

Adversarial CSI Obfuscation Against Device-Free Activity Recognition in RIS-Assisted Systems

GABRIELE RESTUCCIA^{1,2}, SILVIA SCHILLECI^{1,2,3} (Graduate Student Member, IEEE),
ALESSANDRA DINO^{1,2}, DOMENICO GARLISI^{2,3}, FABRIZIO GIULIANO^{2,3},
STEFANO MANGIONE^{1,2} (Member, IEEE), AND ILENIA TINNIRELLO^{1,2}

¹Department of Engineering, University of Palermo, 90128 Palermo, Italy

²CNIT, 43124 Parma, Italy

³Department of Mathematics and Computer Science, University of Palermo, 90123 Palermo, Italy

CORRESPONDING AUTHOR: G. RESTUCCIA (gabriele.restuccia@unipa.it)

This work was supported in part by the Fondo Finalizzato alla Ricerca di Ateneo (FFR)–Anno 2026 of the University of Palermo, Italy, under Project PJ_GEST_FFR_2026 and in part by the Project Secure Ubiquitous Networks and Services (SUNS) - Codice Unico di Progetto (CUP) under Grant B79H26000210007, Grant B72F26000500005, and Grant B79J26001240005.

ABSTRACT Reconfigurable Intelligent Surfaces (RISs) extend 5G New Radio (NR) coverage by redirecting signals toward Non-Line-of-Sight regions, but the same reflected footprint exposes Channel State Information (CSI) to passive eavesdroppers, who can train a Machine Learning classifier on it and perform unauthorized device-free Human Activity Recognition (HAR). Because the information leaks through the propagation environment rather than through transmitted content, encryption and access control are ineffective, and artificial-noise approaches either fail against trained neural networks or violate the strict Error Vector Magnitude (EVM) limits of 3GPP TS 38.104. This paper proposes a transmitter-side adversarial defense that obfuscates the CSI intercepted by a non-adaptive eavesdropper within the EVM budget of the active modulation. The defense reformulates the Carlini & Wagner attack for the radio-frequency domain, replacing additive norm-bounded perturbations with multiplicative amplitude scaling and phase rotation on the complex channel coefficients, and embeds the modulation-dependent EVM limits directly into the adversarial objective. An offline Universal Adversarial Perturbation (UAP) variant removes the per-packet optimization cost and meets the latency budget of Ultra-Reliable Low-Latency Communication services. The instance-specific attack achieves a feasible Attack Success Rate above 90% under QPSK and 16-QAM and above 83% under the 3.5% EVM budget of 256-QAM. The UAP retains 73% to 83% of this effectiveness with no online computation and transfers across receiver positions, reducing a target-position classifier from 83.6% to roughly 33% accuracy. End-to-end physical-layer simulations bound the cost to the legitimate link below 1 dB of Signal-to-Noise-Ratio penalty on Block Error Rate after Low-Density Parity-Check decoding.

INDEX TERMS Adversarial machine learning (AML), channel state information (CSI), device-free human activity recognition (HAR), error vector magnitude (EVM), physical layer privacy, reconfigurable intelligent surfaces (RIS), universal adversarial perturbation (UAP), wireless sensing.

I. INTRODUCTION

THE deployment of Reconfigurable Intelligent Surfaces (RISs) is becoming a key enabler for Beyond 5G (B5G) and 6G networks, especially in controlled indoor environments operating at millimeter-wave (mmWave) frequencies. In these scenarios, physical obstacles and room partitions often place User Equipments (UEs) in Non-Line-of-Sight (NLOS) conditions with respect to the Next Generation Node Base (gNB), leading to severe coverage degradation and reduced link quality [1]. RIS-assisted communications

mitigate this limitation by reconfiguring the electromagnetic propagation environment through passive beamforming. By steering the incident signal toward NLOS regions, RIS can improve coverage and recover part of the capacity loss caused by obstructed propagation paths [2], [3].

This broader and more controllable signal footprint, however, also enlarges the physical-layer exposure surface. Once the RIS is activated to extend coverage, a passive sensing device operated by a malicious actor, positioned in NLOS with respect to the gNB, receives access to channel state

information (CSI) that is implicitly modulated by the activity of people within the environment [4], [5]. By collecting this CSI data and training a Machine Learning (ML) classifier, such as a Convolutional Neural Network (CNN), the adversary can perform unauthorized Human Activity Recognition (HAR), effectively turning the communication infrastructure into an involuntary surveillance mechanism [6].

Unlike conventional data-centric privacy risks, CSI-based inference does not require access to application data and operates opportunistically on signals already present in the environment. Conventional security mechanisms, including access control and application-layer encryption, are therefore insufficient, as they do not prevent information leakage through propagation-dependent measurements [6], [7]. This fundamental tension between communication utility and physical-layer privacy motivates defenses that act directly on the signal representation [8].

Traditional Physical Layer Security (PLS) approaches rely on Artificial Noise (AN) injection [9] or randomized RIS phase configurations [10]. Both strategies, however, share a critical weakness: modern CNNs inherently exhibit substantial robustness against isotropic random noise, such as Gaussian perturbations, maintaining stable predictive accuracy in the presence of stochastic distortions. [11].

In 5G New Radio (NR) systems, increasing noise power to compromise recognition is impractical due to the strict Error Vector Magnitude (EVM) thresholds mandated by 3GPP TS 38.104 [12]. These requirements, which limit the mean-square distance between ideal and received constellation points, ensure that signal distortion does not induce symbol errors or undermine the Quality of Service (QoS) at the legitimate receiver.

Standard Adversarial Machine Learning (AML) methods such as Fast Gradient Sign Method (FGSM) [13] and Projected Gradient Descent (PGD) [14] assume arbitrary additive perturbations bounded in an L_p norm, which ignores communication-standard requirements and limits their direct applicability to wireless physical-layer settings.

To address this privacy-utility trade-off, we propose a physically realizable **adversarial defense** for RIS-assisted networks. We reformulate the Carlini & Wagner (C&W) [15] optimization framework for the radio frequency (RF) domain, replacing unconstrained additive perturbations with multiplicative amplitude scaling (α) and phase rotation (θ) over spatio-temporal CSI tensors. The EVM is explicitly incorporated into the adversarial objective, ensuring 3GPP compliance while degrading the eavesdropper's recognition performance.

We first study a per-sample formulation, in which the perturbation is optimized independently for each CSI instance. While this approach provides maximal obfuscation, generating a dedicated perturbation for every transmitted packet introduces prohibitive computational latency in real-time networks. To overcome this limitation, we extend the formulation to an offline Universal Adversarial Perturbation (UAP)

[16]: a single perturbation mask computed once and applied uniformly across all inputs, which generalizes over the data distribution by exploiting adversarial transferability [17], [18] without requiring any online optimization.

The architecture of the proposed defense is illustrated in Fig. 1. The gNB introduces structured perturbations on the transmitted symbols within the EVM budget of the target modulation scheme. The perturbed signal is forwarded to the RIS, which redirects it toward NLOS areas. The legitimate UE successfully demodulates the signal, while the malicious node obtains an incorrect activity prediction from its HAR classifier.

The main contributions of this paper are summarized as follows:

- We identify a physical-layer privacy threat specific to RIS-assisted 5G NR deployments, where a passive observer in the RIS-enhanced NLOS region exploits the reflected signal for device-free HAR.
- We reformulate the C&W attack for the RF domain, replacing additive L_p -bounded perturbations with multiplicative amplitude scaling and phase rotation on the complex channel coefficients, and embedding the modulation-dependent EVM limits of 3GPP TS 38.104 directly into the adversarial objective.
- We extend the per-sample formulation to an offline UAP that shares a single (α, θ) pair across all CSI realizations under the same EVM budget, removing the per-packet optimization overhead while retaining effectiveness through cross-sample transferability.
- We evaluate the physical C&W as an adversarial defense against a non-adaptive eavesdropper along three independent axes: (i) effect on the legitimate link, with an Signal-to-Noise-Ratio (SNR) penalty below 1 dB on Block Error Rate (BLER) after Low-Density Parity-Check (LDPC) decoding; (ii) physical signature of the perturbation, showing that effectiveness depends on its structure rather than its magnitude; and (iii) spatial transferability of the UAP.

Table 1 clarifies the terminology used in this paper.

The remainder of this paper is organized as follows. After reviewing the related literature on physical-layer privacy and AML in Section II, Section III introduces the considered system and threat models, and Section IV presents the proposed DL-based HAR framework. The formulation of the physical-domain adversarial defense under 5G NR EVM constraints is then detailed in Section V. Sections VI and VII describe the experimental methodology and system configuration, respectively, while Section VIII evaluates the instance-specific adversarial defense performance. Section IX introduces and analyzes the UAP, followed by Section X, which investigates the impact of the defense on legitimate communication reliability. Transferability across different spatial positions is examined in Section XI. Finally, Section XII concludes the paper.

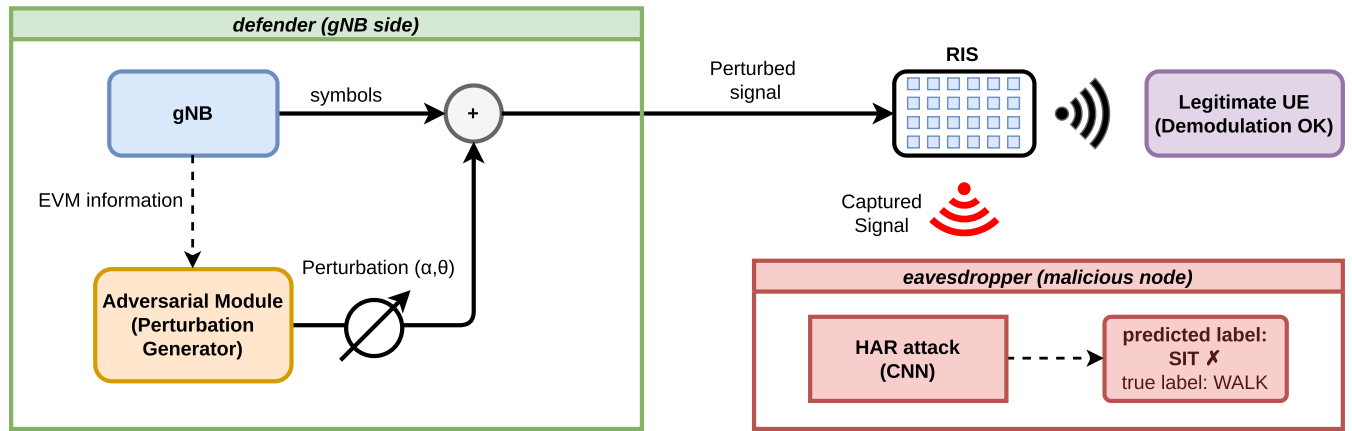


FIGURE 1. Architecture of the proposed defense solution.

TABLE 1. Terminology.

Term	Definition
Eavesdropper	Passive malicious entity that intercepts CSI to perform unauthorized HAR.
HAR classifier	ML model used by the eavesdropper to infer human activities from CSI.
Adversarial perturbation	Controlled modification of CSI (amplitude and phase) designed to mislead the eavesdropper HAR classifier.
Adversarial defense	Proposed CSI obfuscation strategy based on adversarial perturbations.
Attack Success Rate (ASR)	Percentage of perturbed CSI samples that cause the eavesdropper HAR classifier to mispredict the underlying activity.
Adversarial attack	AML procedure, used in our adversarial defense, to generate controlled CSI perturbations that induce misclassification by the eavesdropper's HAR classifier.

II. BACKGROUND & RELATED WORK

Device-free HAR via CSI was first systematically demonstrated in WiFi environments by Yousefi et al. [5], who showed that human motion produces structured, reproducible variations across OFDM subcarriers that can be reliably classified. This result has given origin to a wide range of work on HAR based on ML, ranging from classical signal processing to deep learning classifiers that operate directly on raw CSI tensors. The severity of this privacy threat is underscored by recent studies demonstrating the extraction of fine-grained human behaviors from wireless signals. For instance, adversaries can leverage complementary CSI amplitude and phase to perform highly accurate device-free gesture recognition [19]. Furthermore, the feasibility of passive wireless sensing extends beyond controlled laboratory settings; recent work has shown that router-acquired CSI can be effectively used to detect human activities, such as worker falls, even in complex, unconstrained environments [20].

Liu et al. [4] extended this paradigm to RIS-assisted environments with RISAR, a system that leverages controlled reflection paths to enhance CSI-based HAR in Wi-Fi: by directing the reflected signal toward target areas, the RIS improves spatial resolution and classification accuracy. In RISAR, the RIS is an intentional detection aid. In our scenario, in contrast, the RIS is deployed for 5G NR coverage extension at 28 GHz, and the HAR capability of a passive eavesdropper is a side effect, a threat that is harder to anticipate and cannot be addressed by limiting coverage.

A. PHYSICAL-LAYER PRIVACY AND CSI OBFUSCATION

Since CSI-based inference operates on propagation-dependent measurements rather than transmitted content, data-centric mechanisms such as encryption and access control offer no protection [6], [7]. The most widely studied countermeasure is artificial noise (AN) injection [9], and random RIS phase configurations have been proposed as an alternative [10]. As already mentioned, however, neural networks are inherently robust to random noise [11], [13], [21], which limits the effectiveness of both strategies. Moreover, simply increasing noise power is impractical in 5G NR systems due to the strict EVM limits imposed by 3GPP TS 38.104 [12].

More structured approaches have been proposed for Wi-Fi environments. Cominelli et al. [8] introduced Anti-Sense, which deploys an external relay that continuously re-transmits received frames with a time-varying random delay governed by a Markov-Uniform process, injecting an artificial reflection path that renders the CSI fingerprint statistically inconsistent with any fixed target position. The communication constraint is expressed in terms of Packet Delivery Rate (PDR), which the authors show is not degraded and can even improve, since the relay acts as a signal amplifier. IRShield [22] focuses on motion detection rather than activity classification, using a 256-element binary-phase RIS whose reflection coefficients $r_m(t) \in \{-1, +1\}$ are continuously randomized to induce channel variations that exceed the interceptor's detection threshold. The communication constraint is expressed in terms of Packet Error Rate (PER),

TABLE 2. Comparison of structured CSI obfuscation approaches. “Constr. opt.” indicates whether the defense formulates obfuscation as a constrained optimization problem, and “Eff. vs. trained clf.” whether it remains effective against a well-trained classifier rather than relying on randomization.

Work	Env.	QoS	Device	Constr. opt.	Eff. vs. trained clf.
AntiSense [8]	Wi-Fi	PDR	Relay	No	No
IRShield [22]	Wi-Fi	PER	RIS	No	No
Ours	5G NR	EVM	gNB	Yes	Yes

which remains unchanged since the random configurations simultaneously improve and degrade the channel.

Table 2 contrasts these approaches with the proposed defense. Prior work does not formulate obfuscation as a constrained optimization problem: AntiSense and IRShield perturb the channel through randomized strategies, which a well-trained classifier resists [23]. Both operate in Wi-Fi, where PDR or PER measure communication quality, whereas in 5G NR the binding constraint is the modulation-dependent EVM that bounds the admissible signal distortion [12]. Each also relies on a dedicated relay or RIS, while in our scenario the RIS belongs to the communication infrastructure whose activation creates the privacy threat, so the defense reuses that same infrastructure.

