Magmaw: Modality-Agnostic Adversarial Attacks on Machine Learning-Based Wireless Communication Systems

Jung-Woo Chang*, Ke Sun*, Nasimeh Heydaribeni*, Seira Hidano†, Xinyu Zhang*, Farinaz Koushanfar*
*University of California San Diego †KDDI Research, Inc.

***Abstract*—Machine Learning (ML) has been instrumental in enabling joint transceiver optimization by merging all physical layer blocks of the end-to-end wireless communication systems. Although there have been a number of adversarial attacks on ML-based wireless systems, the existing methods do not provide a comprehensive view including multi-modality of the source data, common physical layer protocols, and wireless domain constraints. This paper proposes Magmaw, a novel wireless attack methodology capable of generating universal adversarial perturbations for any multimodal signal transmitted over a wireless channel. We further introduce new objectives for adversarial attacks on downstream applications. We adopt the widely-used defenses to verify the resilience of Magmaw. For proof-of-concept evaluation, we build a real-time wireless attack platform using a software-defined radio system. Experimental results demonstrate that Magmaw causes significant performance degradation even in the presence of strong defense mechanisms. Furthermore, we validate the performance of Magmaw in two case studies: encrypted communication channel and channel modality-based ML model. Our code is available at https://github.com/juc023/Magmaw.**

I. Introduction

Next-generation (NextG) networks promise to support ultra-reliable and low-latency communication for rapidly evolving wireless devices [28]. Emerging networks are thus challenged to establish new features (e.g., adaptive coding and enhanced modulation) to overcome rapidly changing channel conditions and to achieve more efficient use of spectrum [62], [101]. Machine Learning (ML) overcomes this barrier by revolutionizing the entire wireless network protocol stack [66].

Recent research [18] introduces joint source-channel coding (JSCC), an end-to-end wireless communication system leveraging deep neural networks (DNNs) for both transmitter and receiver. This ML approach jointly optimizes source and channel coding in a cross-layer framework to handle diverse and challenging channel conditions. To effectively cope with the multipath fading effects, the JSCC-encoded data can be further modulated into continuous signal waveforms through orthogonal frequency division multiplexing (OFDM) [98]. The DNN models for JSCC are tailored to specific modalities (e.g., texts, images, etc.), so as to convey semantic information more accurately than traditional communication systems (see §II-B). The advantages of such ML-based communication systems are increasingly recognized by standardization bodies such as the Third Generation Partnership Project (3GPP) [83]. Industry leaders, such as Apple [69], Huawei [87], Nokia Bell Labs [6], Qualcomm [63], and ZTE [50] are also investigating AI-native 6G communications. NVIDIA has established an ML-based, GPU-accelerated communication signal processing framework [38] for 6G applications. These developments underscore the growing consensus that ML-based wireless communications will play a crucial role in shaping the future of 6G technology.

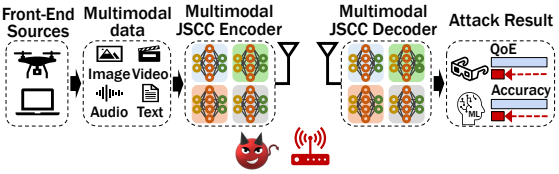


Fig. 1: High-level view of Magmaw.

Unfortunately, ML is vulnerable to adversarial attacks [15], [76], where small, imperceptible changes to input can yield substantial changes in the model's output. The susceptibility of the models to adversarial examples raises serious concerns for the safety of ML adoption in NextG.

Traditional jamming or overshadowing attacks [29], [75], [92], [96] have been dedicated to developing a malicious RF device to disrupt legitimate wireless communications. However, these approaches typically rely on high-power transmissions to cause large-scale disruptions in the spectrum, leading spectrum owners to respond swiftly. Highly effective attacks that use low signal strengths are missing in this literature.

There have been recent works on small signal manipulations designed to target ML-based wireless systems [12], [39], [47], [54], [59]. However, they make unrealistic assumptions about the attacker's capabilities. For example, even though JSCC has a modality-specific structure, they assume that only a single modality (e.g., one-hot vector message or image) is wirelessly communicated. They also assume that the adversary knows which modality is sent by the transmitter. In practice, the above assumptions are not valid for the following reasons: 1) the transmitter typically incorporates data from all modalities into the data blocks and then sends them to the receiver; 2) if the adversary wants to recognize the modality of the signal, it needs to have access to the target ML model that carries out JSCC, and this is not always feasible, and 3) even if the adversary can detect the modality, high latency occurs until perturbations are generated and added to the victim signal.

