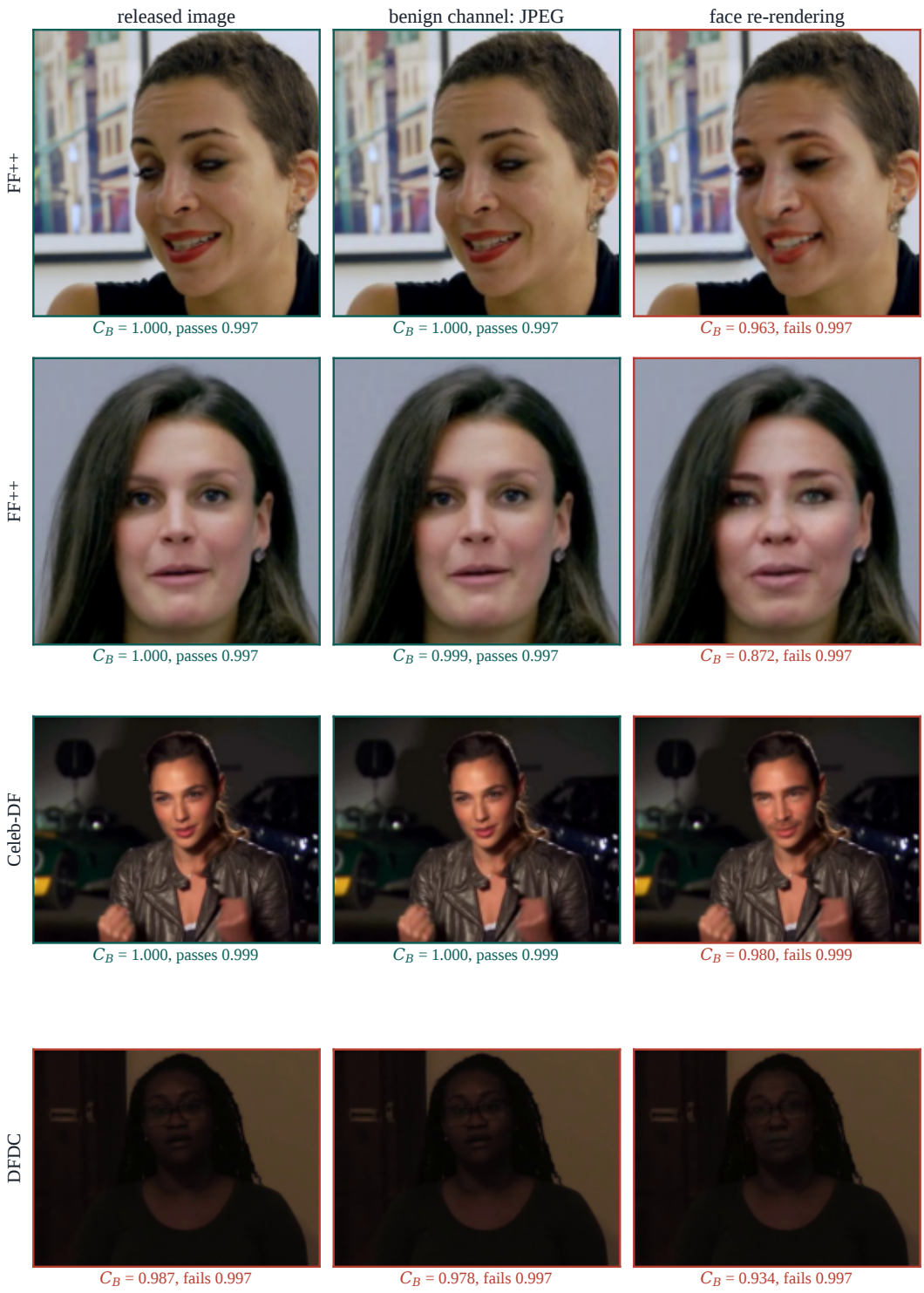

Figure 9: **Verification examples across datasets.** Released images and JPEG-$q85$ copies retain higher sub-band template agreement C_B than SimSwap queries in these examples. Cutoffs use 1% benign training rejection per dataset; green and red borders indicate relation acceptance and rejection.

abbreviates G_{s^-} as G for the wrong key s^-. The panel symbols δ and r denote $\delta_s(k)$ and $r_k(\mathbf{J})$ on the displayed bins. The signed projection measures this alignment with the keyed displacement.

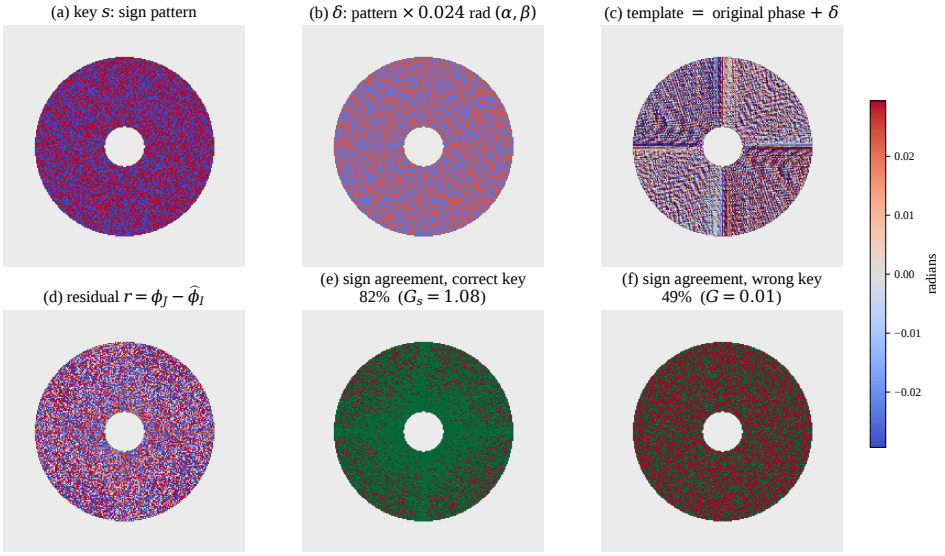

Figure 10: **Keyed displacement and query residual.** One FF++ test asset, showing the central region of the shifted spectrum; gray marks bins outside the active support. Top: the key pattern, displacement, and enrolled template. Bottom: the query residual and its alignment with correct and wrong keys.

