SeLo-Mark: Content-Derived Robust Localized Watermarking for Facial Integrity Verification

Anonymous authors
Paper under double-blind review

Abstract

Image watermarking enables proactive facial forensics by embedding information before publication. Verifying facial content, however, requires more than recovering a payload: the embedded signal must represent the original face while remaining recoverable after image processing. Random payloads provide no such content reference, while geometric transformations disrupt the spatial correspondence required by localized watermarking. We propose SeLo-Mark, a content-derived localized watermarking framework for facial integrity verification. First, a compact semantic reference is derived from the protected face and embedded only in the surrounding unprotected region, leaving the protected facial pixels unchanged. Second, a shared geometric estimate is used to align both watermark recovery and facial-content encoding after geometric processing, so that the recovered reference can be compared with the received face in a common coordinate frame. The semantic reference is obtained by projecting frozen CLIP features with PCA and binarizing them into a 32-bit facial-content code, which is combined with an independent 16-bit tracing field. At reception, the payload is recovered by a reader adapted to rectified images, and facial consistency is assessed by comparing the recovered content code with one recomputed from the received face, without reconstructing the original face. Across four editor families and 22 processing conditions on 2,999 CelebA-HQ images, SeLo-Mark achieves 99.23% AUC and 95.85% F1. On unedited watermarked images, exact 48-bit recovery is increased from 88.38% with direct, unadapted decoding to 99.82% with the full receiver.

1 Introduction

Facial editing can change the depicted identity or facial attributes while leaving an image visually convincing. Proactive forensics embeds evidence before publication to make such changes verifiable after the image has been distributed. For applications where facial content should remain pixel-wise unchanged, we study a protected-region setting: the designated facial pixels are preserved at publication, and verification information is carried by the surrounding image. At reception, the verifier must use the received image alone and distinguish facial edits from benign processing such as compression and resizing.

Robust watermarking is designed to preserve messages through image processing (Zhu et al., 2018; Bui et al., 2025), and localized methods support embedding or recovery within selected regions (Sander et al., 2025; Hu et al., 2025). Successful recovery, however, does not establish that the protected face is unchanged. An editor may alter the face while leaving enough of the external watermark intact for the message to be decoded. Content-derived watermarking makes the embedded evidence relevant to the source face (Wang et al., 2024; Wu et al., 2026; An et al., 2026). It supplies a reference against which the received facial content can be compared, but this comparison must distinguish content changes from variations introduced by routine processing.

Geometric processing makes this distinction harder. Rotation, perspective changes, and cropping affect both the external carrier and the facial observation used to recompute the reference. They can disrupt watermark recovery even when no facial edit has occurred, while changes in position or visible extent can alter the recomputed content code. Cropping may also remove part of the observed

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.

face without changing the content of the retained pixels. Geometric synchronization restores the coordinates required for message extraction (Fernandez et al., 2026). For facial verification, the resulting alignment can also reduce geometric variation in the observation used for comparison.

This motivates a design in which a robust watermark preserves the original reference and the content comparison provides sensitivity to editing. A compact description of the published face fits a short localized payload and provides a comparison target without reconstructing the original appearance. The representation used for comparison should tolerate benign processing while remaining responsive to facial edits. Recovering the published reference and encoding the received face in a common geometric frame helps reduce geometric mismatch in the resulting comparison.

SeLo-Mark realizes this design with a compact facial-content code derived from frozen CLIP features (Radford et al., 2021). A fixed PCA projection and binarization produce a 32-bit reference, which is embedded only in the unprotected region. The protected pixels are copied unchanged into the published image. An independent 16-bit tracing field accompanies the content reference, while only the content bits enter the facial-consistency score. The recovered reference is compared directly with a code recomputed from the received face using the same fixed mapping.

At reception, one geometric estimate is shared by watermark recovery and facial encoding. The watermark branch reads the stored reference, while the facial branch encodes the aligned observation. This makes geometric correction part of content comparison as well as message extraction. Because rectification resamples the watermark signal, we adapt the reader to corrected inputs while keeping the publisher fixed.

Our contributions are:

- A localized facial-verification framework that derives a compact reference from the protected face and carries it only outside the protected region, enabling verification without modifying the protected pixels.
- A shared geometric correction for watermark recovery and facial-content encoding, together with reader adaptation to rectified watermark signals.
- An evaluation across four editor families and 22 processing conditions, with controlled studies of the content reference and receiver geometry. SeLo-Mark achieves 99.23% AUC and 95.85% F1 on CelebA-HQ, with 99.82% exact recovery of the complete 48-bit payload on unedited watermarked images. The models and threshold also transfer to FFHQ without refitting.

2 Related Work

Robust and localized watermarking. Robust watermarking preserves embedded messages through image processing, with learned systems commonly training embedding and recovery together under simulated distortions (Zhu et al., 2018; Bui et al., 2025). Localized approaches add spatial awareness by identifying watermarked regions and reading from the surviving support; masked training also supports embedding in designated carriers (Sander et al., 2025; Hu et al., 2025). These developments provide message transport under spatial constraints. For facial verification, the remaining question is how the recovered message represents the content being protected.

Proactive forensics and content-derived references. Proactive forensics uses the behavior of embedded signals to reveal editing. Semi-fragile designs seek robustness to benign processing and sensitivity to manipulation (Neekhara et al., 2024; Wu et al., 2023). Spatially structured signals extend this idea to localizing edited regions (Zhang et al., 2024; 2025; Wang et al., 2025). A complementary line embeds source information, including facial geometry (Wang et al., 2024; Wu et al., 2026), recoverable appearance (An et al., 2026), and semantic patterns for diffusion-generated images (Arabi et al., 2025). This provides evidence about the published content that can be recovered or compared after editing. Our setting requires that evidence to be carried outside the protected face and compared with a received observation after geometric processing.

Geometric synchronization. Geometric synchronization restores the spatial correspondence needed for decoding through transformation estimation or object-pose normalization (Luo et al., 2023; Zhao et al., 2024; Fernandez et al., 2026). These approaches align the carrier for message recovery. Content-derived facial verification also depends on the geometry of the observation used to

recompute the comparison code, so correction must support both recovery and comparison with the received face.

3 METHOD

3.1 PROBLEM SETUP AND OVERVIEW

In the protected-region setting of Section 1, verification information is carried outside the protected facial region it describes. Let x be an 8-bit RGB image normalized to $[-1, 1]$ and $M \in \{0, 1\}^{H \times W}$ its watermark carrier mask. We set $M = 0$ on the facial protection region and require the published image x^{w} to satisfy

$$(1 - M) \odot (x^{\mathrm{w}} - x) = 0. \tag{1}$$

Protected pixels are copied from the original image; the remaining pixels form the carrier.

The verifier receives only the RGB image y. We consider facial edits followed by transmission processing, assuming that sufficient facial content and part of the watermark carrier remain observable; the editing procedures are not optimized against the verifier.

Figure 1 summarizes publication and verification. A compact binary reference to facial content is embedded in the external carrier, then recovered and compared with a code recomputed from the received face. A shared geometric estimate aligns watermark reading and facial observation before comparison. All components are fixed at test time.

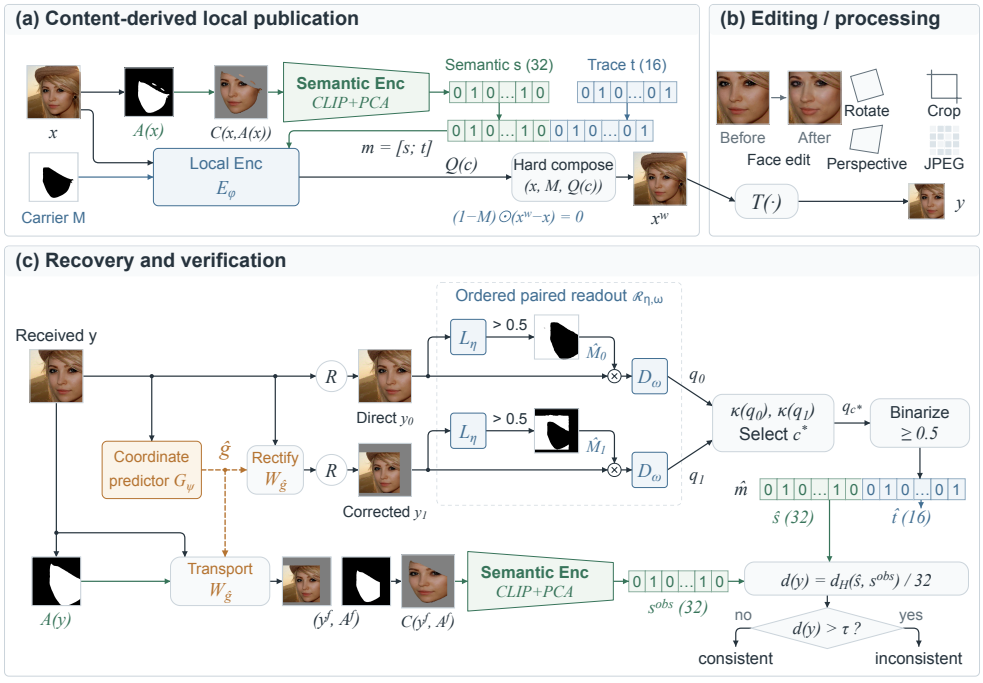

Figure 1: **SeLo-Mark overview. (a) Content-derived local publication.** Frozen CLIP features of the facial observation $C(x, A(x))$ are projected by a fixed PCA mapping and binarized into a 32-bit content reference s. An independent 16-bit tracing field t completes the payload $m = [s; t]$. The local encoder embeds it in carrier M, and hard composition with the original image preserves protected facial pixels. **(b) Editing and processing.** Facial edits and image processing produce the received image y. **(c) Recovery and verification.** A predictor estimates geometry $\widehat{g}$ from y, shared by watermark rectification and joint transport of the facial image–mask pair. Direct and corrected watermark views use the same localizer and adapted reader; prediction confidence selects the recovered payload. The corrected facial observation is encoded with the same CLIP–PCA mapping, independently of the selected watermark view. Normalized Hamming distance compares its code s^{obs} with the recovered content reference $\widehat{s}$. A distance above the calibrated threshold τ flags an edit; the tracing field $\widehat{t}$ is output separately.

3.2 Content-Derived Facial Reference

The reference should tolerate routine processing while responding to facial edits. We derive it by compressing a frozen visual representation into a short binary code.

Let $A(x)$ be the facial observation mask predicted from RGB using SCRFD and BiSeNet (Guo et al., 2022; Yu et al., 2018), and $C(x, A)$ the corresponding masked crop. Mask A selects the content used for the reference; mask M specifies where it is embedded.

A frozen CLIP ViT-L/14 vision encoder F (Radford et al., 2021) extracts an L2-normalized feature $z = F(C(x, A)) \in \mathbb{R}^{1024}$. PCA is fitted on facial observations from the training split using annotated masks. Given the feature mean μ and the 32 leading covariance eigenvectors $U \in \mathbb{R}^{1024\times 32}$, the projected features are binarized at zero:

$$a(x, A) = (z - \mu)^\top U, \qquad s(x, A) = \mathbf{1}[a(x, A) \geq 0] \in \{0, 1\}^{32}. \tag{2}$$

The same mean, projection, and binarization rule are used at publication and reception.

The 32-bit content reference is concatenated with an independent tracing field $t \in \{0, 1\}^{16}$:

$$m = [s(x, A(x)); t] \in \{0, 1\}^{48}. \tag{3}$$

Only the 32 content bits enter the facial-consistency score.

