Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Consistency-Aware Multimodal Media Authentication for Deepfake Detection, Tampering Localization, and Source Verification

Byeongchan Park

pbc866@gmail.com

Soongsil University

Research Article

Keywords: Deepfake detection, Tampering localization, Multimedia authentication, Digital watermarking, Content provenance, Multimodal evidence fusion

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11251699/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Consistency-Aware Multimodal Media Authentication for Deepfake Detection, Tampering Localization, and Source Verification

Byeongchan Park[1]*
[1]Department of Computer Science and Engineering, Soongsil University, Seoul, Korea, Republic of
* ***Corresponding author****: Byeongchan Park*
E-mail: *pbc866@gmail.com*

Abstract
Generative manipulation increasingly complicates media authentication because visual content, embedded watermarks, and provenance metadata may provide incomplete, degraded, or mutually inconsistent evidence regarding content integrity and origin. This study presents a consistency-aware multimodal media authentication method for jointly addressing deepfake detection, tampering localization, and claimed-source verification. The proposed method integrates visual features, latent-domain and pixel-domain watermark evidence, and structured provenance metadata through modality-specific encoding, reliability-aware evidence interaction, and cross-evidence consistency learning. Unlike conventional multimodal fusion approaches that treat heterogeneous inputs as interchangeable information sources, the proposed method explicitly models their different reliability characteristics and authentication meanings. Separate prediction heads estimate manipulation probability, localize altered regions, and classify source claims as supported, contradicted, or insufficiently evidenced.
Experiments conducted on a controlled corpus derived from CelebAMask-HQ, together with cross-dataset evaluation using FaceForensics++ and DeeperForensics-1.0, demonstrate an AUC of 0.969 for deepfake detection, an IoU of 0.816 for tampering localization, and a Macro-F1 of 0.945 for claimed-source verification. Compared with Early Fusion, the proposed method improves detection AUC by 0.017, localization IoU by 0.044, and source-verification Macro-F1 by 0.029. Under conflicting-evidence conditions, the proposed method maintains an average Macro-F1 of 0.921, compared with 0.870 for Early Fusion and 0.759 for Watermark-only verification. Robustness experiments further show that the method maintains an AUC of 0.914 under 20% cropping and 0.884 under full generative regeneration. Ablation results confirm that dual-domain watermark evidence, provenance information, reliability-aware interaction, and cross-evidence consistency learning each contribute to overall performance. These results indicate that effective multimedia authentication requires manipulation analysis and source verification to be treated as related but distinct inference tasks and that consistency-aware integration of heterogeneous evidence improves robustness under degraded, missing, and conflicting conditions.

Keywords: Deepfake detection; Tampering localization; Multimedia authentication; Digital watermarking; Content provenance; Multimodal evidence fusion

1 Introduction

Recent advances in generative models have substantially increased the realism and accessibility of synthetic and manipulated visual content. Facial identity replacement, expression modification, localized generative editing, and full-image regeneration can now be performed while preserving much of the perceptual structure of the original media. This development has expanded the practical importance of multimedia authentication because

visually plausible manipulation may affect only limited semantic regions while leaving most of the surrounding content unchanged. Existing studies have therefore investigated spatial artifacts, frequency-domain characteristics, blending boundaries, facial structures, and generalizable forgery representations for detecting synthetic or manipulated content (Cao et al., 2022; Chen et al., 2021; Li et al., 2020; Qian et al., 2020; Rössler et al., 2019; Shiohara & Yamasaki, 2022; Sun et al., 2022; Wang et al., 2023; Yan et al., 2023; Zhao et al., 2021). More recent studies have further emphasized domain generalization, manipulation localization, and multimodal reasoning to improve robustness against previously unseen manipulation processes (Gao et al., 2025; W. Huang et al., 2025; Z. Huang et al., 2025; J. Li et al., 2025; Liu et al., 2024; Zhao et al., 2025).

Despite these advances, multimedia authentication based exclusively on visual evidence remains fundamentally limited. A visual detector estimates whether observed content contains characteristics associated with learned manipulation patterns, but such a prediction does not directly establish the origin or provenance of the content. Moreover, the discriminative features used by passive detectors may be degraded by compression, resizing, filtering, geometric transformation, or generative reconstruction. Face X-Ray, Self-Blended Images, UCF, and recent domain-generalization approaches have demonstrated that carefully designed representations and training strategies can improve generalization beyond individual forgery methods (Li et al., 2020; J. Li et al., 2025; Shiohara & Yamasaki, 2022; Yan et al., 2023). Nevertheless, the underlying evidence remains content-derived. Consequently, an image classified as authentic by a visual detector cannot automatically be interpreted as originating from a claimed source, while a manipulated image may still retain valid information about its original source.

Active forensic approaches address part of this limitation by intentionally embedding information into media before or during distribution. Deep watermarking and generation-integrated watermarking have been investigated as mechanisms for associating synthetic content with its generating process, source, or content instance. Tree-Ring Watermarks embeds structured signals into the initial noise of diffusion generation, whereas Stable Signature integrates a recoverable signature into the latent diffusion generation process (Fernandez et al., 2023; Wen et al., 2023). Gaussian Shading and subsequent diffusion watermarking methods further demonstrate that source-related information can be introduced while maintaining perceptual quality (Arabi et al., 2025; K. Li et al., 2025; Yang et al., 2024). Localized watermarking approaches such as EditGuard, Watermark Anything, and OmniGuard extend this concept by associating recoverable information with spatial regions and supporting manipulation localization (Sander et al., 2025; X. Zhang et al., 2024; X. Zhang et al., 2025). These developments indicate that embedded signals can complement passive forensic evidence by providing source- or content-associated information that is not directly recoverable from visual artifacts alone.

However, watermark recovery and content authenticity are not equivalent properties. A watermark may remain recoverable even after the associated content has been modified, whereas watermark extraction may fail because of benign signal degradation rather than intentional manipulation. Recent studies have also shown that semantic and generative watermarks can be affected by removal, averaging, transfer, forgery, overwriting, and regeneration attacks (Luo et al., 2026; Müller et al., 2025; Yang et al., 2024). A successfully decoded watermark therefore establishes only that a particular embedded signal remains recoverable under the adopted extraction procedure. It does not independently establish that the observed content is unmodified or that the recovered payload remains correctly associated with the current media instance. This distinction becomes particularly important when multiple watermark domains are considered because latent-domain and pixel-domain

watermark evidence may respond differently to compression, filtering, geometric transformations, and generative reconstruction.

Provenance metadata introduces another form of authentication evidence. Standards and specifications such as C2PA and JPEG Trust define mechanisms for representing source information, processing history, content bindings, assertions, and cryptographically protected provenance records (Bhowmik et al., 2024; Coalition for Content Provenance and Authenticity, 2026; ISO/IEC, 2026; Temmermans et al., 2024). These mechanisms provide explicit information concerning media creation and modification, but cryptographic validation should not be interpreted as a statistical determination that the depicted content itself is authentic. A valid signature can establish properties of a provenance record and its signer under an adopted trust policy, while the interpretation of the associated claims and their relationship to the currently observed content remains a separate issue. Previous research on image provenance has similarly demonstrated that visual relationships, metadata, transformation histories, and graph-based provenance information can provide complementary information for source analysis (Bharati et al., 2019; Moreira et al., 2018; Nguyen et al., 2025; K. Zhang et al., 2025).

The resulting multimedia authentication problem is therefore not simply a matter of accumulating additional evidence sources. Visual features, embedded watermarks, and provenance metadata describe different properties of an image and support different levels of inference. Visual evidence primarily characterizes observable manipulation traces, watermark evidence provides recoverable source- or content-associated signals, and provenance metadata represents explicit source and processing assertions. Consequently, reliable authentication requires not only combining these observations but also determining whether individual evidence sources are available, reliable, mutually compatible, or contradictory.

For example, a manipulated facial region may be correctly detected while an embedded source identifier continues to support the originating generator. Conversely, an image containing no detectable manipulation may be paired with a substituted source claim or an inconsistent provenance record. Missing evidence introduces another important distinction: the absence of a recoverable watermark or provenance record does not necessarily constitute evidence that a source claim is false. These cases require explicit differentiation among evidence that supports a claim, evidence that contradicts it, and conditions in which the available observations are insufficient for a reliable conclusion.

Recent multimodal manipulation studies have demonstrated that cross-modal relationships and consistency learning can improve detection and grounding when heterogeneous inputs must be interpreted jointly (Z. Huang et al., 2025; Y. Li et al., 2025; Shao et al., 2023; Z. Zhang et al., 2025). However, existing multimodal approaches generally focus on semantic relationships between visual and textual modalities or on integrating multiple manipulation cues. Their objective is primarily to improve recognition performance by combining complementary features.

The authentication setting considered in this study is fundamentally different. Visual evidence represents observed content characteristics, watermark evidence represents recoverable embedded associations, and provenance metadata represents structured source and processing assertions. These evidence sources are not interchangeable modalities describing the same semantic object. They have different meanings, reliability characteristics, availability conditions, and failure modes. Consequently, conventional feature concatenation or decision-level fusion does not explicitly distinguish between unavailable evidence, unreliable evidence, and genuinely contradictory evidence.

This distinction is particularly important in practical multimedia authentication. A manipulated image may retain valid evidence of its originating source, whereas visually unmodified content may be paired with an inconsistent source claim. Likewise, watermark

extraction failure can result from benign signal degradation rather than malicious modification, while a successfully recovered watermark does not independently demonstrate the integrity of the current content. Authentication therefore requires an evidence interpretation mechanism that preserves the semantic role of each observation while modeling whether heterogeneous evidence sources support, contradict, or provide insufficient information about a specific claim.

This study addresses this problem through consistency-aware multimodal media authentication using visual, watermark, and provenance evidence. The proposed method combines visual features, latent-domain and pixel-domain watermark observations, and structured provenance metadata while preserving the distinct interpretation and reliability characteristics of each evidence source. Modality-specific encoders transform heterogeneous observations into compatible representations, while reliability-aware evidence interaction estimates their task-dependent usefulness according to evidence availability, extraction quality, and validation state. Cross-evidence consistency learning then models agreement and contradiction among visual observations, recovered watermark information, provenance records, and the evaluated source claim.

The three prediction tasks are explicitly separated. Deepfake detection estimates whether the observed content contains a manipulation within the evaluated manipulation classes. Tampering localization estimates the spatial extent of an alteration when an appropriate ground-truth region is defined. Claimed-source verification determines whether the available evidence supports, contradicts, or is insufficient to evaluate a specified originating-source claim. The source-verification output is therefore distinct from closed-set source classification and from cryptographic signature validation. This formulation allows an image to be simultaneously classified as manipulated and supported with respect to its originating source when authentic source evidence remains valid after subsequent editing. It also allows an apparently unmanipulated image to be associated with a contradicted source claim when reliable source-related evidence is inconsistent with that claim.

The proposed method is evaluated using a controlled experimental corpus derived from CelebAMask-HQ and external deepfake benchmarks including FaceForensics++ and DeeperForensics-1.0 (Jiang et al., 2020; Lee et al., 2020; Liu et al., 2015; Rössler et al., 2019). The evaluation covers conventional distribution transformations, generative editing, partial and full regeneration, watermark degradation and overwriting, and missing or conflicting provenance conditions. In addition to comparisons with single-evidence and conventional multimodal fusion baselines, controlled ablation experiments evaluate the contribution of dual-domain watermark evidence, provenance metadata, reliability-aware interaction, and cross-evidence consistency learning. Cross-dataset experiments additionally examine the behavior of the visual branch and the missing-evidence handling mechanism when watermark and provenance information are unavailable.

The main contributions of this study are summarized as follows:

1. A unified formulation of multimedia authentication with distinct inference outcomes. Deepfake detection, tampering localization, and claimed-source verification are formulated as related but non-equivalent multimedia authentication tasks. The formulation explicitly distinguishes manipulation inference, watermark recovery, provenance validation, and source-claim verification rather than reducing heterogeneous observations to a single binary authenticity decision.
2. Consistency-aware integration of heterogeneous authentication evidence. Visual features, latent-domain and pixel-domain watermark observations, and structured provenance metadata are jointly analyzed through modality-specific representations and reliability-aware interaction. Cross-evidence consistency learning explicitly models agreement, contradiction, and availability among heterogeneous

evidence sources instead of relying solely on feature concatenation or decision-level aggregation.