F Private References and Release-Key Evidence

Because the direct relation test uses a private per-asset reference, the appropriate controls must receive the same privilege. The original-phase control reads the undisplaced reference but uses the same keyed query images; it does not evaluate an independently generated unmarked-release cohort. Throughout this section, the key projection is $G_s^{(\text{template})}$ from Sec. 3.2, normalized by the stored template's own projection rather than by the release projection, so released images retain their natural score spread.

We compare four reference-based readings on the main test queries: template agreement C, original-phase agreement H, SSIM against the stored released image, and ArcFace (Deng et al., 2019) similarity to the enrolled face. Swap-detection AUC uses mismatch as the positive direction; release-discrimination AUC instead treats the enrolled release as positive.

Table 7: **Private-reference comparisons.** Main-test AUC (↑) for clean face-swap detection (left), discrimination between releases of the same person (middle), and discrimination between the release and its original image before publication (right).

	Swap separation (clean) ↑			Same-person release ↑	Release-version control ↑
Verifier	FF++	DFDC	CDF	Enrolled release vs. another release	Released image vs. unreleased original
Template agreement C	1.000	0.978	0.984	–	–
Original-phase reference	1.000	0.977	0.978	1.000	0.000
Released-image SSIM	1.000	1.000	1.000	–	1.000
ArcFace identity reference	1.000	1.000	1.000	0.9999	1.000
Key projection G_s	–	–	–	1.000	1.000

Table 7 shows that private references already provide strong clean-swap and same-person release discrimination. To isolate what the key adds beyond these reference capabilities, the original control

below tests separation of the released image from its visually matched unreleased original under benign processing.

Unreleased original control. This experiment asks whether a verifier can identify the published version while tolerating JPEG compression. For each of 400 held-out enrollments, we compare the original image, the lossless release, and its JPEG-q90 copy. SSIM and ArcFace use the released image as their reference; spectral tests use its private record. JPEG-q90 thresholds are the first percentiles of all 1,144 FF++, 1,580 DFDC, and 1,203 Celeb-DF benign training releases, computed per dataset and reading. Lossless calibration provides a complementary operating point using 100 training releases per dataset. Calibration and test source videos are disjoint, and all readings are available. Reported rates pool the dataset-specific decisions.

Table 8: **Original and release acceptance under channel calibration.** Acceptance rates on 400 held-out enrollments, using cutoffs fitted to separate lossless or JPEG-q90 training releases at a nominal 1% rejection rate. Lower original acceptance and higher release acceptance are desirable.

	Lossless calibration			JPEG-q90 calibration		
Reading	Unreleased original ↓	Lossless ↑	JPEG ↑	Unreleased original ↓	Lossless ↑	JPEG ↑
Original phase H	100%	89.25%	32.75%	100%	100%	99.25%
Template C	100%	89.0%	27.0%	100%	100%	99.25%
Released-image SSIM	0%	100%	0%	100%	100%	99.5%
ArcFace	0%	98.75%	0%	99.25%	100%	99.0%
Key projection G_s	0%	98.5%	96.0%	0%	100%	99.5%

Table 8 illustrates the tension between compression tolerance and version discrimination. Lossless-calibrated SSIM and ArcFace distinguish the original from the release, but reject JPEG-processed releases. JPEG-calibrated reference tests admit processed releases alongside most originals, while G_s rejects every original and accepts 99.5% of JPEG releases. Thus, the keyed signal supplies version evidence that remains useful when reference similarity must tolerate compression.

Release identity and record age are separate questions: later re-enrollment defines a new valid release. Trusted timestamping establishes when a record existed, while an exact hash binds the pristine file. The keyed displacement supplies a distinct, channel-tolerant measurement of release-specific evidence.

Template specificity and privacy. Fig. 11 shows the displaced phase template and its query–record comparisons. Across 24 enrollments, matched pairs score 0.9974 ± 0.0046, versus 0.103 ± 0.091 for 552 mismatches; the lowest matched score is 0.979 and highest mismatch 0.426 (AUC 1.000). Despite this specificity, unit-amplitude inversion reveals facial structure, so the template remains sensitive image-derived data. This diagnostic does not isolate the causes of cross-dataset transfer.

Paired-key evidence and model scores. Table 9 holds each query fixed while changing the projection key. The cosine readings are identical within each pair, while key projections separate the paired keys on all three datasets. Mean learned scores distinguish benign, swapped, and wrong-key query types on FF++. Acceptance rates at calibrated cutoffs are reported separately in App. H.2.