These differences make a direct reimplement of [8] and [22] unsuitable as a baseline, since environment, quality metric, and hardware all diverge from our setting. The mechanism they share with AN injection [9] and random RIS configurations [10] is randomization, and trained networks retain their discriminative features under unstructured perturbations [11], [23]. The proposed defense instead applies a structured, classifier-aware perturbation that reduces the eavesdropper’s accuracy while leaving the signal almost unchanged (as discussed in Section VIII). Its effectiveness follows from the structure of the perturbation rather than its magnitude, which is what randomized standard-compliant strategies cannot provide.

B. ADVERSARIAL MACHINE LEARNING FOR WIRELESS SYSTEMS

1) ADVERSARIAL MACHINE LEARNING

AML studies the vulnerability of trained ML models to small, intentionally designed input perturbations [24]. Let $f : \mathcal{X} \rightarrow \mathbb{R}^K$ be a classifier that maps an input $\mathbf{x} \in \mathcal{X}$ to a vector of K class logits, with predicted label $\hat{y}(\mathbf{x}) = \arg \max_k f_k(\mathbf{x})$. An adversarial example is constructed as

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \delta, \quad \|\delta\|_p \leq \epsilon, \quad (1)$$

where δ is optimized so that the classifier prediction changes while the perturbation remains small under the chosen L_p norm.

Adversarial attacks are commonly divided into *untargeted* and *targeted* attacks. In the first case, the objective is to induce any incorrect prediction, i.e., $\hat{y}(\mathbf{x}_{\text{adv}}) \neq \hat{y}(\mathbf{x})$, while

in the second case the classifier is driven toward a specific false class $t \neq y$. Among optimization-based approaches, the C&W attack [15] formulates adversarial example generation as a continuous optimization problem that explicitly balances perturbation magnitude and misclassifications confidence. In Section V-V-A, we reformulate the original additive attack in terms of physically realizable amplitude and phase perturbations on complex CSI coefficients under EVM constraints.

2) AML APPLICATION IN WIRELESS ENVIRONMENT

The application of AML to wireless classifiers builds on the observation that neural networks used for radio classification and sensing exhibit the same sensitivity to structured perturbations already documented in computer vision [13], [14], [15]. In the wireless domain, however, the feasibility of a perturbation cannot be assessed through input-space norms alone: it also depends on how the perturbation is generated, propagated, and received, and on whether the underlying communication link remains usable.

A representative study in this direction is the work of Flowers et al. [25], who evaluated FGSM-based adversarial attacks against raw-IQ Adaptive Modulation and Coding (AMC) classifiers. Their framework distinguishes attacks according to the injection point: directly at the classifier input, at the transmitter before over-the-air propagation, or through a separate jammer. The *self-protection* setting is particularly relevant to our scenario, as it considers perturbations introduced by the legitimate transmitter to mislead an unauthorized classifier. The same authors also argue that wireless adversarial attacks should be assessed against communication-level metrics such as Bit Error Rate (BER), expressing the perturbation constraint through the power ratio E_s/E_p rather than through an L_p norm alone. Their formulation, however, does not include a modulation-dependent distortion threshold, and the perturbations remain additive in the digital domain, without an explicit model of channel-aware amplitude and phase transformations. To bridge this gap, Kim et al. [26] introduced channel-aware adversarial attacks against deep learning wireless signal classifiers, generating perturbations subject to strict power budgets while explicitly accounting for over-the-air channel effects from the adversary to the receiver.

In the context of Wi-Fi HAR, Huang et al. [23] report that gradient-based adversarial perturbations can fool CSI-based classifiers while remaining small at the signal level. They also observe that random noise injection provides limited protection, since trained neural networks retain discriminative features under unstructured perturbations. This line of work suggests that effective obfuscation requires perturbations that are structured and classifier-aware. To mitigate these vulnerabilities, algorithmic defensive strategies have been introduced. For instance, the SecureSense framework [27] leverages prediction consistency between normal and adversarial examples as a regularization technique to protect device-free systems. Other frameworks rely on robust training paradigms, such as AdvLoc [28], which defends indoor

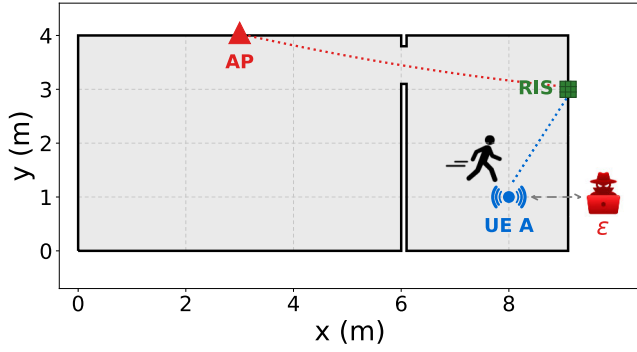


FIGURE 2. Considered indoor scenario, including the gNB, the RIS, and the sensing node.

localization by applying adversarial training directly onto CSI tensors, and Selective Adversarial Training (SAT) [29], which autonomously calibrates perturbation intensities to optimize classifier defense in Wi-Fi HAR. Similar vulnerabilities and corresponding defenses have also been investigated across alternative sensing domains, such as Doppler-based HAR systems [30].

Existing CSI protection and AML frameworks, however, typically operate in the digital domain or under generalized L_p or average power bounds without enforcing standard-compliant physical-layer constraints. Our proposed methodology explicitly addresses this limitation by embedding modulation-dependent EVM boundaries directly into the continuous optimization framework, highlighting a physically realizable transmitter-side defense tailored for 5G NR infrastructure.

III. SYSTEM MODEL AND THREAT ASSESSMENT

We consider a two-room office environment separated by a wall and connected by a door, as illustrated in Fig. 2. The system operates at 28 GHz with a bandwidth of 200 MHz, numerology $\mu = 3$, and subcarrier spacing $\Delta f = 120$ kHz. Coordinates are denoted as $[x; y; z]$, where x is the horizontal axis, y the depth, and z the height. The gNB is installed in the first room at $[3; 4; 2]$ m. The RIS is deployed at $[9; 3; 1]$ m in line of sight (LOS) with the gNB through the door, and is configured to redirect the incident signal toward the shadowed second room, where human activities take place and legitimate UEs are located.

A sensing node “UE A” is placed in the second room under NLOS conditions with respect to the gNB. This node models a malicious observer such as a smart-home or Internet of Things (IoT) device, that passively exploits the communication infrastructure without requiring direct visibility of the transmitter.

We define the receiver sensitivity as the minimum detectable signal (MDS) level, i.e., the point at which the received signal power equals the noise floor (SNR = 0 dB). The noise floor is

$$N_0 = 10 \log_{10} \left(\frac{k_B T_0 B}{1 \text{ mW}} \right) + F, \quad (2)$$

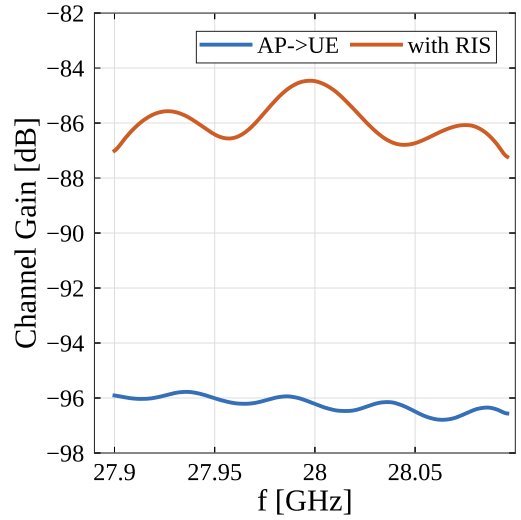


FIGURE 3. Channel gain at the sensing node with and without RIS.

with Boltzmann’s constant k_B , reference temperature $T_0 = 290$ K, the $B = 200$ MHz system bandwidth introduced above, and a receiver noise figure $F = 6$ dB, matching the measured noise figure reported for a 24 GHz CMOS receiver front-end implemented in a mature, low-cost $0.18 \mu\text{m}$ process [31]. This value lies toward the higher end of the 1.7–8.7 dB range reported across CMOS mmWave receiver front-ends operating in the 21–29 GHz band [32], consistent with a low-cost, non-optimized front-end representative of a passive IoT/smart-home sensing device such as UE A, rather than a state-of-the-art high-end receiver. This yields sensitivity of -84.97 dBm, used as the threshold for a usable link.

In the absence of the RIS, the received signal at the sensing node is negligible and the measured CSI is noise-dominated, making ML-based HAR infeasible (quantified in Section III-A). When the RIS is active, a controllable reflected path is introduced, producing CSI with structured multipath components modulated by human motion, as shown in Fig. 3. The RIS therefore creates the physical conditions for passive sensing as a direct side effect of coverage extension.

A. RIS AND PROPAGATION MODELING

The RIS is modeled at element granularity as an $N_r \times N_c = 30 \times 30$ array of unit cells with inter-element spacing $\lambda/2$. Each cell m has a cosine element radiation pattern and a frequency-selective reflection coefficient $\Gamma_m(f) = e^{j\phi_m(f)}$, with ϕ_m restricted to M quantized phase levels. The end-to-end channel frequency response is the coherent sum of the per-element cascaded contributions

$$H(f) = \sum_{m=1}^{N_r N_c} a_m \Gamma_m(f) e^{-j2\pi f(\tau_m^{\text{in}} + \tau_m^{\text{out}})}, \quad (3)$$

where a_m collects the path loss and element-pattern gain, and $\tau_m^{\text{in}}, \tau_m^{\text{out}}$ are the gNB–element and element–UE propagation delays. At 28 GHz, the RIS aperture diagonal is $D \approx 0.22$ m,

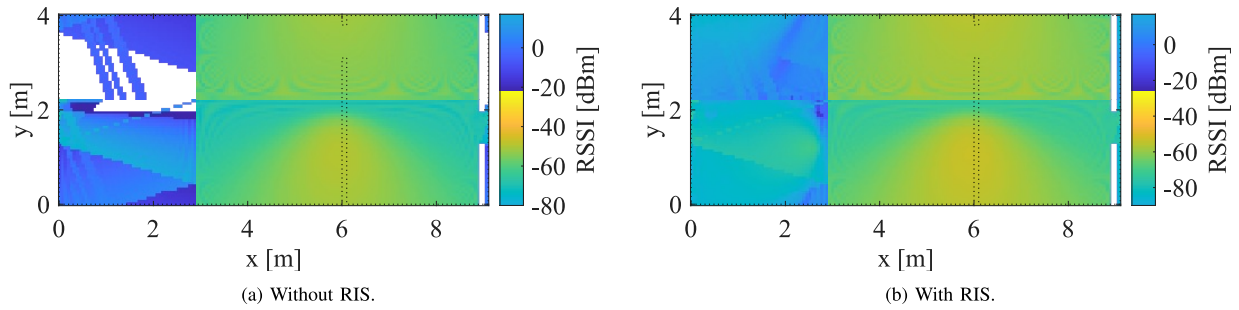


FIGURE 4. Simulated RSSI coverage map of the considered indoor scenario at 28 GHz, without (a) and with (b) the RIS active. The dashed line marks the wall/doorway separating the two rooms. Without the RIS, Room 2 is almost entirely below the receiver sensitivity of -84.97 dBm; with the RIS active, coverage is restored across most of the room.

giving a Fraunhofer distance $d_F = 2D^2/\lambda \approx 9$ m. Since the gNB–RIS and RIS–UE distances in the considered scenario (≈ 6.2 m and ≈ 2 –4 m) are smaller than d_F , both link ends lie within the radiating near-field region of the RIS, where the wavefront is spherical across the aperture. The gNB array ($d_F \approx 0.5$ m) and the single-antenna UE instead operate in the far field.

To assess the coverage conditions underlying this scenario, Fig. 4 reports the simulated RSSI over the full floor plan, with and without the RIS. Without the RIS (Fig. 4a), the small room lies almost entirely below the sensitivity threshold: only a narrow diffraction cone near the doorway reaches usable levels, while the rest of the room remains substantially below -60 dBm, confirming that the received signal is noise-dominated across the sensing area. With the RIS active (Fig. 4b), the reflected path raises the RSSI across most of Room 2 to approximately -20 to -40 dBm, well above the sensitivity threshold, extending usable coverage across the shadowed room rather than only at the specific sensing position considered in the remainder of the paper. This improvement follows directly from the propagation model introduced above. In the absence of the RIS, the only path connecting the gNB to the small room is a heavily obstructed NLOS path, dominated by wall-penetration loss and diffraction around the doorway, which is not captured by the coherent LOS-based model of Eq. 3. When the RIS is active, this model applies: the gNB–RIS and RIS–UE segments are each in LOS, so the end-to-end path loss reduces to two near-field segments combined coherently across the 30×30 RIS elements, bypassing the wall-penetration loss entirely. This substitution of an obstructed single-hop path with a two-hop LOS-LOS path is what produces the coverage improvement observed in Fig. 4 and the channel gain reported at UE A in Fig. 3.

B. EAVESDROPPER MODEL

We consider an eavesdropper $\mathcal{E}$ that controls a CSI-based observer located within the RIS-enhanced coverage region. By collecting channel measurements over time, $\mathcal{E}$ trains a HAR classifier that exploits activity-induced channel variations. Given an observed CSI tensor $\mathbf{x}$, the predicted activity is

$$\hat{y}(\mathbf{x}) = \arg \max_k f_k(\mathbf{x}), \quad (4)$$

where $f_k(\mathbf{x})$ is the classifier output for class k . The objective of $\mathcal{E}$ is to maximize recognition accuracy on activities performed in the monitored area. The feasibility of this threat follows from the coverage results: the RIS turns a communication enhancement mechanism into an enabler of covert sensing, which motivates physical-layer countermeasures acting directly on the intercepted CSI.

We model $\mathcal{E}$ as a physical layer (PHY)-aware passive eavesdropper. The eavesdropper knows the downlink transmission configuration needed to synchronize to the received signal and form the CSI tensors used for HAR, including the carrier frequency, bandwidth, numerology, reference-signal/CSI resource allocation, and active modulation scheme. It remains passive: it only observes the transmitted waveform and does not transmit, jam, alter the RIS, or modify the wireless channel. We further assume the eavesdropper is non-adaptive: its HAR classifier is trained on unperturbed CSI and is not retrained on, or otherwise adapted to, perturbed observations. Adaptive eavesdroppers that retrain on perturbed CSI, or that adopt alternative feature pipelines, are outside the scope of this work and are discussed in Sec. XII and Appendix A. We also assume ideal CSI acquisition, i.e., channel coefficients without estimation noise, synchronization error, or receiver hardware impairments. This gives the adversary the strongest observation quality considered in our model. If the above PHY-layer parameters were unknown or imperfectly recovered, synchronization and alignment errors would reduce the reliability of the CSI input used for HAR. The reported results therefore correspond to a conservative attacker model. Since the perturbation is embedded in the transmitted signal and, as detailed in Section X, is also applied to the reference signals, the eavesdropper’s channel estimate already contains the adversarial perturbation. The CSI used in our evaluation is ray-traced and noise-free; over-the-air acquisition effects are left to the mmWave-testbed validation discussed in Section XII.

IV. DEEP LEARNING-BASED HUMAN ACTIVITY RECOGNITION

The fundamental premise of device-free HAR in RIS-assisted networks is that the presence and movement of a human body within the propagation environment induce specific, reproducible variations in the CSI. In this section, we detail

the transformation of raw CSI measurements into a structured spatio-temporal representation and describe the CNN architecture employed by the adversary to classify human activities.