We propose Magmaw, a new hardware-driven wireless attack framework that creates universal adversarial perturbations (UAPs) to subvert ML-based wireless systems. We show for the first time that modulated multimodal data can be perturbed by adversaries, resulting in failure to restore the original data as well as subversion of downstream services. We consider examples of downstream services such as video classification (VC), which analyzes human activity from video, and audio-

Network and Distributed System Security (NDSS) Symposium 2025
23 - 28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.23336
www.ndss-symposium.org

visual event recognition (AVE) which predicts the event label based on representations over multiple input modalities. Magmaw can cause significant disruptions or threaten user safety in quality-sensitive applications, e.g., remote surgery [3] and autonomous driving [36], as illustrated in Figure 1. Emerging applications (e.g., XR [48]) would suffer even more from the corruption of multiple input modalities.

Magmaw must address four main design challenges. Firstly, we assume that the adversary lacks prior knowledge about the data's modality and the exact channel model. Additionally, the attacker's ability to adjust its transmit signal pattern effectively depends on knowing the channel matrix between the sender and receiver ($\mathbf{H_t}$). However, since $\mathbf{H_t}$ varies due to factors like link distance, mobility, and environment, not having this information makes crafting an effective attack challenging. We solve the above challenges by designing a perturbation generator model (PGM) trained to create input- and channel-agnostic perturbations on surrogate wireless models. We adopt an ensemble learning approach that utilizes surrogate multimodal JSCC models to learn UAPs.

Secondly, previous attacks [12], [39], [47], [54], [59] do not consider that the input signal can be adjusted by physical layer protocols (see §II-C). They only focus on the scenarios where the adversary has prior knowledge of the protocol's full setup. In a practical scenario, the attacker does not know the constellation mapping or how the OFDM system assigns the complex symbols to multiple subcarriers. It is possible to design an attacker that recognizes the protocol from the transmitted signals [72]. However, since wireless protocols change rapidly depending on the channel state, the analyzed output quickly becomes obsolete. A protocol-agnostic attack is required. We address this challenge by incorporating multiple controllable parameters inside the ensemble learning to optimize perturbations generalizable across all modulated data.

Thirdly, an adversarial wireless device may not be precisely synchronized with a legitimate transmitter or receiver in the time or frequency domain, reducing the effectiveness of perturbations. We address de-synchronization issues between the adversarial device and the legitimate transmitter/receiver using our offline training procedure. Specifically, we train the PGM using time shift and phase rotation functions, ensuring that UAPs remain effective even with varying offsets.

Finally, previous studies [12], [54] are vulnerable to adaptive defenses. For instance, a perturbation detector [93] can exploit traces of perturbations to predict whether the input is perturbed. This is because their perturbations are overly rigid and lacking in variability due to overfitting [34]. To craft robust and diverse perturbations, we introduce a discriminator and diversity loss to regularize the learning process explicitly.

After integrating the above solutions, we implement Magmaw on the software-defined radio platform and validate its attack feasibility, as shown in Appendix A. Our experiments show that Magmaw degrades the Peak Signal-to-Noise Ratio (PSNR) by up to 8.04dB and 8.29dB for image and video transmission, respectively, where PSNR is a representative image quality score. For speech transmission, Magmaw prevents receivers from recognizing the speech content, increasing the mean square error (MSE) by up to 3.91× compared to baseline attacks. Furthermore, Magmaw reduces the bilingual evaluation understudy (BLEU) score to 0.338 points for text transmission, indicating that the received text exhibits significant semantic errors and grammatical inaccuracies. Notably, we achieve up to 91.2% attack success rate on the downstream tasks. In our case study, we establish an encryption-based secure image transmission and prove that Magmaw leads to a reduction of up to 5.88dB in PSNR. We also evaluate Magmaw with channel modality-based ML models. Magmaw introduces up to 2.2× more error in the ML results than the baseline.

In summary, we make the following contributions:

- We introduce Magmaw, a novel wireless attack framework implemented over software-defined radio against ML-based multimodal communication systems and underlying downstream applications.
- We adopt an ensemble learning approach on a set of surrogate JSCC models to craft our UAP input- and protocol-agnostic, i.e., oblivious to the modality, constellation, coding rate, OFDM specifications, and channel conditions.
- We evaluate Magmaw against various defense techniques, including adaptive ones. Extensive results from case studies further show Magmaw's efficacy.