3. An evidence-aware three-state formulation for claimed-source verification. Source claims are evaluated as supported, contradicted, or insufficiently evidenced. This formulation prevents missing evidence from being interpreted as direct contradiction and allows manipulated content to retain a valid originating-source association when the available evidence remains consistent with the evaluated claim.
4. A comprehensive robustness evaluation under multimedia transformations and evidence degradation.
   The proposed method is evaluated under conventional signal transformations, geometric transformations, generative regeneration, watermark degradation and overwriting, evidence dropout, and controlled provenance conflicts. Cross-dataset and ablation experiments further quantify the contribution of individual evidence sources and consistency-aware interaction under degraded, missing, and conflicting conditions.

The remainder of this paper is organized as follows. Section 2 reviews deepfake detection, tampering localization, generative media watermarking, and provenance verification. Section 3 describes the proposed dual-domain watermark representation, multimodal evidence representation, reliability-aware interaction, cross-evidence consistency learning, and joint prediction objective. Section 4 presents the datasets, implementation details, baseline methods, attack conditions, and evaluation metrics. Section 5 reports the experimental results, cross-dataset evaluation, robustness analysis, and ablation studies. Finally, Section 6 concludes the paper and discusses its limitations and future research directions.

2 Related Work

This section reviews prior studies related to deepfake detection, tampering localization, generative media watermarking, and provenance verification. The discussion focuses on the type of evidence used by each research direction and the level of inference that can be derived from that evidence. In particular, visual manipulation cues, embedded watermark signals, and provenance metadata are treated as distinct sources of information because they characterize different aspects of media authenticity and may remain valid under different transformation or attack conditions.

The review first examines passive and active approaches for deepfake detection and manipulation localization, with emphasis on generalization to unseen manipulation methods and the use of spatial forensic cues. It then considers generative media watermarking, including latent-domain and pixel-domain approaches, and discusses the distinction between watermark recoverability, tampering evidence, and source attribution. Finally, provenance-based approaches such as C2PA and JPEG Trust are reviewed to clarify the role of structured source claims, cryptographic validation, and content binding. Particular attention is given to recent studies that combine multiple forensic tasks or heterogeneous evidence sources, as these works define the methodological context and comparison requirements for the proposed approach.

2.1 Deepfake Detection and Tampering Localization

Deepfake detection aims to determine whether visual content has been synthetically generated or manipulated by identifying forensic traces introduced during generation, blending, resampling, or post-processing. FaceForensics++ established a widely used benchmark for evaluating multiple facial manipulation methods (Rössler et al., 2019). However, models trained on specific manipulation distributions often experience performance degradation when applied to unseen generators, compression conditions, or acquisition environments.

To improve generalization, several studies have focused on manipulation characteristics that are less dependent on specific generation methods. Face X-Ray models blending boundaries between facial regions (Li et al., 2020), while frequency-aware approaches such as F^3-Net exploit inconsistencies in frequency-domain representations (Qian et al., 2020). Local-relation learning and multi-attentional methods further emphasize localized manipulation characteristics rather than relying solely on global image representations (Chen et al., 2021; Zhao et al., 2021). Self-Blended Images reduces dependence on specific deepfake algorithms by constructing synthetic forgery samples from pristine images (Shiohara & Yamasaki, 2022), while UCF and more recent domain-generalization methods aim to learn common representations that remain informative across unseen manipulation conditions (Gao et al., 2025; W. Huang et al., 2025; J. Li et al., 2025; Yan et al., 2023; Zhao et al., 2025).

Image-level detection alone provides limited information about the spatial extent of manipulation. Tampering localization addresses this limitation by identifying altered regions at the pixel or region level. DH-GAN explicitly investigates image manipulation localization (Liu et al., 2024), while mask-guided supervision incorporates spatial manipulation information into deepfake detection and domain generalization (J. Li et al., 2025). More recent approaches such as SIDA jointly consider detection and localization, reflecting a broader transition from binary authenticity classification toward spatially interpretable forensic analysis (Z. Huang et al., 2025).

Active forensic approaches provide another source of localization evidence by embedding information before manipulation occurs. EditGuard, Watermark Anything, and OmniGuard demonstrate that embedded watermark signals can support manipulation localization and content authentication (Sander et al., 2025; X. Zhang et al., 2024; X. Zhang et al., 2025). However, watermark recoverability should not be directly interpreted as a manipulation mask. A region in which watermark extraction fails may represent an actual modification, signal degradation, originally unwatermarked content, or extraction failure.

Joint forensic approaches have further combined deepfake detection, localization, and source-related inference. All in One integrates these tasks using a robust landmark-identity watermark (Wu et al., 2026), demonstrating that their joint consideration is already an established research direction. The distinction of the present study therefore lies not simply in combining multiple forensic tasks, but in representing and interpreting heterogeneous evidence when its components may be incomplete, degraded, or mutually inconsistent.

Accordingly, the present study does not distinguish itself merely by combining detection, localization, and source-related inference. Instead, it focuses on how heterogeneous authentication evidence should be represented and interpreted when individual sources may be incomplete, degraded, unavailable, or mutually inconsistent. Visual evidence primarily supports manipulation detection and spatial localization, whereas watermark and provenance information provide complementary evidence regarding source and content associations. Rather than forcing all evidence sources to produce a common authenticity decision, the proposed method explicitly models how their availability, reliability, agreement, and contradiction contribute differently to each authentication task.

2.2 Generative Media Watermarking and Provenance Verification

Generative media watermarking provides an active mechanism for associating synthesized content with its generation process, source, or content instance. Existing approaches differ in the stage at which watermark information is embedded and the procedure used for recovery. Tree-Ring Watermarks embeds a structured signal into the initial noise of diffusion generation (Wen et al., 2023), whereas Stable Signature integrates a recoverable signature into the latent diffusion process (Fernandez et al., 2023). Gaussian Shading and subsequent diffusion watermarking methods further demonstrate that source-related information can be

embedded while maintaining perceptual quality (Arabi et al., 2025; K. Li et al., 2025; Yang et al., 2024). These studies establish watermarking as a complementary source of evidence to passive visual forensics.

Recent methods have extended watermarking toward spatial verification and manipulation localization. EditGuard, Watermark Anything, and OmniGuard associate embedded signals with local image regions and support tampering localization or message recovery (Sander et al., 2025; X. Zhang et al., 2024; X. Zhang et al., 2025). However, watermark recovery and content integrity are not equivalent. A watermark may survive semantic manipulation, while extraction failure may result from compression, filtering, geometric transformation, or regeneration. Therefore, the presence of a recoverable watermark does not necessarily imply that the observed image is unmodified, and watermark absence alone cannot be interpreted as direct evidence of manipulation.

Watermark robustness must also be distinguished from resistance to removal, transfer, and forgery. Recent studies have demonstrated black-box forgery, averaging, regeneration, and overwriting attacks against generative watermarks (Luo et al., 2026; Müller et al., 2025; Yang et al., 2024). These results indicate that source verification requires more than successful payload recovery. The recovered information must also be evaluated in relation to the observed content, the claimed source, and other available evidence. This motivates the use of latent-domain and pixel-domain watermark observations as complementary evidence whose reliability can be assessed separately.

Provenance metadata provides a different form of source-related information. C2PA and JPEG Trust define mechanisms for representing content origin, processing history, assertions, content bindings, and cryptographically protected provenance records (Bhowmik et al., 2024; Coalition for Content Provenance and Authenticity, 2026; ISO/IEC, 2026; Temmermans et al., 2024). Related studies on image provenance have investigated relationships among visual content, metadata, transformation histories, and source records (Bharati et al., 2019; Moreira et al., 2018; Nguyen et al., 2025; K. Zhang et al., 2025). Cryptographic validation can establish properties of a signed record and its association with an asset, but it does not independently establish that the observed content is unmanipulated or that every provenance assertion is factually correct. Accordingly, provenance validation and learned source verification should be treated as related but distinct processes.

Table 1 summarizes the principal evidence sources considered in previous research and clarifies their relationship to the problem addressed in this study. The final column identifies the distinction relevant to the proposed formulation rather than treating the corresponding approaches as insufficient within their original scope.

Table 1. Evidence sources and their interpretation in media authentication.

Research direction	Representative studies	Principal evidence or output	Distinction relevant to this study
Passive forgery detection	Face X-Ray; Self-Blended Images (Li et al., 2020; Shiohara & Yamasaki, 2022)	Visual traces associated with manipulation	Manipulation evidence does not independently verify a claimed source
Joint active forensics	All in One (Wu et al., 2026)	Embedded information combined with forensic predictions	Joint detection and localization already exist; the distinction concerns evidence representation and verification
Generation-	Tree-Ring; Stable	Recoverable	Watermark recovery

integrated watermarking	Signature; Gaussian Shading (Fernandez et al., 2023; Wen et al., 2023; Yang et al., 2024)	generation-associated signals	and content integrity require separate interpretation
Localized watermarking	EditGuard; Watermark Anything; OmniGuard (Sander et al., 2025; X. Zhang et al., 2024; X. Zhang et al., 2025)	Watermarked-region maps and recovered messages	Watermark coverage is not automatically equivalent to a manipulation mask
Signed provenance	C2PA; JPEG Trust (Bhowmik et al., 2024; Coalition for Content Provenance and Authenticity, 2026; ISO/IEC, 2026; Temmermans et al., 2024)	Assertions, signatures, provenance records, and content bindings	Record validation differs from assessment of content authenticity and claim consistency

Taken together, previous studies suggest that the principal challenge in multimedia authentication is not simply the availability of multiple evidence sources but the interpretation of their different meanings and reliability characteristics. Visual observations, dual-domain watermark information, and provenance metadata may support compatible, partially compatible, or contradictory conclusions. The proposed method therefore separates explicit provenance validation from learned evidence interpretation and introduces consistency-aware integration to determine whether the available observations support, contradict, or are insufficient to evaluate a claimed source. This distinction forms the methodological basis of the multimodal authentication method introduced in Section 3.

3 Proposed Method

The proposed method jointly analyzes visual content, dual-domain watermark evidence, and provenance metadata for deepfake detection, tampering localization, and claimed-source verification. Its central premise is that these observations describe different properties of multimedia content and therefore should not be treated as interchangeable indicators of authenticity. A watermark may remain recoverable after manipulation, while a valid source claim may coexist with edited content. Conversely, content with no detectable visual manipulation may be paired with an inconsistent source record.

Accordingly, the proposed method preserves separate task outputs and models task-dependent relationships among heterogeneous evidence sources. The processing procedure consists of evidence acquisition, modality-specific encoding, reliability-aware evidence interaction, cross-evidence consistency learning, and task-specific prediction. Watermark extraction and cryptographic provenance validation precede the learned verification stage. Their outputs are provided as explicit observations, and the learned model neither modifies nor replaces cryptographic validation results.

3.1 Overall Method

The overall procedure of the proposed method is illustrated in Fig. 1. Visual content, latent-domain and pixel-domain watermark evidence, and provenance metadata are jointly analyzed to detect manipulation, localize altered regions, and evaluate a claimed source. Separate

outputs are maintained for these tasks because manipulation status and source validity describe different properties of the observed content.