A. CSI DATA REPRESENTATION

In the considered network configuration, the total number of available subcarriers is $S_{BW} = N_{RB} \times 12$, where $N_{RB} = 132$ is the number of resource blocks in the bandwidth [12]. A subset is allocated to the CSI according to the configured density: $S_{CSI,1} = S_{BW}/12 = 132$ subcarriers for Density 1, and $S_{CSI,3} = S_{BW}/4 = 396$ subcarriers for Density 3. To reduce processing complexity at the receiver, we downsample the reference signals and retain $S = 32$ subcarriers uniformly spaced across the bandwidth, which preserves a reliable channel estimate while lowering the computational cost.

For each subcarrier, the raw data include four signal quantities: the real and imaginary part ($\Re\{H\}$ and $\Im\{H\}$), the amplitude and the phase ($|H|$ and $\angle H$) of the complex channel coefficient. We retain only $\Re\{H\}$ and $\Im\{H\}$, which fully characterize the complex baseband signal.

This yields $2 \times 32 = 64$ real-valued features per time step.

Each CSI stream is processed in windows of $T = 10$ samples, mapped to a tensor $\mathbf{X} \in \mathbb{R}^{C \times T \times S}$, where C corresponds to the feature channels, T is the number of consecutive time steps, and S indexes the subcarriers. The entry $\mathbf{X}_{c,t,s}$ represents the c -th component at time index $t \in \{1, \dots, T\}$ and subcarrier index $s \in \{1, \dots, S\}$, preserving the joint temporal and frequency-domain structure of the CSI.

To ensure numerical stability, tensors are normalized by a global scaling factor

$$M_{\max} = \max_{n,t,s} M_{t,s}^{(n)} \quad (5)$$

where $n \in \{1, \dots, N\}$ indexes the training samples,

$$M_{t,s} = \sqrt{\Re\{H\}_{t,s}^2 + \Im\{H\}_{t,s}^2} \quad (6)$$

and the maximum is taken over all training samples.

Each tensor is then normalized as $\mathbf{X}' = \mathbf{X}/M_{\max}$. Computing $M_{\max}$ exclusively on the training set prevents information leakage, and applying a single global factor to both components preserves the complex signal geometry while enforcing a consistent scale across the dataset.

The dataset is generated by ray tracing a sequence of static indoor scenes, each corresponding to one posture of a subject performing the activity. Each simulation run produces $T = 10$ consecutive scene geometries, from which the corresponding channel frequency responses are computed and assembled into a single CSI window. Every run is independent of the others and contributes exactly one window to the dataset, so that no two samples share a common motion sequence.

The dataset comprises 2,997 activity samples partitioned into training (2,097), validation (450), and test (450) subsets using a stratified split to preserve class balance. Each sample belongs to one of three activity classes $\mathcal{A} = \{A_0, A_1, A_2\}$, where A_0 denotes *fall*, A_1 denotes *sit*, and A_2 denotes *walk*.

TABLE 3. HAR classification performance of VGG-11-BN on the held-out test set.

Class	Precision	Recall	F1-score
A_0 (<i>fall</i>)	1.00	1.00	1.00
A_1 (<i>sit</i>)	0.94	0.99	0.97
A_2 (<i>walk</i>)	0.99	0.94	0.97
Macro avg	0.98	0.98	0.98
Weighted avg	0.98	0.98	0.98

B. HAR CLASSIFICATION MODEL: VGG-11-BN BASELINE

To model a high-capability eavesdropper $\mathcal{E}$, we adopt a VGG-11-inspired CNN with batch normalization, adapted to the shape of the CSI tensors. The backbone follows the VGG design [33], with stacked 3×3 convolutions, but with three modifications. First, ReLU activations are replaced by LeakyReLU to mitigate dead-neuron effects on CSI inputs. Second, dropout with probability 0.4 is applied before the classifier to regularize training on the relatively small dataset. Third, the pooling schedule is shortened: since the input has a short temporal dimension $T = 10$, the original VGG-11 schedule would collapse the time axis, so we retain only the first three max-pooling operations.

The resulting feature maps are aggregated through adaptive average pooling into a fixed-size embedding $\mathbf{z} \in \mathbb{R}^{512}$, followed by a linear classifier

$$\mathbf{o} = \mathbf{W}\mathbf{z} + \mathbf{b}, \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{3 \times 512}$ and $\mathbf{b} \in \mathbb{R}^3$ are learned parameters, and the output logits are mapped to class probabilities through a softmax.

C. TRAINING AND EVALUATION

The model is trained for up to 200 epochs with a batch size of 32, using the Adam optimizer (learning rate 10^{-3} , weight decay 10^{-4}) and cross-entropy loss. A ReduceLROnPlateau scheduler reduces the learning rate by a factor of 0.1 after five epochs without validation loss improvement. Model checkpointing retains the configuration with minimum validation loss. Reproducibility is ensured by fixing random seeds throughout the training pipeline.

As reported in Table 3, the model attains an overall accuracy of 97.78% and a macro F1-score of 0.98. The *fall* class (A_0) is recognized with perfect precision and recall, while minor confusion remains between *sit* (A_1) and *walk* (A_2). These results confirm that a high-capability eavesdropper can reliably perform HAR from the intercepted CSI, and define the baseline against which the proposed defense is evaluated.

V. ADVERSARIAL DEFENSE FORMULATION

In this work, AML is used as a **defensive** mechanism against unauthorized CSI-based sensing. The defender exploits the sensitivity of the eavesdropper's HAR classifier to adversarial perturbations, altering the transmitted signal so that it

remains compatible with communication requirements while becoming less reliable for activity inference.

The defense operates in the RF domain. Rather than applying additive perturbations to the input tensor, we perturb the CSI through multiplicative amplitude scaling and phase rotation. These transformations are physically meaningful for wireless signals and act directly on the complex channel representation. To preserve communication quality, the perturbation is bounded by the modulation-dependent EVM limits defined in the 5G NR specifications discussed in Section VI-A.

A. CARLINI-WAGNER ATTACK

The C&W attack is a widely used optimization-based method for generating adversarial examples [15]. Instead of applying a single gradient step or a fixed perturbation rule, it formulates adversarial example generation as a continuous optimization problem that trades off perturbation size and attack success:

$$\min_{\delta} \|\delta\|_p + c g(\mathbf{x} + \delta), \quad (8)$$

where δ is the adversarial perturbation, $\|\delta\|_p$ penalizes its magnitude, $c > 0$ controls the trade-off between distortion and misclassification, and $g(\cdot)$ is a margin-based adversarial loss defined on the classifier logits.

For an untargeted attack, the loss can be written as

$$g_{\text{unt}}(\mathbf{x}') = \max \left(f_y(\mathbf{x}') - \max_{i \neq y} f_i(\mathbf{x}'), -\kappa \right), \quad (9)$$

where $\mathbf{x}' = \mathbf{x} + \delta$, y is the true class, and $\kappa \geq 0$ is a confidence margin. Minimizing (9) reduces the logit of the true class relative to the strongest competing class. The attack succeeds when at least one incorrect class obtains a larger logit than the true class.

For a targeted attack toward class $t \neq y$, the loss is instead commonly defined as

$$g_{\text{tar}}(\mathbf{x}') = \max \left(\max_{i \neq t} f_i(\mathbf{x}') - f_t(\mathbf{x}'), -\kappa \right). \quad (10)$$

In this case, the optimization increases the relative confidence of the target class until its logit exceeds the logits of all other classes by the required margin.

The C&W formulation is useful because it separates the attack objective from the distortion measure and allows different constraints or regularizers to be incorporated into the optimization. Although the original formulation is typically expressed as an additive perturbation bounded by an L_p norm, the same principle can be adapted to other domains by redefining the perturbation model, the distortion metric, and the feasibility constraints according to the structure of the input data.

B. PHYSICALLY CONSTRAINED CARLINI-WAGNER ATTACK

We propose a variant of the C&W attack that operates directly in the physical signal domain. Unlike the standard additive

formulation, we define multiplicative transformations on the complex channel coefficients, preserving the internal consistency of the signal representation and enabling explicit control of the induced EVM.

1) PHYSICAL-DOMAIN PERTURBATION MODEL

Let $\mathbf{X} \in \mathbb{R}^{2 \times T \times S}$ denote the input CSI tensor introduced in Section IV. We define the adversarial perturbation as a multiplicative transformation on the complex CSI coefficients $H_{t,s} = \Re\{H\}_{t,s} + j\Im\{H\}_{t,s}$ as:

$$H'_{t,s} = \alpha_{t,s} e^{j\theta_{t,s}} H_{t,s}, \quad (11)$$

where $\alpha_{t,s} \in \mathbb{R}_+$ and $\theta_{t,s} \in \mathbb{R}$ denote amplitude scaling and phase rotation. To bound the perturbation magnitude, we parameterize:

$$\alpha_{t,s} = 1 + \Delta_\alpha \tanh(u_{t,s}), \quad \theta_{t,s} = \Theta_{\max} \tanh(v_{t,s}), \quad (12)$$

where $u_{t,s}, v_{t,s} \in \mathbb{R}$ are unconstrained optimization variables. For the joint attack, we set $\Delta_\alpha = 0.2$, $\Theta_{\max} = 0.1$ rad, $\kappa = 0$, and $\lambda_{\text{evm}} = 1$. Since $\tanh(\cdot) \in (-1, 1)$, $0.8 < \alpha_{t,s} < 1.2$, ensuring $\alpha_{t,s} > 0$, and $|\theta_{t,s}| < 0.1$ rad. For amplitude-only and phase-only attacks, $\Theta_{\max} = 0$ and $\Delta_\alpha = 0$, respectively.

Operationally, $\alpha_{t,s} e^{j\theta_{t,s}}$ is introduced as a frequency-domain complex multiplier on each allocated resource element, after resource-element mapping and before OFDM modulation in the 3GPP TS 38.211 downlink chain. Because the weight is applied to the full allocation, including the Demodulation Reference Signal (DMRS), the legitimate receiver incorporates it into the channel estimate and equalizes it transparently, as discussed later in Section X. The perturbation therefore only rescales and phase-rotates in-band resource elements, while leaving the allocation and the average transmitted power essentially unchanged. Occupied bandwidth, adjacent-channel leakage, and unwanted-emission requirements are consequently not tightened, with EVM remaining the binding transmitter constraint.

2) EVM-CONSTRAINED ADVERSARIAL OBJECTIVE

To quantify distortion, we replace the standard L_p norm in (8) with an EVM formulation tailored to our perturbation model, which measures the deviation between ideal and perturbed complex symbols and directly reflects baseband signal degradation. Under the multiplicative perturbation model, the per-symbol relative error is

$$d_{t,s} = \alpha_{t,s} e^{j\theta_{t,s}} - 1, \quad (13)$$

whose squared magnitude is

$$|d_{t,s}|^2 = \alpha_{t,s}^2 + 1 - 2\alpha_{t,s} \cos \theta_{t,s}. \quad (14)$$

Assuming unit-average-power symbols, the RMS EVM is

$$\text{EVM}(\alpha, \theta) = \sqrt{\frac{1}{TS} \sum_{t,s} (\alpha_{t,s}^2 + 1 - 2\alpha_{t,s} \cos \theta_{t,s})}. \quad (15)$$

The adversarial perturbation is obtained by solving

TABLE 4. Maximum EVM thresholds defined in 3GPP TS-38.104.

Modulation	max (%)	η_{EVM}
QPSK	17.5	0.175
16-QAM	12.5	0.125
64-QAM	8.0	0.080
256-QAM	3.5	0.035

$$\min_{\alpha, \theta} c g(\tilde{\mathbf{X}}) + \lambda_{\text{evm}} \text{EVM} + \lambda_{\text{vio}} [\text{EVM} - \eta_{\text{EVM}}]_+, \quad (16)$$

$$c, \lambda_{\text{evm}}, \lambda_{\text{vio}} > 0,$$

where $\tilde{\mathbf{X}}$ is the perturbed tensor induced by (11), c is the trade-off parameter between attack effectiveness and distortion as in (8), $g(\cdot)$ is the adversarial margin loss, λ_{evm} is a regularization weight that penalizes overall signal distortion, λ_{vio} is a penalty weight that enforces compliance with the maximum admissible EVM threshold η_{EVM} , and $[\cdot]_+$ denotes the hinge function. The three terms serve distinct roles: $c g(\tilde{\mathbf{X}})$ drives the classifier toward misclassification, $\lambda_{\text{evm}} \text{EVM}$ regularizes the overall distortion budget, and $\lambda_{\text{vio}} [\text{EVM} - \eta_{\text{EVM}}]_+$ acts as a soft constraint that becomes active only when the EVM exceeds the 3GPP admissible threshold.

The formulation applies to both untargeted and targeted settings, differing only in the definition of $g(\cdot)$. Let $\mathbf{z}$ denote the output logits. For ground-truth label y , the untargeted loss is

$$g_{\text{unt}}(\mathbf{z}, y) = \max \left(z_y - \max_{i \neq y} z_i, -\kappa \right), \quad (17)$$

which enforces misclassification. For target label t , the targeted loss is

$$g_{\text{tar}}(\mathbf{z}, t) = \max \left(\max_{i \neq t} z_i - z_t, -\kappa \right), \quad (18)$$

which promotes classification into the target class.

We solve (16) independently for each input sample using gradient-based optimization. Due to the non-convexity of the objective, multiple candidate solutions may be encountered; for each sample, we retain the solution that (i) achieves attack success, (ii) satisfies the EVM constraint, and (iii) minimizes distortion, in that order of priority.

VI. EXPERIMENTAL METHODOLOGY

This section describes the experimental framework used to evaluate the proposed adversarial defense. We define the operating constraints under which perturbations are generated and detail the evaluation protocol adopted in the performance analysis.

A. EVM BUDGETS

We enforce physical plausibility by constraining perturbations through the EVM-based feasibility condition introduced in Section V-B2. The experimental design therefore reduces to selecting appropriate values of the admissible distortion threshold η_{EVM} , which we set according to the 3GPP TS 38.104 specification [12] as reported in Table 4.

We evaluate all four modulation schemes to characterize the attack behavior across distinct distortion regimes, ranging from the permissive budget of QPSK ($\eta_{\text{EVM}} = 0.175$) to the stringent constraint of 256-QAM ($\eta_{\text{EVM}} = 0.035$).

B. ADVERSARIAL ATTACK CONFIGURATIONS

We perform a systematic exploration of the hyperparameter space to assess both effectiveness and compliance with physical constraints, evaluated independently for each EVM constraint. The main optimization parameters, the C&W trade-off parameter c , the learning rate, the number of optimization steps, and the penalty weight λ_{vio} , are varied over the following ranges:

$$\begin{aligned} c &\in \{1, 5, 10, 50\}, \\ \text{lr} &\in \{0.005, 0.01, 0.05\}, \\ \text{steps} &\in \{100, 200, 300\}, \\ \lambda_{\text{vio}} &\in \{10^2, 10^3, 10^4\}. \end{aligned} \quad (19)$$

To isolate the contribution of different physical components, we consider three perturbation modes: joint amplitude and phase (α, θ), amplitude-only (α), and phase-only (θ).

Both untargeted and targeted attacks are evaluated; for the targeted case, all possible target classes excluding the ground-truth label are considered. Attacks are applied only to correctly classified samples, ensuring that reported results reflect adversarial effects rather than baseline classification errors.