II. Background

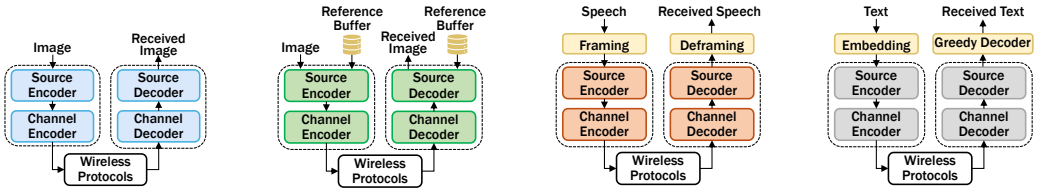
A. Wireless Communication Systems

Current communication standards (e.g., 4G LTE [2], IEEE 802.11 family [40], 5G NR [49]) follow separate source and channel coding designs and require independent optimization of each component. The source encoder transforms the source data into the embedded source bits. The channel encoder adds redundancy to the transmitted signal, allowing the receiver to correct errors caused by noise. However, these conventional systems suffer from dramatic performance degradation due to the cliff effect where the receiver's error correction algorithm cannot recover the transmitted data if channel conditions are worse than a certain threshold [18].

ML-driven wireless systems aim to train a robust JSCC encoder and decoder on wireless channels infused with channel conditions similar to the physical world. The JSCC encoder directly maps the source to complex-valued symbols, and the JSCC decoder recovers its estimate directly from the noisy channel output. To adopt the widely used wireless standards, the JSCC models can be concatenated with OFDM to increase the spectral efficiency and reduce the multipath channel effects [98]. Since multipath fading channels and OFDM blocks can be represented as differentiable layers, ML-based wireless systems are trained end-to-end. As such, JSCC can be built without modifying standard radio hardware (e.g., field test 6G with JSCC on 4G LTE [84]). Furthermore, ML-based wireless communication can significantly save channel bandwidth costs compared to conventional systems while achieving the same end-to-end wireless transmission performance [99].

B. Modality-Specific JSCC Models

Existing JSCC systems [82], [88], [90], [98] adopt modality-specific structures, with each modality requiring a specialized approach for accurate symbol recovery at the receiver. We consider four state-of-the-art JSCC models for image [98], video [82], speech [88], and text transmission [90].



(a) Image JSCC model [98] (b) Video JSCC model [82] (c) Speech JSCC model [88] (d) Text JSCC model [90]

Fig. 2: The modality-specific JSCC model for end-to-end wireless communication system.

Figure 2 depicts the commonly used structures for each modality. The image JSCC is trained to minimize distortion on a frame-by-frame basis. The video JSCC leverages spatiotemporal similarities between successive frames to remove the redundancy. To achieve this, the video JSCC adopts the temporal coding structure σ, which clusters each consecutive sequence of pictures into a group of pictures (GOP). Each frame within the GOP is entered into the video JSCC in coding order rather than display order. This means that the video JSCC encoder compresses frames in a specific order. For a total of P frames included in the GOP, the coding order of each frame is determined by the mapping function $m_\sigma(t)$, where $1 \leq m_\sigma(t) \leq P$. On the other hand, speech signals contain speaker characteristics such as speech rate and tone. The attention mechanism [88] is utilized for speech JSCC to identify the essential features to help accurately recover speech signals at the receiver. The text JSCC is designed to precisely encode context information and cope with semantic distortion based on Transformers [81]. The text features recovered by the receiver are decoded into the text sentence through a greedy decoder [85]. A cross-entropy loss [46] is used to understand semantic meaning while maximizing system capacity.

C. Physical Layer Protocols

Modulation. Wireless standards commonly adopt QPSK, 16-QAM, and 64-QAM to map bits to complex symbols [89]. Therefore, the JSCC-encoded data are mapped to elements in a two-dimensional finite constellation diagram. An adaptive modulation scheme can change the modulation type to balance reliability and spectral efficiency. For example, [1] changes the modulation based on the threshold of the channel state to meet the bit error rate (BER) requirement.

OFDM. To achieve high spectral efficiency, the OFDM transmitter may assign modulated symbols arbitrarily to the subcarriers rather than in a fixed order. Therefore, each subcarrier carries symbol vectors with a different distribution.

Coding Rate. Adaptive encoding is essential to guarantee the reliability of wireless communications [91]. The JSCC encoder estimates the available bandwidth based on the channel state and employs adaptive algorithms to choose an optimal coding rate for efficient real-time streaming.

III. Related Work

A. Conventional Wireless Attacks

Jamming Attacks. RF jamming transmits radio signals indiscriminately across a range of frequencies, causing interference and disrupting communication. Jamming can be broadly categorized as active jamming and reactive jamming [61], [92]. Active jamming continuously emits powerful interference signals, but its continuous operation leaves detectable traces, making it vulnerable to defensive techniques [92]. Reactive jamming [51] adjusts its jamming behavior according to observed signals in the environment. It remains silent when the channel is idle but initiates high-power signal transmission upon detecting activity on the channel. The drawback of these approaches is that spectrum owners may promptly detect the presence of an attack and respond accordingly.