3.3 Localized Reference Publication

The reference is carried by a localized watermark channel based on MaskWM (Hu et al., 2025). Given image x, payload m, and carrier M, the mask-conditioned encoder E_ϕ produces a candidate watermarked image. At reception, localizer L_η estimates the surviving watermark support, which reader D_ω uses to recover the 48 payload scores.

The encoder residual is weighted by a just-noticeable-difference (JND) map, upsampled, and added to x to form candidate c. We restrict embedding to the carrier by composing c with the original image:

$$x^{\mathrm{w}} = (1 - M) \odot x + M \odot Q(c), \tag{4}$$

where Q clips and quantizes values to the normalized 8-bit grid. Equation 4 enforces Equation 1 independently of the learned encoder.

We train the channel with random 48-bit messages, local and full-image carriers, and simulated processing. Using random messages makes channel training independent of how the facial reference is constructed. The objective combines image fidelity, message recovery, and support localization (Appendix A.3). The publisher and localizer are then frozen before receiver adaptation.

3.4 Shared Geometric Correction and Reader Adaptation

Geometric processing affects both sides of the verification comparison: it changes the support from which the external watermark is read and the facial observation from which the received content code is recomputed. SeLo-Mark therefore estimates geometry once and applies the same correction to both branches.

Shared geometric estimation. Predictor G_ψ estimates the source-image locations corresponding to the four corners of the received image. We use SyncSeal's corner representation and pretrained predictor (Fernandez et al., 2026), adapting it to the local watermark channel while keeping the publisher fixed. No additional synchronization watermark is embedded.

Let $\mathcal{N}$ denote the fixed coordinate normalization and $p^\star \in \mathbb{R}^8$ the source locations obtained from a known training transformation. The coordinate predictor is optimized with

$$\mathcal{L}_{\mathrm{geo}} = \mathbb{E}_{(y,p^\star)}\left[\frac{1}{8}\|G_\psi(y) - \mathcal{N}(p^\star)\|_1\right]. \tag{5}$$

The predictor is then frozen. At reception, $\widehat{p} = \mathcal{N}^{-1}(G_\psi(y))$ defines a homography $\widehat{g}$. The corresponding rectification operator $W_{\widehat{g}}$ is used for both the watermark-reading input and the image–mask pair used for facial encoding. Coordinate and interpolation conventions are specified in Appendix A.5.

Reader adaptation after rectification. Rectification restores spatial alignment but also resamples the watermark signal, while cropping may remove part of its support. We therefore adapt the reader to corrected inputs while retaining direct reading for routine views. Only D_ω is updated in this stage; E_ϕ, L_η, and G_ψ remain fixed.

Each publication produces a base view y^{b} with routine or no processing and a cropped view y^{c} carrying the same random message. For $v \in \{\mathrm{b}, \mathrm{c}\}$, the direct and corrected candidates are

$$y_0^v = R(y^v), \qquad y_1^v = R(W_{\widehat{g}^v}(y^v)), \qquad \widehat{M}_j^v = \mathbf{1}[L_\eta(y_j^v) > 0.5], \tag{6}$$

where R resizes to the reading resolution, $\widehat{g}^v$ is estimated from view v, and $j = 0, 1$ indexes direct and corrected candidates. Reading each candidate with its predicted support yields scores q_0^v and q_1^v.

For a minibatch of B publications and $\ell(q, m) = \|q - m\|_2^2/48$, reader adaptation minimizes

$$\mathcal{L}_{\mathrm{read}} = \frac{1}{2B} \sum_{i=1}^{B} \left[\ell(q_{i,0}^{\mathrm{b}}, m_i) + \ell(q_{i,1}^{\mathrm{c}}, m_i) \right]. \tag{7}$$

The first term retains direct decoding under routine processing, while the second exposes the reader to watermark signals resampled by geometric correction. Appendix A.4 gives the training distribution and optimization settings.

3.5 Facial Consistency Verification

At verification time, the watermark branch produces direct and corrected message candidates. Let $c^* \in \{0, 1\}$ index the selected candidate. Confidence is based on the distance of the predicted scores from the binary endpoints:

$$\kappa(q) = -\frac{1}{48} \sum_{j=1}^{48} \min\{q_j^2, (q_j - 1)^2\}, \qquad \widehat{m}_j = \mathbf{1}[q_{c^*,j} \geq 0.5]. \tag{8}$$

The corrected candidate is selected when $\kappa(q_1) > \kappa(q_0) + \epsilon$; otherwise the direct candidate is retained, with $\epsilon = 10^{-6}$. The recovered payload is written as $\widehat{m} = [\widehat{s}; \widehat{t}]$, where $\widehat{s}$ is the recovered 32-bit facial reference.

The facial branch uses the same predicted geometry. The received RGB image and its predicted facial observation mask are rectified as

$$(y^{\mathrm{f}}, A^{\mathrm{f}}) = W_{\widehat{g}}(y, A(y)), \qquad s^{\mathrm{obs}} = s(y^{\mathrm{f}}, A^{\mathrm{f}}). \tag{9}$$

Facial encoding always uses the corrected observation, independently of which watermark candidate is selected.

For a valid facial observation, normalized Hamming distance measures disagreement between the recovered publication reference and the code recomputed from the received face:

$$d(y) = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1}[\widehat{s}_j \neq s_j^{\mathrm{obs}}], \qquad K = 32. \tag{10}$$

An edit is flagged when $d(y) > \tau$, where τ is selected on an independent calibration set and fixed during testing.

4 Experiments

4.1 Experimental Setup

Data and processing. We evaluate on 2,999 CelebA-HQ images (Karras et al., 2018) with disjoint CelebA identities (Liu et al., 2015) across training, validation, calibration, and test splits. A fixed 128-image test subset supports controlled ablations; transfer uses 1,000 FFHQ images (Karras et al., 2019). Images are published at 512×512 through a 256×256 watermark channel.

Each method's watermarked images are edited with SimSwap and CSCS for face swapping (Chen et al., 2020; Huang et al., 2024), StarGAN for attribute changes (Choi et al., 2018), and Stable Diffusion 1.5 inpainting (Rombach et al., 2022). Edited and unedited images undergo the same 22 conditions spanning no processing, regular distortions, geometric transformations, and random 0.75-side cropping. Appendix A gives the full data, editing, processing, and training settings.

Metrics and baselines. We report AUC and class-balanced F1 on available scores, with edits as the positive class. Metrics are computed per editing operation and processing condition, averaged within each editor family, then equally across the four families and conditions. Thresholds maximize Youden's J on calibration data and remain fixed during testing; SeLo-Mark uses $\tau = 6/32$, including on FFHQ. BA measures bit accuracy. Exact32 requires correct recovery of all 32 content bits; Exact48 requires the entire payload, including the 16-bit tracing field. PSNR and SSIM (Wang et al., 2004) measure publication quality. SeLo-Mark results average two receiver-adaptation runs.

The detection comparison covers seven proactive forensic methods, and the recovery comparison additionally includes robust and localized watermarking baselines. LampMark (adapted) uses InsightFace-derived references and a channel adapted to those payloads for detection; LampMark (released) uses the authors' payloads and channel for recovery. Baseline interfaces are specified in Appendix A.6.

4.2 End-to-End Facial Verification

Table 1 reports end-to-end facial-edit detection after transmission processing.

Table 1: Facial-edit detection on CelebA-HQ.

	Image quality ↑		AUC (%) ↑					F1 (%) ↑
Method	PSNR	SSIM	No proc.	Regular	Geometry	Crop	All 22	All 22
FractalForensics	37.88	0.9617	99.99	**99.26**	55.41	49.91	88.83	89.45
SepMark	33.30	0.9124	75.41	75.95	48.31	49.79	69.78	63.84
EditGuard-degrade	32.27	0.7754	**100.00**	78.01	68.60	50.00	75.18	82.04
OmniGuard	**49.51**	0.9937	94.89	66.00	53.16	50.41	64.14	62.34
LIDMark	44.78	**0.9951**	79.57	78.87	52.19	49.60	72.60	63.42
RecoverMark	31.19	0.9351	84.95	80.33	52.04	51.20	74.03	74.25
LampMark (adapted)	31.76	0.9433	99.26	98.66	52.61	50.70	88.05	88.10
SeLo-Mark	39.58	0.9782	99.87	99.22	**99.52**	**98.58**	**99.23**	**95.85**

PSNR is in dB; detection metrics are percentages. Results average the four editor families and processing conditions within each group. F1 uses fixed calibration thresholds. Quality is measured on successfully watermarked images. LampMark (adapted) uses InsightFace/PCA references and a channel adapted to these payloads; its released-payload recovery configuration appears separately in Table 2. OmniGuard uses its localization interface here and its copyright branch in Table 2.

Under regular processing, SeLo-Mark has AUC comparable to FractalForensics and LampMark (adapted). The main difference emerges after geometric transformations and cropping. LampMark's adapted reference and channel give strong detection under regular processing, but this performance does not carry over to the geometric and crop conditions. SeLo-Mark's advantage is thus most apparent when spatial processing disrupts verification. The following experiments examine reference recovery, code stability, and the effects of geometric correction on reading and facial encoding.

4.3 Watermark Channel Robustness

Table 2 evaluates unedited publications using each method's native payload, separating watermark recovery from facial editing.

Most methods recover messages accurately under regular processing. Geometric transformations and cropping separate their performance more clearly, with localized baselines particularly affected when cropping removes part of the carrier. On these unedited publications, SeLo-Mark preserves the reference needed for comparison with the received face. Its recovery advantage occurs under the same conditions as the detection advantage in Table 1.

Complete-payload recovery shows that this stability extends beyond average bit accuracy. The full receiver recovers all 48 bits in 99.82% of cases across the 22 conditions and 98.17% under cropping,

Table 2: Watermark robustness and publication quality on CelebA-HQ.

		Image quality ↑		BA (%) ↑				
Method	Bits	PSNR	SSIM	No proc.	Regular	Geometry	Crop	All 22
TrustMark	100	43.60	0.9938	**100.00**	99.64	75.38	75.24	94.13
FractalForensics	1024†	37.88	0.9617	**100.00**	98.68	57.98	57.02	89.40
SepMark	128	33.30	0.9124	**100.00**	99.97	50.55	49.97	88.69
EditGuard-degrade	64	32.27	0.7754	**100.00**	72.18	49.26	49.82	68.28
OmniGuard	100	42.78	0.9898	**100.00**	99.85	70.47	51.08	91.42
LIDMark	16‡	**44.78**	**0.9951**	**100.00**	99.27	52.81	50.24	88.51
LampMark (released)	128	36.37	0.9801	99.96	99.94	68.94	68.19	92.83
WAM	32	41.21	0.9764	**100.00**	99.85	88.16	85.13	96.93
MaskWM-ED32	32	38.35	0.9719	**100.00**	**100.00**	98.63	55.44	95.76
MaskWM-ED64	64	38.79	0.9693	**100.00**	**100.00**	95.26	51.36	94.93
SeLo-Mark	48	39.58	0.9782	**100.00**	**100.00**	**100.00**	**99.91**	**99.99**

PSNR is in dB. Recovery uses unedited watermarked images and each method's native binary payload. †: structured symbol bits; ‡: 16 ID bits alongside 136 landmark coordinates. All 22 averages the processing conditions equally. LampMark (released) uses the authors' reference payloads and watermark channel. Payload definitions are given in Appendix A.6.

compared with 88.38% overall for unadapted direct decoding. Most cropped inputs therefore retain the complete publication reference. Section 4.5 separates the effects of geometric correction and reader supervision.