VII. SYSTEM CONFIGURATION

A. HARDWARE

All experiments are conducted on a high-performance server running Ubuntu 22.04, equipped with a 120-core Intel Xeon w9-3595X processor, approximately 1 TB of system memory, and four NVIDIA RTX 6000 Ada Generation Graphic Processing Units (GPUs), each with 48 GB of VRAM.

B. SOFTWARE

All components are implemented in Python using the PyTorch deep learning framework, with GPU acceleration enabled via CUDA. The proposed physically constrained C&W attack is implemented by extending the base `Attack` class provided by the `torchattacks` library [34]. Reproducibility is ensured by fixing random seeds across all libraries. Experiments are parallelized across multiple GPUs to efficiently explore the full configuration space. The full implementation is open source and available in our repository [35].

VIII. INSTANCE-SPECIFIC ATTACK: RESULTS AND ANALYSIS

This section presents the experimental results of the instance-specific physically constrained C&W adversarial attack, assessing its effectiveness across different EVM budgets, perturbation modes, and attack configurations. The complete set of optimal hyperparameters and signal-level metrics for

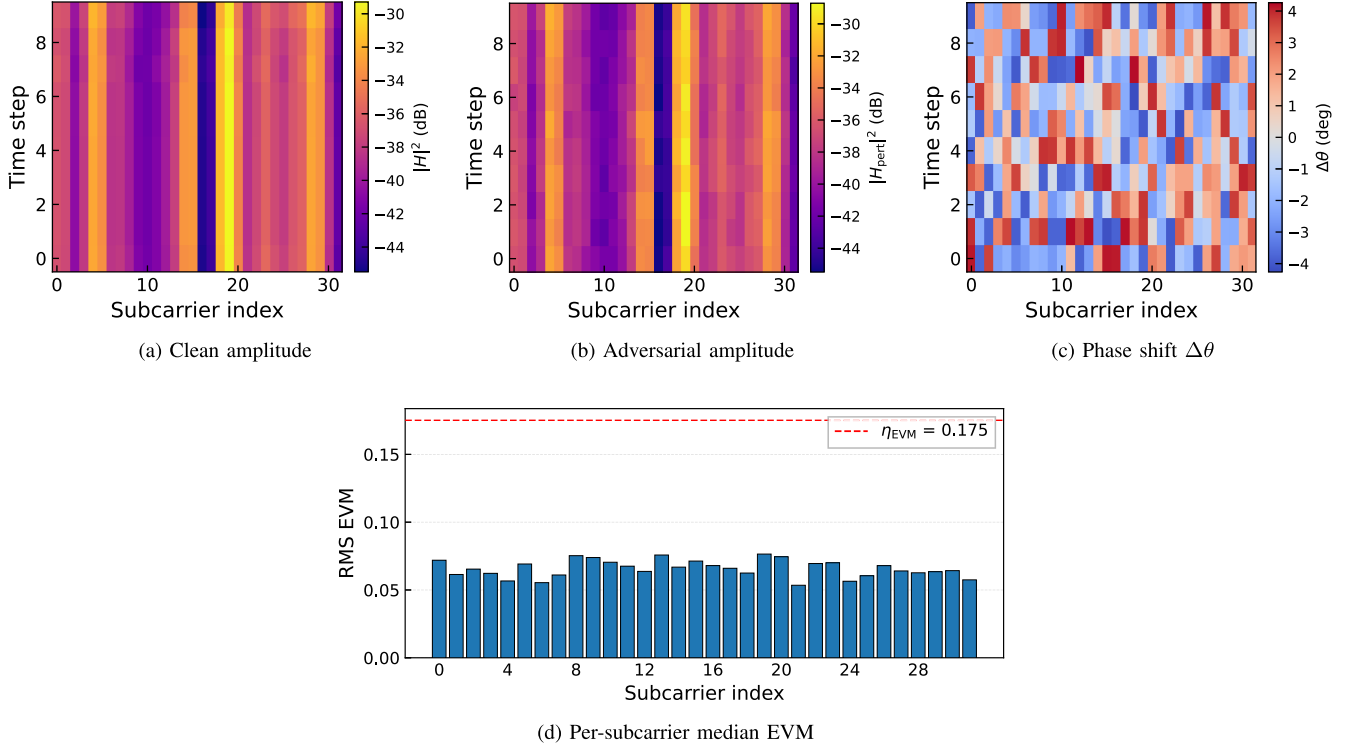


FIGURE 5. Example of a joint (α, θ) targeted adversarial attack under QPSK EVM budget targeting A_0 . The subfigures show clean amplitude, adversarial amplitude, induced phase shift, and per-subcarrier median EVM.

every combination of modulation scheme, perturbation mode, and attack type is reported in Appendix A.

A. METRICS

For each configuration, we report the hyperparameters that maximize the feasible Attack Success Rate (ASR). Starting from the 2997 samples in the dataset, we restrict the evaluation to samples that are correctly classified by the eavesdropper's HAR classifier before the attack. This results in 2958 evaluated samples for the untargeted attack. For targeted attacks, samples whose ground-truth label already corresponds to the target activity are skipped, yielding 1959, 1959, and 1998 evaluated samples for target activities A_0 , A_1 , and A_2 , respectively.

An adversarial attack is counted as successful only if it satisfies the EVM constraint. In the untargeted setting, success occurs when the eavesdropper's HAR classifier mispredicts the ongoing activity. In the targeted setting, success additionally requires the predicted label to match the chosen target class.

We also report the average EVM of successful and feasible attacks, quantifying the fraction of the available distortion budget that is utilized.

1) PHYSICAL METRICS

To assess whether adversarial perturbations preserve the physical structure of the original CSI, we report three

complementary metrics. The relative power variation is

$$\Delta P = 10 \log_{10} \left(\frac{P_{\text{adv}}}{P_{\text{clean}}} \right), \quad (20)$$

where P_{adv} and P_{clean} denote the received average power of the perturbed and clean signals, respectively, computed as

$$P = \frac{1}{TS} \sum_{t,s} (\Re\{H\}_{t,s}^2 + \Im\{H\}_{t,s}^2). \quad (21)$$

The mean absolute phase deviation is

$$|\Delta\theta| = \frac{1}{TS} \sum_{t,s} |\text{wrap}(\theta_{t,s}^{\text{adv}} - \theta_{t,s}^{\text{clean}})|, \quad (22)$$

where $\theta_{t,s} = \text{atan2}(\Im\{H\}_{t,s}, \Re\{H\}_{t,s})$ and $\text{wrap}(\cdot)$ maps phase differences to $[-\pi, \pi]$. Finally, the Pearson correlation coefficient between clean and adversarial amplitude profiles is

$$\rho = \frac{\langle \mathbf{a}_{\text{clean}} - \bar{\mathbf{a}}_{\text{clean}}, \mathbf{a}_{\text{adv}} - \bar{\mathbf{a}}_{\text{adv}} \rangle}{\|\mathbf{a}_{\text{clean}} - \bar{\mathbf{a}}_{\text{clean}}\| \|\mathbf{a}_{\text{adv}} - \bar{\mathbf{a}}_{\text{adv}}\|}, \quad (23)$$

where $\mathbf{a} = \text{vec}(\sqrt{\Re\{H\}^2 + \Im\{H\}^2})$ denotes the flattened amplitude profile. Together, these metrics provide a compact characterization of how perturbations affect the underlying signal while retaining physically meaningful structure.

B. QUALITATIVE ANALYSIS OF ADVERSARIAL PERTURBATIONS

Figure 5 shows one representative joint (α, θ) targeted perturbation generated under the QPSK EVM budget, with target

class A_0 . The figure reports the $S = 32$ subcarriers used by the classifier, following the CSI representation described in Section IV-A.

The clean and adversarial amplitude maps in Fig. 5a and Fig. 5b retain the same dominant time-frequency structure. The main attenuation and enhancement regions across subcarriers remain visible after perturbation, with no evident broadband noise pattern in the adversarial sample. The optimization therefore acts through small, structured changes of the CSI profile while preserving the main channel morphology.

The phase-shift map in Fig. 5c shows a non-uniform rotation over the time-frequency grid. The induced phase shifts vary across both time steps and subcarriers, with values confined to approximately $\pm 4^\circ$. This behavior is consistent with a localized phase perturbation coupled with the amplitude scaling, rather than a global rotation of the whole CSI tensor.

Figure 5d reports the per-subcarrier median RMS EVM induced by the perturbation. All subcarriers remain below the QPSK threshold $\eta_{\text{EVM}} = 0.175$, with median values mostly between 0.06 and 0.08. The distortion is spread over the frequency axis and does not concentrate on a small set of subcarriers. For this sample, the attack satisfies the physical feasibility constraint while using only part of the available EVM budget.

C. ADVERSARIAL ATTACK PERFORMANCE ANALYSIS

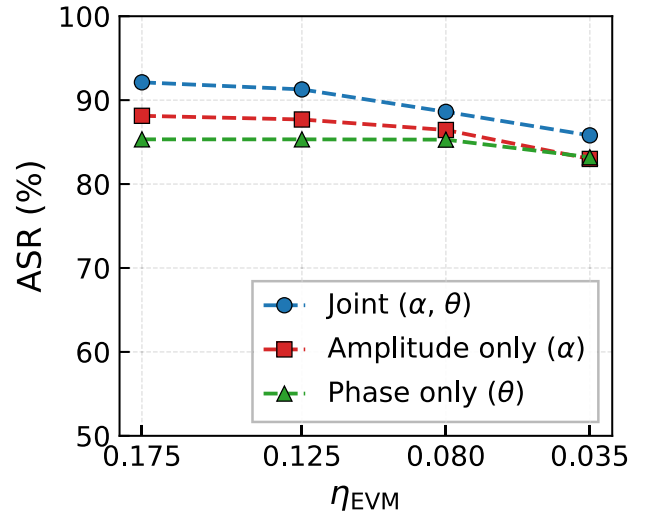
Figure 6 reports the feasible ASR of the proposed defense against the eavesdropper's HAR classifier, for the best grid-search configuration per modulation scheme and perturbation mode. Fig. 6a refers to the untargeted setting, where the ASR is averaged over all perturbed samples. Fig. 6b refers to the targeted setting, where the ASR is additionally averaged over all three target classes. Detailed per-configuration results, including hyperparameters and signal-level metrics, are reported in Appendix A.

In both figures, the feasible ASR decreases as the EVM budget tightens from QPSK to 256-QAM, consistently across all perturbation modes. The joint mode (α, θ) achieves the highest ASR for every modulation scheme, while the three curves converge as the budget becomes more restrictive, with the gap between modes being most pronounced under QPSK and nearly negligible under 256-QAM.

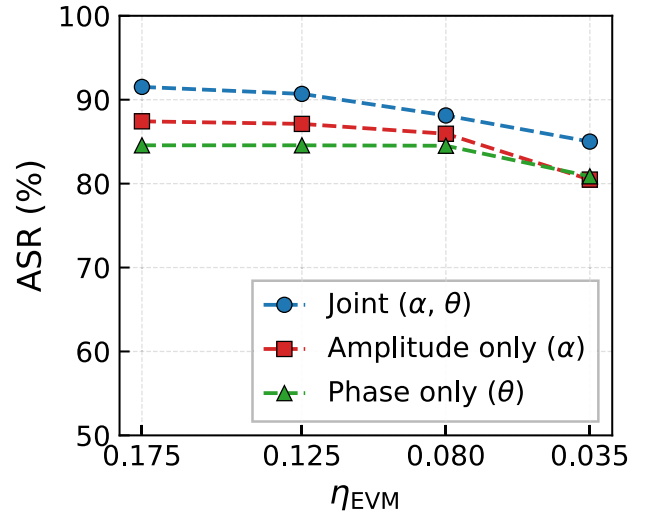
Fig. 7 reports the average ΔEVM of feasible successful perturbations generated by the joint adversarial attack across modulation schemes and attack targets. In all configurations, the consumed EVM lies well below the budget η_{EVM} (dashed), confirming that the constraint is not saturated; notably, targeted attacks toward A_0 exhibit the highest EVM consumption, while A_2 perturbations remain consistently close to zero regardless of modulation.

IX. UNIVERSAL ADVERSARIAL PERTURBATION

Building on the physical perturbation model and the EVM-constrained objective introduced in Section V-B, we extend



(a) Feasible ASR vs. EVM budget for the untargeted attack across perturbation modes.



(b) Average feasible ASR vs. EVM budget for the targeted attack, averaged over all target classes.

FIGURE 6. Impact of the EVM budget η_{EVM} and perturbation mode on feasible ASR for untargeted and targeted attacks.

the adversarial attack for our defense to the universal setting. The instance-specific formulation serves to characterize the perturbation model and to establish an upper bound on the achievable obfuscation, but it is not intended for deployment: solving a separate optimization problem for each transmitted CSI sample is not compatible with the latency budgets of Ultra-Reliable Low-Latency Communication (URLLC) services in 5G and 6G deployments [36]. To address this, we craft a UAP by learning a single pair of parameters (α_u, θ_u) shared across all inputs, computed once offline and applied at line rate as a fixed per-resource-element mask. Its online cost is therefore one complex multiplication per resource element, independent of the payload, and no per-packet optimization is performed at transmission time. The resulting perturbation is

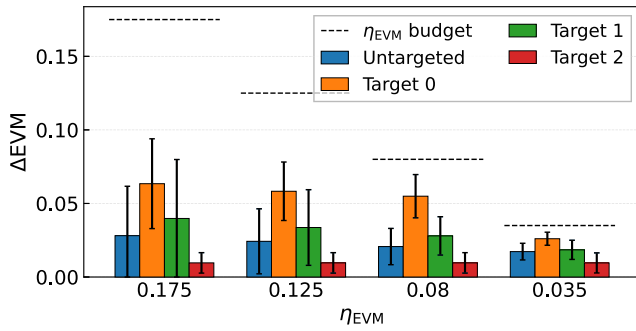


FIGURE 7. Average EVM of feasible successful perturbations generated by the joint attack (α, θ) for untargeted and all targeted classes, across modulation schemes. Dashed lines indicate the EVM budget η_{EVM} for each scheme. Error bars denote one standard deviation.

input-independent and must generalize across the distribution of CSI tensors observed by the HAR classifier.

A. UAP FORMULATION

The UAP is obtained by minimizing the expected margin loss over the dataset $\mathcal{D}$:

$$\begin{aligned} \min_{\alpha_u, \theta_u} \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} [g(\tilde{\mathbf{X}})] \\ + \lambda_{\text{evm}} \text{EVM}(\alpha_u, \theta_u) \\ + \lambda_{\text{vio}} [\text{EVM}(\alpha_u, \theta_u) - \eta_{\text{EVM}}]_+, \end{aligned} \quad (24)$$

where the notation follows Section V-V-B. Since (α_u, θ_u) is shared across all samples, the EVM term is independent of the input tensor $\mathbf{X}$. This preserves the same physical feasibility criterion used in the per-sample formulation in (15), while making the optimization more restrictive because a single perturbation must remain effective over multiple CSI realizations.

For targeted attacks, the margin function is set to $g(\cdot) = g_{\text{tar}}(\cdot)$ to promote prediction of a fixed target class y^* . Samples whose ground-truth label is already y^* are excluded from the evaluation, so that the reported ASR measures only genuine adversarial transfers toward the target class. For untargeted attacks, we use $g(\cdot) = g_{\text{unt}}(\cdot)$ without further modification. As in the per-sample analysis, we evaluate the three perturbation modes defined in Section VI: joint amplitude-phase perturbation (α, θ) , amplitude-only perturbation (α) , and phase-only perturbation (θ) .