Overshadowing Attacks. Cellular networks are vulnerable to overshadowing attacks [86]. Recent works [29], [96] can force the victims to receive the attacker's symbols/subframes by sending high-powered signals to a base station. However, the adversary must have the capability to receive and decode the messages transmitted by the victim. These attacks typically require signal strengths ranging from -3.4dB to +3dB over the benign signal [29]. Since the adversary's signal strength is comparable to or even stronger than the legitimate signals, it becomes easier for the legitimate nodes to identify the attack.

B. Adversarial Attacks on Computer Vision Domains

Adversarial ML has been studied to analyze the robustness of the ML model across multiple areas, such as image classification [21], [64], speech recognition [5], human activity recognition [19], [25], neural video compression [23], [24], etc. Most studies provide an attacker with the capabilities to perform a man-in-the-middle attack where he/she intercepts data in the middle and then injects small perturbations. However, these are not physically feasible and only expose theoretical vulnerabilities. As the demand for physically feasible attacks grows, recent studies [41], [55], [70] define practical methodologies so that attacks can be realized in the real world. SLAP [55] applies a projector to superimpose light onto an object, causing the model to misclassify the object. Compared to wireless domains, physical attacks in vision domains are less susceptible to signal distortion and have relatively fewer domain constraints.

C. Adversarial Attacks on Wireless Domains

There are two types of target wireless systems: (1) wireless networking, which concentrates on efficient dataflow management between networked devices, and (2) wireless communication and sensing for restoring and analyzing radio signals at the physical layer. In this paper, we focus on the second point.

Attackers targeting wireless networking seek to deceive the ML-based network devices into making wrong decisions (e.g., for resource allocation). Certain attacks [42], [56] operated in a white-box setting with complete knowledge of the target ML model. In contrast, Apruzzese *et al.* [8] devised a realistic threat model by assuming a constrained attacker and demonstrated their performance across various ML systems.

TABLE I: A comparison of existing adversarial attacks against ML-based wireless communication and sensing.

Attacks	Type	Channel	Non-WB ML	HW Demo	Input-Agnostic		Protocol-Agnostic			Sync-Free		Defenses			
					Multimodal	H_t	Constellation	Coding Rate	OFDM	Time	Phase	RT	PS	PD	OD
[68]	Offline Attacks	AWGN	✓							✓		□	□	□	□
[4]			✓								✓	□	□	□	□
[59]		Multipath Fading										□	□	□	□
[31]												□	□	□	□
[47]												□	□	□	□
[8]	Online Attacks	-	✓									■	□	□	□
[67]		AWGN	✓							✓		□	□	□	□
[12]			✓								✓	■	■	□	■
[39]		Multipath Fading	✓			✓						■	□	□	□
[43]			✓			✓						■	□	□	□
[54]			✓	✓		✓				✓	✓	■	□	□	□
Ours			✓	✓	✓	✓	✓	✓	✓	✓	✓	■	■	■	■

H_t: a channel matrix between the sender and the receiver; ✓: the item is supported; **WB**: White Box.
RT: Robust Training (Adversarial Training, Defensive Distillation, and Randomized Smoothing); **PS**: Perturbation Subtraction; **PD**: Perturbation Detection; **OD**: Oracle Defense.
■: the attack can compromise the defense; ■: the defense was considered, but the attack was ineffective.; □: not mentioned in the paper.

When attacking wireless communication and sensing, it is crucial to design physically realizable perturbations. Table I summarizes existing attacks in two categories: offline attacks and online attacks. Offline attacks are impractical as they allow attackers unlimited access to inputs and models. Online attacks address this by adding UAPs to victim signals. Several works [43], [67] studied methods for crafting UAPs against radio signal classifiers but aimed to identify theoretical vulnerabilities rather than design physically feasible attacks. Flowers *et al.* [31] identified victim's transmissions by sniffing the signal strength of the target channel, but sniffing does not provide an accurate time offset due to latency and cannot reveal the modality and wireless protocol. Bahramali *et al.* [12] adopted a generative model to produce diverse UAPs, but they made the unrealistic assumption that the target system sends only one-hot vector messages [60]. Their attacks are evaluated individually on each physical layer component rather than on an end-to-end system. RAFA [54] designed a practically feasible UAP in a limited-knowledge setting. They solely target the publicly-known preambles, so their attacks are not applicable to JSCC which transmits unknown data symbols. In addition, the JSCC-encoded data are modulated by various protocols (e.g., modulation, coding rate, and OFDM). Furthermore, due to the lack of diversity in its perturbations, RAFA can be directly mitigated by the adaptive defense with high accuracy (see §IX-B). Additionally, a recent study [45] attacked wireless sensing systems by assuming that an adversary could install malicious firmware on the victim transmitter and change pilot packets. However, we are interested in a more realistic scenario where an adversarial signal is injected into the target channel.