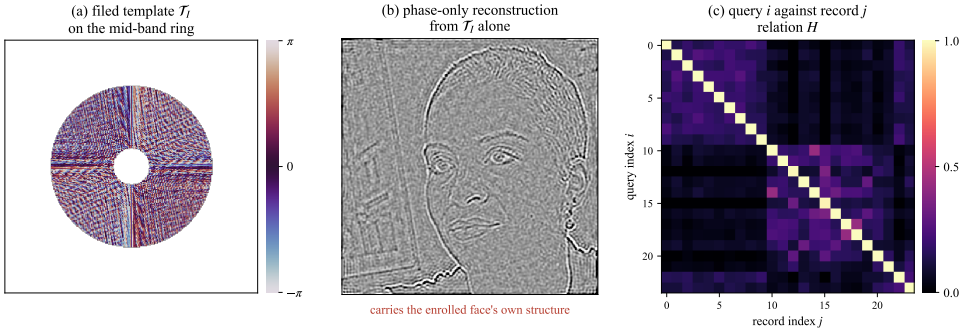

Figure 11: **Enrolled template and asset specificity.** (a) The template $\mathcal{T}_I$ stores original phase after the keyed displacement on the mid-band ring. (b) Unit-amplitude inversion reveals facial structure, illustrating the template's sensitivity. (c) Original-reference agreement H across 24 query–record pairs: matched pairs lie on the diagonal and have higher agreement than mismatched pairs.

Table 9: **Key discrimination and learned record scores.** Top: correct-versus-wrong-key AUC ↑ with each image held fixed, using 1,175, 636, and 209 test pairs on FF++, DFDC, and Celeb-DF; 0.500 denotes chance. Bottom: mean relation and key scores over three seeds on FF++, with 1,175 queries per type. Bottom arrows mark desired score directions.

Quantity	Reading	FF++ ↑	DFDC ↑	CELEB-DF v2 ↑
Cosine controls H/C	Spectral agreement	0.500 / 0.500	0.500 / 0.500	0.500 / 0.500
Key projections $G_s/G_\Delta/G_z$	keyed displacement	1.000 / 1.000 / 1.000	1.000 / 1.000 / 1.000	1.000 / 1.000 / 1.000
Query	Decision input	S_{rel}	S_{key}	
Benign query	both should agree	0.999 ↑	0.999 ↑	
Re-rendered after release	both should fail	0.001 ↓	0.001 ↓	
Wrong key	key must fail	1.000 ↑	0.000 ↓	

G CHANNEL ROBUSTNESS AND SPECTRAL VERIFICATION

Fixed enrollment reference. Enrollment fixes the template phase, original phase and amplitude, and encoded-release projection used for normalization. Processing acts only on the query. Thus, a lossless release has normalized key projection one, while its template agreement can be below one because the template precedes image quantization. Full-ring H and sub-band C_B both compare queries with this retained reference.

G.1 BENIGN CHANNELS ACROSS DATASETS

Table 10 compares full-ring agreement with dataset-trained image-only detectors. Spectral separation varies with geometry, active-bin counts, and release strength; median crops are 256×256 on FF++, 379×284 on Celeb-DF, and 759×685 on DFDC. Distortion-augmented training strengthens image-only detection under several blur and codec conditions. The analytic sub-band alternative instead adjusts the frequencies used for comparison, as examined below.

These AUCs measure face-swap separation. Table 9 and App. F examine release-key evidence; App. H.2 combines both forms of evidence in joint decisions.

G.2 PRESERVED FREQUENCIES AND VERIFICATION PERFORMANCE

Reading preserved evidence. Stored original-image amplitudes emphasize strong enrolled coefficients, supporting separation under JPEG compression and pixelation. Blur and transcoding can still suppress initially strong upper-ring coefficients. A lower sub-band concentrates comparison on frequencies more likely to preserve the enrolled relation.

Table 10: **Benign-channel robustness across datasets** (AUC ↑). Main-test AUCs averaged within each channel family. The $1 - H$ rows use full-ring original-reference agreement; image-only detectors are trained on their respective datasets with distortion augmentation. Table 1 reports the sub-band comparison.

Dataset	Reading	JPEG ↑	Blur ↑	Codec ↑	Pixel ↑	Clean ↑
FF++	$1 - H$ (full ring)	0.994	0.597	0.756	0.996	1.000
	image-only detector	0.972	0.773	0.776	0.976	0.986
DFDC	$1 - H$ (full ring)	0.799	0.568	0.589	0.879	0.977
	image-only detector	0.995	0.974	0.910	0.997	0.997
CELEB-DF	$1 - H$ (full ring)	0.736	0.533	0.550	0.809	0.978
	image-only detector	0.967	0.896	0.831	0.969	0.979

To evaluate the fixed $\kappa \leq 0.15$ setting on further sources, we select 100 additional FF++ videos, excluding sources in the recorded experiment manifests and the pilot evaluation. Two predetermined frames per video yield 197 valid release–SimSwap pairs after three face-detection failures, covering all 100 videos; replacement identities come from training data. The cohort, generation settings, frequency cut, and thresholds are fixed before evaluation.

Separation and operating points. Table 11 compares C and C_B on these paired queries. Both readings use first-percentile agreement thresholds from the same 1,144 benign training releases, calibrated per channel. JPEG uses quality 85; blur uses Gaussian radius 3; reposting applies $0.75\times$ bicubic resizing and JPEG quality 70 before restoring enrollment geometry; H.264 encodes each image independently at CRF 30.

Table 11: **Face-swap detection after image processing.** Restricting spectral comparison to the fixed lower sub-band improves swap rejection under blur and reposting. After H.264 encoding, higher AUC does not translate into better rejection at the calibrated threshold. TPR and FPR report the percentages of swapped and benign queries rejected, respectively.

	Full ring C			Sub-band C_B		
Channel	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓
Clean	1.000	100.0	0.0	1.000	100.0	0.0
JPEG $q85$	0.998	88.8	0.0	1.000	100.0	0.0
Blur	0.592	21.3	12.2	0.922	60.9	5.6
Social repost	0.971	51.3	0.0	1.000	92.9	0.0
H.264 intra	0.849	6.6	0.0	0.973	3.6	0.0

Under blur, the sub-band raises AUC by 0.330; 2,000 paired source-video bootstrap resamples give a 95% interval of $[0.299, 0.353]$, conditional on the fixed training calibration. Together with improved reposting rejection, this supports frequency restriction for preserving useful face evidence. Operating-point transfer remains channel dependent: blur FPR is 5.6% despite the 1% training target, and H.264 AUC improves without higher TPR. Channel selection therefore depends on both benign acceptance and swap rejection at the deployed threshold.