4.4 Content Reference

Table 3 compares three 32-bit references on the 128-image subset while keeping the watermark channel, tracing field, masks, geometry predictor, and adapted reader fixed. Landmark–PCA32 uses 106 InsightFace landmarks; CLIP–RP32 replaces PCA with a fixed random orthogonal projection of the same centered CLIP features. Each variant embeds its own code before editing and uses an independently calibrated threshold.

Table 3: Content-reference comparison in the same watermark channel.

	Code change (bits)		Detection (%)	
Content reference	Unedited ↓	Edited ↑	AUC ↑	F1 ↑
Landmark–PCA32	**2.44**	9.01	94.88	88.56
CLIP–RP32	3.66	**12.69**	98.29	93.17
CLIP–PCA32	2.57	12.05	**99.16**	**95.71**

Code changes compare recomputed codes with the original embedded reference; detection uses recovered references. All columns average the four editor families, 22 conditions, and both runs on shared valid observations. Independently calibrated thresholds are $4/32$, $7/32$, and $6/32$, respectively.

The two pairwise comparisons in Table 3 distinguish the roles of the representation and projection. Compared with Landmark–PCA32, CLIP–PCA32 has similar variation on unedited inputs but responds more strongly to facial edits. Its detection gain thus accompanies greater edit sensitivity rather than lower benign variation. Compared with CLIP–RP32, PCA reduces both benign and edit-induced changes, yet improves AUC and F1. A larger response to editing alone is therefore insufficient: a useful reference must also control variation caused by benign processing. These code changes are measured against the original embedded reference, so they characterize the mapping independently of errors in reading that reference.

Figure 2 shows how benign variation affects the full comparison. Unedited inputs remain concentrated at small distances under regular and geometric processing, whereas cropping shifts them toward the fixed threshold. Together with the high crop-recovery rate, this suggests that retaining the original reference does not make the re-encoded face invariant to processing; stability of the received content code also matters for verification.

On the main CelebA-HQ evaluation, we replace the recovered reference with the original embedded code while fixing facial observations and the threshold (Table 4). AUC and F1 change little. Edited and unedited inputs remain separable without reference-reading errors, supporting verification through comparison with the published facial content.

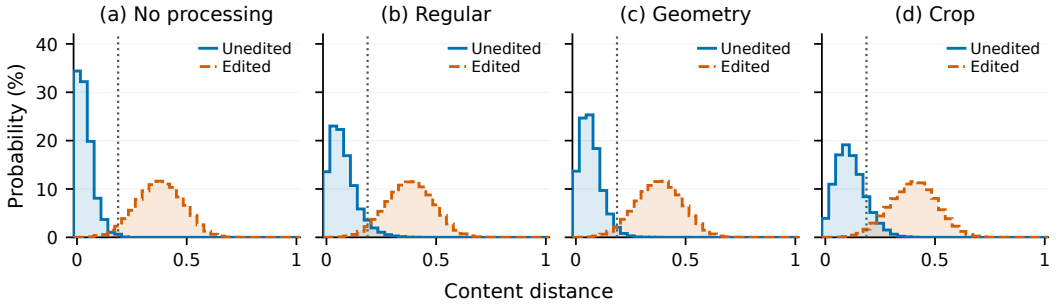

Figure 2: Facial consistency through the external reference. Distances compare the recovered 32-bit code with the code recomputed from the received face. Blue denotes unedited inputs and orange denotes edits. Histograms are normalized within each class and averaged over group conditions, editor families, and both receiver runs. The dotted line is the fixed threshold $6/32$.

Table 4: Verification using recovered or original content references.

Reference used for verification	AUC (%)	F1 (%)
Recovered reference	99.23	95.85
Original embedded reference	99.18	95.78

4.5 Shared Geometry

Shared correction. Table 5 separates the two uses of the predicted geometry on the 128-image subset, keeping the publications and adapted reader fixed. Message correction produces the dominant gain in detection and complete-payload recovery. With the corrected message candidate held fixed, aligning the facial branch further improves detection while leaving the recovered reference unchanged. This smaller gain comes from the observation used to recompute the content code. Correcting only the face has little effect while message recovery remains unreliable. The comparison thus depends on both access to the publication reference and alignment of the received face, with unequal contributions from the two branches.

Table 5: Receiver controls: facial-edit detection and message recovery.

	Receiver views		Detection (%) ↑		Recovery (%) ↑
Configuration	Message	Face	AUC	F1	Exact48
No correction	D	D	93.31	91.78	89.79
Message correction only	C	D	98.79	94.89	99.88
Face correction only	D	C	93.58	91.97	89.79
Both branches corrected	C	C	**98.92**	**95.42**	99.88
Full receiver	D or C	C	**98.92**	95.41	**99.89**

All rows share the watermarked inputs and adapted reader. D/C denote direct/corrected views. Detection pools common valid observations across conditions, weighting editor operations inversely by family size; Exact48 uses all unedited inputs.

Reader adaptation. Rectification restores coordinates but resamples the watermark signal. The direct-view and corrected-view supervision variants in Table 6 use the same geometry predictor at inference. Training on corrected crop views raises crop Exact48 from 80.96% to 99.32%, showing the value of adapting the reader to the resampled signal after alignment. At matched total embedding MSE, this receiver also achieves recovery comparable to the local SyncSeal-marker control without adding a synchronization watermark.

Controlled comparison. Providing FractalForensics with the same geometric frontend substantially narrows its AUC gap to SeLo-Mark (Table 7), showing that geometric recovery explains part of the difference. Further reader adaptation does not consistently improve its detection: at native strength, AUC increases slightly while F1 falls. SeLo-Mark retains higher overall AUC and F1, including when publication PSNR is matched to 40 dB and edits are regenerated. Alongside the reference comparison in Table 3 and branch controls in Table 5, these results support complementary roles for the content reference and receiver design: the code must distinguish edits from benign changes, while the receiver must recover and align the evidence used in that comparison.

Table 6: Reader supervision and synchronization on the separate geometric suite.

	Exact48 (%) ↑	
Configuration	Geometric suite	Crop subset
Direct-view supervision	96.29	80.96
SyncSeal control (local marker)	99.66	98.44
Corrected-view supervision (ours)	**99.86**	**99.32**

Same 128-image ablation subset and geometry predictor. The suite uses rotations of $\pm 5, \pm 15, \pm 30^\circ$, 0.9/0.75-side crops, 0.75/0.5 downsampling, and perspective, each with/without JPEG60. Geometric suite averages all 22 conditions in this separate suite; Crop subset averages its four crop settings. Embedding amplitude is 1.0; the marker control matches total embedding MSE.

Table 7: Detection with shared geometry and matched publication quality.

	Native strength		40 dB PSNR	
Configuration	AUC	F1	AUC	F1
FractalForensics, direct	89.02	89.32	88.23	87.69
+ shared correction	95.81	90.35	94.86	88.41
+ correction + reader adaptation	95.93	89.57	91.94	85.55
SeLo-Mark	**99.05**	**95.54**	**99.08**	**95.46**

Percentages on shared valid inputs from the 128-image subset; main family/condition averaging. Configurations with correction use the same geometric frontend; reader adaptation uses the same 20,000-update budget and transformation schedule. At 40 dB, edits are regenerated and thresholds independently recalibrated. Readers are reused.

4.6 Cross-Dataset Transfer

FFHQ tests transfer of the content reference, geometric receiver, and operating threshold together. The content mapping, publisher, receiver, and threshold are all determined on CelebA-HQ and used without further fitting; both protection and observation masks are predicted from RGB. AUC remains high across all four processing groups (Table 8), while fixed-threshold F1 is lower than on CelebA-HQ. Complete-payload recovery remains high, with most remaining errors concentrated under cropping.

Table 8: SeLo-Mark transfer to FFHQ.

	Detection (%) ↑		Recovery (%) ↑	
Group	AUC	F1	BA	Exact48
No proc.	99.50	91.19	100.00	100.00
Regular	97.60	90.10	100.00	100.00
Geometry	98.49	91.02	100.00	100.00
Crop	96.04	89.37	99.71	94.72
All 22	97.67	90.21	99.97	99.52

The CelebA-HQ models and calibration thresholds remain fixed. Detection uses the same four editor families and processing groups as the main comparison; recovery uses unedited watermarked images.

Additional affine transformations, composed transmission pipelines, regeneration, and runtime measurements are reported in Appendix E.

5 Conclusion

We presented SeLo-Mark, a localized watermarking framework that carries a compact facial-content reference outside the protected facial region. Shared geometric correction aligns watermark recovery and facial encoding, enabling direct comparison between the recovered reference and the received face. Experiments on facial editing and image processing, including transfer to FFHQ, demonstrate that localized watermarking can support content verification while leaving the protected facial pixels unchanged at publication.

AI USE STATEMENT

The authors acknowledge the use of large language models (LLMs) in the preparation of this manuscript. Specifically, LLMs were employed for language editing and writing refinement purposes, including but not limited to grammar correction, sentence restructuring, and stylistic improvements to enhance the clarity and readability of the text. All technical content, experimental design, data analysis, results interpretation, and scientific conclusions presented in this work are entirely the authors' own contributions. The use of LLMs was limited to linguistic assistance and did not involve the generation of research ideas, methodological approaches, or substantive content. All factual claims and technical details have been independently verified by the authors.

ETHICS STATEMENT

This work studies content verification for facial images published with authorization. The watermark score measures consistency with the published facial reference. Its reliability across demographic groups remains to be evaluated.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–F describe the implementation, data splits, inference procedures, evaluation conditions, and additional results.

REFERENCES

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*, pp. 8587–8597, 2026.

Kasra Arabi, R. Teal Witter, Chinmay Hegde, and Niv Cohen. SEAL: Semantic aware image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16196–16205, 2025.

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*, 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. doi: 10.1145/3394171.3413630.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

Pierre Fernandez, Tomáš Souček, Nikola Jovanović, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, and Alexandre Mourachko. Geometric image synchronization with deep watermarking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 9522–9526, 2026. doi: 10.1109/ICASSP55912.2026.11464560.

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

Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Mask image watermarking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146313–146346, 2025. doi: 10.52202/085713-4895.

Ziyao Huang, Fan Tang, Yong Zhang, Juan Cao, Chengyu Li, Sheng Tang, Jintao Li, and Tong-Yee Lee. Identity-preserving face swapping via dual surrogate generative models. *ACM Transactions on Graphics*, 43(5):1–19, 2024. doi: 10.1145/3676165.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5549–5558, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiyang Luo, Michael Goebel, Elnaz Barshan, and Feng Yang. LECA: A learned approach for efficient cover-agnostic watermarking. *Electronic Imaging*, 35(4):376–1–376–6, 2023. doi: 10.2352/EI.2023.35.4.MWSF-376.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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*, pp. 10515–10524, 2024. doi: 10.1145/3664647.3680869.

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*, pp. 7210–7219, 2025. doi: 10.1145/3746027.3754544.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 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*, 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*, pp. 1190–1201, 2023. doi: 10.1145/3581783.3612471.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Computer Vision – ECCV 2018*, pp. 334–349, 2018. doi: 10.1007/978-3-030-01261-8_20.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024.

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3008–3018, 2025.

Chengxin Zhao, Hefei Ling, Sijing Xie, Han Fang, Yaokun Fang, and Nan Sun. SSyncOA: Self-synchronizing object-aligned watermarking to resist crop-paste attacks. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2024. doi: 10.1109/ICME57554.2024.10687716.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *Computer Vision – ECCV 2018*, pp. 682–697, 2018. doi: 10.1007/978-3-030-01267-0_40.

A Implementation and Evaluation Protocols

We provide the data splits, processing settings, and implementation details for the publication and verification pipeline in Sections 3 and 4.