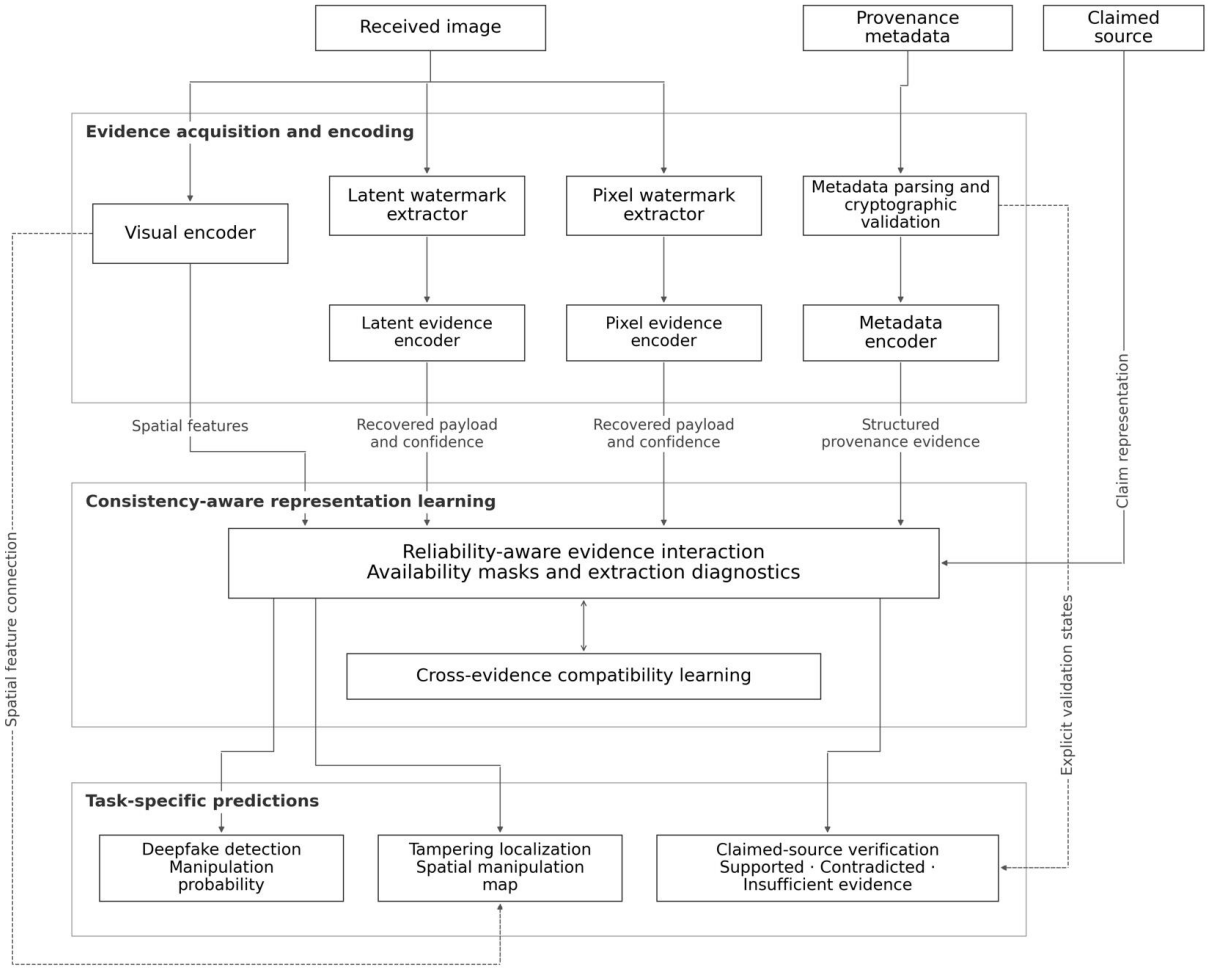


Fig. 1. Overview of the proposed method for joint deepfake detection, tampering localization, and claimed-source verification.

As shown in Fig. 1, the received image is first processed by a visual encoder and two watermark extractors. The visual encoder captures spatial features associated with content manipulation, while the watermark extractors recover latent-domain and pixel-domain evidence together with extraction diagnostics. Available provenance metadata is independently parsed and cryptographically validated to obtain source assertions, content associations, and validation outcomes.

The resulting observations are mapped into modality-specific representations and combined through reliability-aware evidence interaction. Availability indicators distinguish missing evidence from extraction failure, while cross-evidence compatibility learning models agreement and contradiction among available observations. Separate prediction heads subsequently produce an image-level manipulation probability, a spatial manipulation map, and a claimed-source verification outcome.

Let I_0 denote an image produced through a participating generation process, and let I_w denote its watermarked version. The image received for authentication is expressed as

$$\boldsymbol{I = \mathcal{T}(I_w)} \tag{1}$$

where $\mathcal{T}$ represents a distribution transformation, a content manipulation, or a composition of both. Distribution transformations include operations such as compression and resizing. Content manipulations include facial replacement, expression modification, and generative editing. The manipulation label is determined by the operation's defined semantic effect rather than by the presence of any pixel-level change.

The input to the recognition procedure is

$$\boldsymbol{\mathcal{X}} = \left(\boldsymbol{I}, \hat{\mathbf{b}}_{\boldsymbol{L}}, \hat{\mathbf{b}}_{\boldsymbol{P}}, \boldsymbol{M}, \boldsymbol{q}, \mathbf{a}, \mathbf{v}\right), \tag{2}$$

where $\hat{\mathrm{b}}_L$ and $\hat{\mathrm{b}}_P$ are recovered latent-domain and pixel-domain watermark observations, respectively. M denotes available provenance metadata, and q is the source claim being evaluated. The availability vector a indicates which evidence sources are present. The vector v contains explicit provenance validation outcomes, including signature, content-binding, and trust-policy results where applicable.

A source claim must have a defined scope. In this study, q identifies a claimed originating generation source. Verification of this claim does not establish the identity of the person depicted, the current distributor, or the truth of the depicted event. These concepts are kept separate from source attribution.

The outputs comprise an image-level manipulation probability, a spatial manipulation map, and a claim-verification distribution:

$$\boldsymbol{\mathcal{Y}} = \left(\hat{\boldsymbol{y}}_{\boldsymbol{d}}, \hat{\mathbf{Y}}_{\boldsymbol{\ell}}, \hat{\mathbf{p}}_{\boldsymbol{s}}\right), \qquad \hat{\mathbf{p}}_{\boldsymbol{s}} \in \boldsymbol{\Delta^3}, \tag{3}$$

where $\hat{y}_d \in [0,1]$, $\hat{\mathrm{Y}}_\ell \in [0,1]^{H\times W}$, and Δ^3 denotes the probability simplex over three verification outcomes: supported, contradicted, and insufficient evidence.

The third outcome prevents the absence of recoverable evidence from being treated as a contradiction. A claim is supported when the available observations satisfy the specified source-association policy. It is contradicted when reliable evidence is incompatible with the claim. It is assigned insufficient evidence when neither conclusion is justified. These labels describe evidential support under the adopted policy rather than absolute factual certainty.

The dual-domain setting assumes access to the generation process required for latent watermark insertion and to the information required by its extraction procedure. Images outside this participating process may still be analyzed by the visual branch, but unavailable watermark observations are explicitly masked. Such images are evaluated separately from images for which both watermark channels are expected.

The attacker may modify image content, apply distribution transformations, remove or replace metadata, and attempt watermark removal or transfer. The proposed recognition stage does not claim to resolve compromised signing credentials. Such cases require separate treatment through the trust policy and are outside the present threat model.

3.2 Dual Domain Watermark Representation

The two watermark channels encode a common association with the originating source while operating through different insertion and recovery procedures. Rather than treating either watermark as an independent authenticity indicator, the proposed method uses them as complementary authentication evidence whose agreement and reliability can be evaluated after distribution or manipulation.

Let s denote the originating source identifier and u a content-instance identifier. The payloads are constructed as

$$\mathbf{b}_{\boldsymbol{L}} = \mathbf{ECC}_{\boldsymbol{L}}(\mathbf{Serialize}(\boldsymbol{s}, \boldsymbol{u})), \qquad \mathbf{b}_{\boldsymbol{P}} = \mathbf{ECC}_{\boldsymbol{P}}(\mathbf{Serialize}(\boldsymbol{u}, \boldsymbol{h})), \tag{4}$$

where ECC_L and ECC_P are error-correcting encoders. The value h is a digest of a canonical source record containing the source identifier and content-instance identifier. The shared identifier u provides an explicit relationship between the two channels.
The record used to compute h excludes the final image's content hash. This avoids a circular dependency between the embedded payload and the completed image. After both watermark stages, the final image can be bound to a signed provenance record through a separate content-binding operation.
The watermarked image is generated as

$$\boldsymbol{I_w = \mathcal{E}_P(G_\theta(\mathcal{E}_L(\mathbf{z}, \mathbf{b}_L; k_L), c), \mathbf{b}_P; k_P)}, \tag{5}$$

where z is the initial latent variable, c is the generation condition, and G_θ is the image generator. $\mathcal{E}_L$ and $\mathcal{E}_P$ denote the latent-domain and pixel-domain insertion operators, with the corresponding keying information k_L and k_P.
The recognition method does not require both insertion operators to be newly introduced. Existing watermark operators can be adopted, provided their access assumptions, payload capacities, and extraction procedures are reported. The methodological contribution examined here concerns the representation and interpretation of their recovered evidence.
At verification time, each extractor returns soft decoding observations and extraction diagnostics:

$$\boldsymbol{(\ell_k, \delta_k) = \mathcal{D}_k(I; k_k)}, \qquad \boldsymbol{k \in \{L, P\}}, \tag{6}$$

where ℓ_k contains bit-level logits and δ_k contains available diagnostics, such as decoding confidence, error-correction status, or synchronization quality.
Soft observations are retained because hard decoding alone discards uncertainty. Successful error correction is treated as evidence of recoverability, not as proof that the recovered payload belongs to the observed content. Likewise, the digest \(h\) supports record matching but does not independently provide cryptographic authentication.
The resulting watermark representations are

$$\boldsymbol{\mathbf{Z}_k = E_k(\ell_k, \delta_k, \hat{u}_k, \mathbf{r}_k)}, \qquad \boldsymbol{k \in \{L, P\}}, \tag{7}$$

where $\hat{u}_k$ is the decoded instance identifier when available and r_k contains explicit comparison results against the claimed source record. Failed or unavailable fields are accompanied by validity masks.
Identifiers are compared symbolically before their comparison results enter the learned representation. Arbitrary numeric identifiers are not interpreted as continuous quantities. This reduces the opportunity for the network to memorize particular source codes instead of learning evidence compatibility.
The pixel extractor may additionally provide a spatial confidence map when supported by the adopted implementation. This map indicates local watermark recoverability and is used as an auxiliary observation. It is not used directly as the ground-truth manipulation mask.

3.3 Multimodal Evidence Representation

The visual encoder produces spatial features that preserve information required for manipulation localization. Watermark observations represent recoverable embedded associations, whereas provenance metadata provides structured information regarding source assertions, content relationships, declared processing history, and cryptographic validation states. Because these observations have different semantic roles, they are first represented independently before cross-evidence interaction is performed.
Metadata fields that directly disclose experimental labels are excluded from the learned representation. File paths, dataset identifiers, partition indicators, and manipulation labels

introduced solely for evaluation are not provided to the model. Missing fields, unavailable observations, extraction failures, and failed validation states are represented separately so that absence of evidence is not automatically interpreted as contradiction.

The evidence tokens are projected into a compatible feature space while retaining modality identifiers. A reliability gate is then estimated for each evidence branch according to availability and applicable quality or validation descriptors.

The visual encoder produces spatial features that retain information needed for localization:

$$\mathbf{Z_c} = \boldsymbol{E_c}(\boldsymbol{I}) \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{d}}, \qquad \mathbf{Z_m} = \boldsymbol{E_m}\big(\boldsymbol{\phi}(\boldsymbol{M}, \boldsymbol{q}, \mathbf{v})\big), \tag{8}$$

where N is the number of visual tokens and d is the embedding dimension. The metadata transformation ϕ converts selected provenance fields into structured observations, including source-record matches, instance associations, declared edit categories, and validation outcomes.

Metadata fields that directly disclose experimental labels are excluded. File paths, dataset identifiers, split indicators, and manipulation labels introduced solely for evaluation are not provided to the model. Missing fields and failed validation states are represented separately.

The evidence tokens are projected into a common feature space and combined with modality identifiers. A reliability gate is then computed for each evidence branch:

$$\boldsymbol{\rho_k} = \boldsymbol{a_k}\,\boldsymbol{\sigma}[\boldsymbol{g_k}(\mathbf{Pool}(\mathbf{Z_k}), \boldsymbol{\delta_k})], \qquad \boldsymbol{k} \in \{\boldsymbol{c}, \boldsymbol{L}, \boldsymbol{P}, \boldsymbol{m}\}, \tag{9}$$

where a_k denotes availability, σ is the sigmoid function, and g_k is a learned gating function. For visual and metadata branches, δ_k contains the applicable quality and validation descriptors.

The gate represents the estimated usefulness of a branch for inference. It does not change whether a signature is valid, whether a record is trusted, or whether an input is present. These explicit states remain accessible to the prediction heads.

Cross-evidence interaction is implemented through masked attention:

$$\mathbf{H} = \mathbf{softmax}\left(\frac{\mathbf{QK^T}}{\sqrt{\boldsymbol{d}}} + \mathbf{B_\rho} + \mathbf{B_a}\right)\mathbf{V}, \tag{10}$$

where Q, K, and V are projections of the evidence tokens. B_ρ adds reliability-dependent attention biases, and B_a excludes unavailable evidence. A claim token representing the scope and available attributes of q is included in the interaction stage.

