

***LightMark*: Towards Trustworthy Crowd-sourced Social Videos through In-Band Visible Light Watermarking**

RUI XIAO, Shanghai University of Finance and Economics, China and MoE Key Laboratory of Interdisciplinary Research of Computation and Economics, China

HUIYU XU, Zhejiang University, China

MING GAO, Nanjing University of Posts and Telecommunications, China

JINSONG HAN*, Zhejiang University, China

As social media videos become pervasive, they have emerged as a major vehicle for misinformation, most alarmingly through deepfakes that undermine trust in public recordings. Existing passive forgery detectors often fail to generalize across domains and unseen manipulations, motivating a more robust video authentication solution. In this work, we introduce *LightMark*, a novel infrastructure-assisted system that detects video forgery through in-band visible light watermarking, where a video digest continuously evolves with the speaker's face in the live scene. Notably, instead of attaching verification data after recording, *LightMark* performs in-band embedding, which seamlessly integrates the watermark into the video during capture. It achieves this by subtly modulating ambient illumination to create light flickers that are imperceptible to the human eye yet faithfully captured by rolling-shutter cameras commonly found in consumer devices, effectively turning ordinary cameras into simultaneous video recorders and watermark receivers. This design imposes no burden on the audience while enabling reliable detection of forged videos through watermark validation, with failed extraction resulting in an unverifiable outcome. We design and implement a proof-of-concept prototype to demonstrate *LightMark*'s feasibility. Extensive evaluations show that *LightMark* achieves an AUC of 0.99 across diverse video-editing and deepfake attacks, outperforming state-of-the-art baselines while maintaining robustness under varying environmental conditions.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

ACM Reference Format:

Rui Xiao, Huiyu Xu, Ming Gao, and Jinsong Han. 2026. *LightMark*: Towards Trustworthy Crowd-sourced Social Videos through In-Band Visible Light Watermarking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 4, Article 266 (December 2026), 29 pages. <https://doi.org/10.1145/3857964>

1 Introduction

Social media videos have become one of the most dominant forms of online content. The ubiquity of camera-equipped devices enables individuals to effortlessly capture and share videos on platforms such as YouTube, TikTok, and Instagram Reels. However, the ease of sharing and circulation also introduces significant risks: *forged videos* can spread rapidly across social platforms. The recent rise of *deepfake* technologies further exacerbates these threats [35, 69, 80, 81]. Such incidents, when involving institutional speakers such as school principals, hospital representatives, or local government officials, can cause substantial misinformation by spreading false security

*Corresponding author.

Authors' Contact Information: Rui Xiao, Shanghai University of Finance and Economics, China and MoE Key Laboratory of Interdisciplinary Research of Computation and Economics, Shanghai, China, xiaorui@sufe.edu.cn; Huiyu Xu, Zhejiang University, Hangzhou, China, huiyuxu@zju.edu.cn; Ming Gao, Nanjing University of Posts and Telecommunications, Nanjing, China, gaomingppm@njupt.edu.cn; Jinsong Han, Zhejiang University, Hangzhou, China, hanjinsong@zju.edu.cn.



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

© 2026 Copyright held by the owner/author(s).

ACM 2474-9567/2026/12-ART266

<https://doi.org/10.1145/3857964>

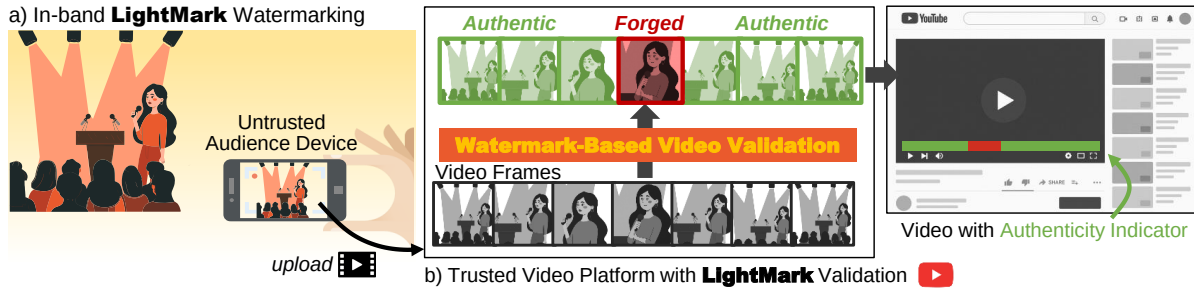


Fig. 1. Figure depicts an application scenario of *LightMark*. (a) *LightMark* performs **in-band visible light watermarking** that is captured seamlessly by untrusted video recording devices from audiences. (b) Upon upload, a trusted video platform conducts **watermark-based validation**. Once verified and published, the platform can display an authenticity indicator that visually distinguishes authentic and forged video segments.

notices, misleading medical guidance, or fabricated policy announcements. The reputation of the individual or institution can also be severely damaged, especially when the forged content depicts inflammatory, discriminatory, or otherwise harmful statements. These risks have also extended to several public figures, such as U.S. House Speaker Nancy Pelosi [8–10, 70], further raising concern over the authenticity of live-recorded videos.

However, protecting a publicly delivered speech from alteration is challenging when anyone in the audience can freely record and publish it online. Existing approaches generally follow two strategies: (a) artifact detection [20, 53, 61, 88] and (b) embedding and verification of meta-information [41, 55, 60, 75]. The artifact-based detector detects visual or temporal inconsistencies in manipulated videos, such as unnatural body motion, lighting, or texture, to identify potential forgeries. Yet, their performance often degrades significantly when applied across different domains or unseen manipulation types [43, 78, 80, 81]. Moreover, the rapid advancement of deepfake technology continues to erode the effectiveness of artifact-based detection [62, 91]. The second strategy focuses on video provenance through meta-information embedding, where cryptographic signatures or immutable markers are inserted at the source to enable verification [41, 75]. Such methods typically rely on *post hoc* embedding, i.e., attaching verification data after the recording process, using trusted devices or controlled distribution channels. However, they are largely ineffective for live public speeches, where anyone in the audience can independently record and distribute videos beyond the control of the original recording device. Without a means to securely bind verifiable information to these crowd-sourced recordings, it remains difficult to establish their authenticity once published.

In light of these limitations, we pose the question: *Can we design a video authentication system that seamlessly embeds verifiable meta-information into audience-recorded videos without imposing any burden on the audience?* To this end, we propose *LightMark*, a novel **infrastructure-assisted** authentication system for live speech events that embeds a visible-light watermark derived from the speaker’s identity and facial dynamics into audience-recorded videos. As the watermark continuously evolves with the live face, forged videos can be detected when the recovered watermark is absent, invalid, or temporally inconsistent. Unlike prior *post hoc* – or equivalently, *out-of-band* – embedding methods that attach verification data after recording, *LightMark* achieves **in-band embedding** by integrating the watermark directly into the video **during the recording**. *LightMark* innovatively realizes in-band watermarking by subtly modulating the background illumination, generating light flickers that are imperceptible to the human eye yet faithfully captured by rolling-shutter cameras commonly found in consumer mobile devices (§2.3). Remarkably, this process works instantly on any audience device without

requiring hardware or firmware modifications, effectively turning cameras into simultaneous video recorders and watermark receivers.

Figure 1 illustrates an example application scenario of *LightMark*. During a public event or speech, the ambient LED lighting encodes the speaker’s features into a dynamic visible-light watermark, generating subtle stripe-like patterns within audience-recorded frames. This process occurs naturally as part of video recording, without any additional steps or modifications to audience devices. When a trusted video platform equipped with the *LightMark* validation service, such as YouTube or TikTok, receives an uploaded video, it can perform watermark-based validation before publication. Once verified, the platform can display an **authenticity indicator** in the form of a segment-level timeline, visually distinguishing verified video segments from those marked as manipulated or unverifiable, thereby providing users with fine-grained feedback beyond a single video-level verdict.

Designing *LightMark* involves two primary challenges. The first is balancing watermark extraction robustness with video quality. Embedding a watermark using visible light introduces subtle stripe patterns in video frames. These patterns must be strong enough for reliable extraction but faint enough to stay invisible to viewers. This creates an inherent dilemma between robustness and visual quality. To address this, *LightMark* co-designs the embedding and extraction pipeline (§4.2). On the embedding side, the in-band watermarking module uses low-power LEDs to generate subtle stripe patterns. On the extraction side, *LightMark* employs a neural-based extraction and restoration module that leverages spatial correlations within each frame to accurately recover these subtle watermarks and, when necessary, restore the original video appearance, ensuring reliable verification and high visual fidelity.

The second challenge lies in achieving strong forgery resistance under strict bandwidth constraints. While remaining seamless and imperceptible to the audience, visible-light modulation inherently offers only a low-bandwidth communication channel. Within this narrow channel, the watermark must still carry cryptographically verifiable information and remain resilient against manipulations such as video slicing and advanced deepfake attacks. To address this, *LightMark* introduces a robust watermark content design (§4.3) that compactly encodes both the speaker’s identity and facial dynamics in real time. These features are transformed into a lightweight cryptographic digest that fits within the limited in-band watermarking channel through structured packetization and secure signing. Our design enables *LightMark* to deliver reliable and verifiable authentication even under severe bandwidth limitations.

We implement a proof-of-concept prototype of *LightMark* using commodity low-power LED illumination sources. *LightMark* is evaluated extensively through real-world experiments covering a wide range of video forgery scenarios, from conventional temporal attacks such as video slicing to advanced deepfake manipulations, resulting in a substantial dataset of 1,800 authentic and 5,520 forged videos. We comprehensively evaluate *LightMark*’s performance under environmental conditions and practical recording factors, including audience device types and locations. Overall, our evaluation shows that *LightMark* achieves an Area Under the Curve (AUC) greater than 0.99 in detecting forgeries across various attack types, significantly outperforming existing state-of-the-art watermarking methods.

- We propose *LightMark*, an in-band video forgery detection system that transforms ordinary rolling-shutter cameras into simultaneous video recorders and watermark receivers.
- We design and implement *LightMark* to achieve strong watermarking robustness while preserving video quality, leveraging low-power embedding and a neural-based watermark extraction framework.
- We conduct extensive real-world experiments demonstrating that *LightMark* attains superior detection performance compared with state-of-the-art baselines.

2 Background

We present the background of *LightMark*.

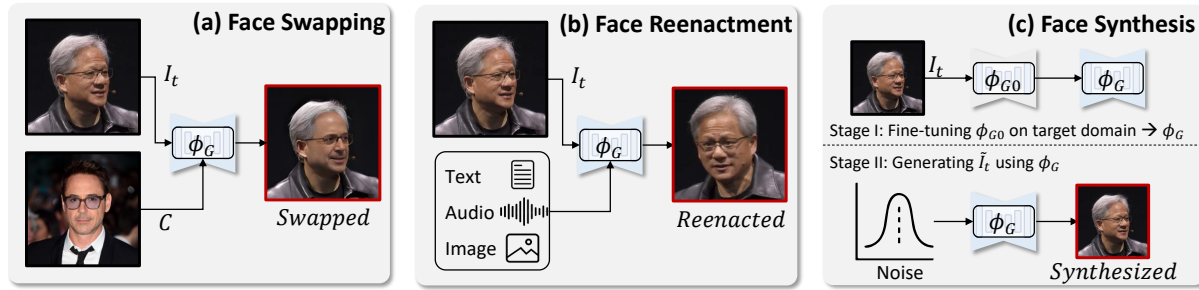


Fig. 2. Figure depicts three main deepfake-based video forgery attacks: (a) **Face swapping**, which replaces the original face in frame I_t with another identity. (b) **Face reenactment**, which manipulates the facial expression or motion of the same identity based on external conditions such as text, audio, or reference images. (c) **Face Synthesis**, which generates entirely new synthetic faces from scratch. When disseminated on social media, these forgeries can create highly realistic yet deceptive videos of public figures, leading to misinformation and erosion of public trust.

2.1 Video Forgery Detection and Limitations

Video forgery and modern manipulation techniques. Video forgery refers to the intentional alteration or generation of video content to misrepresent reality, often for the purpose of misinformation, impersonation, or defamation [81]. Such manipulations can have severe social and political consequences, particularly when used to fabricate speeches or appearances of public figures. Traditional video forgery typically involves non-AI manipulations such as frame removal, frame insertion, or temporal slicing, which modify the temporal continuity of the video. These techniques are often used to omit, repeat, or reorder segments to distort context or meaning.

Recent advances in deep generative models have enabled a new wave of AI-based forgeries, commonly referred to as **deepfakes** [47], can be divided into three categories: face swapping [15, 77], face reenactment [68, 69], and face synthesis [19, 57]. As illustrated in Figure 2, **face swapping** replaces the identity of a person in a video with another individual, whereas **face reenactment** reanimates the expressions or lip movements of the source face to match a target audio or motion sequence. Recent **face synthesis** techniques, such as Stable Diffusion [57], can generate entirely new synthetic faces from scratch. These techniques can fabricate convincing yet entirely falsified appearances of public figures, **enabling large-scale misinformation, identity theft, and reputational damage**, thereby posing serious threats to information integrity and public trust.