A.1 Data and Evaluation

Data splits. Table 9 summarizes the data used at each stage. The CelebA-HQ training, validation, calibration, and test splits have disjoint annotated identities. PCA is fitted on training observations, and the 128-image ablation subset is drawn from the main test set. Cross-dataset evaluation uses 1,000 FFHQ images selected in a fixed order and screened for exact duplicates.

Table 9: Data used for training and evaluation. All splits are from CelebA-HQ except FFHQ.

Split	Images	Identities	Use
Channel / PCA fitting	20,999	4,273	Watermark training and PCA fitting
Receiver adaptation	20,996	4,272	Coordinate and reader adaptation
Validation	2,789	440	Code stability and discrimination
Calibration	128	128	Thresholds and matched-quality controls
Main test set	2,999	648	Detection and recovery comparisons
Ablation subset	128	128	Receiver and geometry controls
FFHQ transfer	1,000	—	Cross-dataset evaluation

Quality and recovery for SeLo-Mark are evaluated on 2,996 CelebA-HQ and 995 FFHQ watermarked images.

Watermarking and editing. PCA fitting masks and CelebA-HQ protection regions are constructed from CelebAMask-HQ facial-part annotations (Lee et al., 2020). The CelebA-HQ carrier is the complement of the supplied protection mask, while facial observations are predicted from RGB. Both masks are predicted on FFHQ. Each editor acts on the watermarked image produced by the method being evaluated, after which transmission processing is applied. Methods share the source–target assignments for face swapping and the random processing parameters associated with each source image. Face swapping uses SimSwap (Chen et al., 2020) and CSCS (Huang et al., 2024); attribute editing uses mask-based StarGAN (Choi et al., 2018) and the Stable Diffusion 1.5 inpainting model (Rombach et al., 2022). The inpainting pipeline uses DDIM sampling (Song et al., 2021).

Table 10: Processing conditions used in the 22-condition evaluation.

Group	Conditions	Count
No processing	Identity	1
Regular: compression	JPEG qualities 80 and 60, both with 4:2:0 chroma subsampling	2
Regular: resizing	Bicubic resize to 0.75 and back; resize to 0.5 and back	2
Regular: filtering	PIL Gaussian blur radii 0.8 and 1.8; Gaussian blur radius 1.2 followed by sharpening; Gaussian filtering (kernel 3, $\sigma = 1$); median filtering (kernel 3)	5
Regular: noise	Gaussian RGB noise with standard deviations $3/255$ and 0.025; per-channel salt-and-pepper noise with total probability 0.05	3
Regular: color	Separate brightness, contrast, and saturation factors uniform in $[0.7, 1.3]$; hue offset uniform in $[-0.1, 0.1]$	4
Geometry	Integer rotation in $[-30, 30]$ degrees; perspective distortion scale in $[0.1, 0.3]$; horizontal flip	3
Crop	Random-position crops with side fraction 0.75, bilinearly resized to 512; also followed by JPEG quality 60 (4:2:0)	2

The five regular-processing groups contain 16 conditions in total. Together with no processing, geometry, and crop, the suite contains 22 equally weighted conditions.

Transmission processing. Table 10 gives the conditions used in the main comparison and FFHQ evaluation. Publication and reception use 8-bit images. The 0.75-side crop retains 56.25% of the image area; the remaining facial content and watermark carrier depend on the crop position and protection mask.

Detection metrics. We compute AUC and F1 from available detection scores separately for each editing operation and processing condition. Edited and unedited examples receive equal total class weight when computing F1. The Male and Young operations are averaged within StarGAN, and smile, eyeglasses, and older-face operations within Stable Diffusion. We then average the four editor families and 22 conditions equally.

Threshold calibration and independent runs. We select each method's threshold on the independent calibration set using Youden's J, as in LIDMark (Wu et al., 2026). For each operation–condition pair, J is the fraction of edited images with scores above τ minus the corresponding fraction of unedited images. We average these values using the same family and condition weights as the detection metrics, then choose the threshold that maximizes the average, taking the larger value in a tie. An edit is flagged when the score exceeds τ.

SeLo-Mark uses two independent receiver-adaptation runs that share the content mapping and publisher. Calibration selects $\tau = 6/32$ for both: more than six disagreeing content bits indicate an edit. We keep the threshold fixed on the main test set and FFHQ, compute metrics separately for each run, and average the results.

A.2 Facial Observations and Content Mapping

Facial masks. We use annotated facial masks for PCA fitting and predict the observation mask from RGB at publication and reception. SCRFD (Guo et al., 2022) selects the face with the highest detection confidence, using a 320×320 input, a confidence threshold of 0.5, and a non-maximum-suppression threshold of 0.4. We enlarge the detected bounding box by a factor of 1.25 and retain the facial classes predicted by BiSeNet (Yu et al., 2018) within it. The 19-class BiSeNet model distributed with SimSwap (Chen et al., 2020) supplies labels $\{1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13\}$. Publication proceeds when a face is detected and its observation mask is nonempty.

These masks serve different purposes: the observation mask selects the content to encode, the protection mask specifies the pixels preserved at publication, and the predicted watermark support selects pixels for message reading.

Semantic input. We resize RGB images to 256×256 by bilinear interpolation and round them to the 8-bit grid. Masks are resized by nearest-neighbor interpolation and binarized at 128. We normalize RGB to $[-1, 1]$, apply the facial mask, and crop to its bounding box with a margin of 15% of the longer side, or at least two pixels. The crop is resized to 224×224 using antialiased bilinear interpolation. Pixels outside the mask are zero in the normalized representation, corresponding to neutral gray in RGB.

Content mapping. We take the L2-normalized, 1024-dimensional pooled output of the frozen CLIP ViT-L/14 encoder (Radford et al., 2021). The mean μ and the leading 32 eigenvectors of the centered covariance are fitted on the 20,999 training observations. We project each centered feature onto these eigenvectors without whitening and binarize the projected coordinates at zero. The feature extractor, mean, projection, and binarization rule are shared by publication and reception.

A.3 Local Channel and Publication

The watermark channel follows MaskWM's architecture (Hu et al., 2025), operates at 256×256, uses GroupNorm (Wu & He, 2018), and carries 48 bits. Stage 1 trains this channel on the CelebA-HQ split used for PCA fitting.

Training objective. We train the channel with random 48-bit messages and a mixture of local and full-image carriers. Let c denote the encoder candidate at the native training resolution. Before quantization, we form the continuous-valued composite $x_{\mathrm{tr}}^{\mathrm{w}} = (1 - M) \odot x + M \odot c$. A sampled processing operation T acts on this composite; geometric operations also transform its carrier mask:

$$(y, M') = T(x_{\mathrm{tr}}^{\mathrm{w}}, M), \qquad q = D_\omega(y, M'), \qquad \widetilde{M} = L_\eta(y). \tag{11}$$

The known mask M' provides the reader's support and the localizer's training target. The per-example losses are

$$\mathcal{L}_{\text{img}} = \frac{\|c - x\|_F^2}{3HW}, \qquad \mathcal{L}_{\text{msg}} = \frac{\|q - m\|_2^2}{48}, \qquad \mathcal{L}_{\text{loc}} = \frac{\|\widetilde{M} - M'\|_F^2}{HW}. \tag{12}$$

The image loss measures the difference between the encoder candidate and the original image before composition. The message and localization losses supervise the recovered scores and support map, respectively. We jointly optimize (ϕ, η, ω) with the batch-averaged objective

$$\mathcal{L}_{\text{wm}} = \frac{1}{B} \sum_{i=1}^{B} \left[\mathcal{L}_{\text{img}}^{(i)} + \lambda_{\text{wm}}(u) \big(\mathcal{L}_{\text{msg}}^{(i)} + 0.5 \mathcal{L}_{\text{loc}}^{(i)} \big) \right], \tag{13}$$

where B is the batch size and u is the training update index.

Optimization and processing. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01 and a batch size of 16 for 100,000 updates, retaining the final checkpoint. The learning rate warms up to 10^{-4} over 2,000 updates and then follows a cosine schedule. The loss weight in Equation 13 is

$$\lambda_{\text{wm}}(u) = \max\{0.2, 20 - 19.8u/5000\}. \tag{14}$$

Full-image carriers are used for the first 500 updates. Thereafter, we sample irregular regions, boxes, the full image, or the annotated facial region with equal probability as carriers. Non-full-image masks are inverted with probability 0.5; for a facial mask, this selects pixels outside the annotated face. Image distortions are enabled at update 1,001 and JND weighting at update 5,001. The channel pretraining JND factor is 1.0.

Once distortions are enabled, one of ten operations is selected uniformly: identity, JPEG frequency masking, JPEG at quality 50 with chroma subsampling factor 2, 5×5 Gaussian filtering with $\sigma = 1$, Gaussian noise with standard deviation 0.05 in $[0, 1]$, 5×5 median filtering, salt-and-pepper noise with total probability 0.1 per channel, integer-angle rotation in $[-90, 90]$ degrees, perspective distortion with scale in $[0.1, 0.5]$, or horizontal flipping. JPEG frequency masking retains the 5×5 low-frequency coefficients of Y and the 3×3 coefficients of U/V in each 8×8 block.

Publication. For publication, we resize the 512×512 image and carrier to the channel's 256×256 resolution. We apply JND weighting to the encoder residual at this resolution, with factor 1.75 and relative RGB weights $(0.75, 0.50, 1.00)$. The residual is then upsampled, added to the original image, and clipped to the valid range. Composition through the original-resolution binary carrier preserves the protected pixels, and the image is quantized and saved as PNG, as in Equation 4.

A.4 Geometry and Reader Adaptation

We use a ConvNeXt V2-Tiny coordinate predictor (Woo et al., 2023), initialized from SyncSeal (Fernandez et al., 2026). Its input is the 512×512 image converted to $[0, 1]$ and resized to 256×256 by antialiased bilinear interpolation. The eight output coordinates represent four point correspondences.

The coordinate predictor and message reader are adapted in sequence on the training subset in Table 9. Each stage uses 20,000 updates, a batch size of 16 with two-image microbatches, and AdamW with weight decay 0.01. The learning rate warms up to 10^{-4} over 1,000 updates and then decays to 10^{-5} with a cosine schedule. We clip the global gradient norm at 1. The first stage updates only the coordinate predictor; the second updates only the message reader, with the publisher, localizer, and coordinate predictor fixed.

In both stages, each training publication carries a random 48-bit message and produces a base view and a cropped view. The base view is unchanged with probability 0.25; otherwise, one of the 13 operations in Table 11 is selected uniformly. For the cropped view, we independently sample an integer side length uniformly from 256 to 460 and a valid top-left position uniformly, then resize the crop to 512×512 by bilinear interpolation. No additional compression or noise is applied to this view.

The base and cropped views receive equal loss weights. Reader adaptation supervises the base view's direct output and the crop's corrected output, as in Equation 7. Training uses continuous-valued publisher outputs; evaluation uses quantized PNG publications.

Table 11: Base-view processing operations for receiver adaptation.

Operation	Parameters
JPEG	Quality 60, chroma subsampling factor 2
Gaussian blur	Kernel 3×3, $\sigma = 1$
Gaussian noise	Standard deviation 0.025 in $[0, 1]$, then clipping
Median filter	Kernel 3×3
Salt-and-pepper	Total probability 0.05 per channel
Resize	Scale 0.5, then restore size with antialiasing
Brightness; contrast; saturation	Separate factors uniform in $[0.7, 1.3]$
Hue	Offset uniform in $[-0.1, 0.1]$
Rotation	Uniform integer angle in $[-30, 30]$ degrees
Perspective	Distortion scale uniform in $[0.1, 0.3]$
Horizontal flip	Left–right reflection

Thirteen alternatives are sampled for the non-identity base view. Brightness, contrast, hue, and saturation are four separate alternatives.