Relation and release evidence. Strong processing can push a released image's key score toward that of its unreleased original while retaining useful relation separation. Thus, frequency restriction alone does not establish version discrimination. The JPEG controls in Table 3 and joint channel tests in App. H.2 evaluate whether both forms of evidence remain usable.

Cropping as a known mask. The verifier reconstructs a spatial reference from stored amplitudes and template phases, applies the known crop mask, and compares Fourier phase with the mean-centered query, zero-padded to enrollment geometry. Comparison uses amplitude weights over the full support and a ± 2-pixel placement search, calibrated on 1,144 benign training crops. A center crop retaining 90% of each dimension yields AUC 0.962 over 200 pairs, accepting 99% of benign queries but rejecting only 6% of swaps. Geometry correction therefore restores benign agreement more readily than reliable swap rejection.

Clip-level aggregation. Averaging C_B within an enrolled video reduces the influence of individual degraded frames. Calibration uses the first percentile of mean scores from 120 benign training videos. On 120 test videos, averaging accepts 98.3% of benign clips and rejects 65% of swaps, with AUC 0.986, versus 5.7% swap rejection per frame. This supports temporal aggregation for independently encoded images; inter-frame coding remains to be evaluated.

H Operating Points and Calibration

We distinguish descriptive test-ROC operating points from training-calibrated decisions, then examine reuse of thresholds across datasets.

H.1 Empirical Detection Operating Points

To resolve differences hidden by near-ceiling AUC, Table 12 reports recall along the empirical test ROC at specified false-positive limits. At empirical FPR at most 10^{-3}, all four additional engines reach recall 1.000 over 1,175 FF++ pairs each. StarGAN-v2 and SD inpainting also reach this operating point on all 636 DFDC and 209 Celeb-DF pairs, extending the observed separation to GAN- and diffusion-based manipulations across datasets. SimSwap recall is 0.999 on FF++ and 0.734 on DFDC, despite AUCs of 1.000 and 0.992. These are descriptive test-ROC operating points, not results at independently calibrated deployment thresholds.

Table 12: **Empirical operating points of the direct relation test.** Template mismatch $1 - C_B$; n counts enrolled pairs, with benign queries as negatives. Each TPR is the largest test-ROC value at or below the target empirical FPR. Targets below $1/n$ use the zero-false-positive segment. Thresholds are read from the test ROC rather than calibrated on training data.

Re-rendering	Dataset	n	AUC ↑	TPR@FPR 10^{-2} ↑	TPR@FPR 10^{-3} ↑
StarGAN-v2	FF++	1,175	1.000	1.000	1.000
StarGAN-v2	DFDC	636	1.000	1.000	1.000
StarGAN-v2	Celeb-DF	209	1.000	1.000	1.000
UniFace	FF++	1,175	1.000	1.000	1.000
InSwapper	FF++	1,175	1.000	1.000	1.000
SD inpainting	FF++	1,175	1.000	1.000	1.000
SD inpainting	DFDC	636	1.000	1.000	1.000
SD inpainting	Celeb-DF	209	1.000	1.000	1.000
SimSwap	FF++	1,175	1.000	0.999	0.999
SimSwap	Celeb-DF	209	0.991	0.938	0.928
SimSwap	DFDC	636	0.992	0.868	0.734

H.2 Joint acceptance with training-only calibration

Table 13 evaluates the confirmed branch of Eq. (11) on clean queries using three fixed FF++-trained record models. Calibration uses 1,144, 1,580, and 1,203 benign training faces on FF++, DFDC, and Celeb-DF, with source-video groups disjoint from testing. The original per-head first-percentile thresholds yield joint calibration rejection of 1.57%, 1.88%, and 1.88%, respectively; the matched comparison below instead controls their conjunction.

Direct versus learned decisions. To compare the decision paths at a common budget, we evaluate full-ring C and release-normalized G_s against learned $S_{\mathrm{rel}}, S_{\mathrm{key}}$ on the same queries. For each method, both thresholds use a common lower-tail quantile, chosen as the largest value between zero and 0.01 for which joint rejection on benign training queries does not exceed 1%. This matches the calibration budget rather than each head's rejection rate; test queries do not select thresholds. Realized calibration rejection is 0.962%, 0.949%, and 0.998% on FF++, DFDC, and Celeb-DF for both methods. Test benign rejection can differ. To assess joint decisions after delivery, we use the 200 FF++ test pairs in Table 4 and calibrate each policy on 100 benign training releases per channel at 1% joint rejection. This comparison covers benign and swapped queries with fixed learned weights. JPEG uses quality 85; reposting applies $0.75\times$ bicubic resizing and JPEG quality 70, followed by restoration to enrollment geometry.

Table 13: **Joint verification with per-head and joint calibration budgets.** Under a matched joint budget, direct thresholds match FF++ swap rejection and reject more transferred swaps; the learned policy slightly increases benign acceptance on FF++ and DFDC. Rates are percentages; learned results are means over three FF++ seeds.