A dedicated compatibility predictor evaluates selected relationships between evidence branches:

$$\hat{\boldsymbol{c}}_{\boldsymbol{ij}} = \boldsymbol{\sigma}\big[\boldsymbol{g}_{\mathbf{cons}}\big(\overline{\mathbf{H}}_{\boldsymbol{i}}, \overline{\mathbf{H}}_{\boldsymbol{j}}, \mathbf{r}_{\boldsymbol{ij}}, \mathbf{e}_{\boldsymbol{q}}\big)\big], \tag{11}$$

where $\overline{\mathrm{H}}_i$ and $\overline{\mathrm{H}}_j$ are pooled branch representations, r_{ij} contains explicit relationship descriptors, and e_q represents the claim.

Compatibility is defined according to the semantic meaning of each evidence relationship rather than by forcing all branches to produce the same authenticity judgment. For example, a correctly declared edit may be compatible with both a positive manipulation prediction and a supported originating-source claim. By contrast, reliably recovered content-instance identifiers referring to different source records provide contradictory evidence. Compatibility labels are therefore derived from controlled source associations and transformation histories, while ambiguous relationships are excluded from direct supervision. This design prevents the model from learning that every edited image must have inconsistent provenance or that every missing observation represents a conflict.

3.4 Joint Detection, Localization, and Source Verification

The interacted multimodal representation is processed by separate task-specific prediction heads because manipulation detection, spatial localization, and source verification correspond to different inference targets. The detection head estimates an image-level manipulation probability, the localization head predicts the spatial extent of manipulation, and the source-verification head estimates evidential support for the evaluated originating-source claim.

The interacted representation is processed by separate task heads:

$$\hat{y}_d = \sigma\big(f_d(\mathbf{Pool}(\mathbf{H}_c), \mathbf{h}_e)\big),$$
$$\hat{\mathbf{Y}}_\ell = \sigma\big(f_\ell(\mathbf{H}_c, \mathbf{Z}_c, \mathbf{h}_e)\big), \qquad \hat{\mathbf{p}}_s = \mathbf{softmax}\Big(f_s\big(\mathbf{H}_q, \mathbf{h}_e, \mathbf{a}, \mathbf{v}\big)\Big), \tag{12}$$

where H_c contains interacted visual tokens, H_q is the claim-token representation, and h_e summarizes the available nonvisual evidence. The localization head retains spatial features through connections to Z_c.

Image-level detection is trained using binary cross-entropy. Localization combines pixel-level binary cross-entropy with a Dice loss:

$$\mathcal{L}_{\mathbf{det}} = \mathbf{BCE}(y_d, \hat{y}_d),$$
$$\mathcal{L}_{\mathbf{loc}} = \mathbf{BCE}\big(\mathbf{Y}_\ell, \hat{\mathbf{Y}}_\ell\big) + \lambda_{\mathbf{Dice}}\left[1 - \frac{2\sum_p Y_{\ell,p}\hat{Y}_{\ell,p} + \epsilon}{\sum_p Y_{\ell,p} + \sum_p \hat{Y}_{\ell,p} + \epsilon}\right], \tag{13}$$

where Y_ℓ is the ground-truth manipulation mask. Localization supervision is applied only when an appropriate mask is available. Routine distribution transformations preserve the manipulation label, while geometric operations also transform the ground-truth mask.

Source verification uses categorical cross-entropy over the three evidence outcomes. Compatibility supervision is applied only to relationships with defined labels:

$$\mathcal{L}_{\mathbf{src}} = -\sum_{r=1}^{3} y_{s,r} \log \hat{p}_{s,r},$$
$$\mathcal{L}_{\mathbf{cons}} = \frac{\sum_{(i,j)\in\mathcal{P}} \mu_{ij}\ \mathbf{BCE}(c_{ij}, \hat{c}_{ij})}{\sum_{(i,j)\in\mathcal{P}} \mu_{ij} + \epsilon}, \tag{14}$$

where $\mathcal{P}$ is the set of supervised evidence relationships and μ_{ij} indicates whether a relationship has an unambiguous compatibility label.

The recognition objective is

$$\mathcal{L}_{\mathbf{total}} = \lambda_d \mathcal{L}_{\mathbf{det}} + \lambda_\ell \mathcal{L}_{\mathbf{loc}} + \lambda_s \mathcal{L}_{\mathbf{src}} + \lambda_c \mathcal{L}_{\mathbf{cons}}. \tag{15}$$

For the initial implementation, the watermark insertion and extraction operators are fixed while the evidence encoders, interaction layers, and task heads are trained. This separation allows the contribution of representation learning to be evaluated without confounding it with changes to watermark strength or payload capacity. Watermark quality and recovery are assessed separately.

Training examples include mutually compatible evidence, controlled contradictions, and missing-evidence conditions. Contradictory examples are generated through source-claim substitution, mismatched content-instance records, and inconsistent watermark-record associations. Evidence dropout is additionally used to simulate unavailable branches. Importantly, removal of an evidence source does not automatically generate an insufficient-evidence label; the resulting verification label depends on whether the remaining observations satisfy the adopted source-verification policy.

Manipulation status and source-support status are varied independently whenever permitted by the experimental construction. Consequently, the training data include manipulated images

with supported originating-source claims as well as unmanipulated images associated with contradicted claims. This design prevents the source-verification head from simply reproducing the manipulation decision.
At inference time, detection and localization thresholds are selected using validation data. Claimed-source verification reports the predicted evidence state together with explicit provenance validation outcomes. Confidence-based rejection, when applied, is treated separately from the insufficient-evidence class because the former represents predictive uncertainty whereas the latter represents a supervised state concerning the sufficiency of available evidence.

4 Experimental Setup

This section describes the experimental protocol used to evaluate the proposed multimodal provenance representation learning method. The experiments are designed to assess three distinct but related tasks: deepfake detection, tampering localization, and claimed-source verification. Because the proposed method relies on heterogeneous evidence from visual content, dual-domain watermarks, and provenance metadata, the evaluation considers not only recognition performance under standard conditions but also the behavior of the method when individual evidence sources are degraded, unavailable, or mutually inconsistent.
The experimental design includes conventional image transformations, generative manipulations, and attacks affecting watermark or provenance evidence. Particular attention is given to conditions in which manipulation status and source-support status do not coincide, such as edited images that retain valid source evidence or unmodified images associated with contradictory source claims. The following subsections describe the datasets and implementation details, followed by the comparison methods, attack conditions, and evaluation metrics used to quantify detection, localization, source verification, watermark recovery, and perceptual quality.

4.1 Datasets and Implementation Details

The experiments were conducted using a controlled corpus derived from CelebAMask-HQ as the primary dataset, together with FaceForensics++ and DeeperForensics-1.0 for cross-dataset evaluation. The controlled corpus was designed to support the joint evaluation of deepfake detection, tampering localization, and claimed-source verification using visual content, dual-domain watermark evidence, and structured provenance information. In contrast, the external datasets were used to evaluate generalization when watermark and provenance evidence were unavailable.
CelebAMask-HQ contains 30,000 high-resolution facial images with semantic masks covering facial and accessory regions. The dataset was divided into 20,000 training images, 5,000 validation images, and 5,000 test images. Identity information associated with CelebA was used to construct identity-disjoint partitions so that images belonging to the same identity appeared in only one subset. This partitioning strategy was adopted to prevent identity leakage between training, validation, and test data.
Each CelebAMask-HQ source image was associated with a controlled source record and processed using the watermarking procedure described in Section 3. A corresponding manipulated sample was subsequently generated for each canonical image. The resulting corpus therefore consisted of 30,000 canonical images and 30,000 manipulated images, yielding a total of 60,000 experimental instances.
Four manipulation categories were considered: face replacement, facial expression or attribute editing, localized generative editing, and generative regeneration. The manipulated samples were evenly distributed across the four categories. The training partition therefore

contained 20,000 canonical and 20,000 manipulated samples, while the validation and test partitions each contained 5,000 canonical and 5,000 manipulated samples.

Table 2. Composition of the controlled experimental corpus.

Partition	Canonical	Manipulated	Face replacement	Expression/ attribute	Generative editing	Regeneration	Total
Training	20,000	20,000	5,000	5,000	5,000	5,000	40,000
Validation	5,000	5,000	1,250	1,250	1,250	1,250	10,000
Test	5,000	5,000	1,250	1,250	1,250	1,250	10,000
Total	30,000	30,000	7,500	7,500	7,500	7,500	60,000

Claimed-source verification conditions were constructed independently of manipulation status. Supported samples contained mutually consistent source claims, watermark evidence, and provenance records. Contradicted samples were generated through source-claim substitution or mismatched associations between recovered watermark information and source records. Insufficient-evidence samples were constructed by withholding selected evidence sources.

This design was intended to prevent source verification from collapsing into manipulation classification. In particular, manipulated samples could retain valid originating-source evidence, whereas unmanipulated samples could be paired with contradictory source claims. As a result, manipulation status and source-support status were varied independently within the controlled experimental corpus.

FaceForensics++ was used as the primary external benchmark for cross-dataset evaluation. Five temporally distributed frames were extracted from each original and manipulated video. The resulting evaluation set contained approximately 25,000 frames covering original sequences and four manipulation methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Video-level separation was performed before frame extraction to prevent frames originating from the same video from being distributed across evaluation subsets.

Because FaceForensics++ samples were not generated through the participating watermarking process used in the controlled corpus, latent-domain watermark, pixel-domain watermark, and provenance observations were treated as unavailable. The corresponding evidence branches were therefore explicitly masked during inference.

DeeperForensics-1.0 was additionally used to evaluate robustness under cross-dataset domain shift. A balanced subset of 10,000 frames was constructed, consisting of 5,000 pristine and 5,000 manipulated samples selected from distinct videos. As with FaceForensics++, no artificial watermark or provenance information was added to these samples. This setting enabled evaluation of the visual representation and the missing-evidence handling mechanism without introducing synthetic source-related evidence into the external benchmark.

All input images were resized to 512 × 512 pixels. When geometric transformations were applied to samples with spatial manipulation annotations, the corresponding ground-truth masks were transformed using the same geometric parameters. JPEG compression, resizing, blur, noise, rotation, and cropping were treated as robustness conditions rather than as additional manipulation classes and are described separately in Section 4.2.

The watermark insertion and extraction operators were kept fixed throughout the experiments. Only the evidence encoders, cross-evidence interaction components, and task-specific

prediction heads were jointly optimized. This experimental separation was adopted so that changes in authentication performance could be attributed to the proposed evidence representation and interaction procedure rather than to changes in watermark embedding or extraction strength.
The hardware and software environment used in the experiments is summarized in Table 3.

Table 3. Hardware and software environment used in the experiments.

Category	Component	Specification / Version
Hardware	CPU	AMD Ryzen 9 7950X, 16 cores / 32 threads
	GPU	NVIDIA GeForce RTX 4090, 24 GB VRAM
	System memory	64 GB DDR5 RAM
System	Operating system	Ubuntu 24.04 LTS, 64-bit
	Programming language	Python 3.12
	GPU computing	CUDA 13.0
Deep learning	PyTorch	2.14.0
	torchvision	0.29.0
Image processing	OpenCV	4.14.0
	Albumentations	2.0.8
Generative processing	Diffusers	0.40.0
	Transformers	5.17.0
Numerical analysis	NumPy	2.5.3
	scikit-learn	1.9.1
Provenance processing	C2PA SDK	Stable release

All experiments were conducted on the same workstation environment. Model selection, threshold determination, and hyperparameter adjustment were performed exclusively using the validation partition. The test partition and external benchmark datasets were reserved for final evaluation and were not used for model selection.
Principal experiments were repeated using three independent random seeds to reduce the influence of initialization and sampling variability, and the reported results represent the mean performance across the three runs.

4.2 Baselines, Attack Conditions, and Evaluation Metrics

The proposed method is evaluated against single-evidence and multimodal baselines to determine whether performance improvements originate from cross-evidence representation learning rather than from the simple availability of additional information. The comparisons include visual-only recognition, watermark-based verification, provenance-based verification, feature-level fusion, decision-level fusion, and the proposed consistency-aware method. Representative active forensic approaches discussed in Section 2 are additionally considered where their evaluation targets are directly comparable. This comparison strategy follows the methodological distinction established in the related work: watermark recovery, manipulation detection, localization, and claimed-source verification are evaluated as related but non-equivalent outcomes.
The principal baseline configurations are summarized in Table 4. The Visual-only configuration uses only image features and therefore represents conventional passive manipulation recognition. The Watermark-only configuration uses the recovered latent- and pixel-domain watermark observations without visual or provenance features. The

Provenance-only configuration uses structured source assertions and validation states. Early Fusion concatenates the available modality representations before prediction, whereas Decision Fusion independently produces predictions from each evidence branch and combines their output scores. The proposed method uses reliability-aware interaction and explicit cross-evidence compatibility learning as described in Section 3.