A.5 Coordinate Conventions and Inference

Correspondences use the canvas boundary corners $b = ((0,0),(512,0),(512,512),(0,512))$ in clockwise order from the top-left. A target point $p_k^\star$ is the location in the source publication corresponding to received corner b_k. For each coordinate component, the inherited output convention is

$$\mathcal{N}(p) = \frac{255p}{512 \cdot 128} - 1, \qquad \mathcal{N}^{-1}(z) = \frac{512 \cdot 128}{255}(z+1). \tag{15}$$

We use this convention in training and inference. Rectification maps the received boundary corners b to the predicted source locations $\widehat{p}$ with a four-point perspective warp.

For message reading, we warp normalized RGB using bilinear interpolation and zero filling in $[-1, 1]$, then resize the 512×512 result to the 256×256 reading resolution. Each candidate uses its own predicted support, binarized at > 0.5. The corrected candidate is selected when its confidence exceeds the direct candidate's by more than $\epsilon = 10^{-6}$; otherwise, the direct candidate is retained. We binarize the selected message scores at ≥ 0.5.

For facial encoding, the same point correspondences warp the received image and its predicted face mask with bilinear interpolation. Both are quantized to the 8-bit grid before semantic preprocessing. The corrected facial observation is used for either message candidate. Verification compares the 32 content bits; the remaining 16 bits form the independent tracing field.

A.6 Baseline Methods

Table 12 summarizes how each comparison method produces an image-level facial-edit score from watermark discrepancies, tamper-localization maps, or embedded facial information.

LampMark configurations. LampMark (adapted), used for detection, replaces the original Face++ API with 106-point InsightFace landmarks and fits the PCA mapping on the training split. We initialize the 256-pixel encoder and decoder from the released weights and adapt the channel to this fixed 128-bit mapping for 20 epochs on 20,982 training images. A separate 512-image validation set selects the checkpoint by minimizing image MSE plus 12 times mean message MSE over five conditions, yielding epoch 7. Training, validation, calibration, and test identities are disjoint. The calibration procedure above selects a detection threshold of $11/128$. We transfer embedding residuals to the common 512-pixel interface, obtaining 2,998 watermarked test images, and generate detection edits from these publications.

LampMark (released), used for recovery and regeneration, retains the authors' 128-bit reference payloads and original channel weights. The released payloads cover 2,997 test images; 2,996 yield finite unprocessed decodes, with 99.96% BA at both the native 256-pixel and shared 512-pixel

Table 12: Comparison methods and facial-edit detection scores.

Method	Core mechanism and detection score
FractalForensics (Wang et al., 2025)	Embeds spatially structured fractal watermarks for detection and localization. We score the fraction of recovered symbols that disagree with the reference pattern.
SepMark (Wu et al., 2023)	Uses a shared encoder, a robust Tracer, and a semi-robust Detector. The detection score is the Detector's bit disagreement with the embedded message.
EditGuard-degrade (Zhang et al., 2024)	Combines a localization watermark with copyright bits. We use the fraction of pixels flagged by discrepancies in the recovered localization watermark.
OmniGuard (Zhang et al., 2025)	Combines watermark embedding with degradation-aware tamper extraction. We use the mean predicted tamper probability.
LIDMark (Wu et al., 2026)	Embeds facial landmark coordinates and source ID bits. We use the mean Euclidean distance between recovered and redetected landmarks.
RecoverMark (An et al., 2026)	Embeds facial content into the background for localization and recovery. We use the mean absolute RGB difference between the recovered and received faces within the predicted facial mask.
LampMark (adapted) (Wang et al., 2024)	Embeds 128-bit references derived from PCA-projected facial landmarks. We use normalized Hamming disagreement between the recovered and recomputed references.
SeLo-Mark	Embeds 32 content bits and 16 tracing bits outside the protection region. We score normalized Hamming disagreement between recovered and recomputed content codes.

resolutions. Recovery metrics use finite decoder outputs. Publication quality is measured separately for the adapted and released configurations because both their payload mappings and channel weights differ.

Detection interfaces. EditGuard-degrade uses a localization-residual threshold of 0.4 per channel and the union of the resulting masks. RecoverMark uses the predicted facial mask defined in Appendix A.2.

Recovery payloads and image quality. Recovery uses the native payloads listed in Table 2, including TrustMark's 100-bit BCH codeword. SepMark retains its 256-pixel publisher, with image quality evaluated at 512 pixels; Tracer provides recovery and Detector provides facial-edit detection. OmniGuard uses its copyright branch for recovery and its localization interface for detection.

B Content-Reference Analyses

We extend the content-reference comparison in Section 4.4 with results by processing group, code statistics, and reference recovery after facial editing.

B.1 Alternative References in the Same Channel

We compare three ways to construct the 32-bit content reference while keeping the 48-bit watermark channel, 16-bit tracing field, protection masks, geometry predictor, and adapted reader fixed. CLIP–RP32 replaces the PCA projection in CLIP–PCA32 with a fixed random orthogonal projection of the same centered CLIP features. Landmark–PCA32 applies PCA to 106 two-dimensional facial landmarks extracted by InsightFace. Feature means and both PCA mappings are fitted on training data; the random projection is obtained by orthonormalizing a Gaussian random matrix once and then kept fixed. Each variant embeds its own reference before applying the same editing and transmission pipeline.

CLIP–PCA32 gives the highest AUC and F1 in every processing group. Landmark–PCA32 is competitive without processing but loses more detection accuracy under geometry and cropping; both CLIP-based references retain stronger separation in these groups. Relative to CLIP–RP32, PCA reduces benign variation while retaining sensitivity to edits, as shown in Table 3. The ordering also

Table 13: Facial-edit detection with different content references in the same 48-bit watermark channel.

Content code	No proc.	Regular	Geometry	Crop	Average
AUC (%) ↑					
Landmark–PCA32	98.57	97.29	87.19	85.28	94.88
CLIP–RP32	99.74	98.18	99.14	97.12	98.29
CLIP–PCA32	**99.75**	**99.20**	**99.43**	**98.08**	**99.16**
F1 (%) ↑					
Landmark–PCA32	92.52	91.05	83.20	74.73	88.56
CLIP–RP32	96.01	93.18	94.95	89.04	93.17
CLIP–PCA32	**97.31**	**95.81**	**96.59**	**92.84**	**95.71**

Detection uses recovered references and the same valid observations for all three codes in the 128-image ablation subset. Groups and averaging follow Appendix A.1; Average weights all 22 conditions equally. Independently calibrated thresholds are $4/32$ for Landmark–PCA32, $7/32$ for CLIP–RP32, and $6/32$ for CLIP–PCA32.

holds when the original embedded codes replace the recovered references, indicating that it is not solely a consequence of watermark-reading errors.

B.2 Code Usage and Cross-Identity Distances

Applying the fixed PCA mapping to all 2,999 test images with annotated facial masks produces 2,993 distinct codes: 2,987 occur once, and six occur twice. Mean per-bit entropy is 0.99824 bits, so individual bits are nearly balanced across this set.

For the codes embedded in the 2,996 test publications, the facial masks are predicted from RGB. Codes from different annotated identities differ by 15.94 bits on average; 0.079% of these pairs differ by at most six bits. By comparison, re-encoding unedited publications after transmission changes an average of 2.46 bits across the 22 processing conditions and both receiver runs. The content reference therefore varies more across these identities than under the tested processing of the same face.

B.3 Reference Recovery after Facial Editing

Facial verification compares the received face with the reference embedded before editing. Table 14 measures how much of that original reference survives each editor and the subsequent transmission conditions.

Table 14: SeLo-Mark reference recovery after facial editing and transmission.

	Content reference (%) ↑		Full payload (%) ↑	
Editor	BA32	Exact32	BA48	Exact48
SimSwap	99.94	99.15	99.94	98.98
CSCS	99.94	99.18	99.94	99.02
StarGAN	99.98	99.69	99.98	99.61
Stable Diffusion	94.67	67.18	94.63	63.12
Mean	98.63	91.30	98.62	90.18

Recovery is measured against the payload embedded before editing. Results average the 22 processing conditions and two receiver runs, with operations averaged within each editor family. Exact32 requires all content bits to be correct; Exact48 additionally requires all tracing bits to be correct.

Editing and subsequent processing. Without subsequent processing, SimSwap, CSCS, and StarGAN preserve all 48 bits, while Stable Diffusion yields 82.70% Exact32 and 78.96% Exact48 (Figure 3). Subsequent cropping reduces Exact32 to 90.77–96.64% for the first three families and to 5.32% for Stable Diffusion. Thus, the reference usually survives the tested swaps and StarGAN edits, whereas inpainting and cropping together can remove much of the information needed for exact recovery. Recovery includes all generated edits, independently of facial-observation availability, and follows the averaging in Appendix A.1.

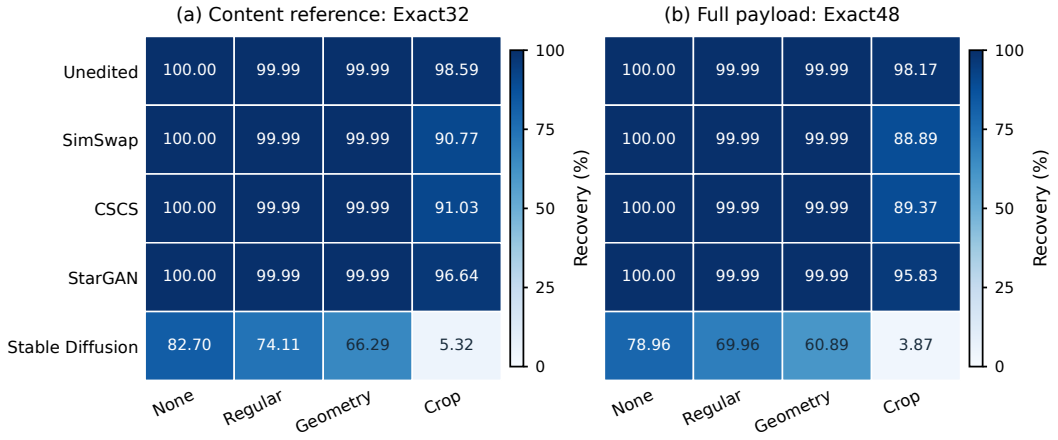

Figure 3: Recovery of the original reference after editing and transmission: (a) all 32 content bits and (b) all 48 payload bits. Colors share a 0–100% recovery scale. "None" has no processing after editing; unedited rows apply processing only. Cells average group conditions, family operations, and both runs.

Content changes and reading errors. The reference-replacement control in Table 4 shows that aggregate detection changes little when reference-reading errors are removed, despite the lower exact recovery after some edits.

C Shared-Geometry and Reader Controls

The following controls distinguish the effects of branch alignment, message-candidate selection, reader supervision, and coordinate-predictor adaptation.

C.1 Full-Receiver Recovery on the Main Test Set

Table 15 compares the unadapted direct reader with the full receiver on the same unedited publications. Direct reading already handles regular processing well. Geometric correction and reader adaptation are most consequential after transformations that displace or remove the carrier, particularly cropping.

Table 15: Exact 48-bit recovery with direct decoding and the full receiver.