Existing deepfake detection and limitations. State-of-the-art deepfake detectors aim to determine whether an image or video has been manipulated, typically by analyzing semantic artifacts [3, 37, 72, 82], spatio-temporal irregularities [18, 22, 26], frequency-domain inconsistencies [25, 54, 66], or mismatches between the audio and visual modalities [59, 90]. Although these models achieve impressive accuracy under laboratory conditions, they often fail in real-world settings due to two fundamental limitations.

- (1) **Cross-domain generalization.** Domain¹ refers to the data and manipulation distribution, including its source dataset, forgery method, and post-processing pipeline [80, 81]. Existing detectors often perform well on the datasets used for training, but their performance drops significantly when tested on videos from unseen sources or generated by different deepfake methods. This domain dependency limits practical applicability, as deepfake styles, rendering artifacts, and post-processing pipelines vary widely in the wild.
- (2) **Sensitivity to video quality.** Existing detectors are vulnerable to compression, noise, and other quality degradations commonly found in online videos. Subtle forgery cues such as blending and frequency artifacts can be easily perturbed by these degradations, making forgeries indistinguishable from real content.

¹We follow the conventional definition of "domain" in prior deepfake detection works [80, 81].

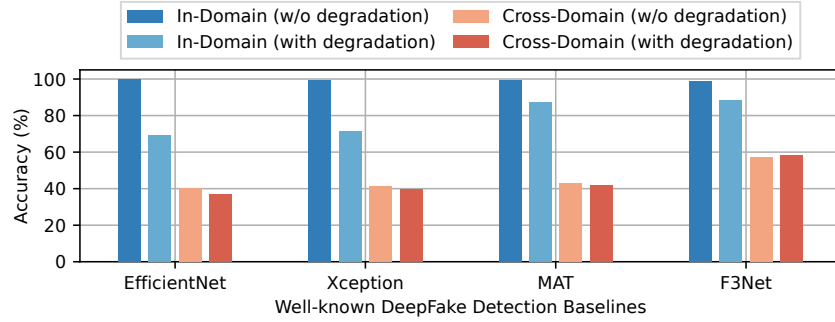


Fig. 3. Figure depicts limitations of passive Deepfake detection methods.

Empirical evaluation. To illustrate these limitations, we evaluate four representative detectors, i.e., F3Net [53], MAT [88], EfficientNet-B4 [67], and Xception [20]. F3Net focuses on abnormal frequency components in forged faces, while the others rely primarily on spatial texture cues. Each model is trained on the FaceForensics++ dataset [58], which contains 1,000 pristine and 5,000 manipulated videos produced by various deepfake methods. For testing, we randomly sample 2,000 frames (256×256) from four datasets, i.e., FaceForensics++, DFDC, Celeb-DF, and FaceShifter [24, 35, 38, 58], allocating 400 frames from each. We adopt the accuracy metric to measure the proportion of correctly identified real and fake frames, following prior work [15, 62, 85, 91].

Findings. As shown in Figure 3, all detectors exhibit performance degradation under real-world distortions. In in-domain evaluations (FaceForensics++), frequency-based models (F3Net and MAT) experience over 10% accuracy loss, while spatial-based detectors (EfficientNet and Xception) drop by more than 20%. When tested across domains (Celeb-DF, DFDC, FaceShifter) with the same degradations, accuracy further declines below 50%. These results indicate that real-world quality degradations disrupt the subtle statistical cues relied upon by current detectors, collapsing the separation between real and fake features in the latent space and rendering them ineffective. Moreover, detectors that overfit to their training dataset, e.g., Xception with nearly 100% accuracy on FaceForensics++, suffer the steepest degradation, exceeding 30%. In short, **existing deepfake detectors lack robustness and generalizability for trustworthy verification of audience-generated videos in the wild.** This motivates the need for a complementary mechanism that can *provide verifiable authenticity when necessary*.

2.2 Watermarking for Content Authentication

Compared with the aforementioned deepfake detectors, watermarking provides a proactive alternative for integrity verification by embedding verifiable provenance directly into the video itself through imperceptible pixel or frequency modifications. The embedded watermark can later be extracted for validation. To better position our work, we distinguish existing approaches by introducing two notions of watermarking: *out-of-band*, where embedding occurs after recording, and *in-band*, where the authentication signal is integrated during the recording process itself, as shown in Figure 4.

Out-of-band Watermarking. Conventional watermarking techniques embed authentication information into multimedia files through digital signal processing in the post-production or transmission stages [7], and are widely used in image, audio, and video authentication. With the rise of generative models, several commercial systems, such as Amazon Titan [5], have further adopted watermarking principles to mark AI-generated content for provenance tracking [39]. However, these schemes typically operate **out-of-band**, i.e., they require access to the digital media after capture for post-processing. As a result, **such methods are unsuitable for embedding**

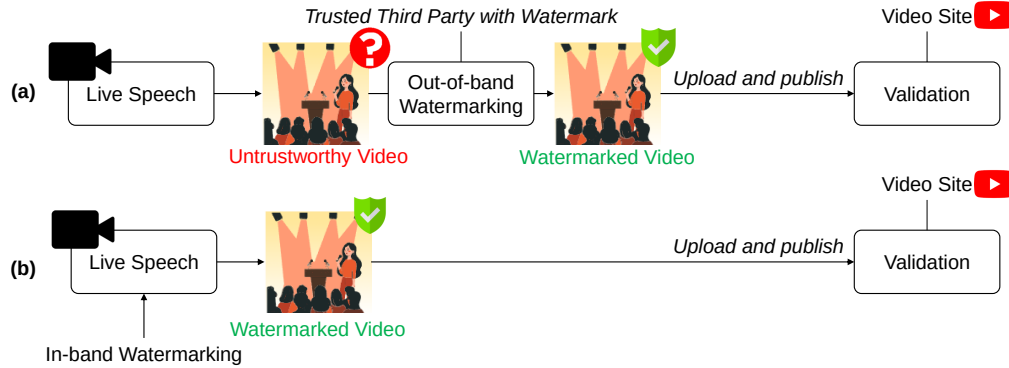


Fig. 4. Figure depicts the comparison between out-of-band and in-band watermarking for video authentication. (a) With **out-of-band watermarking**, a live speech is first recorded into an untrustworthy video, which is later processed by a separate watermarking and verification module before validation. Because watermarking occurs after capture, this approach cannot guarantee the authenticity of audience-generated recordings at the time of creation. (b) In contrast, **in-band watermarking** embeds the authentication signal directly during the recording of the live speech (e.g., through visible light modulation), producing a watermarked video inherently bound to the original event and immediately verifiable upon validation.



Fig. 5. Figure depicts that (a) a **rolling-shutter camera** exposes the image line by line. (b) When illuminated by a modulated light source, each column of pixels integrates a slightly different light intensity over time, resulting in vertical stripe patterns that encode the watermark signal within the captured frame. (c) **Implementation setup of LightMark's in-band watermarking system**, consisting of a DC power supply, an Arduino controller, and an LED light source. The modulated light embeds authentication data into the captured scene, producing watermarked frames for subsequent validation.

trustworthiness into audience's live recordings, which are captured on personal devices and directly uploaded to social platforms without opportunities for post-production embedding or modification.

In-band Watermarking. To enable authenticity for real-time, audience-generated video, we introduce the concept of *in-band watermarking*, which embeds authentication signals during the physical recording process itself. This ensures that the watermark is captured simultaneously with the video by the recording sensor, inherently linking the authentication data to the genuine recording event. Under this definition, a prior study, TalkLock [60], can be regarded as an early instance of in-band watermarking, as it embeds voice authentication data into live video streams using a visible QR code that can later be extracted for verification. However, this approach requires the QR code to occupy a substantial portion of the frame to remain decodable, making it visually intrusive. Speakers might not want to adopt such a design during public speeches, as the QR code distracts the audience and compromises the visual quality of the recording.

2.3 Non-intrusive In-band Watermarking with Visible Light

Drawing inspiration from visible light communication [12, 28, 34, 84], **LightMark leverages the rolling shutter mechanism of CMOS image sensors to realize non-intrusive in-band watermarking**. Rolling shutter is a characteristic feature of the CMOS image sensor in commercial cameras. As shown in Figure 5(a), instead of exposing the entire frame simultaneously, rolling shutter sensors expose image columns sequentially, with each column integrating the light intensity over its exposure period. When the illumination source (e.g., an LED) is modulated at a given frequency (e.g., 1 KHz), this sequential exposure transforms the temporal modulation of light into subtle vertical stripes within the captured frame. Such high-frequency flicker is imperceptible to the human eye, making the process **non-intrusive to both speakers and viewers**. By modulating the illumination at discrete frequencies, the temporal blinking pattern of the light transmitter is thus translated into spatial periodicity in the recorded image, serving as the channel for watermark transmission, as shown in Figure 5(b). Let $\mathcal{F}$ denote the set of modulation frequencies; each symbol can thus encode $\lfloor \log_2 |\mathcal{F}| \rfloor$ bits.

Based on that, *LightMark* enables an in-band watermarking mechanism that allows the camera to **simultaneously capture both visual content and embedded watermark within the same frame**. As shown in Figure 5(c), our prototype employs a low-cost Arduino Uno board [6] connected to a MOSFET driver and an LED light powered by a DC supply. The Uno's microcontroller supports pulse-width modulation (PWM), which generates square waves at designated frequencies for watermarking encoding. Compared with the original frame, the watermarked frame exhibits subtle vertical stripes carrying the authentication signal, which can be effectively removed after validation to preserve visual quality by *LightMark*'s restoration design (detailed in §4.2.3).

3 System and Threat Model

We now present the system and threat model of *LightMark*.

3.1 System Model

LightMark's **goal** is to establish the authenticity and integrity of crowd-sourced, face-centered speech videos. By leveraging the camera's rolling-shutter mechanism for in-band watermarking, *LightMark* enables video-sharing platforms to identify manipulated content before public distribution (e.g., on YouTube, X, and TikTok). Specifically, *LightMark* verifies the speaker's facial identity and dynamics, targeting common misinformation scenarios where attackers forge a person's talking-head video [80, 81], as the face is the major carrier of identity. Meanwhile, non-facial manipulations, such as background changes and overlaid subtitles, are outside the scope of *LightMark* (discussed in §6.3). *LightMark* operates under three key **design requirements**: (1) *Effectiveness*: the watermarking and verification mechanism must reliably detect video forgery with high accuracy, even when recordings originate from various untrusted devices. (2) *Non-intrusiveness*: the watermarking technique should remain visually imperceptible to both audience and speakers, causing minimal disruption or discomfort during live events. (3) *Compatibility with commodity audience recorders*: the watermark embedding and reception should not require specialized camera hardware or software modifications, thus enabling seamless adoption by ordinary audience's recording devices.

3.2 Threat Model

The **attacker's goal** is to forge speech videos and disseminate manipulated content on mainstream platforms to mislead viewers and propagate disinformation. The **attacker's capabilities** include both *traditional forgeries* (e.g., video slicing and frame insertion or deletion) and *AI-based deepfakes* (e.g., face swapping and face reenactment). An attacker may capture a genuine recording of a public speech and subsequently manipulate it to alter the speaker's identity or expressions before releasing it online. We also assume the attacker possesses white-box knowledge of *LightMark*. However, we assume the **attacker cannot**: (1) obtain the private keys held by the legitimate

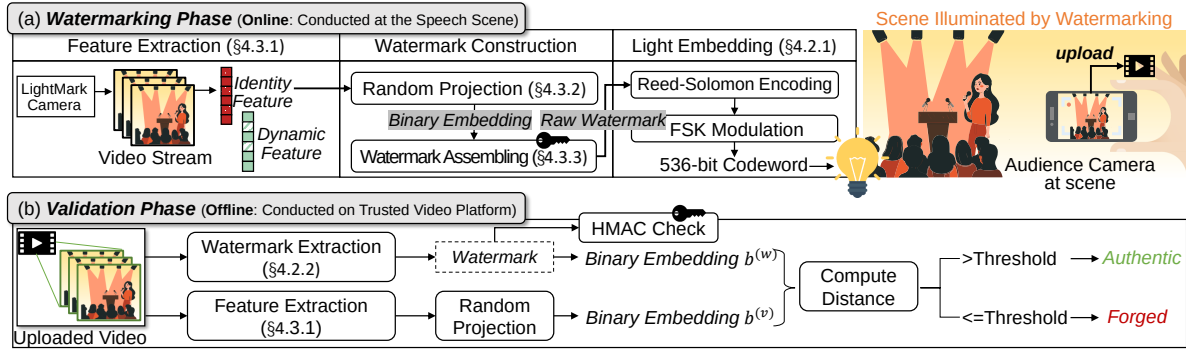


Fig. 6. Figure depicts the overview of *LightMark*, which operates in two phases: (a) **Watermarking Phase**, conducted online at the speech scene, where the *LightMark* camera extracts facial features, generates a binary watermark, and embeds it into the scene illumination via FSK modulation; and (b) **Validation Phase**, conducted offline on a trusted platform, where the uploaded video undergoes watermark extraction and integrity verification to distinguish authentic content from forged segments.

watermark transmitter for generating authenticated watermark messages, and (2) bypass the verification process conducted by trusted platforms (e.g., YouTube).