The UAP is trained for 100 epochs using mini-batches of 128 samples and 3 random restarts for each configuration. The hyperparameters are selected through the same grid search (19) in Sect. VI-VI-B.

The admissible distortion threshold η_{EVM} is set according to the modulation-specific limits reported in Table 4. This search is performed independently for each modulation scheme, perturbation mode, and adversarial attack setting.

Among the explored candidates, we first discard all solutions that violate the feasibility condition $\text{EVM}(\alpha_u, \theta_u) \leq \eta_{\text{EVM}}$. Within the feasible set, we select the perturbation

that maximizes validation ASR. The selected perturbation is then evaluated on a held-out test set. The difference between validation and test ASR is reported as an additional diagnostic of the extent to which the UAP transfers beyond the samples used for model selection.

B. UAP ADVERSARIAL ATTACK EFFECTIVENESS

Figs. 8a and 8b report the ASR of the UAP across the four EVM budgets and the three perturbation modes, for the untargeted and targeted settings respectively. Targeted results are averaged over all target classes excluding the ground-truth label of each sample.

Across both settings, the joint (α, θ) mode attains the highest ASR at every EVM budget, while the amplitude-only and phase-only modes yield comparable but lower performance. The gap between the joint mode and the single-component modes narrows as η_{EVM} decreases, in line with the trend already observed in the instance-specific analysis of Sect. VIII.

In the untargeted setting (Fig. 8a), the joint mode achieves an ASR of approximately 83% at $\eta_{\text{EVM}} = 0.175$ (QPSK) and decreases monotonically to 72% at $\eta_{\text{EVM}} = 0.035$ (256-QAM). The amplitude- and phase-only modes fall in the 64%–78% range across modulation schemes. In the targeted setting (Fig. 8b), the joint mode reaches 88% ASR at $\eta_{\text{EVM}} = 0.175$ and remains above 82% at $\eta_{\text{EVM}} = 0.035$.

Comparing the two settings, the targeted UAP exceeds the untargeted one across all configurations.

C. SOURCE-TARGET CLASS ASYMMETRY

Figure 9 reports the targeted ASR as heatmaps at source-target class granularity, averaged over all modulation schemes. Each cell (i, j) is the fraction of samples whose true activity is class i for which the defender forces the eavesdropper to predict class j ; diagonal entries are omitted, since samples already belonging to the target class are excluded.

The dominant pattern is the accessibility of target A_2 : transitions from A_0 and A_1 reach 95.3% across all perturbation modes, consistent with the aggregate results in Fig. 8b. The stability across modes points to the learned class geometry rather than to a specific perturbation component as the source of this asymmetry.

The joint (α, θ) mode achieves the highest ASR for all non-diagonal pairs (Fig. 9c). Its advantage is most visible on transitions toward A_0 and A_1 , where it reaches 80.3%–82.8%, against 72.6%–79.7% for amplitude-only (Fig. 9a) and 78.2%–78.8% for phase-only (Fig. 9c).

The heatmaps also expose a moderate asymmetry between class pairs: in the joint mode, the transition from A_0 to A_1 reaches 82.8%, while the reverse reaches 80.3%. This reflects a directional structure in the learned CSI feature space, where some class regions are more easily reached by an EVM-constrained UAP than others.

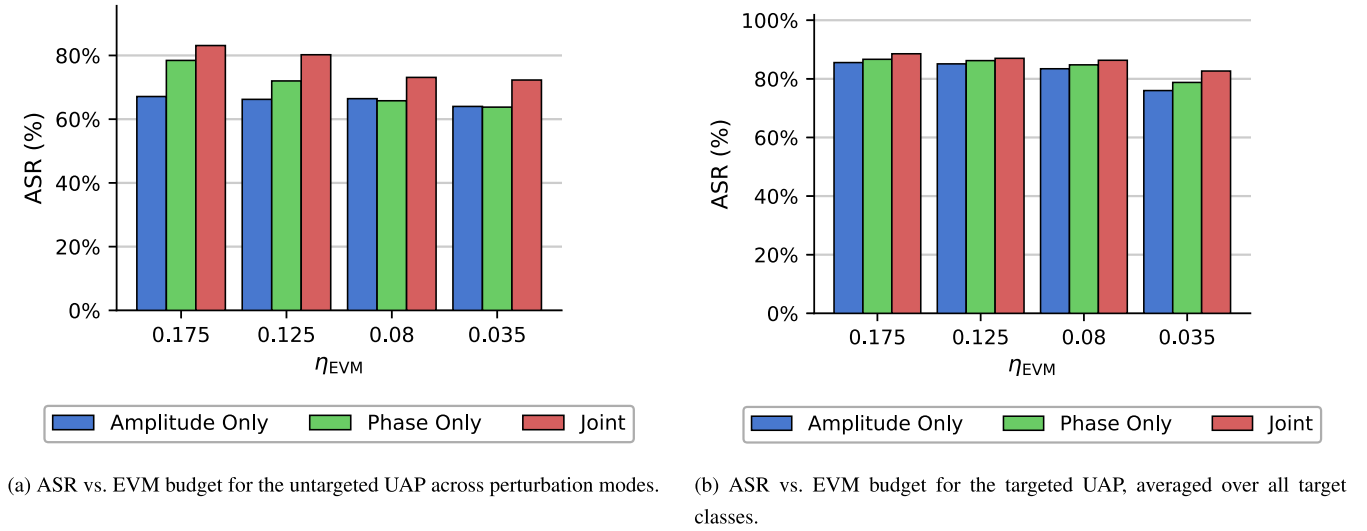


FIGURE 8. ASR of the UAP on the held-out test set as a function of the EVM budget η_{EVM} and perturbation mode for untargeted and targeted attacks.

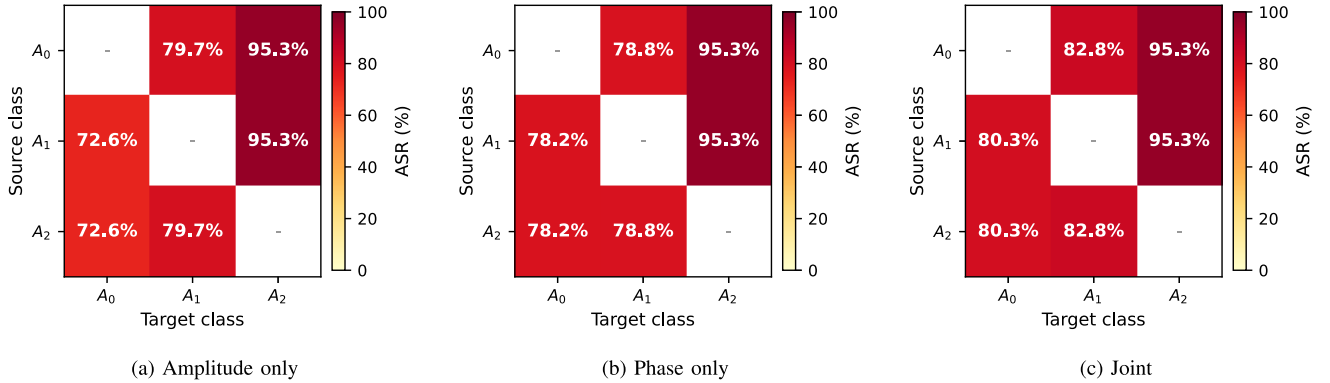


FIGURE 9. Targeted test ASR averaged over all modulation schemes for each attack mode. Rows indicate the true source class and columns the intended target class. Diagonal entries are undefined. The joint attack in panel (c) achieves the highest ASR across all source-target pairs.

D. PHYSICAL DISTORTION ANALYSIS

Figure 10 reports the physical distortion signature of the learned UAPs across modulation schemes and perturbation modes. Each subfigure corresponds to one modulation order, from QPSK to 256-QAM. The four axes of each radar plot are the mean power variation Mean [dB], the mean phase shift, the maximum phase shift, and the complementary Pearson correlation ($1 - \rho$), which captures the structural deviation of the amplitude profile: $1 - \rho = 0$ indicates perfect preservation, $1 - \rho = 1$ complete decorrelation. Axes are normalized to $[0, 1]$ over the global range across all modulations, so values are directly comparable across subplots and metrics. Larger polygons therefore correspond to greater overall physical distortion.

The distortion footprint shrinks as the modulation order increases. The polygons are larger under QPSK and 16-QAM (Figs. 10a and 10b), where the admissible EVM budget is more permissive, and become more compact under 64-QAM and 256-QAM (Figs. 10c and 10d). The trend

is consistent with the modulation-dependent constraints in Table 4 (i.e., tighter EVM thresholds limit the admissible range of both amplitude and phase transformations) and with the ASR curves in Figs. 8a and 8b, where stricter modulations reduce the effectiveness of the UAP without eliminating it.

The three perturbation modes leave distinct physical signatures. Amplitude-only perturbations concentrate on the mean-power axis with limited contribution on the phase-shift axes. Phase-only perturbations show larger values on the mean and maximum phase-shift axes and a more visible contribution on $(1 - \rho)$. The distinction follows directly from the perturbation parameterization, but it carries an operational implication: changes that preserve the power profile can still modify the complex CSI structure in ways that affect the CNN decision.

The joint (α, θ) mode combines power-domain and phase-domain modifications, but does not uniformly maximize all distortion metrics. Its radar profile is more balanced than the



FIGURE 10. Physical distortion signature of the three UAP attack modes across modulation schemes. Each axis represents a normalised distortion metric: mean power change (dB), mean phase shift, maximum phase shift, and complementary Pearson correlation $(1 - \rho)$. All axes are normalised to $[0, 1]$ with respect to the global range across all modulations. Larger polygons indicate greater physical distortion.

single-component modes, reflecting the optimizer's ability to distribute the perturbation across amplitude scaling and phase rotation. Read alongside the higher ASR of the joint mode in Figs. 8a and 8b, this indicates that the gain in adversarial effectiveness comes from the availability of complementary perturbation directions rather than from uniformly larger physical distortion.

The 256-QAM case (Fig. 10d) illustrates the effect of a stringent feasibility constraint. All modes show a compact footprint, particularly on the mean-power and phase-shift axes, yet a residual contribution on $(1 - \rho)$ remains visible. The perturbation still introduces small structural changes in the CSI tensor even when explicit power and phase deviations are tightly bounded, which explains why the UAP retains non-negligible ASR under the most restrictive EVM budget considered.

Overall, adversarial effectiveness is not determined by the aggregate magnitude of physical distortion alone. The joint mode is the most effective, yet its distortion signature is not always the largest along every axis. The location and structure of the perturbation in the complex CSI representation matter alongside the total admissible distortion.

E. COMPARISON WITH MATCHED-EVM RANDOM BASELINES

To assess whether the learned universal perturbation provides an advantage beyond simply injecting distortion at the same power level, we compare the untargeted joint-mode UAP against two fixed, input-independent random baselines. Both baselines are calibrated to match the target EVM budget, η_{EVM} , and are applied unchanged to every test sample.

1) RANDOM AMPLITUDE-PHASE MASK

For the random-mask baseline, the amplitude and phase directions, $\mathbf{U}_\alpha$ and $\mathbf{U}_\theta$, are sampled independently from a uniform distribution,

$$\mathbf{U}_\alpha, \mathbf{U}_\theta \sim \mathcal{U}[-1, 1]. \quad (25)$$

The corresponding amplitude and phase masks are defined as

$$\alpha(s) = 1 + \text{clip}(s\mathbf{U}_\alpha, -\Delta_\alpha, \Delta_\alpha), \quad (26)$$

$$\theta(s) = \text{clip}(s\mathbf{U}_\theta, -\Theta_{\max}, \Theta_{\max}), \quad (27)$$

where the amplitude and phase masks satisfy the same box constraints as the learned UAP. The scale factor s is

TABLE 5. Untargeted joint-mode ASR (%) of the learned UAP compared with universal random baselines under matched EVM constraints. Random baselines show mean $\pm$ standard deviation over 10 independent draws; the learned UAP is deterministic and reported from a single run.

η_{EVM}	UAP	Random mask	Gaussian I/Q
0.175	82.6%	$(64.8 \pm 0.2)\%$	$(64.9 \pm 0.0)\%$
0.125	80.1%	$(61.6 \pm 6.5)\%$	$(64.2 \pm 2.4)\%$
0.080	73.0%	$(54.1 \pm 13.7)\%$	$(50.5 \pm 7.5)\%$
0.035	72.6%	$(33.6 \pm 5.4)\%$	$(31.5 \pm 0.5)\%$

determined via bisection as the largest value satisfying the EVM constraint

$$\sqrt{\frac{1}{N} \sum_{n=1}^N (\alpha_n^2 + 1 - 2\alpha_n \cos \theta_n)} \leq \eta_{\text{EVM}}. \quad (28)$$

2) GAUSSIAN I/Q PERTURBATION

For the additive Gaussian baseline, a universal complex perturbation $\mathbf{Z}$ is sampled with independent standard-normal in-phase and quadrature components,

$$Z_I, Z_Q \sim \mathcal{N}(0, 1). \quad (29)$$

The perturbation is then scaled according to

$$\delta = \eta_{\text{EVM}} \sqrt{\frac{P_{\text{global}}}{\mathbb{E}[Z_I^2 + Z_Q^2]}} \mathbf{Z}, \quad (30)$$

where

$$P_{\text{global}} = \mathbb{E}[I^2 + Q^2] \quad (31)$$

denotes the mean complex-signal power computed over the entire dataset. This normalization ensures that the additive perturbation has an RMS EVM equal to η_{EVM} . The sampled perturbation tensor is fixed and applied identically to every test sample.

3) ATTACK EFFECTIVENESS COMPARISON

Table 5 compares the test ASR of the learned UAP against the two universal random baselines under matched EVM constraints. The learned UAP significantly outperforms both baseline perturbations across all modulation orders. The performance gap is most pronounced under the tightest EVM constraint ($\eta_{\text{EVM}} = 0.035$, corresponding to 256-QAM). Here, the UAP maintains a 72.6% ASR, whereas the random mask and Gaussian baselines degrade to 33.6% and 31.5%, respectively. At the loosest constraint ($\eta_{\text{EVM}} = 0.175$) every Gaussian draw saturates to the same set of predictions and the corresponding standard deviation vanishes, even though the per-draw distortion metrics still differ. This demonstrates that matching the EVM budget with arbitrary noise is insufficient for effective protection; the adversarial gain fundamentally relies on the structurally learned mask.

TABLE 6. Construction and application costs of the proposed UAP and the matched-EVM reference perturbations. Offline costs are averaged over the four EVM budgets of Table 4 with three repetitions each; Model passes denote forward and backward evaluations of the adversary's classifier.

Method	Offline (one-time)		Online (per observation)	
	Time	Model passes	Real ops	Time
Gaussian I/Q	2.8 ms	0	6.4×10^2	$0.7 \mu\text{s}$
Random mask	1.9 ms	0	1.9×10^3	$3.1 \mu\text{s}$
UAP	5.6 s	4.6×10^5	1.9×10^3	$3.1 \mu\text{s}$

4) COMPUTATIONAL COMPLEXITY AND SYSTEM OVERHEAD

A practical physical-layer defense must respect the stringent latency limits of 5G NR specifications (e.g., a 0.5 ms slot duration at a 30 kHz subcarrier spacing) [37]. Table 6 separates the one-time cost of constructing a perturbation from the per-observation cost of applying it, since the two are incurred on entirely different schedules. All measurements are collected using the system setup described in Section VII. Hardware-dependent metrics, such as execution time, should therefore be interpreted with respect to this setup, whereas hardware-independent metrics, such as floating point operation (FLOP) counts, are not affected by the underlying platform.