	Exact48 (%) ↑				
Configuration	No proc.	Regular	Geometry	Crop	All 22
Base decoder (direct)	**100.00**	99.98	81.56	0.00	88.38
Full receiver	**100.00**	**99.99**	**99.99**	**98.17**	**99.82**

Both configurations use the same unedited watermarked images from the main test set. The base decoder precedes receiver adaptation; the full receiver includes geometric correction, reader adaptation, and candidate selection. Processing groups follow Table 10.

C.2 Branch Alignment and Candidate Selection

Branch-correction comparison. For Table 5, we retain the common set of observations with valid scores for all evaluated receiver configurations. Scores are pooled across the 22 processing conditions, with each editing operation weighted by the inverse of the number of operations in its editor family. We then compute AUC and class-balanced F1, using the fixed threshold $6/32$. Metrics are computed separately for each independent run and averaged. Exact48 uses all unedited watermarked inputs from the ablation subset, regardless of facial-observation availability.

Once the message branch is corrected, facial alignment improves AUC by 0.13 percentage points and F1 by 0.54 points, with paired 95% confidence intervals of [0.06, 0.21] and [0.25, 0.82], respectively. This smaller gain improves the comparison with the received face while leaving the recovered reference unchanged. The intervals use 2,000 identity-level bootstrap resamples, retaining all edits, processing conditions, and both receiver runs for each sampled identity.

Direct and corrected message candidates. Table 16 compares always-direct, always-corrected, and confidence-selected reading on unedited publications from the 128-image ablation subset. Within each run, all three strategies use the same adapted reader and watermarked inputs under the main 22 processing conditions.

Table 16: Exact 48-bit recovery (%) by message-candidate strategy on unedited watermarked images.

Message candidate	No proc.	Regular	Geometry	Crop	All 22
Always-direct	**100.00**	**100.00**	91.80	0.00	89.79
Always-corrected	**100.00**	99.98	**100.00**	**98.83**	99.88
Confidence-selected	**100.00**	**100.00**	**100.00**	**98.83**	**99.89**

Processing groups follow Table 10; All 22 averages the 22 conditions equally. Results average two independent runs.

Always-corrected reading accounts for nearly all of the recovery gain, particularly under cropping. Candidate selection retains direct-reading performance under regular processing and corrected-reading performance under geometry and cropping. Its additional average gain is small: 0.018 percentage points over always-corrected reading (paired 95% interval [0.000, 0.053]).

C.3 Reader Supervision and Synchronization

The geometric receiver comparison uses the ablation subset and a separate suite of 22 conditions: rotations of ± 5, ± 15, and ± 30 degrees; crops retaining 0.9 and 0.75 of the side length; downsampling by 0.75 and 0.5; and a perspective transformation, each with and without JPEG quality 60. Table 17 uses embedding amplitude 1.0 relative to the Stage 1 publication residual.

Direct-view and corrected-view supervision differ in whether cropped training inputs are corrected before reading; both use our geometry predictor at inference. The SyncSeal controls add a synchronization watermark to the local channel and adapt the reader to the resulting corrected inputs. The full-image marker control uses SyncSeal's geometry predictor; the local marker control uses ours and matches the total embedding MSE budget.

Table 17: Recovery with different reader supervision and synchronization configurations.

	Exact48 (%) ↑		BA (%) ↑
Configuration	Geometric suite	Crop subset	Geometric suite
Base decoder (direct)	37.36	0.39	89.60
Direct-view supervision	96.29	80.96	99.85
SyncSeal control (full-image marker)	63.99	97.75	96.21
SyncSeal control (local marker)	99.66	98.44	**99.99**
Corrected-view supervision (ours)	**99.86**	**99.32**	**99.99**

Reference payloads on the ablation subset at amplitude 1.0. Geometric suite averages all 22 conditions in this separate suite; Crop subset averages its four crop conditions. This suite differs from Table 10.

Corrected-view supervision improves crop recovery relative to direct-view supervision while using the same geometry predictor at inference. At amplitude 1.0, it also gives recovery comparable to the local SyncSeal-marker control without adding a synchronization watermark. Its Exact48 difference relative to that control is 5.33 percentage points at amplitude 0.75, 0.20 at 1.0, and -0.02 at 1.25; the advantage is therefore most apparent at the lower embedding amplitude.

C.4 Geometry Adaptation Inputs

We examine how geometry adaptation inputs affect alignment and subsequent message recovery. Starting from the same pretrained coordinate predictor, we perform 20,000 updates using watermarked RGB, watermark-free RGB, or RGB with shuffled watermark residuals. Each variant has two independent runs and is evaluated on the same watermarked inputs. Table 18 reports geometric error and recovery averaged over the two runs.

Geometric error and recovery are similar across all three inputs. Starting from the same pretrained predictor, adaptation on watermark-free RGB retains nearly the same performance as adaptation on

Table 18: Geometric accuracy and recovery with different adaptation inputs.

Adaptation input	Corner error ↓	Reprojection error ↓	Exact48 (%) ↑
Watermarked RGB	23.87	13.35	**99.91**
Watermark-free RGB	23.90	13.37	99.88
Shuffled watermark residual	**23.83**	**13.33**	99.89

Errors are measured in pixels on the 512-pixel coordinate frame. Corner error averages the four corner displacements; reprojection error averages displacements on a 9×9 grid. Results average two independent runs on the same 128-image cohort and 22 processing conditions, using watermarked test inputs.

watermarked RGB. Shuffling the residual has little effect either, suggesting that image appearance provides sufficient information for this adaptation.

D Matched Comparisons and Carrier Survival

We examine how shared geometric correction and matched publication quality affect the comparisons in Section 4.5, then relate crop recovery to the surviving carrier area.

D.1 Controlled Baseline Comparisons

Shared geometry and reader adaptation. We evaluate three FractalForensics configurations: direct decoding, decoding after the shared geometric correction, and correction with an adapted reader. All retain FractalForensics' publisher, native structured payload, and 512-pixel reading interface. Correction uses SeLo-Mark's frozen geometry predictor.

Each of two adaptation runs uses SeLo-Mark's 20,996 training images and follows the same image order and transformation schedule for 20,000 updates. Each update uses 16 images with two supervised views per image: a directly processed base view and a corrected cropped view. The two message-MSE losses receive equal weight. The optimizer and learning-rate schedule follow Appendix A.4. FractalForensics uses a fixed native training pattern per image, drawn from patterns disjoint from calibration and test. Only the reader is updated; the publisher and geometric frontend remain fixed. We evaluate the final checkpoint from each run and select each configuration's threshold by Youden's J on the independent 128-image calibration set.

Table 19: FractalForensics with shared geometric correction and matched-budget reader adaptation.

Configuration	No proc.	Regular	Geometry	Crop	Average
AUC (%) ↑					
FractalForensics, direct	**100.00**	**99.30**	55.97	50.93	89.02
+ shared correction	**100.00**	99.14	81.70	88.17	95.81
+ correction + reader adaptation	99.97	98.60	82.18	93.20	95.93
SeLo-Mark	99.69	99.13	**99.24**	**97.88**	**99.05**
F1 (%) ↑					
FractalForensics, direct	**98.80**	**95.79**	66.78	66.67	89.32
+ shared correction	98.75	95.69	74.83	66.67	90.35
+ correction + reader adaptation	96.18	93.21	83.22	66.67	89.57
SeLo-Mark	97.14	95.68	**96.36**	**92.40**	**95.54**

Results use shared valid observations from the 128-image ablation subset at each method's original publication strength. Metrics average processing conditions, the four editor families, and two runs, following Appendix A.1. Average gives equal weight to all 22 conditions; groups follow Table 10.

Shared correction substantially improves FractalForensics' AUC under geometric transformations and cropping. Reader adaptation further increases crop AUC, although crop F1 remains unchanged at the calibrated threshold and regular-processing performance declines. SeLo-Mark retains higher overall AUC and F1 when the methods share geometric preprocessing and the adaptation budget.

Matched publication PSNR. We scale each method's native publication residual to obtain 40 dB PSNR after 8-bit quantization, with per-image deviations below 0.05 dB. Pixels with zero original residual remain unchanged, preserving SeLo-Mark's protected facial region. We regenerate all facial

edits from the matched publications, reuse the fitted readers, and independently recalibrate each configuration at this strength.

Table 20: Facial-edit detection at a matched publication PSNR of 40 dB.

Configuration	No proc.	Regular	Geometry	Crop	Average
AUC (%) ↑					
FractalForensics, direct	**99.91**	98.33	54.91	51.63	88.23
+ shared correction	99.88	98.19	79.64	88.52	94.86
+ correction + reader adaptation	98.10	94.25	77.83	91.63	91.94
SeLo-Mark	99.78	**99.13**	**99.25**	**98.11**	**99.08**
F1 (%) ↑					
FractalForensics, direct	**97.28**	93.64	66.78	66.67	87.69
+ shared correction	95.70	92.97	76.16	66.67	88.41
+ correction + reader adaptation	94.36	88.83	77.74	66.67	85.55
SeLo-Mark	97.20	**95.59**	**96.10**	**92.66**	**95.46**

Each saved publication is matched to 40 dB PSNR within 0.05 dB. All edits are regenerated from these publications. The 128-image test subset and aggregation follow Table 19; each configuration uses its own threshold selected on the independent calibration set at this publication strength.

The matched-quality comparison preserves SeLo-Mark's advantage in overall AUC and F1. Its detection performance changes little from the native-strength comparison, while the FractalForensics reader adapted at native strength is less effective after the residual is rescaled. Shared correction remains beneficial to FractalForensics' AUC under geometry and cropping. Together, the two comparisons show that the overall detection difference persists after controlling both geometric preprocessing and publication PSNR.

D.2 Crop Recovery at Matched Image Quality

The robustness comparison in Table 2 uses each method's native payload and publication strength. Here, we hold publication PSNR fixed to examine crop recovery at comparable embedding distortion. We scale each method's residual until the saved PNG is within 0.05 dB of the target. Table 21 compares WAM-32, MaskWM-ED32, and SeLo-Mark on the calibration set using fixed reference payloads. Bit accuracy uses all 32 bits for the baselines and the first 32 bits of our 48-bit payload, averaged over the two 0.75-side crop conditions.

Table 21: Message recovery at matched image quality.

	Crop BA (%) ↑		
PSNR (dB)	WAM-32	MaskWM-ED32	SeLo-Mark
38.0	87.71	54.70	**99.98**
40.0	85.62	54.21	**99.88**
42.0	81.84	53.17	**99.37**
46.5	65.77	50.39	**90.07**
48.5	54.02	49.87	**81.32**
50.5	49.07	49.88	**72.41**

Image quality is matched by the PSNR of saved PNG images on the calibration set. BA averages the two 0.75-side crop conditions, using all 32 bits for WAM-32 and MaskWM-ED32 and the first 32 bits of the SeLo-Mark reference payload.

As the embedding residual becomes weaker, crop bit accuracy declines for all three methods (Figure 4). SeLo-Mark retains the highest accuracy at every target PSNR, extending the crop-recovery advantage in the native-payload comparison to matched image quality.