4 *LightMark* Design

We now present *LightMark*'s design and implementation.

4.1 *LightMark* Overview

LightMark delivers **real-time, in-band watermark embedding** during live recording and **offline validation** after upload to verify the authenticity of uploaded videos. It consists of two key phases: the *Watermarking Phase* and the *Validation Phase*, as illustrated in Figure 6(a) and (b). The **Watermarking Phase** operates online, i.e., at the speech scene, where in-band watermarking is performed in real time. The goal is to continuously generate and embed a compact video digest that is tightly bound to the speaker's identity and facial dynamics, such that any subsequent alteration (e.g., deepfake generation or video splicing) disrupts this binding and becomes detectable. Specifically, a trusted camera captures the live scene and feeds the video stream to the Feature Extraction module (§4.3.1), which derives both identity and dynamic facial features. These features are then processed by the Random Projection module (§4.3.2) to obtain binary embeddings and subsequently assembled by the Watermark Assembling module (§4.3.3) into a raw watermark that encapsulates the content's semantic and cryptographic information. The resulting watermark is optically embedded into the recording environment through the In-band Watermark Embedding module (§4.2.1), which imperceptibly modulates the illumination of the existing lighting infrastructure. As a result, all cameras present in the scene inherently capture the embedded watermark within their recorded videos.

After the recording is uploaded, a trusted verification service (e.g., a hosting platform such as YouTube) performs the offline **Validation Phase**. It first extracts the embedded watermark from the uploaded video using the Neural-based Watermark Extraction module (§4.2.2) and re-computes the corresponding features from the same video segment. The system then verifies the integrity of the watermark using the embedded hash-based message authentication code (HMAC), and if the verification succeeds, measures the similarity between the decoded binary embedding $b^{(w)}$ and the re-derived embedding $b^{(v)}$ from the video. Based on this comparison, the

platform determines the authenticity of the content and, if necessary, highlights inconsistent or forged segments along the video timeline.

Designing *LightMark* involves two main challenges –

Challenge 1: Balancing Watermark Extraction Robustness with Video Quality. Embedding watermark through visible-light modulation inevitably introduces structured stripe artifacts into video frames. Robust extraction demands that these patterns remain sufficiently strong and temporally stable, yet their visibility must stay below the perceptual threshold to preserve natural appearance. This creates a fundamental physical-perceptual trade-off between signal recoverability and video fidelity. Achieving this balance under varying illumination, camera motion, and scene complexity requires a carefully co-designed embedding and extraction pipeline. We address this through the Watermark Embedding and Extraction module (§4.2).

Challenge 2: Maintaining Forgery Resistance Under Bandwidth Constraints. Visible-light modulation forms a low-bandwidth, high-distortion communication channel subject to motion blur, compression, and exposure variation. Meanwhile, the watermark must remain cryptographically verifiable and resistant to advanced manipulation, including video slicing and advanced deepfake attacks, without inflating the data rate. Designing a watermarking scheme that encodes semantically meaningful and robust content within such limited capacity presents a formidable challenge. We overcome this through the Watermark Content Design (§4.3).

4.2 Watermarking Embedding and Extraction

We now detail the watermark embedding and extraction mechanisms of *LightMark*, which implement the principle of *in-band watermarking*, i.e., embedding verification signals directly into visible illumination rather than applying them post-capture. The key challenge (i.e., Challenge 1) is to achieve both *imperceptibility*, to preserve video quality, and *recoverability*, ensuring that the embedded watermark remains physically extractable from the captured frames for reliable validation. To address this, *LightMark* integrates two core modules. The In-band Watermark Embedding Module (§4.2.1) employs low-power light emitters (5-10 watts LEDs) whose intensity is modulated using frequency-shift keying (FSK) within the camera’s rolling-shutter exposure band, producing subtle stripe patterns that are nearly imperceptible yet physically encoded. The Neural-based Extraction (§4.2.2) and Restoration Module (§4.2.3) subsequently recovers and, if necessary, removes these subtly embedded watermarks from the recorded frames, ensuring accurate verification and high visual fidelity.

We formally model the watermark embedding and extraction process by the following formulation:

$$\mathbf{I}(\mathbf{x}) = \mathbf{I}_0(\mathbf{x}) + \mathbf{W}(\mathbf{x}), \quad (1)$$

where $\mathbf{I}_0$ denotes the watermark-free frame (original content) and $\mathbf{W}$ represents the watermark layer at pixel position $\mathbf{x} = (x, y)$. The **embedding process** generates $\mathbf{W}$ and adds it to $\mathbf{I}_0$ through light modulation, producing the observed frame $\mathbf{I}$ that visually preserves $\mathbf{I}_0$ while optically encoding verification signals. The **extraction process** performs the reverse operation, recovering the watermark layer $\mathbf{W}$ from the observed frame $\mathbf{I}$ by estimating and subtracting the clean background component $\mathbf{I}_0$.

4.2.1 In-band Watermark Embedding. *LightMark* employs a rolling-shutter frequency-shift keying (RS-FSK) modulation scheme to encode data into the LED illumination. Four distinct modulation frequencies between 1020 Hz and 1980 Hz are selected, each representing one of four possible symbols (two bits per symbol). Symbols are transmitted continuously at a fixed rate of 60 symbols per second, resulting in an effective data throughput of 120 bps. To ensure reliable watermark embedding and recovery, the 368-bit raw watermark packet (introduced in §4.3) is encoded using a *Reed-Solomon (RS)* code, producing a 536-bit codeword. This corresponds to an RS(67,46) configuration, consisting of 46 data bytes and 21 parity bytes for error correction. During verification, the extracted watermark is first RS-decoded, and the decoded raw watermark packet is then used for validation against the expected identity and dynamic feature embeddings.

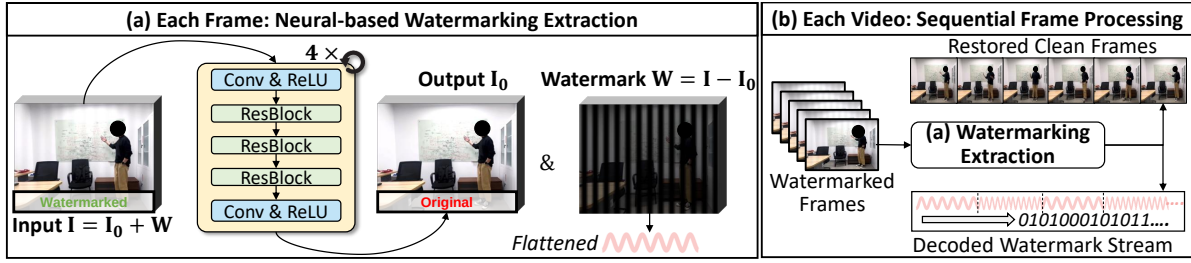


Fig. 7. Figure depicts the overview of *LightMark*'s watermark embedding and extraction framework. (a) For each watermarked frame, the extraction module predicts the clean background I_0 and isolates the watermark layer W . (b) Sequential frame processing over the video yields a restored clean sequence and a continuous demodulated watermark stream.

4.2.2 Neural-based Watermark Extraction. This module takes the recorded video frames as input and outputs both the predicted clean frames I_0 and the extracted watermark layers W , as illustrated in Figure 7(a) and (b). In practical scenarios, especially with handheld audience devices, the captured video often exhibits slight camera motion and scene dynamics, causing I_0 to vary across frames. Consequently, assuming a consistent background is invalid in such scenarios, making the extraction process mathematically ambiguous and inherently **ill-posed** [11]. Moreover, since W is deliberately designed to be extremely subtle to preserve video quality, its recovery becomes even more challenging under dynamic I_0 conditions.

To robustly extract the subtle watermark from such recordings, we develop a progressive deep neural network that predicts the clean ambient frame I_0 from the observed input I , as shown in Figure 7(a). The network iteratively refines its prediction through multiple extraction blocks. Each block contains two convolutional layers, one at the input and one at the output, enclosing three intermediate residual blocks (ResBlocks). Each ResBlock consists of two 3×3 convolutional layers with 32 filters and ReLU activations, along with skip connections to enhance feature propagation. Subtracting the predicted clean frame from the input yields the rolling-shutter watermark layer $W = I - I_0$.

The training objective minimizes the discrepancy between the predicted clean frame and the ground-truth ambient frame I_0 , accounting for both spatial- and frequency-domain differences motivated by the column-wise periodicity inherent to rolling-shutter patterns. The extraction loss function $\mathcal{L}_{\text{ext}}$ is defined as:

$$\mathcal{L}_{\text{ext}} = \|\hat{I}_0 - I_0\|_2 + \|F(\hat{I}_0) - F(I_0)\|_2, \quad (2)$$

where $\hat{I}_0$ denotes the network-predicted clean frame, $\|\cdot\|_2$ represents the Euclidean norm, and F is the 1D discrete Fourier transform applied horizontally. The model is trained on a synthesized dataset derived from the NYU V2 Dataset [64], covering diverse scenes and lighting conditions. It follows the methodology used in similar image restoration tasks such as deraining [17, 56]. The network is optimized using the Adam optimizer with a learning rate of 1×10^{-4} . As shown in Figure 7(b), the module processes video frames sequentially, producing both the restored clean sequence and the demodulated watermark stream for validation.

4.2.3 Video Restoration. While the embedded watermark stripes enable reliable authentication, they may introduce slight visual artifacts. To mitigate this, *LightMark* leverages a byproduct of the neural-based extraction framework described above, which simultaneously reconstructs stripe-free frames while extracting the watermark. As shown in Figure 7(b), this restoration step preserves verification information and ensures high visual fidelity, achieving a balance between security and viewing quality. For forged videos without valid watermark stripes, restoration would produce an output nearly identical to the input. In practice, however, restoration is applied only to authenticated segments. The evaluation in §5.6 further demonstrates the effectiveness of restoration.

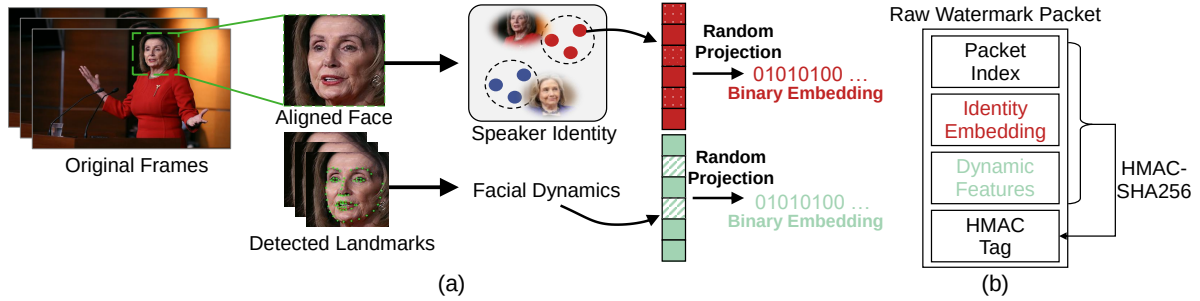


Fig. 8. Figure depicts (a) *LightMark*'s Feature Extraction (§4.3.1), which extracts the speaker's identity and facial dynamics and transforms them into binary embeddings via random projection (§4.3.2), and (b) Watermark Assembling (§4.3.3), where the embeddings are combined with a packet index and an HMAC-SHA256 tag to form the authenticated watermark packet.

4.3 Watermarking Content Design

The goal of this subsection is to bind authentic visual content to a robust watermark, enabling a trusted platform with a verification service to later confirm the authenticity of the video. Given the limited transmission bandwidth (i.e., Challenge 2), *LightMark* improves robustness to forgery by designing a compact watermark that encodes both the speaker's identity and facial dynamic features in real time. This design is aligned with typical misinformation scenarios in which attackers forge a person's talking-head video [80, 81], as the face is the major carrier of identity.² These facial features are then transformed into a lightweight cryptographic digest that fits within the in-band watermarking channel. As shown in Figure 8(a) and (b), *LightMark* processes live video frames to extract identity and dynamic features, which are then randomly projected into a binary embedding. This embedding, together with a cryptographic tag (e.g., HMAC) and control metadata, forms the raw watermark, which is optically modulated and embedded into the video via rolling-shutter-compatible illumination. We describe the design in detail below.

4.3.1 Feature Extraction. This module extracts semantic features that characterize both the speaker's identity and facial dynamics within each video window. In our prototype, each window spans 4.5 seconds and is sampled at 20 frames per second, resulting in a total of 90 frames per window. We first use Dlib [23], a popular facial detection and alignment toolkit, to locate and align the face region in each frame.