Dataset / channel	Policy	n	Benign confirmed ↑	SimSwap confirmed ↓	Wrong key confirmed ↓
Original policy: 1% calibration rejection per head					
FF++	Learned	1,175	99.15 ± 0.34	0.09 ± 0.00	0.00 ± 0.00
DFDC	Learned	636	97.33 ± 0.16	2.04 ± 0.00	0.00 ± 0.00
Celeb-DF	Learned	209	100.00 ± 0.00	8.93 ± 0.28	0.00 ± 0.00
Matched comparison: at most 1% joint calibration rejection					
FF++	Direct C, G_s	1,175	99.57	0.09	0.00
	Learned	1,175	99.69 ± 0.26	0.09 ± 0.00	0.00 ± 0.00
DFDC	Direct C, G_s	636	97.64	0.79	0.00
	Learned	636	98.01 ± 0.36	2.31 ± 0.18	0.00 ± 0.00
Celeb-DF	Direct C, G_s	209	100.00	3.83	0.00
	Learned	209	100.00 ± 0.00	14.04 ± 0.73	0.00 ± 0.00
FF++ delivery: 1% joint calibration rejection per channel					
Lossless	Direct C, G_s	200	99.00	0.00	—
	Learned	200	95.00 ± 0.00	0.00 ± 0.00	—
JPEG $q85$	Direct C, G_s	200	99.00	2.00	—
	Learned	200	99.83 ± 0.29	9.17 ± 0.29	—
Repost	Direct C, G_s	200	97.50	15.50	—
	Learned	200	99.50 ± 0.00	38.83 ± 1.76	—

Implication for the verifier. On these clean queries, direct thresholds provide competitive joint rejection without training, while the record model trades slightly higher benign acceptance for greater transferred-swap acceptance. Both reject every tested wrong-key query. Direct verification therefore provides the starting point when swap rejection is the priority. A similar tradeoff appears under JPEG and reposting: the learned policy retains more processed releases but also accepts more swaps. Both policies lose swap rejection after reposting, reinforcing the need to evaluate joint relation–key verification for each supported channel. These comparisons use full-ring direct scores and do not establish robustness to adaptive attacks.

Other manipulations and operating-point sensitivity. We evaluate StarGAN-v2, UniFace, InSwapper, and SD inpainting on the same $1{,}175$ FF++ release pairs per method, using the $1{,}144$ benign training queries for calibration. At a 1% joint calibration budget, both policies reject all four types of manipulation. Benign test rejection is 0.43% for direct thresholds and 0.31% for the learned policy, averaged over three seeds. Extending StarGAN-v2 and SD inpainting to all 636 DFDC and 209 Celeb-DF releases, both policies again reject every generated query using the existing target-dataset thresholds and fixed FF++ model weights. Benign rejection is 2.36% for direct verification and 1.99% for the learned policy on DFDC, and zero observed for both on Celeb-DF. These clean generative manipulations therefore support cross-dataset verification without showing a learned advantage in manipulation rejection.

A sweep of 101 joint calibration budgets from 0% to 10% reveals an operating-point tradeoff under reposting. At a 10% training budget, the learned policy reaches test TPR 93.83% and FPR 2.33%; the direct policy at its 1% budget reaches 84.50% and 2.50%, respectively. This exploratory comparison shows that the learned path can be useful at other operating points, but deployment budget selection requires separate validation.

H.3 Cross-Dataset Threshold Transfer

To distinguish weight transfer from threshold transfer, Table 14 holds the FF++-trained record model fixed and varies the calibration dataset. An empirical quantile fitted to one dataset's benign test scores is applied unchanged to the other datasets. Off-diagonal rejection rates range from zero observed

to approximately 125 times the target. Dataset and enrollment strength vary together for DFDC; the matrix does not isolate their effects or measure manipulation recall. These diagnostic test-set quantiles are separate from the training-only calibration used for the joint policy.

Table 14: **Cross-dataset threshold transfer.** Benign rejection rates from the fixed FF++ record model using S_{rel}. Each row fits a test-score quantile on the marked dataset (✓) and applies it unchanged to the others. Scores pool three seeds over 1,175, 209, and 636 distinct benign faces, respectively. Entries measure empirical rejection, subject to finite-sample resolution; zeros denote no observed rejections.

Fitted on	Target FPR	FPR realized when reused on ↓		
		FF++	CELEB-DF	DFDC
FF++	10^{-2}	1.0×10^{-2} ✓	0	9.7×10^{-2}
	10^{-3}	1.1×10^{-3} ✓	0	6.4×10^{-2}
CELEB-DF	10^{-2}	7.0×10^{-2}	1.1×10^{-2} ✓	1.5×10^{-1}
	10^{-3}	3.9×10^{-2}	1.6×10^{-3} ✓	1.3×10^{-1}
DFDC	10^{-2}	0	0	1.0×10^{-2} ✓
	10^{-3}	0	0	1.0×10^{-3} ✓

H.4 CALIBRATION SENSITIVITY AND FINITE-SAMPLE RESOLUTION

Fig. 12(a–b) separates the target rejection rate from the amount of calibration data. Relation recall remains 0.999 across the tested targets, while increasing the calibration cohort from 25 to 100 faces reduces variability in benign rejection. Panel (c) reports the record model's key-loss sensitivity.

With n independent negative faces, empirical FPR changes in increments of $1/n$; zero observed errors does not establish a lower population error rate. The paired-key tests contain 1,175, 636, and 209 negative faces, and repeated training seeds do not increase these counts. All joint-policy cutoffs use empirical training quantiles without tail extrapolation. Because either failed criterion prevents confirmation, a joint rejection budget must be calibrated on the combined decision, as in the matched comparison above.