D.3 Carrier Retention under Cropping

A 0.75-side crop retains 56.25% of the image area, but the fraction of the watermark carrier that survives depends on the crop position and protection mask. We measure this fraction in the original

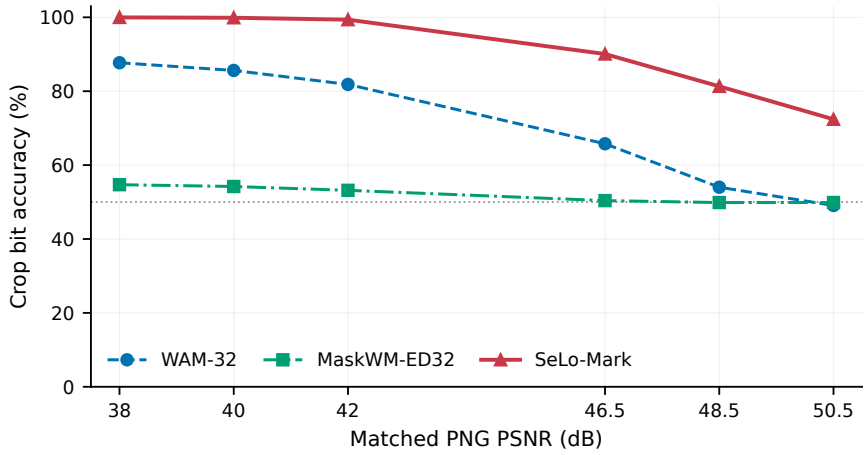

Figure 4: Crop recovery as embedding distortion is reduced. Markers show six PNG PSNR targets matched within 0.05 dB on the same 128 calibration images; higher PSNR means less distortion. BA averages the two 0.75-side crop conditions, with and without JPEG60, using the 32-bit baseline payloads and the first 32 bits of SeLo-Mark's 48-bit reference payload. SeLo-Mark averages both receiver runs. The horizontal dotted line indicates 50% random-guessing accuracy.

image, before resizing the crop. For the binary carrier M and crop window $\mathcal{C}$, it is

$$\rho_{\text{carrier}} = \frac{\sum_{p \in \mathcal{C}} M(p)}{\sum_{p} M(p)}. \tag{16}$$

Table 22 groups unedited watermarked test images by ρ_{carrier} and reports recovery within each group.

Table 22: Message recovery grouped by the fraction of the original carrier retained after cropping.

Carrier retained (%)	Images	BA (%)	Exact48 (%)
$[0, 25)$	1	92.71	0.00
$[25, 35)$	375	99.79	94.33
$[35, 45)$	2,164	99.94	98.81
$[45, 55)$	453	99.87	98.51
$[55, 65)$	3	100.00	100.00

Each of the 2,996 watermarked test images is evaluated using the same crop window with and without JPEG60, in both receiver runs. BA and Exact48 average the resulting four observations per image.

Most publications retain between 35% and 45% of their original carrier, with high exact recovery in this group. Recovery is also high in the 25–35% group, showing that the reference often survives when cropping removes more than half of its embedding support.

E Transfer, Additional Processing, and Runtime

We examine transfer to FFHQ, recovery under additional image-processing conditions, and the computational cost of verification.

E.1 Transfer to FFHQ

Each FFHQ image supplies its own content code, embedded before editing using the fixed CelebA-HQ publisher. The content mapping, adapted receiver, and threshold also remain fixed, and both protection and observation masks are predicted from RGB. Table 8 reports detection and unedited-image recovery under the protocol in Appendix A.1.

Watermarked FFHQ images have a mean PSNR of 39.09 dB and SSIM of 0.9792. Pixels inside the predicted protection region remain unchanged by construction.

Figure 5 shows detection and threshold sensitivity on FFHQ, with the CelebA-HQ operating threshold $\tau = 6/32$ marked in both panels.

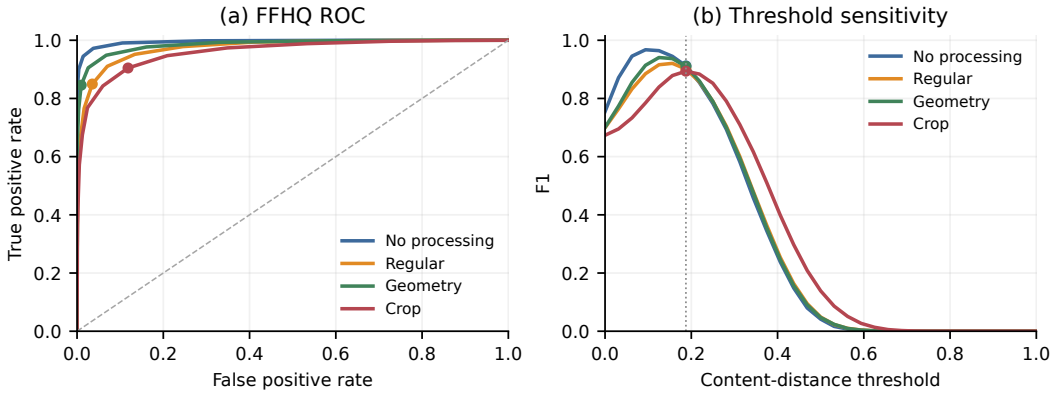

Figure 5: Detection on FFHQ with the CelebA-HQ system held fixed. (a) ROC curves and (b) F1 as the content-distance threshold varies, grouped by transmission processing. At each threshold, rates and F1 are computed for each operation–condition pair and then averaged across conditions within the group, the four editor families, and two independent runs. Colors identify the same four groups in both panels. Dots mark the CelebA-HQ calibration threshold $6/32$, also shown by the vertical line in (b). Reported FFHQ results use this fixed threshold.

AUC remains high across the four processing groups, but their F1 curves peak at different thresholds. The single CelebA-HQ operating point therefore gives lower F1 than the group-specific optima in several groups, even though the scores continue to distinguish edited from unedited faces. The reported transfer results retain that source-domain threshold.

E.2 Additional Geometric and Transmission Conditions

In addition to the 22-condition suite, we evaluate recovery under a 20° horizontal shear, translation combined with scaling, and three synthetic transmission pipelines (Table 23). The two affine configurations are not explicitly sampled during channel training or receiver adaptation. The pipelines combine compression, resampling, and contrast adjustment, whose individual operation types occur in training. We use the same 2,996 unedited watermarked test images and frozen receivers as in the main recovery comparison.

Table 23: Exact 48-bit recovery under additional geometric transformations and synthetic transmission pipelines.

Condition	Transformation or processing sequence	Exact48 (%)
Shear	Horizontal shear of 20° on the 512×512 canvas	72.50
Translation + scaling	Scale 0.8; translation $(40, -24)$ pixels on the 512×512 canvas	99.92
Repeated JPEG	JPEG qualities $80 \rightarrow 60 \rightarrow 40$	98.53
Simulated screenshot	Resize $512 \rightarrow 640$ (bilinear) $\rightarrow 512$ (bicubic), then contrast factor 1.08	99.97
Simulated social sharing	Resize $512 \rightarrow 384$ (Lanczos), JPEG quality 70, then resize to 512 (bilinear)	99.73

Both affine transforms use bilinear interpolation and neutral-gray padding. JPEG uses 4:2:0 chroma subsampling. Results average both receiver runs for each condition.

The receiver preserves nearly all complete payloads under translation with scaling and the three composed transmission pipelines. Shear is more demanding, with lower exact recovery despite using the same receiver. These results extend the recovery evaluation beyond the individual conditions in the main suite.

E.3 Message Recovery after Regeneration

On the 128-image ablation subset, we compare unchanged inputs, VAE reconstruction, and Stable Diffusion 1.5 image-to-image regeneration (Rombach et al., 2022) at strengths 0.1, 0.3, and 0.5. VAE reconstruction uses the posterior mode. Diffusion uses 50 DDIM steps (Song et al., 2021), an empty prompt, guidance scale 1.0, and $\eta = 0$ at 512×512. Methods share per-image random

seeds, and outputs are saved as 8-bit PNGs. These conditions are evaluated separately from the main transmission suite.

Table 24 reports bit accuracy and exact recovery of the original payload, using each method's native payload and publication strength.

Table 24: Message recovery after regeneration: BA / exact-message recovery (%).

Method	Unchanged	VAE	SD 0.1	SD 0.3	SD 0.5
TrustMark	100.00 / 100.00	99.62 / 99.22	61.22 / 0.00	51.34 / 0.00	50.49 / 0.00
FractalForensics	100.00 / 96.88	86.00 / 0.00	72.57 / 0.00	62.27 / 0.00	59.31 / 0.00
SepMark	100.00 / 100.00	100.00 / 100.00	97.89 / 21.09	84.36 / 0.00	69.41 / 0.00
EditGuard-degrade	100.00 / 100.00	47.60 / 0.00	46.02 / 0.00	47.73 / 0.00	47.97 / 0.00
OmniGuard	100.00 / 100.00	100.00 / 100.00	84.07 / 0.00	61.31 / 0.00	55.10 / 0.00
LIDMark	100.00 / 100.00	98.88 / 85.16	80.71 / 7.81	60.99 / 0.78	54.30 / 0.00
LampMark (released)	100.00 / 100.00	99.99 / 98.44	99.66 / 85.16	86.77 / 1.56	73.59 / 0.00
WAM	100.00 / 100.00	95.97 / 50.00	82.13 / 12.50	58.06 / 0.00	50.24 / 0.00
MaskWM-ED32	100.00 / 100.00	96.58 / 65.62	83.33 / 26.56	61.30 / 0.00	52.32 / 0.00
MaskWM-ED64	100.00 / 100.00	78.50 / 8.59	60.49 / 0.78	50.62 / 0.00	49.29 / 0.00
SeLo-Mark	100.00 / 100.00	99.28 / 83.59	94.03 / 40.62	72.40 / 0.78	54.15 / 0.00

Payloads follow Table 2; SeLo-Mark averages both runs. TrustMark evaluates 100 encoded bits for BA and its 61-bit decoded message for exact recovery; LIDMark uses 16 ID bits, OmniGuard its copyright branch, and SepMark its Tracer. LampMark uses the authors' released 128-bit reference payloads and released watermark channel in this recovery-only experiment.

VAE reconstruction and mild diffusion regeneration preserve much of SeLo-Mark's bit information, but a few incorrect bits are enough to prevent complete recovery. LampMark (released) recovers more complete messages at diffusion strength 0.1. At higher strengths, exact recovery falls sharply for both methods and reaches zero at 0.5. High bit accuracy under mild regeneration thus indicates partial payload retention; complete-message recovery is more demanding.

E.4 Receiver Runtime

We time FP32 inference on an NVIDIA RTX 4090 at batch size one, using 32 images under no processing, rotation, and 0.75-side cropping, with three repetitions after warm-up. Facial mask prediction and geometric message reading account for most of the receiver latency (Table 25); face rectification itself adds little overhead.

Table 25: Receiver latency over 288 timed calls.

Component	Median (ms)	95th percentile (ms)
Geometry estimation and message reading	39.03	39.65
Facial mask prediction	44.88	47.22
Face rectification	1.92	1.98
CLIP–PCA content encoding	20.02	20.58
Full receiver	105.81	108.67

Timings use synchronized GPU execution on a shared workstation. The full-receiver timing includes coordination between stages; model loading, image-file decoding, host-to-device transfer, and transmission processing are excluded.

F Detection Breakdowns and Visual Examples

We expand the main detection comparison by editor family and operating threshold, then illustrate how reference recovery and facial comparison behave on individual images. LampMark uses the adapted configuration described in Appendix A.6.

F.1 Detailed Facial-Edit Detection Results

Detection across editor families. Table 26 separates the main comparison by editor family. SeLo-Mark has the highest AUC and F1 for both face-swapping methods and both attribute-editing families, so its aggregate performance is not driven by a single editor.

Table 26: Facial-edit detection by editor family on CelebA-HQ.