The aligned faces are then processed by the MediaPipe Face Landmarker model [4], a widely used lightweight facial tracking framework, to obtain **dynamic facial features**. Specifically, we extract 11 normalized landmark distances (e.g., 0-17, 40-17, 270-17) and 5 blendshape activations (indices 6, 7, 8, 9, 10, 11, 12, 23, 25, 50, 51). These indices are selected to capture mouth opening, eye blinking, and cheek movement, i.e., facial dynamics that are expressive, speaker-dependent, and sensitive to manipulation. Each feature signal is linearly interpolated to fill tracking gaps and resampled to a fixed 90-sample temporal trace. The resulting 16 feature channels, each with 90 samples, are concatenated to form a 1,440-dimensional motion descriptor representing the speaker's facial dynamics. For **identity features**, the mid-frame of each window is passed through ArcFace [32], a state-of-the-art face recognition model, to produce a 512-dimensional L2-normalized embedding. Together, these dynamic and identity features provide semantically rich and temporally stable representations that capture both appearance

²Accordingly, *LightMark* assumes that the speaker's face is sufficiently visible to support meaningful impersonation. If the face is absent, too small, or heavily occluded, the impersonation threat is also weakened. Meanwhile, edits outside the face region, such as background changes or overlaid text, may not be detected by *LightMark* (discussed in §6.3).

and fine-grained motion cues, which are difficult to manipulate consistently in forged videos. The evaluation results in §5.2 demonstrate their effectiveness in generating robust and discriminative watermarks.

4.3.2 Random Projection. This submodule takes the extracted feature vectors as input and outputs binary embeddings. For each 4.5 s window, the implementation seeds a deterministic cosine locality-sensitive hash (LSH) family and projects both the 512-D identity feature and the 1,440-D dynamic feature onto 150 random hyperplanes each. The sign of each projection determines the binary output, resulting in two 150-bit hashes that compactly encode the speaker’s identity and facial dynamics. This random projection effectively preserves feature similarity in Hamming space, **enabling efficient comparison during validation**.

During the validation phase, the same random projection is applied to features extracted from the uploaded video. The similarity between the embedded watermark and the recovered features is quantified using the Hamming distance:

$$D = \sum_{i=1}^n \mathbf{1} \left[b_i^{(v)} \neq b_i^{(w)} \right],$$

where $b^{(v)}$ and $b^{(w)}$ denote the binary embeddings from the video and watermark, respectively. In our implementation, distances smaller than 32 for identity features and 45 for dynamic features are deemed authentic, whereas larger distances indicate potential forgery, as illustrated in Figure 6(b).

4.3.3 Watermark Assembling. This submodule takes the binary embeddings as input and generates the raw watermark. Each packet consists of (i) the binary embedding (225 bits), (ii) a sequence number (15 bits) serving as the packet index, and (iii) an HMAC-SHA256 tag (truncated to 128 bits) computed over the concatenation of the binary embedding and the sequence number. The 128-bit tag length provides sufficient cryptographic strength for integrity verification while maintaining low transmission overhead. To ensure complete identity representation, one half of the 150-bit identity hash is multiplexed with the 150-bit dynamic hash, alternating which half of the identity hash is transmitted on even and odd sequence numbers.

The resulting 368-bit raw watermark packet is then encoded by Reed-Solomon coding into a 536-bit codeword for robust transmission, as described in §4.2.1, and the encoded watermark is subsequently optically embedded into the video stream using FSK modulation compatible with rolling-shutter capture. We adopt a *pipelined transmission* strategy to ensure **continuous video authentication**: each 4.5 s segment of the video transmits the watermark generated from the preceding 4.5 s segment. With *LightMark*’s visible-light link operating at 120 bps (providing 540 bits over 4.5 s), this capacity perfectly accommodates one 536-bit RS-encoded watermark, enabling seamless real-time embedding without latency. Note here that this pipelined design naturally supports **arbitrary recording start times and durations**. Each watermark packet is self-contained and carried by the subsequent segment, while its embedded packet index identifies the corresponding preceding 4.5 s segment, inherently providing segment-level synchronization. As a result, ***LightMark* can authenticate partial recordings** on a per-segment basis, allowing audience members to begin recording at any time and stop at arbitrary points while still obtaining verifiable video segments.

5 Evaluation

We evaluate *LightMark* through comprehensive real-world experiments, demonstrating its feasibility.

5.1 Experimental Setup

We first introduce *LightMark*’s experimental setup.

Prototype Implementation. As shown in Figure 5(c), the in-band watermarking device of *LightMark* is implemented using a low-cost Arduino Uno board [6] that drives an LED through a MOSFET circuit. The Uno’s hardware pulse-width modulation (PWM) generates square waves for light modulation, where 4-FSK modulation

is realized by tuning the timer register to produce signals between 1020-1980 Hz with a 50% duty cycle. Each watermark symbol lasts 1/60 s, achieving a watermark rate of 120 bps. The LED is powered by a DC supply. The Arduino is connected to a laptop with a camera that extracts facial features from the live video and converts them into a watermark for visible light embedding.

Dataset Collection. We evaluate *LightMark* by collecting videos across 12 representative speech environments, as listed in Table 1. The resulting test dataset includes **1,800 authentic video samples** (150 per environment), each lasting 4.5 s and captured under varied lighting and background conditions to ensure diversity. We recruited eight adult volunteers for this speech recording process. All participants provided informed consent prior to the study. No personally identifiable information was recorded, and all data were anonymized before analysis. We conduct this study upon the approval of our institution’s Institutional Review Board. To create forged data, we apply both non-AI attacks (§5.2.2) and AI-based forgeries (§5.2.3), generating around 270 and 190 manipulated samples per environment, respectively, for a total of **5,520 forged video samples**. Note that *LightMark* does not need to be trained on environment or forgery type to test. Therefore, **all evaluated videos are cross-domain** with respect to *LightMark*.

All authentic recordings are captured on a Huawei P40 Pro smartphone using the mcpro24fps app [31, 45]. We further evaluate *LightMark* under different recording configurations, including camera models, transmission rates, distances, and motion speeds (§5.4). Videos are recorded at 60 frames per second (FPS) with an exposure time of 1/1200 s and auto ISO under well-lit conditions, at a resolution of 1920×1080 pixels. Frames are uniformly resized to 512×512 pixels for processing.

Performance Metrics. We consider a video sample to be a *positive* instance if *LightMark* identifies it as forged, and a *negative* instance otherwise. We conservatively treat unverifiable cases as forged videos, since *LightMark* cannot provide authenticity evidence for such recordings. Following prior studies [15, 62, 85, 91], we evaluate performance primarily using **Accuracy** and **Area Under the Curve (AUC)**. Accuracy measures the overall proportion of correctly classified authentic and forged samples, while AUC quantifies the model’s discriminative capability between authentic and forged content across varying decision thresholds. We additionally report the **False Negative Rate (FNR)**, defined as the fraction of forged videos incorrectly classified as authentic, and the **False Positive Rate (FPR)**, defined as the fraction of authentic videos incorrectly identified as forgeries.

5.2 Overall Performance

We evaluate *LightMark* against both conventional non-AI-based and AI-based attacks, comparing it with state-of-the-art deepfake detection models.

5.2.1 Baseline Selection. To comprehensively evaluate the performance of *LightMark*, we compare it with representative state-of-the-art passive deepfake detectors spanning major methodological families:

- **Spatial-based methods:** Xception [20], EfficientNet-B4 [67], and MAT [88], which identify visual inconsistencies such as texture anomalies, blending seams, and warping artifacts in spatial domains.
- **Frequency-based methods:** F3Net [53], SRM [44], and SPSL [40], which capture spectral cues, such as aliasing, upsampling traces, and abnormal frequency power distributions, introduced during manipulation.
- **Blending-based:** SBI [61], which synthesizes self-blended images from pristine frames to mimic typical forgery artifacts (e.g., boundary blending and statistical inconsistencies) for generalizable detection.
- **Reconstruction-based methods:** RECCE [13], which reconstructs or inpaints facial regions and detects discrepancies in the reconstruction residuals to distinguish authentic from forged content.

These baselines are widely adopted in the deepfake detection literature, covering both traditional artifact-based and reconstruction-driven paradigms. *LightMark* differs fundamentally by embedding verifiable watermarks into the content, providing a proactive integrity guarantee rather than post-hoc detection.

Table 1. Detection performance of *LightMark* under non-AI forgery attacks. FNR_1 and FNR_2 denote false negative rates for forged samples generated by frame insertion/deletion and video clipping/slicing, respectively.

Type	Scene	FNR_1	FNR_2	FPR	ACC	AUC
Static	Office Room	0.00%	0.00%	1.30%	99.53%	99.35%
	Poster Board Area	0.00%	0.00%	2.48%	99.07%	98.76%
	Bookshelf Corner	0.00%	0.00%	1.49%	99.46%	99.25%
	Display Screen Zone	0.00%	0.00%	3.16%	98.83%	98.42%
	Staircase Area	0.00%	0.00%	0.73%	99.77%	99.64%
	Glass Partition	0.00%	0.00%	1.86%	99.30%	99.07%
Handheld	Cafe Environment	0.00%	0.00%	2.04%	99.53%	99.36%
	Canteen Space	0.00%	0.00%	2.99%	98.83%	98.50%
	Indoor Plant Zone	0.00%	0.00%	2.61%	99.07%	98.69%
	Brick Wall Background	0.00%	0.00%	2.04%	99.30%	98.98%
	Decorative Area	0.00%	0.00%	11.45%	95.90%	94.27%
	Workshop Area	0.00%	0.00%	3.25%	98.83%	98.39%

5.2.2 Detecting Non-AI Attacks. We first evaluate *LightMark* under conventional video forgery attacks involving editing operations and test its robustness across 12 recording environments, summarized in Table 1. The first six environments use tripod-mounted phones, while the remaining six use handheld phones. During handheld recording, participants held the phones at a typical audience viewpoint without tripod stabilization, simulating natural crowd-recording conditions, with camera movement covering a typical handheld range of 0–2 m/s. Each scene contains approximately 150 authentic video samples, and 270 forged samples are generated for each scene by applying two types of temporal manipulations, as detailed below. **(1) Frame insertion/deletion:** The adversary inserts or removes consecutive frames to subtly alter timing. We test two magnitudes, i.e., 0.5 s (12–15 frames at 30 fps) and 1.0 s (25–30 frames), with insertion or deletion positions uniformly randomized within each video. **(2) Video clipping/slicing:** The adversary reorders segments to modify temporal order. By default, two segments of random lengths (1–3 s) are swapped. All manipulations are applied to videos recorded under varying environmental conditions to evaluate *LightMark*'s effectiveness and robustness.

As shown in Table 1, *LightMark* achieves uniformly high performance across both static and handheld scenarios. In all cases, the system reports zero false negatives ($\text{FNR}_1/\text{FNR}_2 = 0\%$), indicating that every instance of frame insertion, deletion, or segment reordering is successfully detected. This confirms that *LightMark*'s verification process reliably cross-checks the embedded watermark with the visual content to expose temporal inconsistencies in forged videos. Meanwhile, *LightMark* shows low false positives, typically below 3% across most environments, demonstrating that natural variations such as camera shake, ambient lighting, and background motion rarely trigger false alarms. A slight degradation is observed in the decorative area due to complex reflections, where the false positive rate rises to 11.45%; nevertheless, the system still achieves 95.9% accuracy and 94.3% AUC. This robustness stems from *LightMark*'s identity and dynamic feature extraction, which allows authentic signatures to be consistently retrieved from both the audience's device and the embedding infrastructure. The results also show that videos captured by commodity rolling-shutter cameras generally provide sufficient watermark signals for verification, with few cases being marked as unverifiable. Overall, *LightMark* attains an average accuracy of 99.0% and AUC of 98.6% across 12 scenes, confirming strong resilience under diverse visual conditions. These results

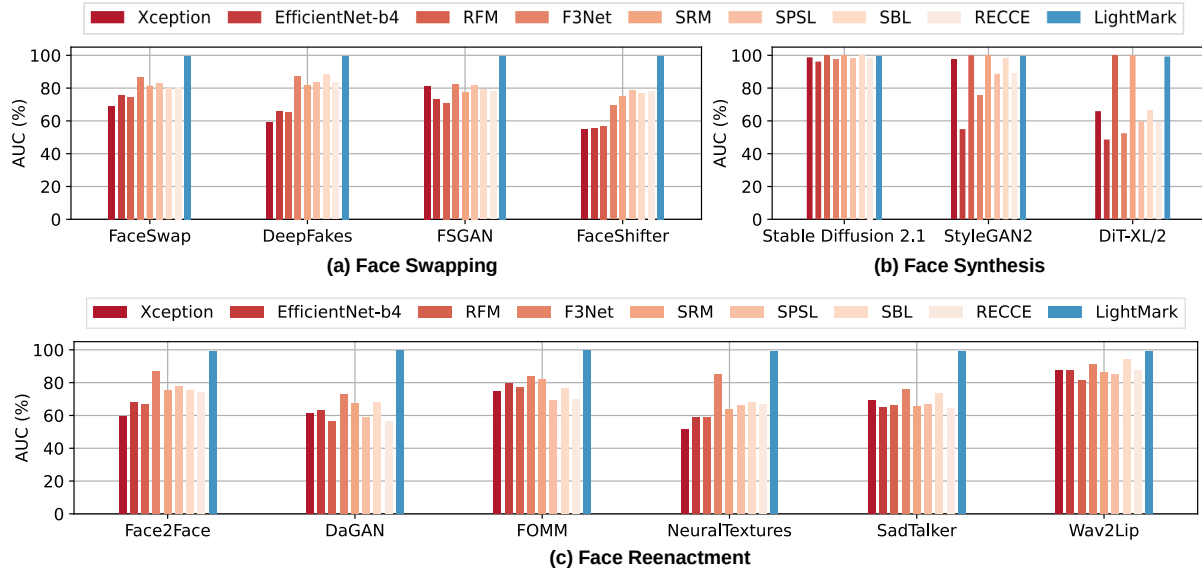


Fig. 9. Figure depicts *LightMark*'s detection performance under AI-based video forgery attacks: (a) face swapping, (b) face synthesis, and (c) face reenactment.

demonstrate *LightMark*'s effectiveness in ensuring temporal integrity and providing reliable protection against non-AI video forgery attacks.