Table 4. Baseline configurations used for comparison.

Method	Visual	Latent WM	Pixel WM	Provenance	Cross-evidence consistency
Visual-only	✓	–	–	–	–
Watermark-only	–	✓	✓	–	–
Provenance-only	–	–	–	✓	–
Visual + Watermark	✓	✓	✓	–	–
Early Fusion	✓	✓	✓	✓	–
Decision Fusion	✓	✓	✓	✓	–
Proposed	✓	✓	✓	✓	✓

Robustness is evaluated under conventional signal transformations, geometric transformations, generative manipulation, and evidence-targeted attacks. Conventional transformations are applied independently at predefined severity levels to measure degradation trends rather than performance at a single operating point. JPEG compression is evaluated at quality factors of 90, 70, 50, and 30. Image resizing is performed using scale factors of 0.75, 0.50, and 0.25, followed by restoration to the input resolution. Gaussian blur is applied with standard deviations of 0.5, 1.0, and 2.0, and additive Gaussian noise is evaluated with standard deviations of 5, 10, and 20 on the 8-bit image scale. Geometric robustness is measured using rotations of ±5°, ±10°, and ±15° and cropping ratios of 5%, 10%, and 20%.

Generative attack conditions include localized generative editing and image regeneration. Local editing modifies selected facial or surrounding regions while preserving the remainder of the image, whereas regeneration reconstructs a larger portion or the complete image through a generative model. These conditions are evaluated separately because local editing primarily affects spatial consistency, whereas regeneration may simultaneously alter visual characteristics and embedded watermark evidence. The current recognition formulation already distinguishes such semantic manipulations from routine distribution transformations.

Watermark-targeted conditions include removal, degradation, and conflicting-evidence attacks. Removal attacks use denoising, filtering, recompression, and generative regeneration to reduce watermark recoverability. Overwriting is evaluated by introducing a second watermark association after generation, while transfer or mismatch conditions pair recovered watermark information with an incompatible source or content-instance record. These attacks are intended to determine whether the method relies only on successful watermark decoding or also evaluates whether the recovered information is consistent with the observed image and provenance claim. This distinction is important because successful watermark recovery is treated as evidence of recoverability rather than proof that the recovered payload belongs to the current content.

Provenance robustness is evaluated using complete, missing, partially available, and conflicting evidence conditions. Missing-evidence experiments independently remove latent watermark, pixel watermark, or provenance observations and additionally evaluate combinations in which multiple branches are unavailable. Conflicting conditions include

substituted source claims, mismatched content-instance identifiers, inconsistent watermark-record associations, and provenance records whose source assertions do not agree with the available watermark evidence. Missing information is not treated as direct evidence of manipulation or contradiction; instead, the model may produce the insufficient-evidence outcome defined in Section 3.

Table 5. Robustness conditions used in the experiments.

Category	Transformation / Attack	Evaluation levels
Compression	JPEG compression	QF = 90, 70, 50, 30
Scaling	Resize and restore	0.75×, 0.50×, 0.25×
Filtering	Gaussian blur	σ = 0.5, 1.0, 2.0
Noise	Gaussian noise	σ = 5, 10, 20
Geometry	Rotation	±5°, ±10°, ±15°
Geometry	Cropping	5%, 10%, 20%
Generative	Local generative editing	Facial / local regions
Generative	Regeneration	Partial / full-image
Watermark	Removal / degradation	Filtering, recompression, regeneration
Watermark	Overwriting / mismatch	Secondary or inconsistent payload
Provenance	Missing evidence	Single- and multi-branch removal
Provenance	Conflicting evidence	Source, instance, or record mismatch

Evaluation metrics are selected separately for each task. Deepfake detection is measured using area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, and F1-score. AUC is used as the principal threshold-independent detection metric, while F1-score is reported to characterize the balance between manipulated and canonical samples.

Tampering localization is evaluated using Intersection over Union (IoU), Dice coefficient, and pixel-level F1-score between the predicted manipulation map and the ground-truth mask. Localization metrics are reported only for samples with spatially defined manipulation masks. For global regeneration conditions, detection results are reported independently unless a meaningful full-image manipulation mask is defined.

Claimed-source verification is evaluated as a three-class problem consisting of supported, contradicted, and insufficient-evidence outcomes. Overall accuracy and macro-averaged precision, recall, and F1-score are reported together with the corresponding confusion matrix. Macro-F1 is used as the principal metric because it assigns equal importance to all three evidence outcomes and is less affected by differences in class frequency.

Watermark recovery is evaluated independently using bit accuracy and bit error rate (BER). Where the selected watermark implementation produces a decoding confidence score, recovery rate and confidence are additionally reported. Perceptual distortion introduced by watermark embedding is measured using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS). These metrics are reported separately from authentication performance so that improved recognition cannot be attributed to increased watermark visibility.

Table 6. Evaluation metrics for the experimental tasks.

Evaluation target	Metrics	Principal metric
Deepfake detection	AUC, Accuracy, Precision, Recall, F1	AUC
Tampering localization	IoU, Dice, Pixel F1	IoU / Dice
Source verification	Accuracy, Macro-Precision, Macro-Recall, Macro-	Macro-F1

	F1	
Watermark recovery	Bit accuracy, BER, Recovery rate	BER
Perceptual quality	PSNR, SSIM, LPIPS	LPIPS / SSIM

For all comparison methods, the same training, validation, and test partitions are maintained wherever applicable. Input resolution, augmentation policy, and evaluation conditions are also kept consistent. Threshold-dependent results use thresholds selected exclusively from the validation set. For experiments repeated across three random seeds, the reported values represent the mean performance across the three runs. Together, these comparisons are intended to determine not only whether the proposed method improves individual task performance, but also whether consistency-aware learning provides additional robustness when the available visual, watermark, and provenance evidence becomes degraded, incomplete, or contradictory.

5 Results and Discussion

This section evaluates the proposed consistency-aware multimodal authentication method from four complementary perspectives. First, overall performance is examined for deepfake detection, tampering localization, and claimed-source verification. Second, cross-dataset experiments evaluate behavior when active source-related evidence is unavailable. Third, robustness experiments analyze degradation under conventional transformations, generative manipulation, watermark-targeted attacks, and incomplete or conflicting provenance evidence. Finally, controlled ablation experiments quantify the contribution of individual evidence sources, reliability-aware interaction, and cross-evidence consistency learning.

The analysis is intended not only to determine whether multimodal evidence improves recognition accuracy but also to examine whether the proposed method remains reliable when heterogeneous authentication evidence becomes degraded, unavailable, or mutually inconsistent.

5.1 Detection, Localization, and Source Verification

This section evaluates the proposed method from three perspectives: deepfake detection, tampering localization, and claimed-source verification. These tasks are treated as related but distinct prediction problems rather than as a single authenticity decision. Evidence supporting the originating source may remain valid even after content manipulation, whereas an unmanipulated image may still be associated with a source claim that conflicts with the available watermark or provenance evidence.

First, deepfake detection performance was evaluated on the controlled test set derived from CelebAMask-HQ. Table 7 compares the Visual-only, Watermark-only, Provenance-only, Visual + Watermark, Early Fusion, Decision Fusion, and proposed methods.

Table 7. Deepfake detection performance on the controlled test set.

Method	AUC	Accuracy	Precision	Recall	F1-score
Visual-only	0.918	0.864	0.871	0.852	0.861
Watermark-only	0.812	0.753	0.768	0.729	0.748
Provenance-only	0.731	0.684	0.701	0.652	0.676
Visual + Watermark	0.944	0.891	0.898	0.883	0.890
Early Fusion	0.952	0.903	0.909	0.896	0.902
Decision Fusion	0.947	0.896	0.902	0.889	0.895
Proposed	0.969	0.926	0.931	0.919	0.925

The improvement over Early Fusion and Decision Fusion is particularly relevant to the objective of this study. All three methods have access to the same major evidence sources, but only the proposed method explicitly models evidence reliability and cross-evidence compatibility. The observed performance difference therefore suggests that the advantage does not arise solely from increasing the number of input modalities. Instead, the manner in which heterogeneous authentication evidence is interpreted contributes to the final manipulation decision.

The proposed method achieved an AUC of 0.969, an accuracy of 0.926, and an F1-score of 0.925, showing the highest performance among the compared methods. Compared with the Visual-only method, AUC increased by 0.051, accuracy by 0.062, and F1-score by 0.064. The results show that incorporating watermark and provenance evidence in addition to visual features improves manipulation recognition compared with relying exclusively on image-derived characteristics.

Early Fusion and Decision Fusion also benefited from multiple evidence sources, achieving AUC values of 0.952 and 0.947, respectively. However, the proposed method improved AUC by 0.017 and F1-score by 0.023 over Early Fusion, while the corresponding improvements over Decision Fusion were 0.022 and 0.030. These results indicate that explicitly modeling evidence reliability and cross-evidence consistency provides additional discriminative information beyond simple feature concatenation or output-level aggregation.

Next, tampering localization performance was evaluated using localized manipulation samples for which ground-truth manipulation masks were available. Table 8 presents the IoU, Dice coefficient, and pixel-level F1-score.

Table 8. Tampering localization performance on localized manipulation samples.

Method	IoU	Dice	Pixel F1
Visual-only	0.714	0.833	0.827
Visual + Watermark	0.748	0.856	0.851
Early Fusion	0.772	0.871	0.866
Decision Fusion	0.758	0.862	0.857
Proposed	0.816	0.899	0.894

The proposed method achieved an IoU of 0.816, a Dice coefficient of 0.899, and a Pixel F1-score of 0.894. Compared with Visual-only, IoU increased by 0.102, while an improvement of 0.044 was observed over Early Fusion. The Dice coefficient increased by 0.066 compared with Visual-only and by 0.028 compared with Early Fusion.

Although non-visual evidence does not directly provide pixel-level localization information, the consistency relationships between the observed content and source-related evidence contribute auxiliary information for learning manipulation-sensitive representations. In particular, the proposed method retains spatial visual features while incorporating interactions with non-visual evidence, enabling manipulation localization without reducing the representation to an image-level authenticity decision.

Table 9 further analyzes localization performance according to manipulation type.

Table 9. Tampering localization performance according to manipulation type.

Manipulation type	IoU	Dice	Pixel F1
Face replacement	0.851	0.919	0.915
Expression / attribute editing	0.781	0.877	0.871
Localized generative editing	0.816	0.899	0.895

Face replacement achieved the highest localization performance, with an IoU of 0.851, followed by localized generative editing with 0.816. Expression and attribute editing showed a relatively lower IoU of 0.781. Face replacement generally introduces broader and more distinguishable spatial changes, whereas expression and attribute editing preserve much of the original facial structure and alter relatively limited regions. Consequently, the manipulation boundaries are less explicit in the latter case, making precise localization more difficult.

These results also indicate that nonvisual evidence can contribute to localization without being interpreted as a direct spatial manipulation mask. The localization head remains grounded in spatial visual representations, while watermark and provenance observations provide complementary context through consistency-aware interaction. This distinction is important because local watermark recoverability and actual manipulation boundaries are not equivalent quantities.

Claimed-source verification was then evaluated using three outcomes: supported, contradicted, and insufficient evidence. Overall accuracy and macro-averaged metrics were used to evaluate performance across the three evidence conditions.

Table 10. Claimed-source verification performance.

Method	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Watermark-only	0.842	0.847	0.838	0.841
Provenance-only	0.874	0.879	0.868	0.872
Visual + Watermark	0.889	0.893	0.884	0.888
Early Fusion	0.917	0.920	0.913	0.916
Decision Fusion	0.906	0.910	0.901	0.905
Proposed	0.946	0.948	0.943	0.945

The proposed method achieved an accuracy of 0.946 and a Macro-F1 of 0.945. Compared with Watermark-only, whose Macro-F1 was 0.841, the proposed method improved performance by 0.104. Compared with Provenance-only, the improvement was 0.073. Early Fusion and Decision Fusion achieved Macro-F1 scores of 0.916 and 0.905, respectively, whereas the proposed method exceeded them by 0.029 and 0.040.

Provenance-only achieved higher performance than Watermark-only because structured provenance records provide explicit source-related information that can be directly compared with the evaluated claim. However, provenance evidence alone is insufficient when records are missing, incomplete, or inconsistent. By jointly considering watermark recovery states, provenance validation results, visual representations, and claim information, the proposed method more effectively distinguishes whether the available evidence supports, contradicts, or is insufficient to evaluate a source claim.