IV. Threat Model

A. Attack Scenario

Magmaw is targeted towards radio signals created by front-end sources that are used to transmit the multimodal source to back-end user(s). The attacker deploys commercial off-the-shelf (COTS) hardware (e.g., software-defined radios) to send the attack signals. We focus on vulnerabilities unique to ML in wireless environments, leading to the failure of the receiver's JSCC decoder to correctly decode the received packet. Note that we exclude the jamming effect [61], a brute-force solution that disrupts all communication within the medium.

Multiple transmitters and receivers can share the spectrum. As described in Appendix A, the standard Wi-Fi protocol ensures only one device uses the wireless channel at a time within a cell to avoid collision. Magmaw can thus inject adversarial signals to target different transmitter-receiver pairs sequentially. Magmaw can also be positioned in a selective attack [9] that only targets a specific wireless device, leaving other devices unaffected. Specifically, Magmaw can identify the victim by sniffing the MAC address in the packet, and launch the attack whenever the victim device transmits the packet. Please note that all the above cases are equivalent to applying Magmaw's adversarial perturbation to a single transmitter-receiver pair (as shown in Appendix Figure 20).

B. Adversary's Goal

Magmaw aims to transmit well-crafted perturbations over the target wireless channel to prevent legitimate receivers from recovering the source data and performing target downstream tasks. To ensure stealthiness, Magmaw sends adversarial signals with a small magnitude. As a result, the victim cannot differentiate between adversarial perturbations and natural noise from wireless channels. Following the previous studies [12], [68], we utilize a perturbation-to-signal ratio (PSR) metric to compare the power of the perturbation at the receiver with the received legitimate signal power. The PSR is set to be [-20,-10] dB [12], [68] so that the perturbation is not distinguishable from the expected natural noise in the channel.

C. Adversary's Capability and Knowledge

We envision a constrained attacker [7] with limited knowledge of ML-based wireless systems as described below.

Wireless System. We assume that the adversary has no prior knowledge about the ML model architecture/parameters, but knows the category of target models (e.g., autoencoder which is the de facto model for JSCC) and the physical layer techniques being used (e.g., OFDM modulation which is specified in the communication standard). This is a realistic assumption for the following reasons: 1) standard documentation usually describes the core technology and is open to the public, and 2) specialized operations (see §II-B) for each modality have already been widely known in the ML community. The adversary trains surrogate ML-based JSCC models using a large amount of publicly available data. Note that the attacker cannot access the target JSCC model or observe the output.

Knowledge about Input and Protocols. We assume that the adversary does not know the modality and the constellation mapping method due to the following reasons: 1) all the application-layer source data, regardless of modalities, need to multiplex the transmitter radio and wireless channel, 2) the transmitter can adapt several types of modulation techniques

according to channel conditions. Additionally, the JSCC model can dynamically adjust the coding rate in real time based on the current channel conditions, so the adversary has no prior knowledge about the number of OFDM symbols encoded by the JSCC model in the transmitted signal. However, we assume that the adversary can refer to the possible coding rates specified in the standards documents. Lastly, we do not assume that the adversary knows how the transmitter maps the OFDM symbol to the subcarriers.

Target Wireless Channel. We consider a real-world attack scenario where the attacker cannot have access to the channel matrix between the transmitter and the receiver, i.e., $\mathbf{H_t}$. In addition, we do not assume that the adversary is synchronized with either the transmitter or the receiver, leading to random time and frequency offsets. Furthermore, we assume that the attacker can determine the carrier frequency used by the targeted channel. The attacker can overhear the victim's signals by arbitrarily adjusting its waveform bandwidth and carrier frequency using a software-defined radio [52], [102].

Attacker's Wireless Channel. The attacker employs a single antenna to send the adversarial signal. We denote the channel matrix for the attacker as $\mathbf{H_a}$. According to the Wi-Fi protocol, the receiver periodically sends beacons to wireless devices within the range [13]. An attacker can overhear this transmission and estimate the channel matrix from the receiver to itself. Due to the principle of reciprocity, this channel is the same as $\mathbf{H_a}$. In contrast to recent work [54], we relax the assumption that the adversary knows the exact channel matrix between the attacker and the receiver. We make a weaker assumption that the adversary has limited information, i.e., the distribution of the channel between the attacker and the receiver.