5.2.3 Detecting Deepfake Attacks. We next evaluate *LightMark* against modern AI-driven video forgery techniques, encompassing both *face swapping* and *reenactment*-based manipulations. All evaluations are conducted on a dataset comprising 460 authentic videos, randomly selected across all recording scenes to ensure generality. For each deepfake method, we generate 230 forged samples for testing. We compare *LightMark* with representative passive detector baselines introduced in §5.2.1, with the officially released weights trained on FaceForensics++ dataset [58, 61, 80, 81].

Face Swapping. We test four representative face-swapping methods, i.e., FaceSwap [2], DeepFakes [1], FSGAN [49], and FaceShifter [35], which replace the speaker's facial identity while preserving expressions and background. As shown in Figure 9(a), *LightMark* achieves high detection performance with an average AUC above 0.99, substantially outperforming all eight baselines, whose AUCs typically range between 0.55 and 0.87. As discussed in §2.1, passive detectors that rely on visual artifacts or texture inconsistencies often suffer from poor cross-domain generalization and sensitivity to video quality degradation. In contrast, *LightMark* explicitly encodes the speaker's physical identity in its in-band watermark. Consequently, any identity substitution disrupts the correspondence between the visual content and the watermark, allowing *LightMark* to reliably flag such manipulations.

Face Synthesis. We evaluate three representative high-quality face synthesis models spanning major generative paradigms: (1) **Stable Diffusion 2.1** [57], a latent diffusion text-to-image model; (2) **StyleGAN2** [33], a high-fidelity GAN-based image generator; and (3) **DiT-XL/2** [50], a large-scale Diffusion Transformer model for image synthesis. As shown in Figure 9, these models are challenging cases for passive baselines, especially DiT-XL/2, which yields the lowest AUC for 6 out of 8 baselines. In contrast, *LightMark* maintains high detection performance

Table 2. *LightMark*'s BER performance across transmission distances. With both Neural-based Watermark Extraction and Reed-Solomon encoding, *LightMark* maintains error-free decoding up to 11 m.

Method	1 m	2 m	3 m	4 m	5 m	6 m	7 m	8 m	9 m	10 m	11 m
w/o Neural-based Extraction	0.000	0.010	0.075	0.036	0.126	0.445	0.313	0.472	0.230	0.471	0.391
w/o Reed-Solomon Encoding	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.001	0.010	0.007
<i>LightMark</i> (Full)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3. *LightMark*'s BER performance under varying camera angles.

Method	0°	15°	30°	45°
w/o DNN-based Extraction	0.126	0.251	0.213	0.185
w/o Reed-Solomon Encoding	0.003	0.006	0.006	0.015
<i>LightMark</i> (Full)	0.000	0.000	0.000	0.004

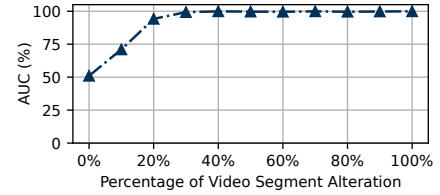


Fig. 10. Figure depicts AUC of *LightMark* under different levels of video alteration.

because fully synthesized videos do not contain a valid rolling-shutter watermark and are therefore readily identified as forged.

Face Reenactment. We further evaluate four state-of-the-art reenactment techniques, i.e., Face2Face [69], DaGAN [27], FOMM [63], and NeuralTextures [68], which manipulate the speaker's facial expressions or lip movements while retaining their identity. We also examine two speech-driven **lip-synchronization methods**, i.e., SadTalker [87] and Wav2Lip [51], where mouth movements are synthesized to match manipulated audio. Lip-synchronization can thus be viewed as a specialized form of reenactment in which facial dynamics are driven by audio rather than another facial video. Across all these attacks, *LightMark* maintains consistently high detection accuracy, achieving an average AUC of 0.992 and surpassing the best baseline (F3Net) by over 12%. This is because *LightMark*'s in-band watermarking encodes fine-grained temporal dynamics reflecting authentic motion patterns captured by the rolling shutter, making temporal inconsistencies between the visual content and the embedded watermark detectable.

Overall, *LightMark* demonstrates robust and consistent deepfake detection performance across all attack types, achieving AUCs above 0.99 for both identity and dynamic forgeries. The few practical failure modes mainly correspond to unverifiable or false-alarm cases under degraded recording conditions. In contrast, prior passive detectors suffer significant degradation across domains and manipulation methods, highlighting the benefit of *LightMark*'s in-band watermark-based validation.

5.3 Module Evaluation

We evaluate the different modules of *LightMark* to justify its design.

5.3.1 Performance of Watermarking Embedding and Extraction module. This module (§4.2) forms the *physical-layer foundation* of *LightMark*'s in-band watermarking pipeline, responsible for optically embedding and recovering the encoded watermark signals. To assess its performance, we conduct experiments under varying transmission distances and camera angles. This module comprises two key components: a neural-based watermark extraction network and a Reed-Solomon (RS) error-correction scheme. We compare the performance of the complete *LightMark* system with versions that disable each component individually.

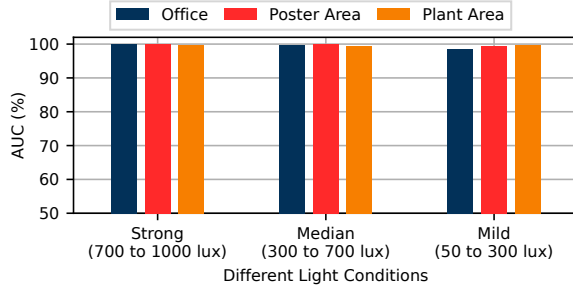


Fig. 11. Figure depicts *LightMark*'s performance under varying ambient light conditions.

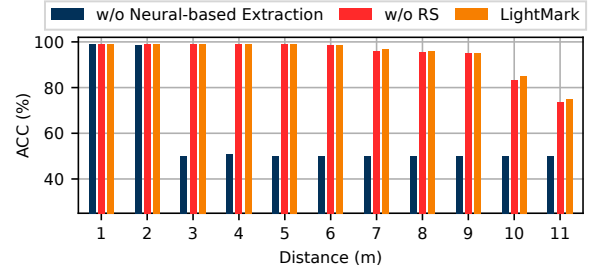


Fig. 12. Figure depicts *LightMark*'s performance under varying distances.

Table 2 presents the results across distances from 1 m to 11 m, using a 10-watt LED as transmitter. As this module is essentially a visible-light communication link, we use bit error rate (BER) as the metric. Without RS encoding, the system begins to exhibit minor decoding errors beyond 8 m as light attenuation and rolling-shutter noise accumulate. In contrast, removing the neural-based watermark extraction leads to a rapid BER increase beyond 3 m, confirming that the learned extraction network is essential for recovering watermark patterns under diverse illumination and background conditions. By integrating both the neural-based extraction and RS encoding, the full *LightMark* pipeline achieves error-free decoding up to 11 m, demonstrating strong robustness against light attenuation and validating the effectiveness of the end-to-end visible-light watermark embedding and recovery design. Furthermore, as shown in Table 3, *LightMark* remains highly stable across camera angles, maintaining a BER of only 0.004 even at a 45° view, highlighting its resilience to recording viewpoint variations in real-world deployments. These results evaluate the physical-layer watermark recovery reliability. We further report the corresponding end-to-end authentication accuracy under varying distances in §5.4.2.

5.3.2 Watermark Robustness. As discussed in §4.3, *LightMark* derives each watermark from the speaker's identity and dynamic facial features within a 4.5 s time window. This experiment evaluates the robustness of the embedded watermark against content manipulation, reflecting the reliability of the overall watermarking design. Specifically, we apply temporal slicing attacks by replacing a controlled portion of each 4.5 s video window with frames from another video segment while keeping the remaining frames unchanged. The replacement ratio increases from 0% to 100% in 10% increments. For each setting, we measure AUC to quantify the inconsistency between the embedded watermark and the manipulated content. As shown in Figure 10, *LightMark* achieves an AUC of about 0.94 even when only 20% of the video window is replaced, corresponding to 0.9 s of manipulated content, where the inconsistency already provides sufficient evidence of forgeries. When the replacement ratio exceeds 30%, corresponding to 1.35 s, the AUC further increases to around 0.99. We further examine the sensitivity of *LightMark* to the validation thresholds. Specifically, we measure the *threshold robustness interval*, defined as the range of thresholds under which both the true negative rate (TNR) and true positive rate (TPR) remain above 0.99. The selected thresholds are 32 for identity features and 45 for dynamic features. For identity features, *LightMark* maintains this performance for thresholds from 18 to 48; for dynamic features, it maintains this performance for thresholds from 40 to 63. The selected thresholds lie well within these stable intervals, indicating that *LightMark*'s performance is not sensitive to small threshold variations. These results demonstrate that *LightMark*'s watermark is robust against both forgery and visual perturbations, preserving a verifiable linkage to the original speaker even under minor tampering. Further watermark enhancements are discussed in §6.3.

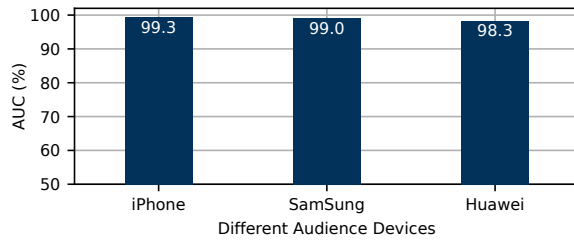


Fig. 13. Figure depicts *LightMark*'s performance with different audience recording devices.

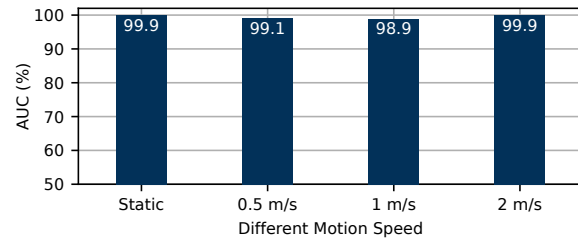


Fig. 14. Figure depicts *LightMark*'s performance under different movement conditions of the audience device.

5.4 Differing Experimental Conditions

We further evaluate *LightMark*'s performance across diverse factors. All experiments are conducted under three representative scenes, i.e., office, poster board area, and indoor plant, and we use AUC as the evaluation metric.

5.4.1 Varying Ambient Light Conditions. We evaluate *LightMark*'s robustness under different ambient illumination levels during recording. Specifically, we consider three lighting ranges: strong light (700-1000 lux), moderate light (300-700 lux), and dim light (50-300 lux), with *LightMark*'s modulated light source set to 10 watts. As shown in Figure 11, *LightMark* consistently achieves high AUC values across all environments, reaching 0.99 under strong light and remaining above 0.98 even under dim conditions. These results demonstrate that *LightMark* is robust to ambient light variation in real-world recording scenarios. This robustness arises from *LightMark*'s neural-based watermark extraction network, which effectively captures subtle watermark patterns by leveraging global spatial correlations of the watermark within each frame.

5.4.2 Varying Distance. In §5.3.1, we evaluate the BER of the recovered watermark under varying distances. Here, we further examine how this link-level reliability translates into end-to-end authentication accuracy, where a longer distance may affect both watermark decoding and facial feature extraction. As shown in Figure 12, removing the neural-based extraction module causes the accuracy to drop rapidly beyond 3 m, approaching random guessing as the watermark becomes difficult to recover from weak rolling-shutter patterns. In contrast, the full *LightMark* pipeline consistently achieves the best performance across all distances, maintaining high accuracy within 9 m and remaining effective at 10–11 m. The degradation at longer distances is mainly caused by reduced facial resolution, which is further investigated in §5.4.5.

5.4.3 Different Audience Recording Devices. To evaluate *LightMark*'s compatibility with different audience recording devices, we test videos captured by iPhone 14 Pro, Samsung S22, and Huawei P40 Pro, representing iOS, Android, and HarmonyOS platforms, respectively. Their corresponding rolling-shutter readout durations are 11.7 μ s, 11.3 μ s, and 4.9 μ s. Video recordings are conducted using the Protake app for the iPhone and the mcpro24fps app for Huawei and Samsung devices [45, 65]. As shown in Figure 13, the resulting AUC values are 0.993, 0.990, and 0.983, respectively. These results confirm that *LightMark*'s watermark decoding is generalizable across commercial devices equipped with rolling-shutter cameras and remains largely unaffected by variations in camera optics or internal image signal processing pipelines.