From a multimedia authentication perspective, the three-state formulation is important because it distinguishes contradictory evidence from insufficient evidence. Conventional binary source verification may force a decision even when relevant source-related observations are missing. In contrast, the proposed formulation allows the verification process to preserve uncertainty arising from evidence availability while still identifying explicit conflicts when reliable observations contradict the evaluated claim.

To further examine the classification behavior of the proposed method, Table 11 presents the confusion matrix for the three source-verification outcomes.

Table 11. Confusion matrix for claimed-source verification (%).

Ground truth / Prediction	Supported	Contradicted	Insufficient evidence
Supported	95.7	1.6	2.7
Contradicted	1.9	94.8	3.3
Insufficient evidence	2.8	3.7	93.5

The proposed method correctly classified 95.7% of supported cases and 94.8% of contradicted cases. The correct classification rate for insufficient-evidence cases was 93.5%. Among the three outcomes, insufficient evidence showed the highest degree of confusion, with 3.7% of these samples classified as contradicted. This result indicates that distinguishing between a genuine conflict among evidence sources and a lack of sufficient evidence remains a more challenging aspect of source verification.

Finally, the relationship between manipulation status and source-verification outcomes was analyzed. Table 12 presents the joint distribution of the two decisions.

Table 22.

Manipulation status	Source verification	Number of samples	Ratio (%)
Unmanipulated	Supported	3,300	33.0
Unmanipulated	Contradicted	850	8.5
Unmanipulated	Insufficient evidence	850	8.5
Manipulated	Supported	1,650	16.5
Manipulated	Contradicted	2,500	25.0
Manipulated	Insufficient evidence	850	8.5
Total		10,000	100.0

Among the overall test cases, 1,650 samples, corresponding to 16.5%, were categorized as Manipulated + Supported. This result demonstrates that evidence supporting the originating source can remain valid even when the observed content has been manipulated. Conversely, 850 samples, corresponding to 8.5%, belonged to the Unmanipulated + Contradicted category, indicating that the absence of detected manipulation does not necessarily validate the claimed source.

These results support the separation of deepfake detection, tampering localization, and claimed-source verification into distinct prediction tasks rather than reducing them to a single authenticity score. The proposed method consistently outperformed single-evidence baselines and conventional fusion approaches across the three tasks, indicating the contribution of reliability-aware interaction and cross-evidence consistency learning. Section 5.2 next evaluates whether the learned representation generalizes to external deepfake datasets when watermark and provenance evidence are unavailable.

5.2 Cross-Dataset Generalization

To evaluate whether the proposed method generalizes beyond the controlled CelebAMask-HQ-based experimental corpus, cross-dataset experiments were conducted using FaceForensics++ and DeeperForensics-1.0. Unlike the controlled corpus, these external datasets do not originate from the participating generation process adopted in this study and therefore do not contain the latent-domain watermark, pixel-domain watermark, or structured provenance evidence used during multimodal authentication. Accordingly, all unavailable evidence branches were explicitly masked, and the evaluation focused on image-level manipulation detection using the visual representation and the proposed missing-evidence handling mechanism.

Table 13 presents the cross-dataset detection results. The Visual-only configuration was evaluated under the same external-data conditions to determine whether the proposed joint learning strategy retains useful manipulation representations when active source-related evidence is unavailable.

Table 13. Cross-dataset deepfake detection performance under unavailable watermark and provenance evidence.

Dataset	Method	AUC	Accuracy	F1-score
FaceForensics++	Visual-only	0.872	0.801	0.793
FaceForensics++	Proposed	0.901	0.836	0.829
DeeperForensics-1.0	Visual-only	0.831	0.768	0.759
DeeperForensics-1.0	Proposed	0.862	0.801	0.794

On FaceForensics++, the proposed method achieved an AUC of 0.901, an accuracy of 0.836, and an F1-score of 0.829. The Visual-only configuration achieved an AUC of 0.872, indicating an improvement of 0.029 for the proposed method. Although the cross-dataset AUC was lower than the 0.969 obtained on the controlled test set, the degradation is expected because the external benchmark contains manipulation processes, video acquisition characteristics, and post-processing conditions that were not represented by the controlled training distribution.

A similar tendency was observed on DeeperForensics-1.0. The proposed method achieved an AUC of 0.862, an accuracy of 0.801, and an F1-score of 0.794, whereas the Visual-only configuration obtained an AUC of 0.831. The AUC improvement of 0.031 indicates that the representation learned through joint training retains useful manipulation-related characteristics even when watermark and provenance observations are unavailable. Nevertheless, the lower absolute performance on DeeperForensics-1.0 compared with FaceForensics++ indicates increased sensitivity to domain shift and real-world perturbations.

Importantly, the external benchmark experiments do not represent the complete multimodal authentication setting because FaceForensics++ and DeeperForensics-1.0 do not contain compatible watermark payloads or provenance records from the participating generation process. Instead, these experiments evaluate whether the proposed method remains operational when all active source-related evidence is unavailable. Explicit evidence masking prevents unavailable watermark and provenance observations from being incorrectly interpreted as contradictory evidence and allows inference to depend on the remaining visual observations.

The cross-dataset experiments therefore complement rather than replace the controlled multimodal evaluation. The controlled corpus evaluates interactions among visual, watermark, and provenance evidence, whereas the external benchmarks evaluate visual generalization and missing-evidence handling under domain shift. The observed performance degradation on external datasets further indicates that cross-domain visual generalization remains a limitation of the current method and motivates future investigation of more domain-invariant manipulation representations.

5.3 Robustness under Generative Manipulations

This section analyzes the robustness of the proposed method under conventional signal transformations, generative manipulations, watermark attacks, and missing or conflicting provenance evidence. Whereas Section 5.1 evaluated detection, localization, and source-

verification performance under standard conditions, this section examines whether the proposed method maintains stable performance when the quality or availability of individual evidence sources is degraded. In particular, conventional distribution transformations such as JPEG compression and resizing are distinguished from generative attacks that directly affect the content or source-related evidence. This evaluation follows the experimental design that separately considers conventional distortions, generative editing, regeneration, watermark removal, overwriting, and missing or conflicting metadata.

First, the variation in deepfake detection performance under conventional image transformations was analyzed. JPEG compression, resizing, Gaussian blur, Gaussian noise, rotation, and cropping were applied, and Table 14 presents the performance measured at the highest severity level for each transformation

Table 14. Deepfake detection robustness under conventional transformations.

Condition	Severity	Visual-only AUC	Early Fusion AUC	Proposed AUC	Proposed F1
No attack	–	0.918	0.952	0.969	0.925
JPEG compression	QF = 30	0.861	0.921	0.945	0.899
Resizing	0.25×	0.836	0.906	0.928	0.881
Gaussian blur	$\sigma = 2.0$	0.848	0.914	0.936	0.889
Gaussian noise	$\sigma = 20$	0.842	0.909	0.932	0.884
Rotation	±15°	0.821	0.895	0.921	0.872
Cropping	20%	0.809	0.886	0.914	0.863

Under the no-attack condition, the proposed method achieved an AUC of 0.969. Under JPEG compression at QF 30, the AUC decreased to 0.945, corresponding to a reduction of 0.024. Under the same condition, the Visual-only method decreased from 0.918 to 0.861, representing a larger reduction of 0.057. This result indicates that when visual characteristics are degraded by compression, the use of multiple evidence sources enables the proposed method to maintain more stable detection performance.

Under resizing, Gaussian blur, and Gaussian noise, the proposed method achieved AUC values of 0.928, 0.936, and 0.932, respectively. Larger reductions were observed under rotation and cropping, where the AUC values decreased to 0.921 and 0.914. Cropping produced the largest degradation because it not only removes spatial image information but may also affect watermark synchronization and the spatial correspondence of manipulated regions.

Table 15 presents the detailed variation in detection AUC according to transformation severity.

Table 15. Detection AUC according to transformation severity.

Attack	Level 1	Level 2	Level 3	Level 4
JPEG	QF 90: 0.964	QF 70: 0.958	QF 50: 0.951	QF 30: 0.945
Resize	0.75×: 0.955	0.50×: 0.942	0.25×: 0.928	–
Gaussian blur	σ 0.5: 0.958	σ 1.0: 0.948	σ 2.0: 0.936	–
Gaussian noise	σ 5: 0.956	σ 10: 0.944	σ 20: 0.932	–
Rotation	±5°: 0.953	±10°: 0.937	±15°: 0.921	–
Cropping	5%: 0.949	10%: 0.932	20%: 0.914	–

Across all transformation conditions, AUC gradually decreased as attack severity increased. JPEG compression showed a relatively moderate decrease from 0.964 at QF 90 to 0.945 at

QF 30. In contrast, cropping produced the largest reduction, decreasing from 0.949 at 5% cropping to 0.932 at 10% and 0.914 at 20%. Rotation also caused a substantial decrease, from 0.953 at ±5° to 0.921 at ±15°, indicating that spatial misalignment has a comparatively strong influence on detection performance.

Next, performance degradation under generative editing and regeneration was analyzed. Local generative editing modifies only selected facial or surrounding regions, while partial regeneration reconstructs a broader portion of the image through a generative model. Full regeneration was defined as the strongest condition, in which the entire input image is regenerated.

Table 16. Performance under generative manipulation and regeneration.

Condition	Detection AUC	Detection F1	Localization IoU	Source Macro-F1
No additional attack	0.969	0.925	0.816	0.945
Local generative editing	0.953	0.904	0.798	0.927
Partial regeneration	0.921	0.866	0.742	0.889
Full regeneration	0.884	0.821	–	0.842

Under local generative editing, the detection AUC was 0.953 and the F1-score was 0.904, while the localization IoU was 0.798. Under partial regeneration, the AUC decreased to 0.921 and the F1-score to 0.866, while localization IoU decreased to 0.742. Full regeneration produced the largest degradation, with an AUC of 0.884 and an F1-score of 0.821.

A similar trend was observed for source verification. Macro-F1 decreased from 0.945 under the no-attack condition to 0.927 under local generative editing, 0.889 under partial regeneration, and 0.842 under full regeneration. This behavior can be attributed to the fact that regeneration affects not only visual characteristics but also watermark recoverability and the relationship between content and source evidence.

To examine the effect of generative attacks on watermark recovery in greater detail, Table 17 reports the BER of latent-domain and pixel-domain watermarks together with the overall recovery rate.

Table 17. Watermark recovery and source verification under watermark-targeted attacks.

Attack condition	Latent WM BER ↓	Pixel WM BER ↓	WM recovery rate ↑	Source Macro-F1 ↑
No attack	0.006	0.011	0.988	0.945
JPEG QF 30	0.009	0.061	0.957	0.932
Blur $\sigma = 2.0$	0.012	0.084	0.941	0.919
Denoising	0.018	0.113	0.913	0.903
Partial regeneration	0.071	0.186	0.824	0.889
Full regeneration	0.148	0.287	0.681	0.842
Watermark overwriting	0.022	0.196	0.876	0.871

Under the no-attack condition, the BER values of the latent-domain and pixel-domain watermarks were 0.006 and 0.011, respectively, while the overall watermark recovery rate was 0.988. Under JPEG compression at QF 30, the latent-domain BER increased only slightly to 0.009, whereas the pixel-domain BER increased to 0.061. Similarly, under Gaussian blur and denoising, the latent-domain BER remained relatively low at 0.012 and 0.018, while the pixel-domain BER increased to 0.084 and 0.113.

These results indicate that the two watermark channels exhibit different degradation characteristics under conventional signal transformations. In particular, the pixel-domain watermark is more sensitive to compression, filtering, and denoising, whereas latent-domain evidence remains relatively stable under the same conditions.

Under partial regeneration, the latent-domain and pixel-domain BER values increased to 0.071 and 0.186, respectively, while the overall recovery rate decreased to 0.824. Under full regeneration, the BER values increased further to 0.148 and 0.287, and the recovery rate decreased to 0.681. Correspondingly, the source-verification Macro-F1 decreased to 0.842.

However, degradation in watermark recovery did not result in an equivalent reduction in source-verification performance. For example, under partial regeneration, the watermark recovery rate decreased to 0.824, whereas the source-verification Macro-F1 remained at 0.889. This result reflects the fact that the source-verification head does not rely exclusively on watermark recovery but also considers provenance validation and visual evidence.

Under the watermark-overwriting condition, the recovery rate remained at 0.876, whereas the source-verification Macro-F1 decreased to 0.871. This finding demonstrates that watermark recoverability alone is insufficient for reliable source verification. The consistency between the recovered payload, provenance record, and claimed source is therefore more important than the mere presence of a recoverable watermark.