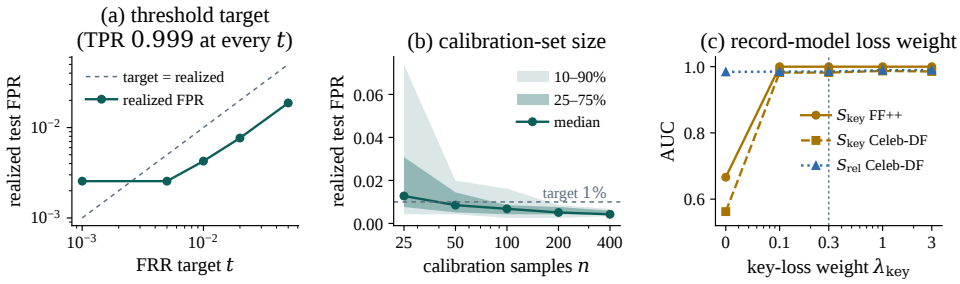

Figure 12: **Calibration and record-model sensitivity** (FF++ unless noted). (a) Target benign false-rejection rate (FRR), denoted t in this panel; (b) number n of calibration samples, with shaded percentile ranges; and (c) key-loss weight λ_{key}.

I SENSITIVITY OF THE RELEASE MECHANISM

To isolate the release mechanism from learned compensation, the strength and ring sweeps use analytic readings with no fitted model weights. The key-loss sweep separately evaluates the record model. For every thresholded quantity, the operating point is recalibrated independently at each setting using benign-training quantiles (n=200 training assets per dataset and setting). The same recorded per-asset SimSwap replacement identities are reused throughout so that settings remain directly comparable.

Fig. 13 shows stable relation and wrong-key separation over the tested strength and ring ranges, with positive gaps between JPEG-processed releases and unreleased originals. Ring effects differ across datasets, where crop geometry and strength also vary. Table 15 details the FF++ results; Fig. 14 measures the cost of additional wrong keys.

Table 15: **Effect of ring location on visual quality and verification** (FF++). The tested rings preserve high visual fidelity and clean face-swap separation, while JPEG robustness and the release–original key-projection gap vary with ring location. Ring bounds $[\kappa_{\min}, \kappa_{\max}]$ are the inner and outer frequency radii, normalized so that zero is DC and one is the Fourier-plane corner. Strength is fixed at the deployed value, with calibration repeated for each setting.

Ring $[\kappa_{\min}, \kappa_{\max}]$	PSNR (dB) ↑	Face-swap detection AUC ↑		Key evidence	
		Clean	JPEG-q90	Release–original key-projection gap ↑	Correct vs. wrong key AUC ↑
[0.05, 0.45]	50.8	1.000	1.000	0.65	1.000
[0.10, 0.45]	51.2	1.000	0.998	0.79	1.000
[0.15, 0.45]	51.3	1.000	0.995	1.06	1.000
[0.20, 0.45]	51.3	1.000	0.990	1.26	1.000
[0.10, 0.30]	51.2	1.000	1.000	0.94	1.000
[0.10, 0.35]	51.2	1.000	0.999	0.89	1.000
[0.10, 0.40]	51.2	1.000	0.999	0.84	1.000
[0.10, 0.50]	51.2	1.000	0.997	0.77	1.000

Key-projection gap between releases and originals. This quantity subtracts the highest key projection among unreleased originals from the first percentile of projections among JPEG-q90 releases. Thus, a positive gap means that at least 99% of the evaluated releases score above every evaluated original, leaving room for a threshold that separates them. It measures separation from unreleased originals, whereas the final AUC column measures separation from wrong keys.

For the record model, removing the key loss reduces key separation toward chance; positive weights from 0.1 to 3 yield similar separation while leaving the relation head largely unchanged. Fig. 12(c) reports this diagnostic.

J Deployment Interface and Cost

The publisher retains the spectral record and the metadata needed to associate a query with its enrolled face. We measure the cost of enrollment and verification with that association already resolved; App. K discusses operational conditions and record protection.

Runtime and storage. We benchmark the accompanying implementation on 20 native 256^2 FF++ test face crops, after three warm-up iterations, using an Intel Xeon Gold 6248R CPU with one PyTorch thread. Enrollment, including the RGB edit, uint8 conversion, and nine release-normalization scalars, takes 66.7 ms per crop; this excludes quality-gate checks and file serialization. Direct C_B comparison with an in-memory record takes 5.9 ms. Loading the query and NPZ, extracting all 75 features with eight alternative keys, and running the approximately 10.3k-parameter MLP takes 149.9 ms. These stages exclude face detection, alignment, and record lookup.

Table 16 reports the CPU verification workloads alongside peer measurements on an A5000. Hardware and processing paths are specified separately for each timing.

Storage depends on the retained support and numerical precision. On 150 FF++ faces, the codec experiment serializes phase, amplitude, and support-mask payloads: float16 arrays require 69.6 KiB, 8-bit phase and 4-bit log-amplitude require 28.0 KiB, and restricting the quantized payload to $\kappa \leq 0.15$ requires 2.3 KiB. These payload variants preserve detection and correct-versus-wrong-key AUCs to three decimals on the tested assets; this codec comparison measures ranking performance.

These codec payloads exclude dimensions, keys, strengths, geometry, calibration metadata, and nine float64 release projections (72 bytes). The sub-band codec retains evidence for restricted comparison,