5.4.4 Under Movement Conditions. *LightMark*'s functionality extends to scenarios involving camera motion, as its physical-layer watermarking transmission operates independently of scene context (e.g., background subtraction) and instead relies on spatial correlations within each single frame. We evaluate performance under different

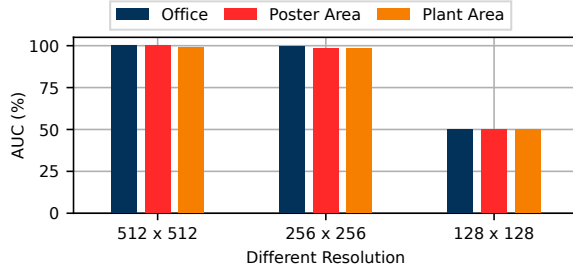


Fig. 15. Figure depicts *LightMark*'s performance under varying video resolutions.

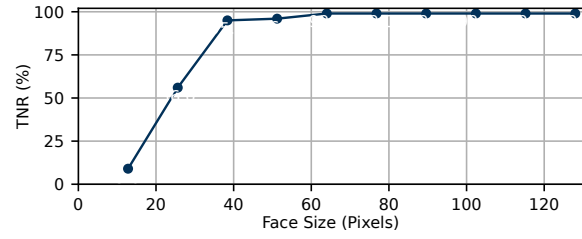


Fig. 16. Figure depicts *LightMark*'s performance under varying face sizes.

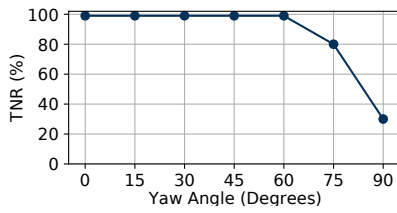


Fig. 17. Impact of face pose.

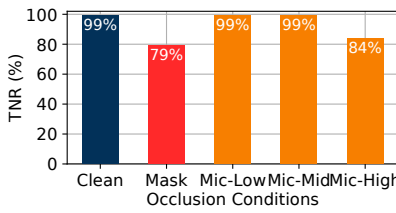


Fig. 18. Impact of face occlusion.

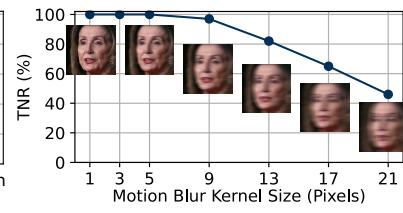


Fig. 19. Impact of motion blur.

motion conditions, including static recording and handheld walking at 0.5 m/s, 1 m/s, and up to 2 m/s. As shown in Figure 14, *LightMark* achieves AUCs with variations smaller than 1% across these conditions, demonstrating that both the in-band watermarking and extraction process remain stable under moderate camera movements. Therefore, *LightMark* shows strong suitability for live speech recording scenarios where handheld or minor audience motion is common.

5.4.5 Performance under Visual Degradation. We evaluate *LightMark* under various visual degradation conditions that may affect facial feature extraction and watermark binding.

Impact of Video Resolution. We evaluate the impact of audience-uploaded video resolution and face size on *LightMark*'s validation performance. Lower spatial resolutions reduce visual detail, making watermark extraction and content verification more challenging. To quantify this effect, we downsample authentic and forged videos to three resolutions (i.e., 512×512, 256×256, and 128×128) and measure the corresponding AUC values. As illustrated in Figure 15, the AUC drops to around 0.5 at the extremely low resolution of 128×128, where fine-grained facial and watermark details become indistinguishable. However, *LightMark* maintains a high AUC of 0.985 at 256×256, which is above the typical lower bound of standard audience-uploaded video resolutions on major social media platforms. These results demonstrate that *LightMark* remains highly suitable for real-world audience uploads and sharing scenarios.

Impact of Face Size. We further evaluate the impact of face size, as *LightMark* relies on the speaker's face for feature extraction. We vary the face size by resizing the entire image. We measure TNR on authentic samples because excessively small faces mainly destabilize identity and landmark features in authentic videos, causing them to be falsely flagged as forged. The results are shown in Figure 16. *LightMark* maintains a consistently high TNR when the face size is larger than about 40 pixels, indicating that feature extraction remains reliable under moderate scale reduction. The TNR degrades noticeably only when the face becomes smaller than 30–40 pixels, where facial landmarks and identity cues are no longer sufficiently stable for reliable validation.

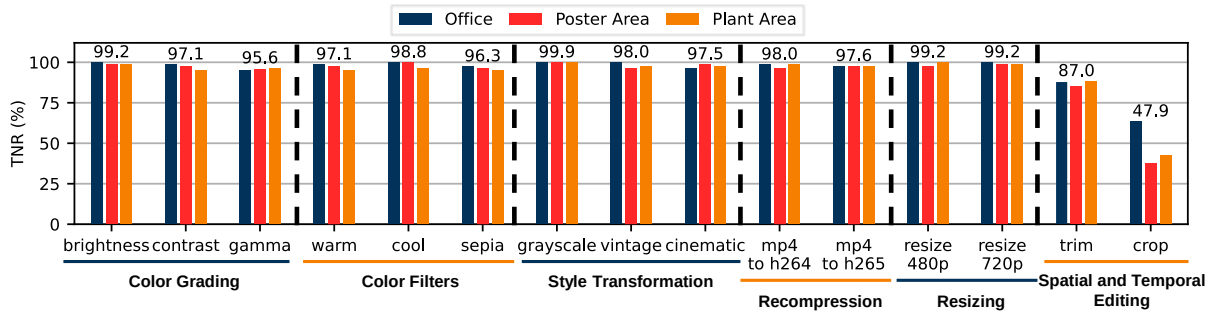


Fig. 20. Figure depicts *LightMark*'s True Negative Rate (TNR) under different benign edits.

Impact of Face Pose. We evaluate robustness under different viewing angles between the audience camera and the speaker. Specifically, we vary the yaw angle from frontal view to side view and show TNRs under each angle in Figure 17. *LightMark* achieves TNR values above 0.99 for yaw angles up to 60° and maintains a TNR of 0.80 even at 75° . The performance drops substantially under a pure side view, i.e., 90° , where facial features become insufficient for reliable feature extraction. These results show that *LightMark* supports typical head turns during speech events, but does not reliably handle pure side views where facial cues become insufficient.

Impact of Face Occlusion. Facial occlusion can occur in public speech scenarios, such as from podium-mounted microphones or face masks. We evaluate *LightMark* under these occlusion conditions by comparing a clean baseline, a face mask condition, and three microphone conditions with increasing occlusion severity. Mic-Low, Mic-Mid, and Mic-High represent minor lower-face overlap, mouth-region occlusion, and mouth/nose occlusion, respectively. We report TNRs in Figure 18. *LightMark* maintains high TNRs under microphone occlusions, achieving 0.99 for both Mic-Low and Mic-Mid conditions, while the more severe Mic-High condition reduces the TNR to 0.84. The face mask condition results in a TNR of 0.79, as the occluded lower-face region affects both facial identity consistency and dynamic feature extraction consistency. Nevertheless, *LightMark* remains effective under partial occlusion because it also leverages upper-face identity cues and non-mouth facial dynamics. These results indicate that while extensive occlusions of the face would reduce reliability, *LightMark* can tolerate typical moderate face occlusions, such as podium microphones, during speech.

Impact of Motion Blur. Motion blur may occur in videos due to speaker movement. We evaluate *LightMark*'s robustness under different levels of motion blur by controlled blur with increasing severity. The blur severity is controlled by varying the blur kernel size, where larger kernels correspond to stronger motion blur. Figure 19 shows the corresponding TNRs together with illustrative blur patterns. *LightMark* maintains reliable performance under mild blur, achieving TNRs above 0.97 with kernel sizes up to 9. The performance gradually degrades under stronger blur and drops to 0.65 at a kernel size of 17, where severe blur removes important facial details and affects reliable feature extraction. Note that speaker movements outside the illuminated region differ from motion blur, as they result in missing watermark signals rather than degraded facial features; such segments are treated as unverifiable by *LightMark*'s verification mechanism. Overall, while severe blur that significantly degrades facial details remains challenging, *LightMark* remains robust under typical moderate motion blur.

5.4.6 Robustness to Benign Video Edits. We evaluate *LightMark*'s robustness to common benign video edits that frequently arise during ordinary video sharing and post-processing. We use TNR as the metric in this experiment, since all edited videos here remain benign and should therefore still be recognized as negative. Figure 20 presents the results. We consider three categories of benign edits:

- **Visual appearance transformations**, including **color grading** operations that increase or decrease brightness, contrast, and gamma by 20%, together with **color filters** (warm, cool, and sepia) and **style transformations** (grayscale, vintage, and cinematic), all implemented using OpenCV.
- **Platform-level post-processing**, including common platform-driven operations such as video **recompression** into widely used codecs (MP4 to H.264 and MP4 to H.265) and **resizing** to 480p and 720p.
- **Spatial and temporal editing**, including random trimming and cropping.

Overall, *LightMark* shows **strong robustness to most benign edits**. Visual appearance transformations generally maintain high TNR. This is because *LightMark* mainly relies on the shading pattern induced by rolling shutter for watermark embedding and recovery, which is much less sensitive to color or style changes. Platform-level post-processing is also well tolerated. In particular, recompression largely preserves the rolling-shutter stripe pattern, and the resizing results are consistent with those in §5.4.5, where moderate resolution reduction still preserves strong validation performance.

Trimming, i.e., temporal editing, causes only moderate degradation, since it mainly affects the boundary segments near the beginning or end of a clip. Cropping, i.e., spatial editing, has a larger impact because it removes part of the spatial watermark pattern, leaving fewer stripe periods and fewer effective samples for demodulation within each FSK symbol. Meanwhile, it is expected that spatial and temporal edits would induce more noticeable degradation. This is because *LightMark* is designed to protect the spatial, and especially the temporal, consistency of the recorded content, which forgery attacks for misinformation typically seek to disrupt.

5.5 System Runtime Performance

We further evaluate the runtime performance of *LightMark* to quantify its efficiency.

Audience-side overhead. *LightMark* performs *in-band* watermark embedding directly within the audience's recording device and is compatible with various camera models (§5.4.3). As a result, the *audience requires no additional hardware, computation, or resource overhead*. This design ensures seamless integration with audience recording devices, preserving an experience identical to ordinary video capture.

On-scene Watermarking Phase. We benchmark the on-scene watermarking system on a CPU-based platform (Intel Ultra 7 265K). The total processing time per 4.5 s video window is below 3.2 s, including 3.1 s for feature extraction and 6.6 ms for watermark construction and Reed-Solomon encoding. With a pipelined implementation, the system achieves real-time operation, continuously generating watermarks without latency.

Cloud-side Validation Phase. For validation, we deploy the decoding pipeline on an NVIDIA RTX 3090 GPU, representative of trusted hosting platforms (e.g., YouTube) with sufficient server-side capability. The total processing time for one 4.5 s video sample is 1.15 s, including 0.36 s for watermark extraction and 0.79 s for subsequent feature extraction and matching, corresponding to a validation throughput of 3.91 seconds of video per second. These results show that the verification process can efficiently authenticate uploaded videos. We also note that platform-scale deployment does not require exhaustively verifying all uploaded videos. In practice, *LightMark* is most relevant to authenticity-sensitive crowd-sourced recordings, and verification can be selectively triggered for designated events, rapidly spreading videos, or content flagged by users or moderation systems. This selective verification model would substantially reduce overhead and support deployment at platform scale.

Overall, *LightMark* exhibits low computational overhead and high responsiveness across all stages of the pipeline, demonstrating its practicality for real-world deployment.

5.6 User Study on Perceptibility of Watermarking

We conducted a small-scale user study to examine whether *LightMark*'s in-band watermarking is perceptible to the audience and whether it affects the visual quality of uploaded videos. The study includes two parts: (1) on-scene observation and (2) video playback evaluation.

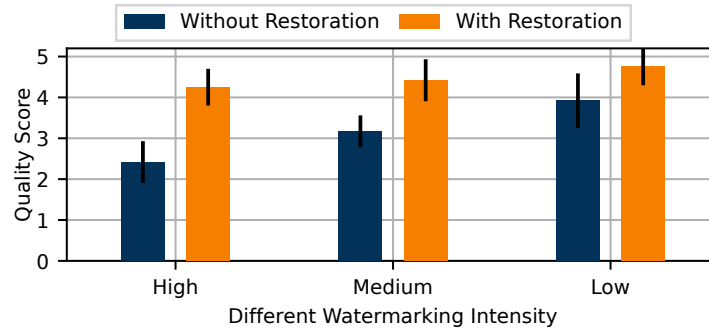


Fig. 21. Figure depicts the performance of *LightMark*'s restoration module under varying watermarking intensities.

(1) On-scene perceptibility. We invited the volunteers to attend short speech sessions where the watermarking system was active or inactive in random order. Participants were seated about 3-5 m from the speaker under typical indoor lighting. Across all sessions, no participant reported visible flickering, indicating that the visible light-based watermarking remains imperceptible to the human eye, consistent with the well-established finding that flicker frequencies above approximately 1 kHz are beyond human visual perception.