Next, the effect of missing individual evidence branches on source verification was analyzed.

Table 18. Source verification under missing-evidence conditions.

Available evidence	Accuracy	Macro-F1	Insufficient-evidence recall
All evidence	0.946	0.945	0.935
Without latent WM	0.932	0.930	0.927
Without pixel WM	0.928	0.925	0.922
Without provenance	0.901	0.897	0.909
Without both WMs	0.878	0.872	0.896
Visual only	0.741	0.728	0.846

When all evidence branches were available, source-verification accuracy and Macro-F1 were 0.946 and 0.945, respectively. Removing the latent-domain watermark reduced Macro-F1 to 0.930, while removing the pixel-domain watermark resulted in a Macro-F1 of 0.925. The relatively limited degradation in these cases indicates that when one watermark branch is unavailable, the remaining watermark and provenance evidence can partially compensate for the missing information.

Removing provenance evidence produced a larger performance reduction, with Macro-F1 decreasing to 0.897. When both watermark branches were removed, Macro-F1 decreased further to 0.872, and when only visual evidence was available, the score decreased to 0.728. These results indicate that provenance and watermark information play important roles in evaluating claimed-source consistency.

The insufficient-evidence recall was 0.935 when all evidence sources were available and remained above 0.909 when an individual evidence branch was removed. Although it decreased to 0.846 under the Visual-only condition, the method was still able to distinguish limited source evidence from direct contradiction rather than systematically treating evidence absence as a contradicted claim.

Finally, source-verification performance was evaluated under conflicting-evidence conditions, which are more challenging than simple evidence absence. Four conflict types were considered: source-claim substitution, instance-ID mismatch, watermark-record mismatch, and provenance-watermark conflict.

Table 19. Source verification under conflicting-evidence conditions.

Conflict condition	Watermark-only Macro-F1	Early Fusion Macro-F1	Proposed Macro-F1
Source-claim substitution	0.772	0.881	0.928
Instance-ID mismatch	0.801	0.893	0.936
Watermark-record mismatch	0.748	0.862	0.917
Provenance-watermark conflict	0.713	0.844	0.904
Average	0.759	0.870	0.921

The proposed method achieved Macro-F1 scores of 0.928 for source-claim substitution, 0.936 for instance-ID mismatch, and 0.917 for watermark-record mismatch. Even under the most difficult provenance-watermark conflict condition, the method maintained a Macro-F1 of 0.904.

Across the four conflicting-evidence conditions, the proposed method achieved an average Macro-F1 of 0.921, compared with 0.870 for Early Fusion and 0.759 for Watermark-only. Thus, the proposed method outperformed Early Fusion by 0.051 and Watermark-only by 0.162 in average Macro-F1.

The largest degradation was observed under provenance-watermark conflict. The Watermark-only method achieved a Macro-F1 of 0.713, while Early Fusion decreased to 0.844. In contrast, the proposed method maintained a Macro-F1 of 0.904. This result indicates that when multiple evidence sources are available but imply different source associations, simply concatenating their representations is insufficient. Explicit modeling of compatibility among evidence sources is required to distinguish contradictory evidence from reliable source support.

Overall, the proposed method exhibited relatively limited degradation under conventional transformations such as JPEG compression, blur, and noise, while larger performance reductions were observed under regeneration, which simultaneously alters visual characteristics and watermark evidence. Nevertheless, when individual evidence branches were missing or mutually inconsistent, the proposed method maintained higher performance than single-evidence and conventional fusion baselines by exploiting relationships among the remaining observations.

The different behaviors observed under missing and conflicting evidence further support the design choice of treating these conditions separately. When evidence is unavailable, the insufficient-evidence outcome allows the method to avoid an unjustified source decision. In contrast, when multiple reliable evidence sources support incompatible source associations, cross-evidence consistency learning enables explicit recognition of contradiction. Section 5.4 further investigates the contribution of individual components through ablation experiments that progressively remove visual evidence, dual-domain watermark information, provenance metadata, reliability-aware interaction, and consistency learning.

The robustness results demonstrate a clear distinction between conventional multimedia transformations and transformations that simultaneously alter content semantics and embedded evidence. Compression, resizing, and moderate geometric operations primarily degrade signal quality, whereas strong generative regeneration can modify visual forensic

traces, watermark recoverability, and content-level associations at the same time. The latter condition therefore represents a substantially more challenging authentication scenario.
The practical implication is that robustness should not be evaluated solely according to whether a single detector or watermark survives a transformation. Multimedia authentication requires examining whether the remaining evidence is still sufficient and mutually consistent for the specific inference being made. The proposed method addresses this requirement by retaining separate evidence states and allowing the contribution of individual branches to vary according to their availability and reliability.

5.4 Ablation Study

This section analyzes the contribution of individual evidence sources and learning components to the overall performance of the proposed method. The ablation experiments progressively remove visual evidence, latent-domain and pixel-domain watermark information, provenance metadata, reliability-aware interaction, and cross-evidence consistency learning. All configurations are trained and evaluated using the same data partitions, optimization settings, and computational budget to ensure that performance differences originate from the evaluated components rather than from differences in training conditions.
The first experiment investigates the contribution of each evidence source. Table 20 presents the performance obtained when visual, watermark, and provenance evidence are progressively combined. Deepfake detection is evaluated using AUC, tampering localization using IoU, and claimed-source verification using Macro-F1.

Table 20. Ablation of individual evidence sources.

Configuration	Detection AUC	Localization IoU	Source Macro-F1
Visual only	0.918	0.714	0.728
Visual + Latent WM	0.936	0.744	0.842
Visual + Pixel WM	0.931	0.751	0.827
Visual + Dual-domain WM	0.944	0.768	0.888
Visual + Provenance	0.947	0.759	0.897
Visual + Dual-domain WM + Provenance	0.958	0.784	0.923
Full proposed method	0.969	0.816	0.945

The Visual-only configuration achieved an AUC of 0.918, an IoU of 0.714, and a source-verification Macro-F1 of 0.728. Adding latent-domain watermark evidence increased AUC to 0.936 and source Macro-F1 to 0.842, while adding pixel-domain watermark evidence resulted in an AUC of 0.931 and a Macro-F1 of 0.827. The latent-domain watermark therefore provided a larger contribution to source verification, whereas the pixel-domain watermark produced a slightly greater improvement in localization.
Combining both watermark domains further increased AUC to 0.944, localization IoU to 0.768, and source Macro-F1 to 0.888. Compared with either watermark branch individually, this result indicates that the two watermark domains provide complementary evidence rather than redundant information.
The Visual + Provenance configuration achieved a source Macro-F1 of 0.897, exceeding the configurations that used only one watermark domain. This result is consistent with the structured source information provided by provenance records. However, the combination of visual, dual-domain watermark, and provenance evidence further increased Macro-F1 to

0.923, indicating that provenance information alone does not replace embedded source evidence.

The full proposed method achieved the highest performance, with an AUC of 0.969, localization IoU of 0.816, and source Macro-F1 of 0.945. The additional gains over simple multimodal evidence inclusion indicate that performance improvement is not solely attributable to access to a larger number of evidence sources.

The second ablation experiment isolates the effects of reliability-aware interaction and cross-evidence consistency learning. Table 21 compares conventional multimodal fusion with progressively added interaction components.

Table 21. Ablation of evidence interaction and consistency learning.

Configuration	Detection AUC	Localization IoU	Source Macro-F1
Early Fusion	0.952	0.772	0.916
+ Availability masking	0.956	0.779	0.923
+ Reliability-aware gating	0.961	0.793	0.932
+ Cross-evidence interaction	0.965	0.805	0.939
+ Consistency learning (Proposed)	0.969	0.816	0.945

Early Fusion achieved an AUC of 0.952, localization IoU of 0.772, and source Macro-F1 of 0.916. Introducing explicit availability masking increased Macro-F1 to 0.923, indicating that distinguishing unavailable evidence from observed but unreliable evidence contributes to source verification.

Reliability-aware gating further improved AUC to 0.961 and source Macro-F1 to 0.932. This improvement indicates that dynamically estimating the usefulness of individual evidence branches is more effective than treating all available evidence as equally reliable.

Cross-evidence interaction increased localization IoU from 0.793 to 0.805 and source Macro-F1 from 0.932 to 0.939. Finally, the addition of consistency learning produced the highest performance across all three tasks, with an AUC of 0.969, IoU of 0.816, and Macro-F1 of 0.945.

The contribution of consistency learning becomes more apparent under conflicting-evidence conditions. Table 22 compares the proposed model with variants in which the compatibility objective or reliability mechanism is removed.

Table 22. Ablation under conflicting-evidence conditions.

Configuration	Source-claim substitution	Instance-ID mismatch	WM-record mismatch	Provenance-WM conflict	Average Macro-F1
Full model w/o reliability gating	0.897	0.906	0.881	0.862	0.887
Full model w/o consistency loss	0.903	0.914	0.892	0.871	0.895
Full model w/o compatibility predictor	0.911	0.920	0.899	0.879	0.902
Proposed	0.928	0.936	0.917	0.904	0.921

Removing reliability-aware gating reduced average Macro-F1 from 0.921 to 0.887. Removing the consistency loss resulted in an average Macro-F1 of 0.895, while removing the dedicated compatibility predictor reduced the score to 0.902.

The largest differences were observed under provenance-watermark conflict. The proposed method achieved a Macro-F1 of 0.904, whereas removing reliability gating, consistency loss, and the compatibility predictor resulted in scores of 0.862, 0.871, and 0.879, respectively. These results indicate that explicit modeling of evidence reliability and compatibility becomes particularly important when individually available evidence sources support incompatible conclusions.

The contribution of the dual-domain watermark design was further examined by separately removing the latent-domain and pixel-domain branches. Table 23 reports watermark recovery and source-verification performance under representative attack conditions.

Table 23. Ablation of dual-domain watermark evidence under attack conditions.

Configuration	No attack Macro-F1	JPEG QF 30	Regeneration	Overwriting	Average
Pixel WM only	0.827	0.781	0.694	0.716	0.755
Latent WM only	0.842	0.824	0.751	0.769	0.797
Dual-domain WM	0.888	0.861	0.803	0.821	0.843
Dual-domain WM + Provenance	0.923	0.901	0.854	0.866	0.886
Proposed	0.945	0.932	0.842	0.871	0.898

The Pixel-WM-only configuration achieved an average Macro-F1 of 0.755, while the Latent-WM-only configuration achieved 0.797. Their performance difference became particularly apparent under JPEG compression and regeneration, where latent-domain evidence remained more stable than pixel-domain evidence.

Combining the two watermark domains increased the average Macro-F1 to 0.843. The addition of provenance evidence further increased the score to 0.886. The complete method achieved the highest overall average of 0.898, indicating that the benefit of dual-domain watermarking is further enhanced when recovered watermark evidence is interpreted jointly with provenance and visual information.

An additional ablation was performed on the multi-task objective to determine whether jointly learning detection, localization, and source verification contributes to the shared representation. The results are shown in Table 24.

Table 24. Ablation of the joint learning objective.

Training objective	Detection AUC	Localization IoU	Source Macro-F1
Detection only	0.947	–	–
Detection + Localization	0.958	0.792	–
Detection + Source Verification	0.961	–	0.928
Localization + Source Verification	–	0.798	0.931
Joint tasks without consistency loss	0.965	0.805	0.939
Full joint objective	0.969	0.816	0.945

Training only the detection task resulted in an AUC of 0.947. Jointly training detection and localization increased detection AUC to 0.958 and produced an IoU of 0.792. Detection combined with source verification achieved an AUC of 0.961 and a source Macro-F1 of 0.928.

When all three prediction tasks were jointly trained without the consistency objective, the model achieved an AUC of 0.965, an IoU of 0.805, and a Macro-F1 of 0.939. The full objective further improved these values to 0.969, 0.816, and 0.945, respectively. This result indicates that the performance gain is associated not only with multi-task supervision but also with the explicit consistency constraint imposed across heterogeneous evidence sources.

Overall, the ablation experiments reveal three characteristics of the proposed method. First, visual, watermark, and provenance observations provide complementary information, and removal of any major evidence source reduces overall authentication performance. Second, dual-domain watermark evidence provides greater resilience than either watermark domain alone, particularly under signal degradation and generative regeneration. Third, reliability-aware interaction and cross-evidence consistency learning provide measurable improvements beyond conventional multimodal fusion, with their contribution becoming more pronounced under missing or conflicting evidence conditions.