Once constructed, the UAP is applied as a fixed multiplicative mask and requires neither model inference nor gradient computation. Each complex multiplication amounts to four real multiplications and two real additions, and the rotation terms (α, θ) are precomputed, so no transcendental function is evaluated at run time. Applying the mask to the 320 coefficients of one CSI observation therefore costs 1.9×10^3 real operations and $3.1 \mu\text{s}$, which is 5.6×10^4 times fewer operations than the 1.07×10^8 FLOPs of a single forward pass through the adversary's classifier. This is identical to the random-mask baseline, which applies the same operator; Gaussian I/Q noise is faster at $0.7 \mu\text{s}$, but all three methods remain at microsecond-scale latency. Deployment requires storing only the 320 complex mask weights, or 2.5 KiB: the transmitter neither stores nor evaluates the 36.9 MB victim model at run time.

The offline stage is where the three methods differ. Constructing the UAP for the 100-epoch schedule used in our experiments takes 5.58 ± 0.06 s, consumes 820 J above idle, and uses 160 MiB of GPU memory, performing 2.5×10^5 forward and 2.1×10^5 backward model evaluations for a total of 4.9×10^{13} FLOPs. The reference perturbations require no model evaluation at all: the random mask is drawn and calibrated to the EVM budget by bisection in 1.9 ms, whereas the Gaussian baseline is dominated by a single pass over the reference set to obtain the global signal power used for EVM scaling, the draw itself costing a further 0.04 ms.

Construction cost is independent of the EVM constraint, as the mask dimension is unchanged across operating points: over the four budgets of Table 4 the per-budget means differ

by 1.3%, comparable to the ± 0.06 s run-to-run spread, which is why a single averaged figure is reported rather than a per-modulation breakdown. This cost is incurred once per adversary model, not once per CSI observation: the mask is reused unchanged across subsequent transmissions and is regenerated only when the target classifier changes. An instance-specific perturbation, by contrast, would have to be re-optimized for every observation and would require the CSI as measured by the adversary together with runtime white-box access to its classifier, neither of which is available at the gNB.

X. COMMUNICATION IMPACT OF THE UAP DEFENSE

We assess the cost of the proposed defense to the legitimate link using a 3GPP-based link-level simulation and measure the resulting loss in communication reliability.

A. CHANNEL SYSTEM AND RECEIVER MODEL

Let $s_{t,s}$ denote the transmitted QAM symbol at time index t and subcarrier s , normalized to unit average power. The baseband-equivalent propagation channel between the gNB and the legitimate UE is represented by the complex coefficient $H_{t,s}$, extracted from the same dataset employed in Section VI.

Following the perturbation model introduced in Section IX, the gNB applies the complex weight

$$\rho_{t,s} = \alpha_{u,t,s} e^{j\theta_{u,t,s}}, \quad (32)$$

where $\alpha_{u,t,s}$ and $\theta_{u,t,s}$ denote the amplitude and phase perturbations generated by the UAP, respectively. Since the perturbation is applied to the transmitted waveform, it affects both Physical Downlink Shared Channel (PDSCH) data symbols and reference signals, including the DMRS. The effective channel observed at the receiver is therefore

$$H'_{t,s} = \rho_{t,s} H_{t,s}, \quad (33)$$

and the received signal is

$$y_{t,s} = H'_{t,s} s_{t,s} + n_{t,s}, \quad (34)$$

where $n_{t,s} \sim \mathcal{CN}(0, \sigma^2)$ denotes additive white Gaussian noise.

The PDSCH transmission is configured according to NR DMRS configuration type 1 with mapping type A, transmitted on a single layer, one front-loaded DMRS symbol, and one additional DMRS position. Under this configuration, the UE obtains channel observations from the DMRS resource elements distributed within the scheduled Physical Resource Block (PRB). For the p -th pilot resource element, the received DMRS signal is expressed as

$$y_{t,s,p}^{\text{DMRS}} = \rho_{t,s,p} H_{t,s,p} x_{t,s,p}^{\text{DMRS}} + n_{t,s,p}, \quad (35)$$

where $H_{t,s,p}$ denotes the propagation channel coefficient and $\rho_{t,s,p}$ represents the multiplicative perturbation introduced by the UAP. Since the perturbation is applied to both data and

reference signals, the UE does not need to estimate $\rho_{t,s,p}$ separately and instead estimates the effective channel

$$H'_{t,s,p} = \rho_{t,s,p} H_{t,s,p}. \quad (36)$$

The channel estimator follows the standard least-squares principle by removing the known DMRS sequence from the received pilot observations,

$$\tilde{H}_{t,s,p} = y_{t,s,p}^{\text{DMRS}} (x_{t,s,p}^{\text{DMRS}})^* = H'_{t,s,p} + \varepsilon_{t,s,p}, \quad (37)$$

where $\varepsilon_{t,s,p} \sim \mathcal{CN}(0, \sigma^2)$ denotes the estimation noise contribution. In the considered link-level model, the channel coefficient is represented at PRB granularity; therefore, the available DMRS observations within the same PRB are combined through averaging, yielding

$$\begin{aligned} \hat{H}_{t,s} &= \frac{1}{N_p} \sum_{p=1}^{N_p} \tilde{H}_{t,s,p} \\ &= H'_{t,s} + \varepsilon_{t,s}, \\ \varepsilon_{t,s} &\sim \mathcal{CN}\left(0, \frac{\sigma^2}{N_p}\right). \end{aligned} \quad (38)$$

No additional time-frequency interpolation is required in the considered model. The reason is not that the perturbation makes the physical channel frequency-flat, but rather that the channel representation and perturbation are defined at the same PRB granularity. Consequently, the channel coefficient estimated from the DMRS is directly associated with the corresponding data resource elements. Furthermore, the RIS reflection configuration is kept fixed for at least one slot, preventing additional channel variations between the pilot and data transmissions within the same scheduling interval.

The estimated effective channel is then used for coherent detection through a bias-corrected MMSE equalizer,

$$\begin{aligned} \hat{s}_{t,s} &= \frac{1}{\mu_{t,s}} \frac{\hat{H}_{t,s}^*}{|\hat{H}_{t,s}|^2 + \sigma^2} y_{t,s}, \\ \mu_{t,s} &= \frac{|\hat{H}_{t,s}|^2}{|\hat{H}_{t,s}|^2 + \sigma^2}, \end{aligned} \quad (39)$$

which satisfies $\mathbb{E}[\hat{s}_{t,s}] = s_{t,s}$. The residual effective noise variance is given by

$$\sigma_{\text{eff},t,s}^2 = \frac{\sigma^2}{|\hat{H}_{t,s}|^2}. \quad (40)$$

The receiver has no explicit knowledge of $\rho_{t,s}$. Instead, the perturbation is embedded into the estimated effective channel, and only the residual channel estimation error affects the reliability of the soft information. The resulting log-likelihood ratios are computed from the Euclidean distances between the equalized symbols and the ideal constellation points, de-interleaved, and processed by the LDPC decoder using normalized Min-Sum Belief Propagation to recover the transmitted transport block.

To ensure that the effective channel remains trackable by a standard NR receiver, the optimization constrains $\alpha_{u,t,s} \in$

TABLE 7. SNR at 10% BLER and induced penalty for untargeted joint UAP (95% CI).

Modulation	Metric	Value (dB)
QPSK	Baseline	4.06 [4.05, 4.08]
	With UAP	4.19 [4.17, 4.21]
	Penalty	+0.13 [+0.11, +0.16]
16-QAM	Baseline	9.27 [9.26, 9.29]
	With UAP	9.37 [9.36, 9.38]
	Penalty	+0.10 [+0.08, +0.12]
64-QAM	Baseline	13.81 [13.80, 13.82]
	With UAP	13.84 [13.82, 13.85]
	Penalty	+0.02 [+0.01, +0.04]
256-QAM	Baseline	18.20 [18.19, 18.21]
	With UAP	18.21 [18.19, 18.22]
	Penalty	+0.00 [-0.01, +0.02]

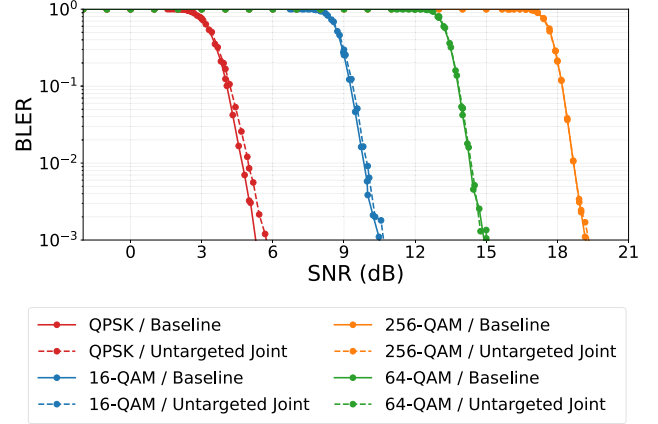
[0.8, 1.2] and $|\theta_{u,t,s}| \leq 0.5$ rad. In addition, a two-dimensional total-variation regularizer with $\lambda_{TV} = 10^{-2}$ penalizes abrupt variations of the perturbation across adjacent PRBs and consecutive slots, limiting rapid amplitude and phase fluctuations. Finally, the perturbation is normalized according to $\mathbb{E}[|\rho|^2] = 1$, preserving the average transmitted power of the baseline system.

B. EVALUATION SETUP AND METRICS

We evaluate PDSCH reliability under QPSK, 16-QAM, 64-QAM, and 256-QAM. Each transport block includes a CRC-24A and follows the interleaving procedure in [38]; a block error is declared upon cyclic redundancy check (CRC) failure. Blocks use independent channel snapshots normalized to unit mean power, so each operates at the nominal SNR. We use a regular rate-1/2 LDPC code rather than an NR base graph, an abstraction applied identically to the baseline and perturbed links.

At each SNR, we simulate until observing 1000 block errors or reaching 20000 blocks. This error-count stopping rule bounds relative BLER uncertainty more uniformly than a fixed block count. Each curve uses 50 SNR points: a 1 dB coarse grid refined to 0.25 dB within ± 2.5 dB of the 10% BLER crossing. The refinement resolves the waterfall and avoids estimating the penalty from interpolation across a single coarse interval.

The primary metric is post-decoding BLER. We define the defense penalty as the additional SNR required by the perturbed link to reach a BLER of 10^{-1} relative to the baseline, as reported in Table 7. Each BLER estimate has a 95% Clopper–Pearson interval. Because the penalty is the difference between two interpolated crossings, its confidence interval is obtained by binomially resampling all measured points and repeating the interpolation over 4000 resamples. The two links use common random numbers (i.e., identical information bits, noise, and channel realizations) which positively correlate their errors and make the reported intervals conservative.


FIGURE 11. BLER across modulation schemes for the baseline channel and the untargeted joint UAP adversarial attack.

C. BLER EVALUATION

Figure 11 and Table 7 show that the perturbed link closely tracks the baseline for every modulation. The curves exhibit only a small rightward shift near the waterfall, which decreases as the modulation order increases and the admissible η_{EVM} tightens. The waterfall shape is preserved, and no error floor appears over the simulated range, indicating no qualitative change in LDPC decoding behavior.

The penalty remains at or below 0.13 dB across all modulation orders and is statistically indistinguishable from zero for 256-QAM. As (36) and (39) show, the UE estimates the effective channel including the multiplicative perturbation and subsequently performs coherent equalization using this estimate. Since the same perturbation affects both DMRS and PDSCH symbols, the receiver does not need explicit knowledge of $\rho_{t,s}$, and the induced channel modification is largely compensated during demodulation. The adversary’s classifier instead operates directly on the channel estimate and remains sensitive to the perturbation. Thus, in the tested configurations, the defense preserves the baseline block-error trend with only a small reliability cost.

XI. ADVERSARIAL DEFENSE TRANSFERABILITY ANALYSIS

The defender does not necessarily know the classifier or the training data used by the eavesdropper $\mathcal{E}$. In this case, the perturbation is optimized on a surrogate HAR model and applied against a different model trained for the same task, a property known as adversarial transferability [17], [18].

In the considered RIS-assisted scenario, the dominant source of mismatch is the receiver position. Each receiver location induces a different CSI distribution because the propagation geometry, multipath structure, and activity-induced channel variations depend on the position of the sensing receiver. Therefore, even when the activity classes and classifier architecture are kept fixed, the models trained at different receiver locations learn receiver-specific decision boundaries.

We therefore evaluate spatial transferability by generating the UAP from our VGG-like reference CNN trained

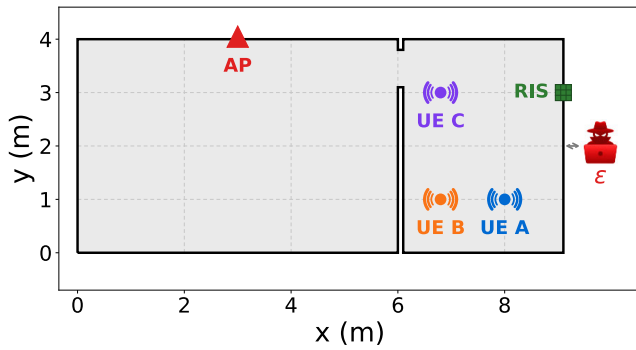


FIGURE 12. Receiver locations used to evaluate spatial transferability. The UAP is optimized on the model trained at UE A and applied to the models trained at UE B and UE C.

on the CSI collected at UE A and applying it to target classifiers trained independently on the CSI collected at different positions. This setup isolates receiver-position-induced distribution shifts from changes in classifier architecture, and captures the effect of different propagation geometries, multipath conditions, receiver-dependent CSI statistics, and the resulting decision boundaries learned from each receiver-specific dataset.

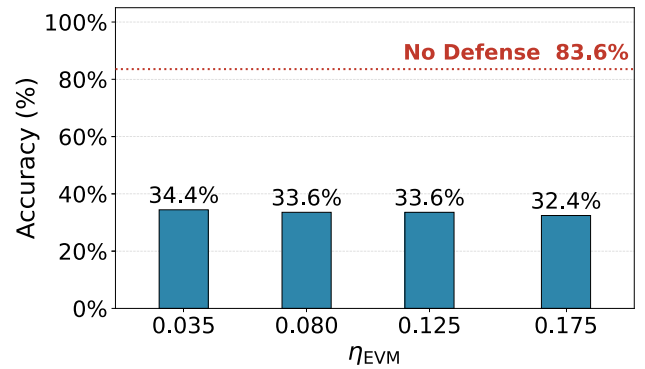
To extract reference positions, a qualitative assessment is conducted by considering three representative UE positions within the region where the RIS enhances coverage (see Fig. 4). The three receiver positions in Fig. 12 instantiate this setting. These positions are uniformly distributed across the area, excluding locations in the immediate vicinity of the RIS.