(2) Uploaded video quality. Visible-light watermarking may introduce subtle stripe-like patterns in the recorded video, although their visibility is minimized through careful intensity control. To further restore video quality after watermark validation, *LightMark* applies the video restoration module described in §4.2.3. We asked 12 volunteers to rate the quality of the recorded videos on a scale from 1 to 5, where the unwatermarked original video was assigned a baseline score of 5. Note that this study was conducted using non-human videos to ensure ethical compliance while maintaining equivalent lighting and watermarking conditions. We evaluated three watermarking intensity levels by adjusting the transmitting power and camera exposure, representing low-, medium-, and high-intensity watermarking settings. The results are shown in Figure 21. Stronger watermarking intensity leads to more noticeable perceptual degradation before restoration, with the quality score decreasing from 3.9 to 2.5 as the intensity increases. After applying the restoration module, the perceived quality is substantially improved across all intensity levels, with scores above 4.3 even under high-intensity watermarking. These results demonstrate that *LightMark* can preserve visual fidelity while maintaining robust watermark transmission under different watermarking intensities. Representative frame comparisons in Figure 22 further demonstrate that the restoration process effectively eliminates the already subtle visible-light watermark after validation, ensuring the visual quality of the posted videos.

6 Discussion

We now discuss potential countermeasures against *LightMark*, its limitation, and potential extensions.

6.1 Deployment Considerations

Deployment Considerations We discuss the deployment considerations of *LightMark*.

LED Setup. *LightMark* is particularly suitable for indoor speech events, such as public speeches, institutional announcements, and conferences. In these scenarios, *LightMark* can be deployed in two practical modes with flexible LED placement. (1) Occasional deployment: for temporary events, an additional dedicated LED transmitter can be placed near the stage, without modifying the venue's existing lighting infrastructure. The LED is not restricted to be positioned close to the speaker, as *LightMark*'s robust watermark extraction module supports reliable recovery under low-SNR conditions. Deployers can optimize the light placement and optical power



Fig. 22. Figure depicts representative frames showing the subtle vertical stripe patterns introduced by visible-light watermarking and the corresponding restored frames produced by the video restoration module (§4.2.3).

according to the venue layout. This mode is suitable for one-off talks and keeps installation overhead low. (2) Regular deployment: for venues that frequently host speeches or recordings, some of the existing lights in the speaking area can be configured to support *LightMark*, enabling long-term operation with minimal additional setup. This mode is suitable for auditoriums, lecture halls, and other regularly used venues. For large venues, only the lighting around the speaking area needs to be configured, rather than the entire venue lighting infrastructure.

Compatibility with Video Sharing Pipelines. Social media pipelines may introduce additional processing, including frame-rate conversion, with potential frame dropping, and transcoding. *LightMark* performs authentication on the uploaded video before such post-verification transformations; thus, these operations do not affect the authentication result. While frame drops during recording may make the affected segment unverifiable, modern smartphones typically maintain stable frame rates during recording, thereby maintaining the robustness of *LightMark* in practice. Finally, *LightMark* is compatible with common video encoding pipelines, as demonstrated by our evaluation under different compression settings in §5.4.6.

Impact of Anti-flicker Mechanisms. Modern cameras often apply image signal processing (ISP)-level anti-flicker processing to improve visual quality, which may raise concerns about watermark recovery in *LightMark*. Since our experiments are conducted using commodity smartphones with default camera pipelines, the evaluated videos already include such ISP processing. Our results demonstrate that *LightMark* can reliably recover the watermark under practical camera settings with anti-flicker processing enabled.

Notification of involuntary watermarking. From an ethical and operational perspective, *LightMark* should be accompanied by a clear notice to attendees, since it actively embeds authentication signals into audience recordings rather than merely restricting recording behavior.

6.2 Security Analysis on Attacker’s Potential Countermeasures to *LightMark*

Potential adversaries may attempt to bypass or mimic *LightMark*’s watermarking mechanism. We discuss three representative strategies and explain why they are ineffective.

Source-level forgery by Watermark Synthesis. We consider a stronger attack scenario where an attacker attempts to generate a forged video together with synthesized light-modulation patterns before watermark validation. *LightMark* remains secure against such source-level counterfeiting under the cryptographic assumption that the secret key is protected. Even if state-of-the-art generative models can synthesize visually realistic rolling-shutter stripe patterns, such patterns alone are insufficient for successful forgery, as generating a valid HMAC-protected watermark message without the secret key is computationally infeasible. Therefore, the forged video will fail *LightMark*’s watermark authentication and content-consistency checks.

Replay Attacks. An attacker may replay previously watermarked videos to masquerade as genuine recordings. However, such attempts can be detected, as *LightMark*'s watermarks are session-specific and tied to real-time packet order. Any replay or reordering of packets disrupts the expected sequence, causing verification to fail.

Scene jamming. An adversary could attempt to jam the watermarking link by introducing conspicuous flickering light at the scene to corrupt or obscure the watermark, i.e., a form of **optical denial-of-service**. Such interference may cause verification failures for legitimate recordings; however, it is easily noticeable to human observers and is counterproductive, as it deviates from the attacker's original goal of creating convincing misinformation through forged videos.

6.3 Limitation of *LightMark* and Potential Extension

Scope of Authenticated Content. *LightMark* focuses on detecting face-related forgery, aligning with typical misinformation scenarios in which attackers forge a person's talking-head video [80, 81], which has attracted significant research attention in recent years. Currently, *LightMark* cannot detect edits that do not alter the speaker's face region, such as the addition or removal of background objects or overlaid text. The current design also targets single-speaker speech events due to the limited watermark capacity. These limitations can potentially be alleviated by increasing the transmission rate. *LightMark* currently operates at a 120 bps visible-light link, which can be doubled to 240 bps by halving the symbol duration. The increased capacity could enable *LightMark* to embed additional information into the watermark, such as features related to non-facial content, multi-speakers, or even speech, to further improve its coverage. Moreover, *LightMark* can also be combined with complementary detectors to handle other types of manipulations.

Dependence on Rolling-shutter Cameras. *LightMark* relies on the rolling-shutter mechanism of commodity cameras to capture the in-band watermark. Therefore, *LightMark* primarily targets common recording devices, such as phones and consumer mirrorless cameras equipped with rolling-shutter sensors. This also means that cameras with global shutters, while a minority, are not supported by *LightMark*. Such recordings would be treated as *unverifiable* due to the absence of a watermark, rather than falsely authenticated. In practice, global-shutter cameras remain less common among commodity audience recording devices. On the other hand, professional media, which are more likely to use global-shutter cameras, usually distribute videos through trusted publication channels rather than crowd-sourced uploads, and thus reasonably fall outside *LightMark*'s primary scenario.

Watermarking Efficiency. The current implementation introduces 3.2 seconds of processing latency for each 4.5-second video segment. Although this latency does not interrupt continuous watermark embedding due to the pipelined design, it can be further optimized to improve system efficiency. This latency is mainly caused by facial feature extraction. Future work will explore lightweight face feature extractors with improved inference speed, leveraging the continuous advances in efficient face analysis models [16, 21]. Moreover, the current prototype performs inference on a CPU; deploying the system with commodity edge accelerators can further reduce inference latency and improve practical deployment efficiency.

Outdoor and Uncontrolled Environments. In this paper, we primarily focus on indoor speech scenarios. Outdoor deployment introduces additional challenges, such as strong sunlight interference and other uncontrolled illumination variations. Addressing these challenges requires further investigation of robust light modulation with extended sensing range, and adaptive illumination under diverse outdoor conditions.

7 Related Work

We now present closely related work with *LightMark*.

7.1 Video Forgery Detection

Video forgery detection methods can be broadly classified into *passive* and *active* approaches. Passive approaches detect video forgery by analyzing visual inconsistencies in space, time, or frequency domains [18, 22, 25, 26, 54, 66], as well as semantic or cross-modal mismatches such as lip-audio desynchronization [3, 37, 59, 72, 82, 90]. While effective in controlled settings, these detectors often fail under unseen manipulations, compression, or environmental variations due to their reliance on artifact-level features [80, 81].

Active methods enhance robustness by embedding auxiliary meta-information into the media for later verification. Traditional approaches use *out-of-band* watermarking, inserting authentication data after capture [7], which is unsuitable for live audience-generated videos lacking post-processing. In this paper, *LightMark* pursues *in-band* schemes, where authentication embedding and video capture occur simultaneously, inherently binding the watermark to the genuine recording event. TalkLock [60] is an important early example of this principle, which embeds authentication data into live video via a sequence of QR codes to protect speech integrity. While effective, the QR code is visually intrusive and unsuitable for natural recordings such as public speeches. Building on the insight of in-band authentication, *LightMark* avoids explicit visual markers, such as QR codes, by carefully designing an imperceptible LED-to-camera channel, thereby achieving unobtrusive watermarking. Moreover, *LightMark* faces a formidable watermarking challenge: the optical channel provides limited bandwidth, while the embedded watermark must remain cryptographically verifiable and robust against sophisticated video forgeries. To address this challenge, *LightMark* carefully designs a watermark representation that encodes semantically meaningful facial identity and dynamics within the constrained channel capacity, enabling robust authentication.

7.2 Light-based Communication

Light has been widely exploited as a communication medium, including LED-to-photodiode links [46, 89], screen-camera channels [30, 36, 48, 52, 73, 86], visible light backscatter [71, 74, 76, 79]. Among them, **rolling-shutter visible light communication (VLC)**, pioneered by RollingLight [34], enables LED-to-camera communication by exploiting spatial stripe patterns generated by temporally modulated LEDs in commodity rolling-shutter cameras. Subsequent studies have mainly focused on improving communication throughput through advanced modulation schemes and receiver designs [14, 29, 42, 83]. Since these systems target data transmission, their optimization objectives focus on communication performance rather than preserving video fidelity.

LightMark repurposes the LED-to-camera channel for *secure, in-band watermarking*. Unlike communication-oriented rolling-shutter VLC systems, *LightMark* must not only achieve reliable watermark transmission but also preserve visual quality. A closely related work, CORE-Lens [42], also considers visual interference in VLC. However, CORE-Lens relies on background-specific training and is evaluated only under a limited set of predefined backgrounds and short-range scenarios, restricting its applicability to *LightMark*'s target scenario, i.e., live speech recordings with unseen and dynamic environments. *LightMark* innovatively introduces a co-designed watermark embedding and extraction framework that jointly optimizes watermark robustness and visual fidelity, enabling reliable authentication under varying environments while maintaining natural video appearance.

8 Conclusion

We presented *LightMark*, a novel video forgery detection system based on *in-band* watermarking through visible light modulation. *LightMark* can reliably detect forged videos through watermark validation, with failed extraction resulting in an unverifiable outcome. We design and implement *LightMark*, as well as perform a real-world evaluation over diverse speech scenes. *LightMark* achieves high AUC and accuracy in detecting both editing-based manipulations and AI-generated forgeries, including face swapping and reenactment. Moreover, *LightMark* is fully non-intrusive, as its watermarking remains imperceptible in both the recording process and the restored

video. We envision *LightMark* as a complementary solution to existing passive detection methods, offering a practical step toward mitigating the risks of visual misinformation.

Acknowledgments

We sincerely thank our anonymous reviewers and AEs for their valuable feedback. This paper is supported by the National Key R&D Program of China (2023YFA1009500), the National Natural Science Foundation of China (NSFC) under grants No.62372400 and 62602402, Natural Science Foundation of Jiangsu Province of China (Grant No. BK20240615), and the Fundamental Research Funds for the Central Universities (2025110528-0).