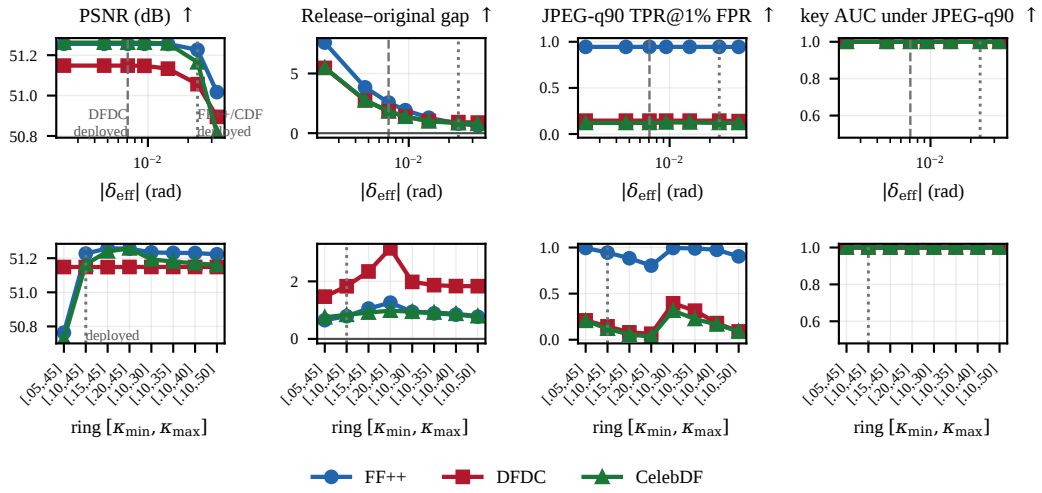

Figure 13: **Effects of enrollment strength and frequency-ring location.** Varying the embedding strength (top) and ring location (bottom) reveals their effects on visual fidelity, release–original separation, and verification performance. Columns show PSNR, the key-projection gap between releases and originals, relation TPR after JPEG-q90, and correct-vs-wrong-key AUC. TPR uses thresholds calibrated to 1% benign training rejection. Dashed and dotted lines indicate the settings used in our experiments.

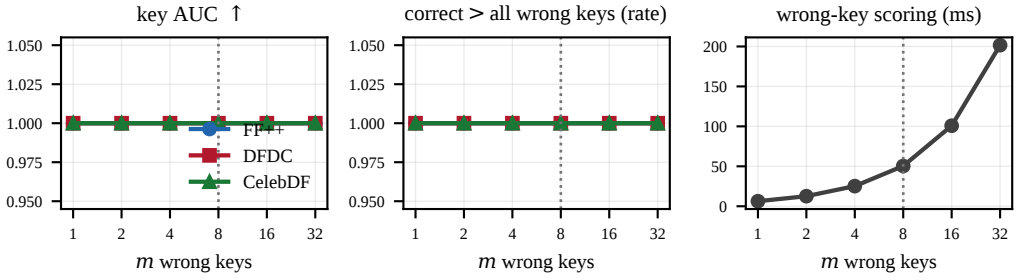

Figure 14: **Wrong-key count.** Key separation remains stable from m=1 to 32, while scoring cost grows approximately linearly. The evaluated configuration uses m=8.

Table 16: **Runtime and records for the evaluated verification paths.** Means over 20 256^2 crops after warm-up. Peer timings use an A5000; our timings (‡) use a Xeon Gold 6248R CPU and the listed records. CPU timings cover relation-only comparison or the learned path, excluding the image-only detector. Compact codecs are evaluated separately.

	Time (ms) ↓		Memory (MiB) ↓		Record used
Method	Embed	Verify	Weights	GPU	Spectral record
StegaStamp	9.3	2.6	208	375	none
FaceSigns	4.4	1.7	72	102	none
SepMark	12.5	17.8	550	653	none
WAM	14.9	17.0	360	450	none
KAD-Net	17.0	12.5	563	650	none
VINE-R	79.3	10.7	3833	4569	none
LIDMark	6.0	5.9	68	93	none
Ours, relation C_B	66.7‡	5.9‡	0	0	in-memory arrays
Ours, record model	66.7‡	149.9‡	0.04	0	356 KiB NPZ

whereas the model needs full-ring features. Runtime measurements use the 356 KiB research NPZ or its in-memory arrays; compact-codec loading and direct full-ring joint latency remain unmeasured.

The optional image-only review signal has 91 MiB of weights and used 185 MiB of GPU memory with 6.9 ms inference in the A5000 benchmark. Its cost is separate from the CPU record verifier.

K Verification Limits and Record Privacy

This section examines how regeneration and targeted attacks affect verification, then discusses operating conditions and the protection of private records.

K.1 Regeneration and Adaptive Attacks

An attacker may seek to make a manipulated face pass verification or cause a genuine release to be rejected. To explore these deployment risks, we examine two routes: diffusion regeneration that may erase release-specific evidence, and targeted perturbations that directly alter verification scores. These tests probe whether the evidence remains useful when processing is deliberately chosen to undermine verification.

Experimental settings. We evaluate 200 FF++ test pairs in Table 17. SD-Turbo regenerates both the release and its re-rendering at strengths 0.2, 0.3, and 0.4, either over the whole crop or only the face region, without accessing the record. Strength controls how much noise is added before denoising: larger values allow more extensive regeneration and greater departure from the input. These values are diffusion settings, not percentages of pixels changed. Targeted attacks use an L_∞ budget of $\epsilon = 8/255$ on images in $[0, 1]$. PGD (Madry et al., 2018) takes 40 steps of $\epsilon/8$ against the listed score; NES (Ilyas et al., 2018) uses 24 score queries followed by one signed step. Transfer attacks use a DFDC-trained detector. Because these attacks lower suspiciousness for swaps and raise it for genuine releases, the resulting AUC measures ranking under this two-sided attack, rather than swap-evasion success at a fixed threshold.

Scores and attack targets. The three columns measure template mismatch $1 - C_B$, image-only score S_{df}, and record-model score S_{fake}. Record-model AUCs average three fixed FF++ checkpoints, using pre-sigmoid logits to avoid numerical saturation. Attacks optimize template or image-only scores; the record-model column measures attack transfer to a learned head that is not directly targeted. AUC 0.500 denotes chance and 0.000 an inverted ranking.