V. System Model

Figure 3 illustrates the core processing blocks in the victim communication link along with the Magmaw attacker.

ML-based Transmitter. We consider OFDM-based JSCC over a multipath fading channel with L_t paths. The multimodal source data are transmitted using N_s OFDM symbols with L_{fft} OFDM subcarriers. Note that N_s has different values depending on the modality and the coding rate. For channel estimation, the sender transmits a preamble (according to the publicly available wireless communication standards) on the subcarriers. We denote the source data as x_t^Q with modality $Q \in \{\mathcal{I}, \mathcal{V}, \mathcal{S}, \mathcal{T}\}$ at time step t, where $\mathcal{I}, \mathcal{V}, \mathcal{S}, \mathcal{T}$ denote the image, video, speech, and text, respectively. We describe a JSCC encoder for processing a modality Q with a given coding rate λ and modulation scheme C as a function $E_{Q,C,\lambda}(x_t^Q, \mathcal{B}_t^Q)$, where $\mathcal{B}_t^Q$ is the transmitter's reference buffer used for the video JSCC model, as illustrated in Figure 2 (b). We define the reference buffer $\mathcal{B}_t^Q$ containing the previously decoded frame $\tilde{x}^{\mathcal{V}}(\cdot)$ as:

$$\mathcal{B}_t^Q = \begin{cases} \{\tilde{x}^{\mathcal{V}}_{m_\sigma(1)}, \cdots, \tilde{x}^{\mathcal{V}}_{m_\sigma(t-1)}\}, & \text{if } Q = \mathcal{V}, \\ \emptyset, & \text{if } Q \neq \mathcal{V}. \end{cases} \tag{1}$$

Recall that $m_\sigma(t)$ is a function that finds the coding order of the t-th image in the given GOP structure σ. $\mathcal{B}_t^{\mathcal{V}}$=$\emptyset$ when t=1. This is because the first frame is coded by the image JSCC. Note that $\tilde{x}^{\mathcal{V}}(\cdot)$ is reconstructed as the output of a video JSCC decoder that takes encoded video sequence $E_{\mathcal{V},C,\lambda}(x_t^{\mathcal{V}}, \mathcal{B}_t^{\mathcal{V}})$ as input. λ is the coding rate to control the number of symbols.

Then, a constellation mapping method $M_C(\cdot)$ moves symbols to the nearest constellation points in a finite constellation diagram C. The modulated symbol, $Y_t^Q \in \mathbb{C}^{N_s \times N_{fft}}$, can then be obtained as:

$$Y_t^Q = M_C(E_{Q,C,\lambda}(x_t^Q, \mathcal{B}_t^Q)). \tag{2}$$

Without loss of generality, we assume the target transmitter/receiver uses a single antenna following the 802.11a/g/n Wi-Fi standard [89]. We split Y_t^Q into a number of signal vectors with dimension of N_{fft}. Afterwards, an OFDM transmitter allocates divided signals on each subcarrier. Each OFDM symbol passes through an inverse discrete Fourier transform (IFFT), then a cyclic prefix (CP) is added and transmitted to the receiver over a multipath fading channel.

ML-based Receiver. The receiver obtains the complex-valued symbols from the channel output by removing the CP and applying FFT with an OFDM receiver. The received signal of the k-th subcarrier in the i-th OFDM symbol is given by:

$$\hat{Y}_t^Q[i,k] = \mathbf{H_t}[k] Y_t^Q[i,k] + W[i,k], \tag{3}$$

where $\mathbf{H_t} \in \mathbb{C}^{N_{fft} \times N_{fft}}$ is the frequency-domain channel matrix, which is a diagonal matrix, and $W \in \mathbb{C}^{N_s \times N_{fft}}$ is the frequency-domain AWGN matrix.