Among them, UE A yields the highest clean accuracy for the eavesdropper, reaching 97.78%, compared with 83.6% at UE B and 82.7% at UE C. This result can be attributed to the fact that UE A captures a larger fraction of the reflected propagation paths identified through ray tracing, as illustrated in Fig. 4. In addition, the region surrounding UE A is the closest to the attacker and is therefore particularly relevant from a security perspective. This region represents the most vulnerable portion of the considered area and constitutes the most significant location for evaluating the attack. Accordingly, the surrogate model is constructed at UE A, which corresponds to the strongest sensing position among the three considered locations.

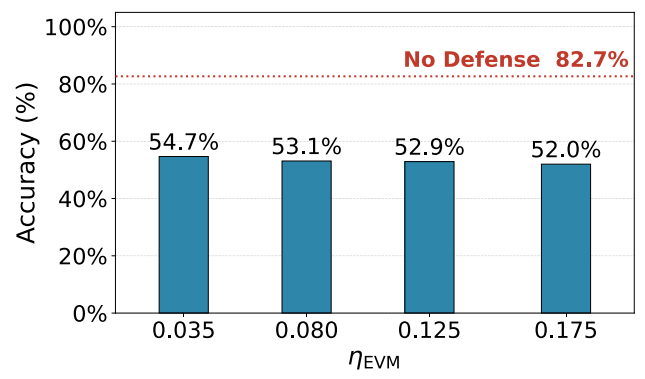
The UAP generated at UE A is then applied without re-optimization to the models trained at UE B and UE C, which represent unknown receiver positions that the eavesdropper may occupy. The weaker degradation observed at UE C compared with UE B is consistent with a larger mismatch between the CSI distribution used to craft the UAP and the distribution observed by the target eavesdropper.

A. UAP TRANSFERABILITY RESULTS

We report HAR accuracy rather than ASR in this section. The goal of the spatial-transfer analysis is to determine whether the defense degrades the eavesdropper's classifier when applied at a different receiver position, and the natural



(a) UE B.



(b) UE C.

FIGURE 13. Transferability of the learned UAP across receiver positions. The dotted line denotes the eavesdropper's HAR accuracy without defense, while the bars report the accuracy obtained when applying a UAP optimized for UE A. Lower values indicate a more effective defense, since the malicious HAR classifier is less accurate. Results are shown for UE B and UE C across all considered EVm budgets.

quantity for this is the resulting classification accuracy: lower accuracy means a more effective defense, regardless of how individual samples are reclassified.

Figure 13 reports the eavesdropper's HAR accuracy when the untargeted joint UAP optimized on UE A is applied to the UE B and UE C classifiers, for all four EVm budgets. At UE B (Fig. 13a), accuracy drops from 83.6% to 32.4%–34.4%, approaching the 33.3% random-guess baseline of a three-class classifier. At UE C (Fig. 13b), the same UAP reduces accuracy from 82.7% to 52.0%–54.7%: still a substantial drop of about 28–30 percentage points, but well above chance, suggesting that the CSI distribution shift between UE A and UE C is larger than between UE A and UE B.

The dependence on the EVm budget is consistent at both positions and in the expected direction: a larger budget yields a slightly more effective defense, in line with the trend observed in Sections VIII–IX. The variation is small, however, around two percentage points at each position, while the gap between UE B and UE C exceeds 18 percentage

points across all budgets. Under spatial transfer, the position-induced distribution shift is therefore the dominant factor, with the admissible distortion playing a secondary role.

An additional cross-model transferability evaluation, where the target classifier architecture differs from the surrogate used to generate the UAP, is provided in Appendix B.

XII. CONCLUSION AND FUTURE WORK

This paper addresses the physical-layer privacy threat introduced when a RIS extends 5G NR coverage into NLOS regions. A passive eavesdropper located within the reflected footprint can perform device-free HAR from the intercepted CSI, while data-centric privacy mechanisms remain ineffective. To mitigate this threat, we propose a transmitter-side adversarial defense that reformulates the C&W attack for the RF domain by replacing additive L_p -bounded perturbations with multiplicative amplitude and phase transformations of the complex channel coefficients. The optimization explicitly incorporates the modulation-dependent EVM limits defined by 3GPP TS 38.104. An offline UAP further removes the per-packet optimization cost, making the defense compatible with URLLC latency constraints.

All results reported below are obtained against non-adaptive eavesdroppers, i.e. classifiers trained on unperturbed CSI. The instance-specific attack achieves an ASR above 90% under QPSK and 16-QAM and remains above 83% under the 3.5% EVM constraint of 256-QAM. The offline UAP preserves most of this effectiveness without online optimization, achieving an ASR between 73% and 83% in the joint mode. The perturbation also transfers across receiver positions, reducing HAR accuracy from approximately 82–84% to 33–53%, depending on the target receiver. Finally, the physical-layer evaluation shows that the perturbation structure, rather than its magnitude alone, determines the attack effectiveness, while the impact on the legitimate communication remains below 1 dB of SNR loss on BLER after LDPC decoding.

The principal direction for future work is robustness against adaptive eavesdroppers that retrain their classifier on perturbed CSI. We note that the defender and the eavesdropper are not symmetric in this respect: regenerating a feasible perturbation is an offline computation performed at the gNB at no online cost, whereas retraining the eavesdropper's classifier requires *labeled* perturbed CSI, and therefore renewed physical access to the monitored environment over a sustained collection period. Perturbation regeneration and mask rotation are consequently a natural design direction, and are to be evaluated together with the broader security assessment. A related direction is to extend the evaluation to a broader range of eavesdropper models, including handcrafted and Doppler-domain features, recurrent architectures, temporal convolutional networks, transformer-based models, ensembles, and adversarially trained classifiers. A further direction is the joint optimization of the gNB precoder and the active RIS phase profile, allowing the obfuscation to be embedded into the propagation

channel while relaxing the dependence on the transmitter EVM budget. Additional extensions include multi-user scenarios and experimental validation on a mmWave testbed.

APPENDIX A: INSTANCE-SPECIFIC ADVERSARIAL ATTACK - DETAILED RESULTS

Table 8 reports the optimal hyperparameters and the full set of performance metrics for each combination of modulation scheme, perturbation mode, and attack type for our adversarial defense, for the instance-specific formulation introduced in Section V. All metrics are computed on correctly classified, EVM-feasible samples under the best hyperparameter configuration per setting.

A. HYPERPARAMETER ANALYSIS

Effective configurations do not systematically require the largest C&W trade-off parameter or the most aggressive settings. Many of the best solutions are obtained with moderate values of c and with 200–300 optimization steps, indicating that feasibility and misclassification are achieved through a sustained optimization process rather than by excessively weighting the classification objective. The violation penalty λ_{vio} tends to increase in stricter regimes and in more constrained targeted configurations, where the optimizer must remain within a narrower feasible region. This behavior is consistent with the role of λ_{vio} : larger values discourage solutions that improve the adversarial loss at the cost of violating the modulation-dependent EVM constraint.

B. SIGNAL-LEVEL METRICS

The retained configurations satisfy the corresponding EVM thresholds in all cases. The average power variation ΔP remains close to 0 dB, and the Pearson correlation between original and perturbed CSI is consistently high, often above 0.99. These values confirm that the selected perturbations preserve the global signal profile and do not act as broad signal degradation. The grid search therefore identifies structured, classifier-aware amplitude and phase transformations that remain small at the signal level but are sufficient to alter the HAR decision.

C. CLASS-DEPENDENT TARGETED DIFFICULTY

Target A_2 is consistently the easiest to induce, reaching feasible ASR values close to 95.25% across most modulation schemes and perturbation modes. The near-constant value across modulations follows from the learned class geometry. The perturbation required to reach A_2 is very small, with mean EVM ≈ 0.0096 , well below even the tightest 256-QAM budget, $\eta_{\text{EVM}} = 0.035$. The modulation-dependent feasibility constraint is therefore not binding for this target, and tightening it from QPSK to 256-QAM does not change the selected solution.

This is visible in Table 8. For the A_2 joint attack, the signal-level signature, including ΔP , $|\Delta\theta|$, Pearson correlation, and consumed EVM, is identical across QPSK–64-QAM and nearly identical at 256-QAM. This indicates that the

TABLE 8. Optimal hyperparameters and attack performance across modulation schemes, perturbation modes, and attack types. All metrics are computed on correctly classified, EVM-feasible samples under the best hyperparameters per configuration.

Mod.	Mode	Attack	c	lr	Steps	λ_{viol}	ASR (%)	95% CI	EVM	ΔP (dB)	$ \Delta\theta $ (°)	Pearson
QPSK	Joint (α, θ)	Untargeted	5.0	0.0500	300	10^2	92.12	[91.1, 93.0]	0.0281	-0.0480	0.6881	0.9973
		Targeted $\rightarrow A_0$	5.0	0.0500	300	10^3	87.70	[86.2, 89.1]	0.0634	-0.0418	1.4146	0.9906
		Targeted $\rightarrow A_1$	5.0	0.0500	300	10^2	91.63	[90.3, 92.8]	0.0398	-0.1041	0.9444	0.9956
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0096	0.0006	0.2410	0.9998
	Ampl. (α)	Untargeted	1.0	0.0500	300	10^2	88.13	[86.9, 89.3]	0.0259	-0.0231	0.0000	0.9966
		Targeted $\rightarrow A_0$	5.0	0.0500	300	10^2	81.57	[79.8, 83.2]	0.0640	-0.0280	0.0000	0.9873
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^2	85.45	[83.8, 86.9]	0.0369	-0.0687	0.0000	0.9947
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0122	0.0004	0.0000	0.9995
	Phase (θ)	Untargeted	1.0	0.0100	300	10^2	85.33	[84.0, 86.6]	0.0164	0.0000	0.8922	1.0000
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^2	77.34	[75.4, 79.1]	0.0403	0.0000	2.0015	1.0000
		Targeted $\rightarrow A_1$	5.0	0.0500	300	10^2	81.11	[79.3, 82.8]	0.0338	0.0000	1.7600	1.0000
		Targeted $\rightarrow A_2$	1.0	0.0050	300	10^2	95.25	[94.2, 96.1]	0.0114	0.0000	0.6358	1.0000
16-QAM	Joint (α, θ)	Untargeted	1.0	0.0500	300	10^2	91.28	[90.2, 92.2]	0.0243	-0.0334	0.5874	0.9979
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^2	86.83	[85.3, 88.3]	0.0583	-0.0329	1.2957	0.9914
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^2	89.99	[88.6, 91.2]	0.0337	-0.0810	0.7880	0.9967
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0096	0.0006	0.2410	0.9998
	Ampl. (α)	Untargeted	10.0	0.0500	300	10^3	87.69	[86.5, 88.8]	0.0250	-0.0195	0.0000	0.9971
		Targeted $\rightarrow A_0$	5.0	0.0500	200	10^4	81.01	[79.2, 82.7]	0.0624	-0.0255	0.0000	0.9879
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^2	85.09	[83.4, 86.6]	0.0352	-0.0621	0.0000	0.9953
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0122	0.0004	0.0000	0.9995
	Phase (θ)	Untargeted	1.0	0.0100	300	10^2	85.33	[84.0, 86.6]	0.0164	0.0000	0.8922	1.0000
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^3	77.34	[75.4, 79.1]	0.0403	0.0000	2.0017	1.0000
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^2	81.11	[79.3, 82.8]	0.0336	0.0000	1.7494	1.0000
		Targeted $\rightarrow A_2$	1.0	0.0050	300	10^2	95.25	[94.2, 96.1]	0.0114	0.0000	0.6358	1.0000
64-QAM	Joint (α, θ)	Untargeted	1.0	0.0500	200	10^3	88.61	[87.4, 89.7]	0.0207	-0.0123	0.5156	0.9986
		Targeted $\rightarrow A_0$	5.0	0.0500	300	10^3	83.15	[81.4, 84.7]	0.0549	-0.0221	1.2194	0.9923
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^2	85.96	[84.4, 87.4]	0.0280	-0.0425	0.6600	0.9977
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0096	0.0006	0.2410	0.9998
	Ampl. (α)	Untargeted	5.0	0.0500	300	10^3	86.44	[85.2, 87.6]	0.0227	-0.0126	0.0000	0.9978
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^2	79.17	[77.3, 80.9]	0.0531	-0.0198	0.0000	0.9884
		Targeted $\rightarrow A_1$	1.0	0.0500	300	10^3	83.41	[81.7, 85.0]	0.0313	-0.0494	0.0000	0.9963
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0122	0.0004	0.0000	0.9995
	Phase (θ)	Untargeted	1.0	0.0100	300	10^2	85.29	[84.0, 86.5]	0.0164	0.0000	0.8899	1.0000
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^2	77.23	[75.3, 79.0]	0.0401	0.0000	1.9837	1.0000
		Targeted $\rightarrow A_1$	1.0	0.0500	200	10^2	81.06	[79.3, 82.7]	0.0336	0.0000	1.7452	1.0000
		Targeted $\rightarrow A_2$	1.0	0.0050	300	10^2	95.25	[94.2, 96.1]	0.0114	0.0000	0.6358	1.0000
256-QAM	Joint (α, θ)	Untargeted	1.0	0.0500	200	10^3	85.80	[84.5, 87.0]	0.0173	-0.0057	0.4233	0.9992
		Targeted $\rightarrow A_0$	1.0	0.0100	300	10^3	78.10	[76.2, 79.9]	0.0260	-0.0101	0.5719	0.9980
		Targeted $\rightarrow A_1$	1.0	0.0100	300	10^3	81.67	[79.9, 83.3]	0.0185	-0.0201	0.4467	0.9991
		Targeted $\rightarrow A_2$	1.0	0.0050	100	10^2	95.25	[94.2, 96.1]	0.0096	0.0006	0.2404	0.9998
	Ampl. (α)	Untargeted	1.0	0.0050	300	10^3	83.00	[81.6, 84.3]	0.0141	-0.0078	0.0000	0.9991
		Targeted $\rightarrow A_0$	1.0	0.0500	300	10^3	68.15	[66.1, 70.2]	0.0262	-0.0145	0.0000	0.9952
		Targeted $\rightarrow A_1$	1.0	0.0100	200	10^3	78.05	[76.2, 79.8]	0.0221	-0.0268	0.0000	0.9984
		Targeted $\rightarrow A_2$	1.0	0.0050	200	10^2	95.25	[94.2, 96.1]	0.0121	0.0004	0.0000	0.9995
	Phase (θ)	Untargeted	1.0	0.0500	200	10^3	83.20	[81.8, 84.5]	0.0151	0.0000	0.7243	1.0000
		Targeted $\rightarrow A_0$	10.0	0.0500	300	10^4	70.60	[68.5, 72.6]	0.0291	0.0000	1.2462	1.0000
		Targeted $\rightarrow A_1$	1.0	0.0500	200	10^3	76.88	[75.0, 78.7]	0.0254	0.0000	1.1688	1.0000
		Targeted $\rightarrow A_2$	1.0	0.0050	300	10^2	95.15	[94.1, 96.0]	0.0113	0.0000	0.6313	1.0000

deterministic grid search re-selects the same minimal-distortion configuration in each case. Targets whose distortion is closer to the budget follow the expected trend. For A_0 , the

consumed EVM decreases from 0.063 at QPSK to 0.026 at 256-QAM as the constraint tightens, with a corresponding drop in ASR.

The constancy observed for A_2 is therefore explained by its easier reachability in the learned feature space together with a non-binding EVM constraint. This asymmetry shows that the target decision regions are not equally reachable through EVM-compliant transformations, and reflects the separability and local arrangement of the activity signatures learned from the CSI tensors.