	SimSwap		CSCS		StarGAN		Stable Diffusion	
Method	AUC	F1	AUC	F1	AUC	F1	AUC	F1
FractalForensics	89.03	90.02	88.56	89.68	88.04	88.05	89.68	90.06
SepMark	67.15	61.38	64.68	56.08	62.97	50.39	84.32	87.51
EditGuard-degrade	75.20	82.69	74.90	81.20	74.92	81.60	75.71	82.65
OmniGuard	69.31	69.04	66.81	67.34	67.63	68.38	52.82	44.61
LIDMark	62.64	44.95	76.17	69.94	65.55	53.87	86.05	84.92
RecoverMark	68.53	68.10	78.12	78.22	70.64	71.79	78.83	78.89
LampMark (adapted)	86.34	83.68	88.40	89.61	88.29	88.74	89.16	90.40
SeLo-Mark	**98.64**	**93.69**	**99.51**	**97.00**	**99.25**	**96.06**	**99.52**	**96.64**

All values are percentages. Each family averages its editing operations and the 22 processing conditions; SeLo-Mark also averages its two receiver runs. Thresholds remain fixed at their calibration values.

Operating curves and threshold sensitivity. Figure 6 shows ROC curves and SeLo-Mark's F1 as the content-distance threshold varies. Cropping shifts the F1 peak above the calibrated threshold $6/32$, consistent with the upward shift in unedited content distances in Figure 2. This explains why cropping can retain high AUC while giving lower F1 at the shared operating point.

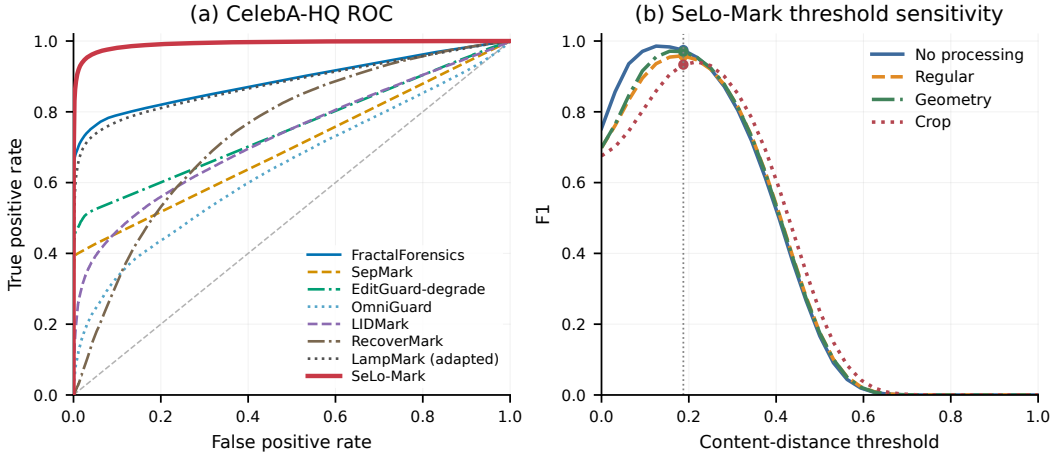

Figure 6: Detection on the CelebA-HQ main test set. (a) ROC curves for all eight methods: each operation–condition ROC is interpolated on a common false-positive-rate grid, then averaged using the main comparison's family, condition, and run weights. (b) SeLo-Mark F1 versus content-distance threshold, averaged within each processing group. Each panel has its own legend. Dots and the vertical line in (b) mark the fixed threshold $6/32$ selected on the independent calibration set.

Comparison on shared valid inputs. This comparison uses the intersection of valid inputs across all methods for each editing operation and processing condition. Each retained source image has both edited and unedited scores available for every method, including both SeLo-Mark runs. Models and calibrated thresholds remain unchanged.

The method ordering and overall scores change little on shared valid inputs (Table 27). Relative to FractalForensics, SeLo-Mark gains 10.39 AUC points (paired 95% interval [10.27, 10.51]) and 6.41 F1 points ([6.14, 6.63]). The detection advantage is therefore also present when all methods are evaluated on the same observations.

Table 27: Facial-edit detection on inputs shared by all methods.

Method	AUC (%) ↑	F1 (%) ↑
FractalForensics	88.84 [88.74, 88.93]	89.44 [89.35, 89.54]
SepMark	69.81 [69.29, 70.33]	63.87 [62.89, 64.85]
EditGuard-degrade	75.18 [75.16, 75.21]	82.05 [81.91, 82.19]
OmniGuard	64.15 [63.97, 64.34]	62.35 [62.00, 62.71]
LIDMark	72.60 [72.14, 73.02]	63.38 [62.49, 64.25]
RecoverMark	73.98 [73.35, 74.67]	74.22 [73.84, 74.62]
LampMark (adapted)	88.06 [87.91, 88.20]	88.12 [87.89, 88.35]
SeLo-Mark	**99.23** [99.15, 99.31]	**95.85** [95.58, 96.09]

Brackets give 95% percentile intervals from 1,000 identity-cluster bootstrap resamples. The same identity resample is used across methods, conditions, operations, and receiver runs; SeLo-Mark averages both runs within each resample. The valid-input intersection is defined separately for each operation–condition pair.

F.2 Qualitative Examples

Figure 7 contrasts benign compression with face swapping: the payload remains readable in both, while the content comparison distinguishes the edits. Figure 8 shows how correction restores payload recovery after cropping, and Figure 9 extends the examples to all four editor families. The examples use one adapted reader and the fixed threshold $6/32$; an alert is raised when the content distance exceeds this value.

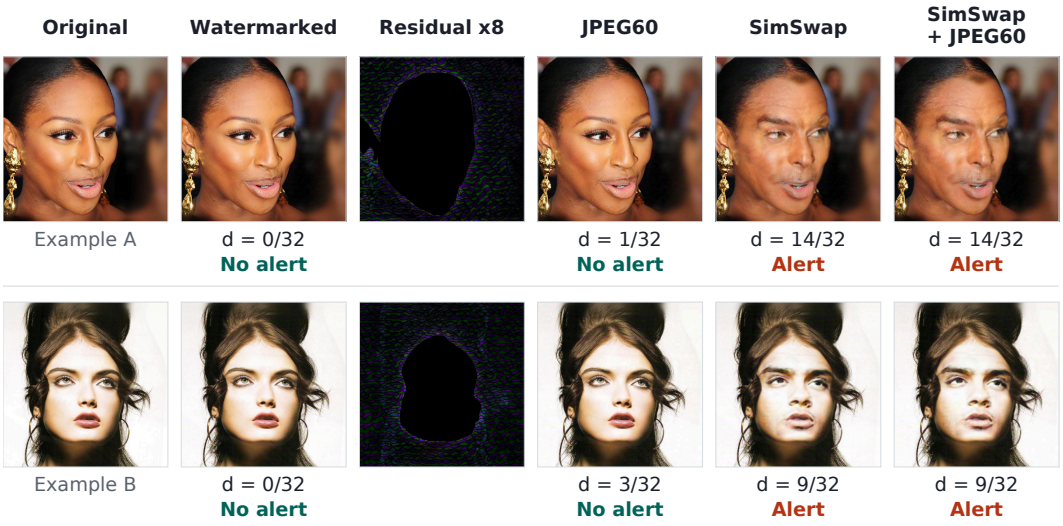

Figure 7: Verification after editing and processing. Each row shows an original image, its watermarked version, the $8\times$ embedding residual, and the JPEG60, SimSwap, and SimSwap+JPEG60 views. The unprocessed watermarked images and all processed views recover the complete 48-bit payload. At threshold $6/32$, the unedited views produce no alert, while the illustrated swaps are detected.

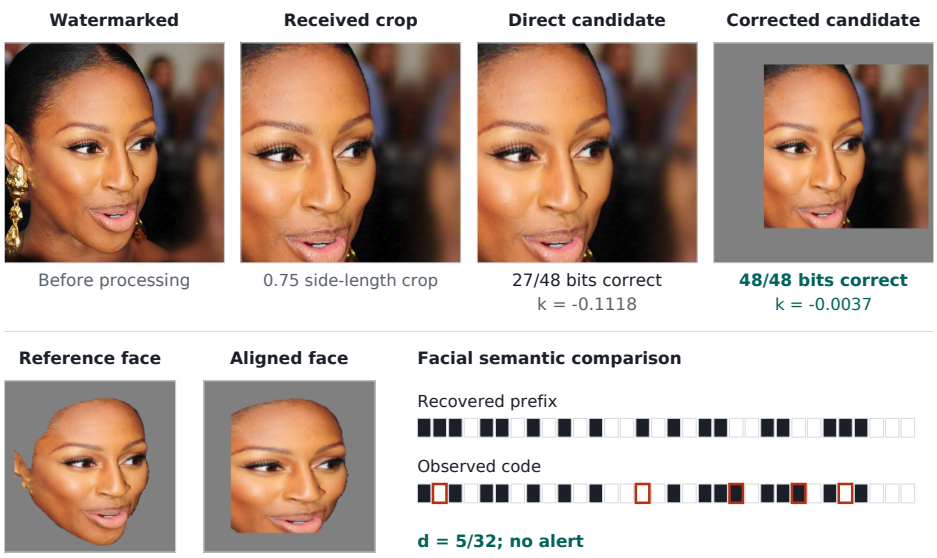

Figure 8: Recovery after a 0.75-side crop. The same adapted reader recovers 27/48 bits directly and 48/48 after correction; decoding confidence selects the corrected view. Five content bits disagree with the code from the aligned face (outlined), giving a normalized distance below the $6/32$ threshold.

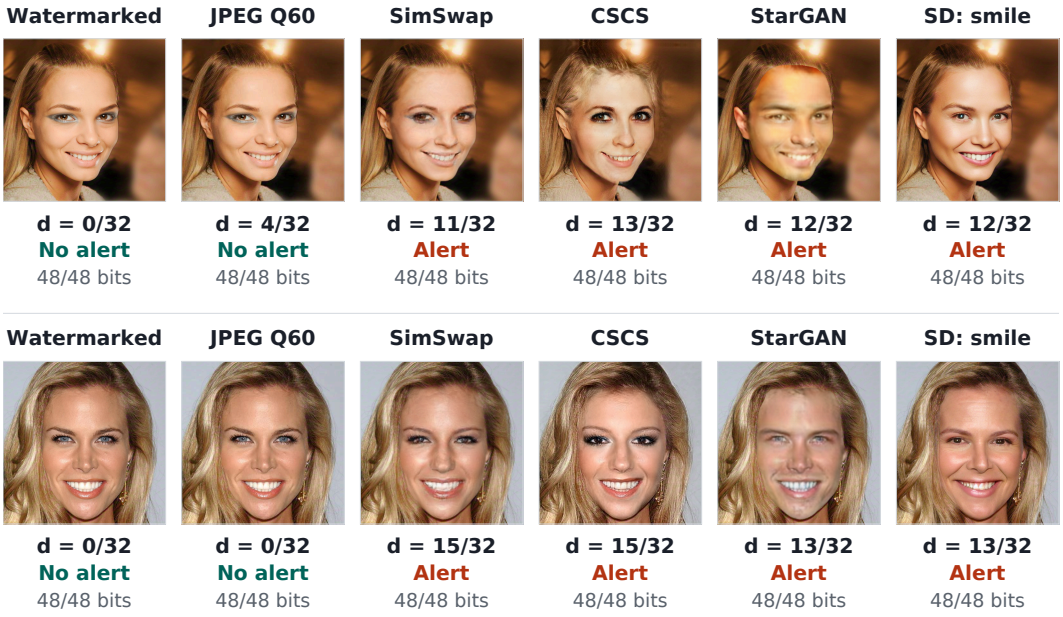

Figure 9: Four editor families applied to two watermarked inputs. Panels show JPEG60, SimSwap, CSCS, StarGAN (Male flip), and Stable Diffusion (smile). Labels give content distance, the decision at $6/32$, and recovered bits. All illustrated edits are detected with 48/48 payload bits recovered.