These findings indicate that the observed performance gains cannot be attributed solely to an increased number of input modalities. Rather, the improvements result from combining complementary evidence sources while explicitly representing their availability, reliability, and mutual consistency. The multi-task ablation further shows that the joint objective contributes to shared representation learning, while the consistency objective provides additional gains beyond multi-task supervision alone.

From the perspective of multimedia authentication, this result supports the central assumption of the study: heterogeneous evidence should not be collapsed into a common authenticity indicator. Authentication performance depends not only on what evidence is available but also on how individual observations relate to one another and to the specific claim being evaluated.

6 Conclusion

This study presented a consistency-aware multimodal media authentication method for jointly addressing deepfake detection, tampering localization, and claimed-source verification. Rather than treating visual characteristics, embedded watermarks, and provenance metadata as interchangeable indicators of authenticity, the proposed method represents them as heterogeneous evidence sources with distinct semantic roles, availability conditions, and reliability characteristics. Visual features provide manipulation-related spatial information, latent-domain and pixel-domain watermarks provide complementary source and content associations, and structured provenance metadata provides explicit information regarding source claims, processing history, and validation states. Reliability-aware interaction and cross-evidence consistency learning are then used to determine how these observations jointly contribute to different authentication tasks.

The experimental analysis demonstrated that consistency-aware integration of heterogeneous evidence provides more reliable authentication than relying on individual evidence sources or conventional feature- and decision-level fusion. The proposed method achieved an AUC of 0.969 for deepfake detection, an IoU of 0.816 for tampering localization, and a Macro-F1 of 0.945 for claimed-source verification. Compared with Early Fusion, the corresponding improvements were 0.017, 0.044, and 0.029. The dual-domain watermark configuration improved robustness compared with either watermark domain alone, while provenance information contributed substantially to claimed-source verification. Reliability-aware interaction and cross-evidence consistency learning provided additional improvements, particularly when individual evidence sources were unavailable, degraded, or mutually inconsistent.

Robustness experiments further demonstrated that conventional distribution transformations produced relatively limited performance degradation, whereas generative regeneration and watermark-targeted attacks caused larger reductions because they can simultaneously affect visual characteristics and embedded evidence. Nevertheless, the combination of complementary evidence sources allowed the proposed method to maintain useful authentication performance under challenging conditions. These observations also demonstrate that robustness in multimedia authentication should not be interpreted solely as the survival of an individual watermark or visual detector but as the ability to retain sufficient and mutually consistent evidence for the specific inference being made.

An important finding of this study is that manipulation detection and source verification should not be interpreted as equivalent authenticity decisions. Manipulated content may retain valid evidence of its originating source, whereas visually unmanipulated content may be associated with an unsupported or contradictory source claim. Similarly, successful watermark recovery indicates that an embedded signal remains recoverable but does not independently establish that the recovered information remains correctly associated with the observed content. By explicitly separating manipulation detection, tampering localization, watermark recovery, provenance validation, and claimed-source verification, the proposed method provides a more structured interpretation of multimedia authenticity than a single binary authenticity decision.

Several limitations remain. The current evaluation assumes a participating generation process in which latent-domain watermark insertion and the corresponding extraction procedure are available. The watermark operators are fixed during training, and their robustness therefore depends on the characteristics of the selected embedding methods. Provenance verification additionally assumes that cryptographic validation results are available as input and does not address compromised signing credentials or failures of the underlying trust infrastructure. Strong generative regeneration may remove or substantially alter multiple evidence sources simultaneously, reducing the information available for reliable source verification. Furthermore, external benchmark datasets do not provide the same controlled watermark and provenance information as the primary experimental corpus, limiting cross-dataset evaluation primarily to visual recognition and missing-evidence conditions.

Future work will extend the proposed approach to a wider range of generative models, multimedia content domains, and previously unseen manipulation processes. Further research will investigate video-level authentication using temporal evidence, stronger adaptive watermark removal and transfer attacks, and verification of previously unseen source claims. Another direction is the integration of uncertainty calibration and evidence-aware rejection mechanisms to improve decision reliability when multiple evidence sources are simultaneously degraded. Deployment-oriented studies will additionally consider heterogeneous multimedia formats, distributed provenance information, and practical media-processing pipelines. Finally, closer integration with standardized provenance mechanisms such as C2PA and JPEG Trust may enable more systematic evaluation of relationships among cryptographically validated records, embedded content signals, and learned forensic evidence in practical multimedia authentication environments.

References

Arabi, K., Witter, R. T., Hegde, C., & Cohen, N. (2025). SEAL: Semantic aware image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 16196–16205). https://doi.org/10.1109/ICCV51701.2025.01503

Bharati, A., Moreira, D., Brogan, J., Hale, P., Bowyer, K. W., Flynn, P. J., Rocha, A., & Scheirer, W. J. (2019). Beyond pixels: Image provenance analysis leveraging metadata.

In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1692–1702). https://doi.org/10.1109/WACV.2019.00185

Bhowmik, D., Caldwell, S., Delgado, J., Ebrahimi, T., Fotos, N., Gu, X., Hu, Z., Kang, X., Pereira, F., Rosenthol, L., Temmermans, F., & Zhou, H. (2024). An international standard for assessing trustworthiness in media. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP)* (pp. 3799–3805). https://doi.org/10.1109/ICIP51287.2024.10647585

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4113–4122).

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., & Ji, R. (2021). Local relation learning for face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(2), 1081–1088. https://doi.org/10.1609/aaai.v35i2.16193

Coalition for Content Provenance and Authenticity. (2026). *Content Credentials: C2PA technical specification* (Version 2.4).

Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). The Stable Signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 22466–22477). https://doi.org/10.1109/ICCV51070.2023.02053

Gao, Q., Zhang, B., Wu, J., Luo, W., Teng, Z., & Fan, J. (2025). Leveraging facial landmarks improves generalization ability for deepfake detection. *Pattern Recognition, 164*, 111528. https://doi.org/10.1016/j.patcog.2025.111528

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4360–4369). https://doi.org/10.1109/CVPR46437.2021.00434

Huang, W., Valsecchi, M., & Multerer, M. (2025). Anisotropic multiresolution analyses for deepfake detection. *Pattern Recognition, 164*, 111551. https://doi.org/10.1016/j.patcog.2025.111551

Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., & Cheng, G. (2025). SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 28831–28841). https://doi.org/10.1109/CVPR52734.2025.02685

ISO/IEC. (2026). *ISO/IEC 21617-1:2026 Information technology—JPEG Trust—Part 1: Core foundation* (2nd ed.). International Organization for Standardization.

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2889–2898). https://doi.org/10.1109/CVPR42600.2020.00296

Lee, C.-H., Liu, Z., Wu, L., & Luo, P. (2020). MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5549–5558). https://doi.org/10.1109/CVPR42600.2020.00559

Li, J., Hu, Y., Liu, B., She, H., & Li, C.-T. (2025). Deepfake detection with domain generalization and mask-guided supervision. *Pattern Recognition, 165*, 111622. https://doi.org/10.1016/j.patcog.2025.111622

Li, K., Huang, Z., Hou, X., & Hong, C. (2025). GaussMarker: Robust dual-domain watermark for diffusion models. In *Proceedings of the 42nd International Conference*

on Machine Learning (ICML). *Proceedings of Machine Learning Research, 267*, 34688–34701.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-Ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5001–5010). https://doi.org/10.1109/CVPR42600.2020.00505

Li, Y., Yang, Y., Tan, Z., Liu, H., Chen, W., Zhou, X., & Lei, Z. (2025). Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9242–9252). https://doi.org/10.1109/CVPR52734.2025.00863

Liu, W., Cun, X., & Pun, C.-M. (2024). DH-GAN: Image manipulation localization via a dual homology-aware generative adversarial network. *Pattern Recognition, 155*, 110658. https://doi.org/10.1016/j.patcog.2024.110658

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 3730–3738). https://doi.org/10.1109/ICCV.2015.425

Luo, Z., Song, Q., Li, H., Rocha, A., & Wan, R. (2026). Adversarially robust multimedia watermarking via data-centric optimization. *Pattern Recognition, 180*, 114080. https://doi.org/10.1016/j.patcog.2026.114080

Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., Flynn, P. J., Rocha, A., & Scheirer, W. J. (2018). Image provenance analysis at scale. *IEEE Transactions on Image Processing, 27*(12), 6109–6123. https://doi.org/10.1109/TIP.2018.2865674

Müller, A., Lukovnikov, D., Thietke, J., Fischer, A., & Quiring, E. (2025). Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 20937–20946). https://doi.org/10.1109/CVPR52734.2025.01950

Nguyen, T. D., Azizpour, A., & Stamm, M. C. (2025). Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3040–3050). https://doi.org/10.1109/CVPR52734.2025.00289

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision—ECCV 2020* (pp. 86–103). Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-58610-2_6

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). https://doi.org/10.1109/ICCV.2019.00009

Sander, T., Fernandez, P., Durmus, A. O., Furon, T., & Douze, M. (2025). Watermark Anything with localized messages. In *International Conference on Learning Representations (ICLR)*.

Shao, R., Wu, T., & Liu, Z. (2023). Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6904–6913). https://doi.org/10.1109/CVPR52729.2023.00667

Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18720–18729). https://doi.org/10.1109/CVPR52688.2022.01816

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., & Ji, R. (2022). Dual contrastive learning for general face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(2), 2316–2324. https://doi.org/10.1609/aaai.v36i2.20130

Temmermans, F., Caldwell, S., Bhowmik, D., & Ebrahimi, T. (2024). JPEG Trust: An international standard facilitating the assessment of trustworthiness of digital media assets. *Proceedings of SPIE, 13137*, 131370B. https://doi.org/10.1117/12.3031171

Wang, X., & Tian, Y. (2026). Towards provenance-aware diffusion: Key-free watermarking with Gaussian shading. *Expert Systems with Applications, 306*, 130822. https://doi.org/10.1016/j.eswa.2025.130822

Wang, Z., Bao, J., Zhou, W., Wang, W., & Li, H. (2023). AltFreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4129–4138). https://doi.org/10.1109/CVPR52729.2023.00402

Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-Rings Watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems, 36*. https://doi.org/10.52202/075280-2529

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

Wu, X., Liao, X., Zhang, J., Chen, M., Wu, Y., & Guo, J. (2026). Versatile and harmless deepfake proactive forensics via conditional watermarking. *Information Sciences, 742*, 123030. https://doi.org/10.1016/j.ins.2025.123030

Xu, R., Hu, M., Lei, D., Li, Y., Lowe, D., Gorevski, A., Wang, M., Ching, E., & Deng, A. (2025). InvisMark: Invisible and robust watermarking for AI-generated image provenance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 909–918). https://doi.org/10.1109/WACV61041.2025.00098

Yan, Z., Zhang, Y., Fan, Y., & Wu, B. (2023). UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 22412–22423). https://doi.org/10.1109/ICCV51070.2023.02048

Yang, P., Ci, H., Song, Y., & Shou, M. Z. (2024). Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-1804

Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., & Yu, N. (2024). Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12162–12171). https://doi.org/10.1109/CVPR52733.2024.01156

Ye, X., Huang, H., Zhou, X., Wang, Y., & Ma, K.-K. (2026). CMUA++: A cross-model universal active defense framework for combating deepfake. *Pattern Recognition, 180*, 114364. https://doi.org/10.1016/j.patcog.2026.114364

Zhang, K., Kong, C., Wang, S., Rocha, A., & Li, H. (2025). Image provenance analysis via graph encoding with Vision Transformer. *IEEE Transactions on Information Forensics and Security, 20*, 4422–4437. https://doi.org/10.1109/TIFS.2025.3559787

Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., & Zhang, J. (2024). EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11964–11974).

Zhang, X., Tang, Z., Xu, Z., Li, R., Xu, Y., Chen, B., Gao, F., & Zhang, J. (2025). OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3008–3018).

Zhang, Z., Wang, Y., Cheng, L., Zhong, Z., Guo, D., & Wang, M. (2025). ASAP: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4005–4014). https://doi.org/10.1109/CVPR52734.2025.00379

Zhao, C., Wang, C., Song, Z., Hu, G., Wang, L., & Miao, D. (2025). Multi-definition deepfake detection via semantics reduction and cross-domain training. *Pattern Recognition, 163*, 111469. https://doi.org/10.1016/j.patcog.2025.111469

Zhao, C., Li, L., Ren, Y., & Feng, G. (2026). RI-Mark: Robust and imperceptible watermarking for diffusion models. *Expert Systems with Applications, 328*, 132903. https://doi.org/10.1016/j.eswa.2026.132903

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2185–2194).

Acknowledgment

This research received no external funding.