Table 17: **Detection under regeneration and targeted attacks** (AUC ↑). Stronger regeneration reduces separation, while template-targeted PGD also degrades the learned record model. Ranges cover three regeneration strengths, from weaker (0.2) to stronger (0.4).

		Face-swap detection AUC ↑		
Method	Setting	Template mismatch $1 - C_B$	Image-only S_{df}	Record model S_{fake}
No attack	Original test queries	1.000	0.996	0.999
Diffusion regeneration: SD-Turbo				
SD-Turbo, 0.2–0.4	Whole crop	0.665–0.893	0.835–0.973	0.725–0.893
SD-Turbo, 0.2–0.4	Face region only	0.783–0.954	0.873–0.988	0.771–0.947
Adaptive attacks ($\epsilon = 8/255$)				
NES (24 queries)	Score-only access	0.971	0.992	0.922
Transfer attack	DFDC-trained detector	0.969	0.000	0.936
White-box PGD	Image-only S_{df}	0.964	0.000	0.926
White-box PGD	Template $1 - C_B$	0.000	0.994	0.081
White-box PGD	$1 - C_B$ and S_{df}	0.124	0.000	0.797

Findings. The verification components provide complementary evidence but share vulnerabilities. Attacks against the image-only detector leave spectral comparison informative, whereas attacks against the template can also undermine the record model because it uses related spectral features. Adding a learned head therefore does not by itself protect the system from manipulation of its underlying evidence. Stronger diffusion regeneration also weakens separation, indicating that substantially regenerated copies require separate channel evaluation before automated verification. For deployment, these findings motivate testing attacks across verification components. Accordingly, processing identified as unsupported should yield *unavailable*. Successful attacks, however, can satisfy the verification criteria without triggering human review. The component AUCs identify failure modes; evaluating their effect on final verdicts additionally requires testing the complete relation–key policy.

Future directions. These results motivate exploring evidence distributed across frequency bands and spatial scales, with the aim of preserving discrimination under regeneration and delivery loss. Candidate designs should be evaluated against attacks targeting the complete relation–key policy, measuring both benign acceptance and manipulation rejection.

K.2 Phase-Ring Transplantation

An attacker holding the released image can copy its ring phase into a re-rendered face while retaining the latter's magnitudes. Phase is copied from the public release rather than recovered from a private template or key. However, the evaluated attack is supplied with the enrolled active-bin mask; its access assumption therefore includes support information from the record. Table 18 evaluates its effect on template-based and image-only detection.

Evaluation settings. We test 200 pairs per dataset, with thresholds calibrated for 1% FPR on that dataset's benign training queries. AUC compares swapped queries with paired benign releases; TPR and FPR use the original thresholds, with the same benign baseline before and after transplantation. The calibration target does not guarantee a 1% test FPR. The image-only detector uses fixed FF++ weights. "Either flag" means that the template or image-only test flags the query; this binary OR rule has no continuous score for AUC and is separate from the relation–key decision policy.

Table 18: **Detection before and after phase-ring transplantation.** Transplantation reduces template-score separation and detection at the original thresholds. FF++ retains substantial ranking ability despite low TPR, whereas DFDC approaches chance. TPR and FPR are percentages.

	Template mismatch $1 - C_B$			Image-only S_{df}			Either flag (OR)	
Dataset	AUC ↑	TPR ↑	FPR ↓	AUC ↑	TPR ↑	FPR ↓	TPR ↑	FPR ↓
Before transplantation								
FF++	1.000	99.5	2.0	0.996	99.5	14.0	99.5	15.5
Celeb-DF	0.990	95.0	1.5	0.687	18.0	4.5	95.0	6.0
DFDC	0.985	98.5	13.0	0.751	34.5	7.0	99.0	14.5
After transplantation								
FF++	0.892	6.0	2.0	0.995	99.5	14.0	99.5	15.5
Celeb-DF	0.717	16.5	1.5	0.613	14.5	4.5	26.5	6.0
DFDC	0.579	15.0	13.0	0.724	30.0	7.0	32.0	14.5

Findings. Copied release phase can satisfy spectral criteria despite changed facial appearance. The learned joint policy confirms 84% of attacked FF++ queries in the three-seed mean-score evaluation, so key evidence does not prevent this bypass. Image-only detection helps but weakens across datasets; stored-amplitude comparison does not restore separation. Bypasses can receive a *confirmed* verdict and escape human review.

Future directions. To resist evidence transplantation, a complementary direction is to couple the release signal to local facial content and verify consistency across facial regions. Evaluation should measure whether these dependencies improve rejection of transplanted evidence while preserving acceptance under supported processing; their effectiveness remains to be established.

K.3 Verification Conditions and Interpretation

Channel evaluation must assess both benign acceptance and manipulation rejection: processing can erase discriminative evidence while releases still pass. Failed comparisons identify discrepancies, not causes; unsupported processing or unresolved geometry requires an unavailable outcome.

Evaluation uses individual face crops with known enrollment geometry; scores do not localize facial edits. End-to-end multi-face detection and association, public signed receipts, and C2PA integration remain unevaluated. The optional image-only review flag supplies separate evidence under Eq. (11).

Future directions. Deployment studies could combine channel-support and face-association confidence to identify when to return *unavailable*. End-to-end evaluation should measure verification errors, automated coverage, and reviewer accuracy and workload, including attacks that are confirmed and escape referral.

K.4 Privacy of Spectral Records

Phase-only reconstruction reveals coarse facial structure (Fig. 11), making private records sensitive image-derived data. Access to templates, keys, and scores can therefore expose information beyond the reported verification outcome.

Future directions. Future work could reduce recoverable facial information while retaining relation and key evidence, jointly evaluating verification performance, reconstruction, and identity leakage. Reporting protocols should authenticate outcomes with minimal disclosure and assess leakage through repeated verification requests.