References

- [1] Accessed: 2025/10/30. DeepFakes. <https://github.com/deepfakes/faceswap>.
- [2] Accessed: 2025/10/30. FaceSwap. <https://github.com/MarekKowalski/FaceSwap>.
- [3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*. 1–7.
- [4] Google AI. [n.d.]. MediaPipe. <https://ai.google.dev/edge/mediapipe/solutions/guide>.
- [5] Amazon. 2025. Overview of Amazon Titan models. <https://docs.aws.amazon.com/bedrock/latest/userguide/titan-models.html>.
- [6] Arduino. 2023. Arduino Uno Rev3. <https://store.arduino.cc/products/arduino-uno-rev3>.
- [7] Md. Asikuzzaman and Mark R. Pickering. 2018. An Overview of Digital Video Watermarking. *IEEE Trans. Circuits Syst. Video Technol.* 28, 9 (2018), 2131–2153.
- [8] BBC. 2019. Nancy Pelosi clip shows misinformation still has a home on Facebook. <https://www.bbc.com/news/technology-48405661>.
- [9] BBC. 2022. Deepfake presidents used in Russia-Ukraine war. <https://www.bbc.com/news/technology-60780142>.
- [10] BBC. 2022. False claims of ‘deepfake’ President Biden go viral. <https://www.bbc.com/news/62338593>.
- [11] Sai Bi, Xiaoguang Han, and Yizhou Yu. 2015. An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Trans. Graph.* 34, 4 (2015), 78:1–78:12.
- [12] Rens Bloom, Marco Zúñiga, and Chaitra Pai. 2019. LuxLink: creating a wireless link from ambient light. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems, SenSys’19*. ACM, 166–178.
- [13] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE, 4103–4112.
- [14] Chun-Ling Chan, Hsin-Mu Tsai, and Kate Ching-Ju Lin. 2017. POLI: Long-Range Visible Light Communications Using Polarized Light Intensity Modulation. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys’17*. ACM, 109–120.
- [15] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *MM ’20: The 28th ACM International Conference on Multimedia*. ACM, 2003–2011.
- [16] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In *Conference on Biometric Recognition, CCBR 2018*, Vol. 10996. Springer, 428–438.
- [17] Xiang Chen, Jinshan Pan, and Jiangxin Dong. 2024. Bidirectional Multi-Scale Implicit Neural Representations for Image Deraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. IEEE, 25627–25636.
- [18] Yunzhuo Chen, Naveed Akhtar, Nur Al Hasan Haldar, and Ajmal Mian. 2025. Deepfake Detection with Spatio-Temporal Consistency and Attention. *CoRR abs/2502.08216* (2025).
- [19] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. Computer Vision Foundation / IEEE Computer Society, 8789–8797.
- [20] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. IEEE Computer Society, 1800–1807.
- [21] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. CVF / IEEE, 5202–5211.
- [22] Xiangling Ding, Wenjie Zhu, and Dengyong Zhang. 2022. DeepFake Videos Detection via Spatiotemporal Inconsistency Learning and Interactive Fusion. In *Proceedings of the IEEE International Conference on Sensing, Communication, and Networking*. 425–433.
- [23] dlib C++ Library. [n.d.]. dlib. <https://dlib.net/>.
- [24] Brian Dolhansky, Russ Howes, Ben Pfau, Nicole Baram, and Cristian Canton-Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. *CoRR abs/1910.08854* (2019).
- [25] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the International Conference on Machine Learning*. 3247–3258.

- [26] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal Inconsistency Learning for DeepFake Video Detection. In *Proceedings of the ACM Multimedia Conference*. 3473–3481.
- [27] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. 2022. Depth-Aware Generative Adversarial Network for Talking Head Video Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE, 3387–3396.
- [28] Jiawei Hu, Yanxiang Wang, Hong Jia, Wen Hu, Mahbub Hassan, Brano Kusy, Ashraf Uddin, and Moustafa Youssef. 2023. Iris: Passive Visible Light Positioning Using Light Spectral Information. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3 (2023), 97:1–97:27.
- [29] Pengfei Hu, Parth H. Pathak, Xiaotao Feng, Hao Fu, and Prasant Mohapatra. 2015. ColorBars: increasing data rate of LED-to-camera communication using color shift keying. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, CoNEXT'15*. ACM, 12:1–12:13.
- [30] Wenjun Hu, Hao Gu, and Qifan Pu. 2013. LightSync: unsynchronized visual communication over screen-camera links. In *The 19th Annual International Conference on Mobile Computing and Networking, MobiCom'13*. ACM, 15–26.
- [31] Huawei. 2023. HUAWEI P40 Pro. <https://consumer.huawei.com/en/phones/p40-pro/>.
- [32] InsightFace. [n. d.]. ArcFace. <https://www.insightface.ai/research/arcface>.
- [33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, 8107–8116.
- [34] Hui-Yu Lee, Hao-Min Lin, Yu-Lin Wei, Hsin-I Wu, Hsin-Mu Tsai, and Kate Ching-Ju Lin. 2015. RollingLight: Enabling Line-of-Sight Light-to-Camera Communications. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys'15*. ACM, 167–180.
- [35] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *CoRR* abs/1912.13457 (2019).
- [36] Tianxing Li, Chuankai An, Xinran Xiao, Andrew T. Campbell, and Xia Zhou. 2015. Real-Time Screen-Camera Communication Behind Any Scene. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys'15*. ACM, 197–211.
- [37] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 46–52.
- [38] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, 3204–3213.
- [39] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip S. Yu. 2025. A Survey of Text Watermarking in the Era of Large Language Models. *ACM Comput. Surv.* 57, 2 (2025), 47:1–47:36.
- [40] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 772–781.
- [41] Xiaomei Liu, Bolin Zhang, Yuchen Wen, Xin Tang, and Haibo Su. 2022. An Improved Image Authentication Method Using QR Code Watermarking Approach. In *Artificial Intelligence and Security - 8th International Conference, ICAIS 2022*, Vol. 13340. Springer, 337–348.
- [42] Ziwei Liu, Tianyue Zheng, Chao Hu, Yanbing Yang, Yimao Sun, Yi Zhang, Zhe Chen, Liangyin Chen, and Jun Luo. 2022. CORE-lens: simultaneous communication and object recognition with disentangled-GAN cameras. In *The 28th Annual International Conference on Mobile Computing and Networking, MobiCom'22*. ACM, 172–185.
- [43] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing Face Forgery Detection With High-Frequency Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 16317–16326. doi:10.1109/CVPR46437.2021.01605
- [44] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing Face Forgery Detection With High-Frequency Features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 16317–16326.
- [45] mcpro24fps. 2023. Professional manual video camera app. <https://www.mcpro24fps.com/>.
- [46] Muhammad Sarmad Mir, Borja Genovés Guzmán, Ambuj Varshney, and Domenico Giustiniano. 2021. PassiveLiFi: rethinking LiFi for low-power and long range RF backscatter. In *The 27th Annual International Conference on Mobile Computing and Networking, MobiCom'21*. ACM, 697–709.
- [47] Yisroel Mirsky and Wenke Lee. 2022. The Creation and Detection of Deepfakes: A Survey. *Comput. Surveys* 54, 1 (2022), 7:1–7:41.
- [48] Viet Nguyen, Yaqin Tang, Ashwin Ashok, Marco Gruteser, Kristin J. Dana, Wenjun Hu, Eric Wengrowski, and Narayan B. Mandayam. 2016. High-rate flicker-free screen-camera communication with spatially adaptive embedding. In *35th Annual IEEE International Conference on Computer Communications, INFOCOM'16*. IEEE, 1–9.
- [49] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*. IEEE, 7183–7192.
- [50] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 4172–4182.

- [51] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *MM '20: The 28th ACM International Conference on Multimedia*. ACM, 484–492.
- [52] Kun Qian, Yumeng Lu, Zheng Yang, Kai Zhang, Kehong Huang, Xinjun Cai, Chenshu Wu, and Yunhao Liu. 2021. AIRCODE: Hidden Screen-Camera Communication on an Invisible and Inaudible Dual Channel. In *18th USENIX Symposium on Networked Systems Design and Implementation, NSDI'21*. USENIX Association, 457–470.
- [53] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *16th European Conference Computer Vision, ECCV 2020*, Vol. 12357. Springer, 86–103.
- [54] Mengyu Qiao, Runze Tian, and Yang Wang. 2025. Towards Generalizable Deepfake Detection with Spatial-Frequency Collaborative Learning and Hierarchical Cross-Modal Fusion. *CoRR* abs/2504.17223 (2025).
- [55] Amna Qureshi, David Megias, and Minoru Kuribayashi. 2021. Detecting Deepfake Videos using Digital Watermarking. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2021*. IEEE, 1786–1793.
- [56] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. 2019. Progressive Image Deraining Networks: A Better and Simpler Baseline. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. Computer Vision Foundation / IEEE, 3937–3946.
- [57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE, 10674–10685.
- [58] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*. IEEE, 1–11.
- [59] Davide Salvi, Honggu Liu, Sara Mandelli, Paolo Bestagini, Wenbo Zhou, Weiming Zhang, and Stefano Tubaro. 2023. A Robust Approach to Multimodal Deepfake Detection. *Journal of Imaging* 9, 6 (2023), 122.
- [60] Irtaza Shahid and Nirupam Roy. 2023. "Is this my president speaking?" Tamper-proofing Speech in Live Recordings. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services, MobiSys 2023*. ACM, 219–232.
- [61] Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting Deepfakes with Self-Blended Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE, 18699–18708.
- [62] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. 2023. BlendFace: Re-designing Identity Encoders for Face-Swapping. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 7600–7610.
- [63] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*. 7135–7145.
- [64] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *12th European Conference on Computer Vision, ECCV'12*, Vol. 7576. Springer, 746–760.
- [65] Apple Apps Store. 2023. Protake = Mobile Cinema Camera. <https://apps.apple.com/us/app/protake-mobile-cinema-camera/id1498431506>.
- [66] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5052–5060.
- [67] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019 (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 6105–6114.
- [68] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: image synthesis using neural textures. *ACM Trans. Graph.* 38, 4 (2019), 66:1–66:12.
- [69] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. IEEE, 2387–2395.
- [70] The New York Times. 2023. 'A.I. Obama' and Fake Newscasters: How A.I. Audio Is Swarming TikTok. <https://www.nytimes.com/2023/10/12/technology/tiktok-ai-generated-voices-disinformation.html>.
- [71] Ambuj Varshney, Andreas Soleiman, Luca Mottola, and Thimo Voigt. 2017. Battery-free Visible Light Sensing. In *Proceedings of the 4th ACM Workshop on Visible Light Communication Systems, VLCS@MobiCom'17*. ACM, 3–8.
- [72] Harsh Kumar Verma and Isha Kumar. 2024. Enhancing Deepfake Detection with a Hybrid CNN-BiLSTM Approach. In *Proceedings of the International Conference on Soft Computing and Machine Intelligence*. 351–356.
- [73] Anran Wang, Zhuoran Li, Chunyi Peng, Guobin Shen, Gan Fang, and Bing Zeng. 2015. InFrame++: Achieve Simultaneous Screen-Human Viewing and Hidden Screen-Camera Communication. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys'15*. ACM, 181–195.
- [74] Purui Wang, Lilei Feng, Guojun Chen, Chenren Xu, Yue Wu, Kenuo Xu, Guobin Shen, Kuntai Du, Gang Huang, and Xuanzhe Liu. 2020. Renovating road signs for infrastructure-to-vehicle networking: a visible light backscatter communication and networking approach. In *The 26th Annual International Conference on Mobile Computing and Networking, MobiCom'20*. ACM, 6:1–6:13.
- [75] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. 2021. FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking. In *MM '21: ACM Multimedia Conference*. ACM, 3546–3555.

- [76] Yue Wu, Purui Wang, Kenuo Xu, Lilei Feng, and Chenren Xu. 2020. Turboboosting Visible Light Backscatter Communication. In *Proceedings of the 2020 Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, SIGCOMM'20*. ACM, 186–197.
- [77] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. 2022. Designing One Unified Framework for High-Fidelity Face Reenactment and Swapping. In *European Conference Computer Vision, ECCV 2022*, Vol. 13675. Springer, 54–71.
- [78] Huiyu Xu, Yaopeng Wang, Zhibo Wang, Zhongjie Ba, Wenxin Liu, Lu Jin, Haiqin Weng, Tao Wei, and Kui Ren. 2024. ProFake: Detecting Deepfakes in the Wild against Quality Degradation with Progressive Quality-adaptive Learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024*. ACM, 2207–2221.
- [79] Kenuo Xu, Chen Gong, Bo Liang, Yue Wu, Boya Di, Lingyang Song, and Chenren Xu. 2022. Low-Latency Visible Light Backscatter Networking with RetroMUMIMO. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, SenSys'22*. ACM, 448–461.
- [80] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. 2024. DF40: Toward Next-Generation Deepfake Detection. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.
- [81] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.
- [82] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 8261–8265.
- [83] Yanbing Yang, Jie Hao, and Jun Luo. 2017. CeilingTalk: Lightweight Indoor Broadcast Through LED-Camera Communication. *IEEE Trans. Mob. Comput.* 16, 12 (2017), 3308–3319.
- [84] Yanbing Yang, Jiangtian Nie, and Jun Luo. 2017. ReflexCode: Coding with Superposed Reflection Light for LED-Camera Communication. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, MobiCom'17*. ACM, 193–205.
- [85] Hao Zeng, Wei Zhang, Changjie Fan, Tangjie Lv, Suzhen Wang, Zhimeng Zhang, Bowen Ma, Lincheng Li, Yu Ding, and Xin Yu. 2023. FlowFace: Semantic Flow-Guided Shape-Aware Face Swapping. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*. AAAI Press, 3367–3375.
- [86] Kai Zhang, Chenshu Wu, Chaofan Yang, Yi Zhao, Kehong Huang, Chunyi Peng, Yunhao Liu, and Zheng Yang. 2018. ChromaCode: A Fully Imperceptible Screen-Camera Communication System. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, MobiCom'18*. ACM, 575–590.
- [87] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, 8652–8661.
- [88] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-Attentional Deepfake Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 2185–2194.
- [89] Tian Zhao, Kevin Wright, and Xia Zhou. 2016. The darkLight rises: visible light communication in the dark. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, MobiCom'16*. ACM, 2–15.
- [90] Yipin Zhou and Ser-Nam Lim. 2021. Joint Audio-Visual Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14780–14789.
- [91] Yixuan Zhu, Wenliang Zhao, Yansong Tang, Yongming Rao, Jie Zhou, and Jiwen Lu. 2024. StableSwap: Stable Face Swapping in a Shared and Controllable Latent Space. *IEEE Trans. Multimed.* 26 (2024), 7594–7607.