Given the FFT output of the pilot signals, the channel estimation and equalization are performed to compensate the channel-induced transformation. We adopt a least squares (LS) algorithm to predict the channel state information. After equalizing all of the divided signals with the channel equalizer $R(\cdot)$, we quantize the phase and amplitude of the signal on each subcarrier with $M_C(\cdot)$. Finally, we employ the decoder $D_{Q,C,\lambda}(\cdot)$ to reconstruct an estimate $\hat{x}_t^Q$ of the original signal. We express the entire process after OFDM receiver as follows:

$$\begin{aligned} \hat{x}_t^Q =& D_{Q,C,\lambda}(M_C(\mathcal{R}(\hat{Y}_t^Q)), \hat{\mathcal{B}}_t^Q) \\ =& \mathcal{F}_{Q,C,\lambda}(\hat{Y}_t^Q, \hat{\mathcal{B}}_t^Q), \end{aligned} \tag{4}$$

where $\hat{\mathcal{B}}_t^Q$ is the receiver's decoded frame buffer for the video JSCC. $\hat{\mathcal{B}}_t^{\mathcal{V}}$=$\{\hat{x}^{\mathcal{V}}_{m_\sigma(1)}, \cdots, \hat{x}^{\mathcal{V}}_{m_\sigma(t-1)}\}$, where $\hat{\mathcal{B}}_t^{\mathcal{V}}$=$\emptyset$ when t=1. $\hat{\mathcal{B}}_t^Q$=$\emptyset$ for other modalities. For simplicity, we denote all processes after the OFDM receiver as $\mathcal{F}_{Q,C,\lambda}(\cdot)$.

VI. Attack Construction

The framework of Magmaw is illustrated in Figure 3. Our attack methodology follows a hardware/algorithm co-design to ensure Magmaw is robust against various signal distortions.

A. Our Attack Formulation

General Attack Formulation. Our adversary aims to find an input-agnostic perturbation $\boldsymbol{\delta}^s \in \mathbb{C}^{N_s \times N_{fft}}$, with a magnitude bounded by the attacker's power budget $\epsilon \in \mathbb{R}$. When $\boldsymbol{\delta}^s$ is injected into the victim wireless channel, the receiver obtains the frequency-domain channel output $\bar{Y}_t^Q$ as:

$$\bar{Y}_t^Q[i,k] = \mathbf{H_t}[k] Y_t^Q[i,k] + \mathbf{H_a}[k] \boldsymbol{\delta}^s[i,k] + W[i,k], \tag{5}$$

where $\bar{Y}_t^Q[i,k]$ and $\boldsymbol{\delta}^s[i,k]$ represent the frequency-domain perturbed response and the adversarial perturbation at the k-th subcarrier of the i-th OFDM symbol, respectively. $\mathbf{H_a}$ is

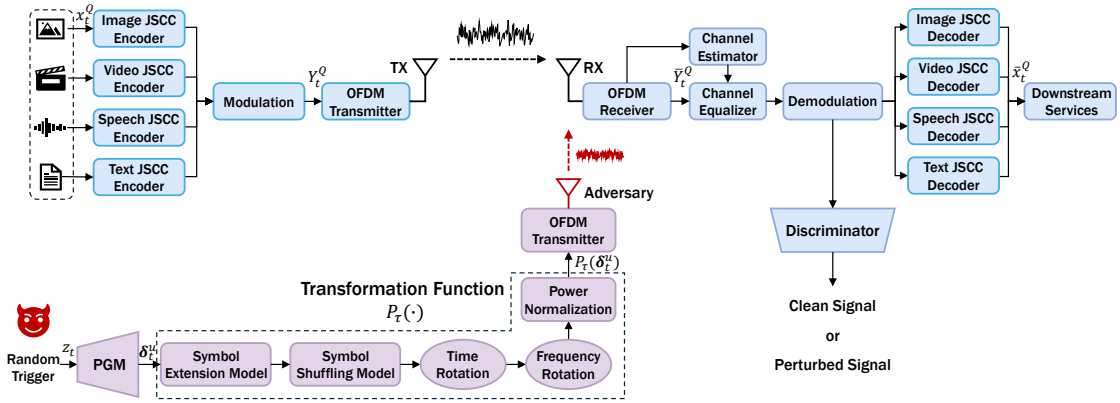


Fig. 3: Overview of Magmaw. During the PGM training, the attacker employs the surrogate JSCC model (blue modules).

the channel matrix between the attacker and the receiver. The attacker can obtain $\mathbf{H_a}$ by leveraging channel reciprocity. However, the attacker does not have access to the target wireless system and therefore does not know modality Q, modulation scheme C, Y_t^Q, $\mathbf{H_t}$, and JSCC models. One method to address such a lack of knowledge is to utilize a set of surrogate models with different configurations (i.e., Q, C, λ) and diverse channel matrix $\mathbf{H_t}$. Specifically, we seek to generate Y_t^Q from a set of surrogate JSCC models, train UAPs using ensemble learning, and transfer the learned UAPs to the target system. During this offline UAP training, we randomly sample $\mathbf{H_t}$ from multipath fading model to make $\boldsymbol{\delta}^s$ channel-agnostic.