APPENDIX B: CROSS-MODEL TRANSFERABILITY ANALYSIS

This section extends the transferability analysis to a ResNet-inspired one-dimensional CNN (“ResNet1D”), assessing whether the UAP generated from the VGG-11-BN surrogate remains effective against a target classifier with a different convolutional structure and residual connections.

The protocol follows Section XI. The UAP is generated once using the VGG-11-BN model trained at UE A and is applied without re-optimization to CSI samples observed at UE B and UE C. At each receiver, the HAR classifier is trained independently on the corresponding receiver-specific dataset. The only change is the target architecture: ResNet1D replaces VGG-11-BN.

The ResNet1D model uses the same CSI tensor representation, $X \in \mathbb{R}^{C \times T \times S}$, but reshapes it before the convolutional layers as

$$X \in \mathbb{R}^{C \times T \times S} \longrightarrow X_{\text{1D}} \in \mathbb{R}^{(CS) \times T},$$

so that CSI features across subcarriers are treated as input channels and the convolution operates along the temporal axis. The network uses a one-dimensional convolutional stem followed by three residual stages with (2, 2, 2) basic blocks. Each block includes two kernel-size-3 one-dimensional convolutions, batch normalization, LeakyReLU activations, and a projection shortcut when the feature dimension changes. The classifier head applies adaptive average pooling, dropout, and a final linear layer.

Compared with VGG-11-BN, the target model changes both the convolutional operator and the network topology: it uses one-dimensional temporal convolutions instead of two-dimensional convolutions over the time-subcarrier grid, and it introduces residual skip connections. The experiment therefore tests transferability across receiver position and architecture at the same time.

Figure 14 reports the results for UE B. Without defense, the ResNet1D classifier reaches 89.1% accuracy. After applying the transferred UAP, the accuracy decreases to 36.2%, 31.8%, 33.8%, and 36.0% for $\eta_{\text{EVM}} = 0.035, 0.080, 0.125$, and 0.175 , respectively. The largest reduction is obtained at $\eta_{\text{EVM}} = 0.080$, with a drop of 57.3 percentage points. The resulting accuracy is close to the random-guessing level for the three-class HAR task, showing that the perturbation remains effective at UE B despite the architectural change.

Figure 15 reports the corresponding results for UE C. Without defense, the ResNet1D classifier reaches 78.4% accuracy. Under the transferred UAP, the accuracy decreases to 49.3%, 50.2%, 49.8%, and 49.6% for $\eta_{\text{EVM}} = 0.035, 0.080$,

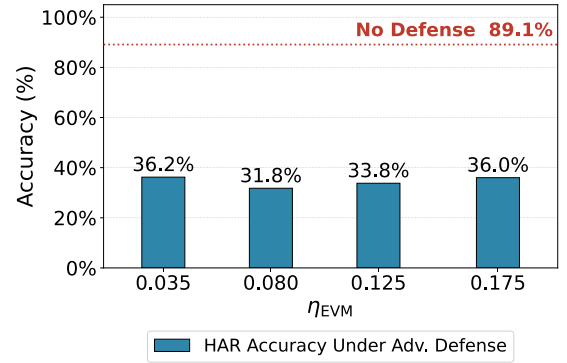


FIGURE 14. Transferability analysis with a ResNet1D target classifier at UE B. The UAP is crafted on the VGG-11-BN surrogate at UE A and evaluated without re-optimization on the ResNet1D classifier trained at UE B. The dashed line reports the no-defense ResNet1D accuracy.

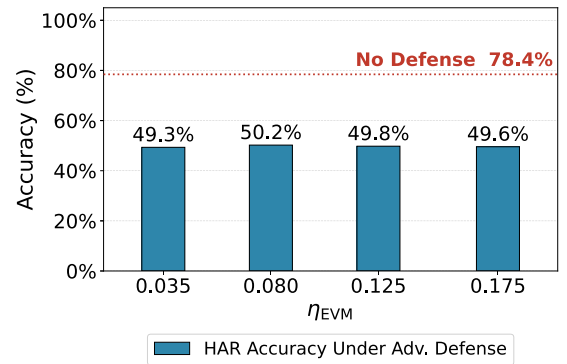


FIGURE 15. Transferability analysis with a ResNet1D target classifier at UE C. The UAP is crafted on the VGG-11-BN surrogate at UE A and evaluated without re-optimization on the ResNet1D classifier trained at UE C. The dashed line reports the no-defense ResNet1D accuracy.

0.125, and 0.175, respectively. The maximum reduction is 29.1 percentage points, obtained at $\eta_{\text{EVM}} = 0.035$. The degradation is smaller than at UE B, consistently with the receiver-position mismatch observed in Section XI.

These results confirm the effectiveness of the UAP in a cross-model setting. The stronger degradation at UE B and the more moderate effect at UE C preserve the spatial trend observed with VGG-11-BN targets, while extending the evaluation to a ResNet-inspired one-dimensional CNN with skip connections.

A. AGGREGATE RESULTS, LIMITATIONS AND SCOPE OF THE TRANSFERABILITY ANALYSIS

Taken together, the spatial analysis of Sec. XI and the cross-model analysis above evaluate the transferred UAP against four independently trained target classifiers, namely VGG-11-BN and ResNet1D at UE B and UE C, each across the four EVM budgets, for a total of 16 transfer configurations covering receiver-position and architectural mismatch simultaneously. Averaged over all of them, the eavesdropper’s HAR accuracy under the defense is 42.7% (standard deviation 8.9), against a mean clean accuracy of 83.5% across the same classifiers, corresponding to a mean reduction of approximately 41 percentage points toward the

33.3% chance level of the three-class task. The residual accuracy is determined primarily by the receiver-position mismatch rather than by the admissible distortion or the target model family.

The evaluation does not exhaust the design space available to an unknown eavesdropper. All transferability experiments use the same underlying CSI representation and classifiers trained on the considered sensing dataset. An eavesdropper could instead rely on different feature pipelines, such as handcrafted statistics, Doppler-domain descriptors, time–frequency representations, or alternative temporal contexts. The classifier may also belong to a different model family, including recurrent networks, temporal convolutional networks, transformer-based architectures, ensembles, or models trained with adversarial augmentation. These choices may change the decision boundaries and the features exploited for HAR, and could therefore affect the response to the same perturbation.

For this reason, the reported results should be read as evidence of spatial and initial cross-model transferability, rather than as a complete characterization against arbitrary eavesdropper designs. A broader security evaluation should include cross-feature transfer, stronger distribution shifts, and adaptive eavesdroppers retrained on perturbed CSI. These extensions are left for future work.

REFERENCES

- [1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, “6G wireless systems: Vision, requirements, challenges, insights, and opportunities,” *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.
- [2] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, “Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities,” *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 990–1002, Sep. 2020.
- [3] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.
- [4] J. Liu et al., “RISAR: Reconfigurable intelligent surfaces-assisted human activity recognition with commercial Wi-Fi devices,” *IEEE Internet Things J.*, vol. 11, no. 24, pp. 40478–40491, Dec. 2024.
- [5] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, “A survey on behavior recognition using WiFi channel state information,” *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.
- [6] R. L. Cigno, F. Gringoli, M. Cominelli, and L. Ghiro, “Integrating CSI sensing in wireless networks: Challenges to privacy and countermeasures,” *IEEE Netw.*, vol. 36, no. 4, pp. 174–180, Jul. 2022.
- [7] M. Harvanek, J. Bolcek, J. Kufa, L. Polak, M. Simka, and R. Marsalek, “Survey on 5G physical layer security threats and countermeasures,” *Sensors*, vol. 24, no. 17, p. 5523, Aug. 2024.
- [8] M. Cominelli, F. Gringoli, and R. Lo Cigno, “AntiSense: Standard-compliant CSI obfuscation against unauthorized Wi-Fi sensing,” *Comput. Commun.*, vol. 185, pp. 92–103, Mar. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0140366421004916>
- [9] H. Niu et al., “A survey on artificial noise for physical layer security: Opportunities, technologies, guidelines, advances, and trends,” *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 341–381, 2026.
- [10] X. Gu, W. Duan, G. Zhang, Q. Sun, M. Wen, and P.-H. Ho, “Physical layer security for RIS-aided wireless communications with uncertain eavesdropper distributions,” *IEEE Syst. J.*, vol. 17, no. 1, pp. 848–859, Mar. 2023.
- [11] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, “Robustness of classifiers: From adversarial to random noise,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
- [12] 3GPP, 5G; NR; Base Station (BS) Radio Transmission and Reception, document TS 38.104 V19.2.0, Oct. 2025.
- [13] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, *arXiv:1412.6572*.
- [14] A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *stat*, vol. 1050, p. 9, Nov. 2017.
- [15] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
- [16] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 86–94.
- [17] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: From phenomena to black-box attacks using adversarial samples,” 2016, *arXiv:1605.07277*.
- [18] A. Demontis et al., “Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks,” in *Proc. 28th USENIX Secur. Symp. (USENIX Secur.)*, 2019, pp. 321–338.
- [19] Z. Cai et al., “Device-free wireless sensing for gesture recognition based on complementary CSI amplitude and phase,” *Sensors*, vol. 24, no. 11, p. 3414, May 2024.
- [20] R. Guo, H. Li, D. Han, and R. Liu, “Feasibility analysis of using channel state information (CSI) acquired from Wi-Fi routers for construction worker fall detection,” *Int. J. Environ. Res. Public Health*, vol. 20, no. 6, p. 4998, Mar. 2023.
- [21] C. Szegedy et al., “Intriguing properties of neural networks,” 2013, *arXiv:1312.6199*.
- [22] P. Staat et al., “IRShield: A countermeasure against adversarial physical-layer wireless sensing,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 1705–1721.
- [23] P. Huang, X. Zhang, S. Yu, and L. Guo, “IS-WARS: Intelligent and stealthy adversarial attack to Wi-Fi-based human activity recognition systems,” *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 6, pp. 3899–3912, Nov. 2022.
- [24] X. Wang, J. Li, X. Kuang, Y.-A. Tan, and J. Li, “The security of machine learning in an adversarial setting: A survey,” *J. Parallel Distrib. Comput.*, vol. 130, pp. 12–23, Aug. 2019.
- [25] B. Flowers, R. M. Buehrer, and W. C. Headley, “Evaluating adversarial evasion attacks in the context of wireless communications,” *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1102–1113, 2020.
- [26] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, “Channel-aware adversarial attacks against deep learning-based wireless signal classifiers,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3868–3880, Jun. 2022.
- [27] J. Yang, H. Zou, and L. Xie, “SecureSense: Defending adversarial attack for secure device-free human activity recognition,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 823–834, Jan. 2024.
- [28] X. Wang, X. Wang, S. Mao, J. Zhang, S. C. G. Periaswamy, and J. Patton, “Adversarial deep learning for indoor localization with channel state information tensors,” *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18182–18194, Oct. 2022.
- [29] Y. Pan, Z. Zhou, W. Gong, and Y. Fang, “SAT: A selective adversarial training approach for Wi-Fi-based human activity recognition,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12706–12716, Dec. 2024.
- [30] Z. Yang, Y. Zhao, and W. Yan, “Adversarial vulnerability in Doppler-based human activity recognition,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.
- [31] Y.-H. Chen, H.-H. Hsieh, and L.-H. Lu, “A 24-GHz receiver frontend with an LO signal generator in 0.18- μ m CMOS,” *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 5, pp. 1043–1051, May 2008.
- [32] A. Verma, P. K. Yadav, S. Ambulker, M. Goswami, and P. K. Misra, “A 36.7 mW, 28 GHz receiver frontend using 40 nm RFCMOS technology with improved figure of merit,” *Anal. Integr. Circuits Signal Process.*, vol. 107, no. 1, pp. 135–144, Apr. 2021.
- [33] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.
- [34] H. Kim, “Torchattacks: A PyTorch repository for adversarial attacks,” 2020, *arXiv:2010.01950*.
- [35] G. Restuccia. (2026). *Adversarial CSI Obfuscation Against Device-Free Activity Recognition in RIS-Assisted Systems*. [Online]. Available: https://github.com/gabrielicacion/adv_cnn_ris
- [36] X. You et al., “Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts,” *Sci. China Inf. Sci.*, vol. 64, no. 1, 2021, Art. no. 110301.
- [37] 3GPP/NR; Physical Channels and Modulation (Release 17), 3rd Generation Partnership Project (3GPP), document (TS) 38.211, 2022. [Online]. Available: <https://www.3gpp.org>

[38] 3rd Generation Partnership Project (3GPP), *NR; Multiplexing and Channel Coding*, 3rd Generation Partnership Project (3GPP), document 3GPP TS 38.212, Jun. 2026, version 19.4.0, Release 19.



applications to smart water networks and next-generation networked infrastructures.

GABRIELE RESTUCCIA received the Ph.D. degree in information and communication technologies from the University of Palermo in 2024. Since 2024, he has been affiliated with CNIT-CONLab. He has been a Visiting Researcher in Germany and USA. He is currently an Assistant Professor with the Department of Engineering, University of Palermo, Italy. His research interests include edge AI, the IoT, adversarial machine learning, cybersecurity, digital twins, and local agentic AI solutions for critical cyber-physical systems, with



DOMENICO GARLISI has been affiliated with CNIT-CONLab since 2024. He is currently an Assistant Professor with the University of Palermo and a Visiting Researcher with UCLA. He also works on smart water distribution and physical-layer security. He has participated in numerous national and international projects, leading two as the Principal Investigator. His research interests include wireless networks, software-defined radio, the IoT, and localization systems.



interference management, cognitive radio, network testbeds, smart water distribution systems, and the integration of large language models into network optimization.

FABRIZIO GIULIANO received the M.S. degree in telecommunications engineering and the Ph.D. degree in electronics and telecommunications engineering from the University of Palermo, Palermo, Italy, in 2009 and 2014, respectively. He is currently an Assistant Professor with the Department of Mathematics and Computer Science, University of Palermo. He contributed to several national and European projects in the ICT and Telecommunication Sector. His research interests include wireless and IoT networks, low-power wide-area networks,



SILVIA SCHILLECI (Graduate Student Member, IEEE) received the M.Sc. degree (cum laude) in electronics engineering from the University of Palermo in 2024, where she is currently pursuing the Ph.D. degree in information and communication technologies. She has experience with cognitive radio systems and 5G jamming, and is involved in simulation frameworks for GNSS spoofing detection and mitigation. Her research interests include physical layer security, wireless signal analysis, and reconfigurable intelligent surfaces.



STEFANO MANGIONE (Member, IEEE) received the M.Sc. degree (cum laude) in electronic engineering from the University of Palermo, Italy, in 2000. Since 2001, he has been an Assistant Professor of signal theory and communication systems. His current research interests include underwater communication systems, physical layer aspects of non-terrestrial networks, and low-complexity passive analysis of wireless mobile communication systems.



ALESSANDRA DINO received the M.Sc. degree in electronics engineering and the Ph.D. degree in ICT from the University of Palermo, Italy, in 2022 and 2026, respectively. She is currently a Researcher with CNIT-CONLab. Her research primarily focuses on physical-layer vulnerabilities in 5G networks and, more broadly, the security and performance of wireless networks.



ILENIA TINNIRELLO received the Ph.D. degree in telecommunications engineering from the University of Palermo in 2004. She has been a Visiting Researcher in South Korea and Singapore and currently coordinates European Digital Innovation Hub i-Nest. She is also a Full Professor with the University of Palermo. Her research involves reconfigurable wireless networks, protocol design, and prototyping.