Using Equation (4), the receiver in the surrogate model then feeds this perturbed signal $\bar{Y}_t^Q$ to the remaining physical layer elements to reconstruct the source with modality Q as:

$$\bar{x}_t^Q = \mathcal{F}_{Q,C,\lambda}(\bar{Y}_t^Q, \bar{\mathcal{B}}_t^Q), \tag{6}$$

where $\bar{\mathcal{B}}_t^Q$ is the perturbed decoded frame buffer to be used in the video JSCC model. $\bar{\mathcal{B}}_t^{\mathcal{V}}=\{\bar{x}_{m_\sigma(1)}^{\mathcal{V}}, \cdots, \bar{x}_{m_\sigma(t-1)}^{\mathcal{V}}\}$, where $\bar{\mathcal{B}}_t^{\mathcal{V}}=\emptyset$ when t=1. $\bar{\mathcal{B}}_t^Q=\emptyset$ for other modalities.

As mentioned before, we aim to find the adversarial signals in a limited-knowledge setting (§IV-C). A representative way to handle this is to exploit the fact that adversarial examples exhibit good transferability between different ML models [12], [54], [58]. By adopting the attack transferability, we first train a surrogate JSCC model for each modality using publicly available datasets that have different distributions from the target model's training data. Then we use an ensemble learning approach to find a modality-agnostic adversarial perturbation $\boldsymbol{\delta}^s$ by solving the following optimization problem:

$$\arg\max_{\boldsymbol{\delta}^s}[\sum_{w\in\Psi^s}\mathcal{L}(w)],\ \text{s.t.}\ \|\boldsymbol{\delta}^s\|_2<\epsilon, \tag{7}$$

where Ψ^s is a set of all wireless signals that can be created by physical layer elements. $\mathcal{L}(w)$ is the loss function of ML-based JSCC model when w is sampled from Ψ^s.

However, this attack formulation is not suitable for making the UAPs physically realizable for the following reasons. First, having a single $\boldsymbol{\delta}^s$ as the UAP allows the receiver to estimate the perturbation signal using OFDM pilot signals, resulting in low robustness and persistence of adversarial attacks. Second, the adversary has no prior knowledge of the number of OFDM symbols in the target signal and thus is unable to define $\boldsymbol{\delta}^s$ as a matrix of the same size as the transmitted signal. Third, the video JSCC model has a network structure that forms a temporal chain between all video frames within the same GOP, so the model encodes current source data based on previous encoding results. This constructs the inter-frame dependency within a video sequence and it should be considered in crafting the UAPs. Fourth, the adversary does not know the distribution of the channel inputs carried by each OFDM subcarrier. Finally, when the perturbation signal overlaps with the benign signal, time or phase offsets may occur.

Practical Attack Formulation. To address the problem of Equation (7), we construct a Perturbation Generator Model (PGM) $G(z_t) = \boldsymbol{\delta}_t^u$ that generates a UAP signal by receiving a random trigger z_t at time step t. We adopt a ResNet-based generator [27]. The adversary changes z_t and injects a new perturbation signal into the target channel each time. Compared with using a single $\boldsymbol{\delta}^s$ as the UAP, the adversary creates an extremely large set of perturbations, which makes it difficult for the receiver to predict the perturbations. The following equation holds for frequency-domain complex-valued symbols at the receiver in the attacker's surrogate models:

$$\bar{Y}_t^Q[i,k] = \mathbf{H_t}[k]Y_t^Q[i,k] + \mathbf{H_a}[k]P_\tau(\boldsymbol{\delta}_t^u)[i,k] + W[i,k], \tag{8}$$

where $\boldsymbol{\delta}_t^u \in \mathbb{C}^{N_g \times N_{fft}}$ denotes a UAP which contains N_g data symbols. Since the attacker does not know the number of target symbols, N_g may not be equal to N_s. We define a novel transformation function P_τ which enables the PGM-generated wireless signals to model the distribution of real wireless data. The transformation function consists of several steps: 1) symbol extension model, 2) symbol shuffling model, 3) time rotation, and 4) frequency rotation. The symbol extension model concatenates multiple PGM-generated perturbations such that the symbol-extended perturbations can perturb all OFDM symbols of the target radio signal. The symbol shuffling model makes our attack robust against unknown target symbols by randomly shuffling symbols between the OFDM subcarriers of the adversarial signal. The time and phase rotation changes the offset of the adversarial signal during offline training so that the adversarial signals are agnostic to random time and phase shifts in the real world. We also incorporate the power normalization into the transformation to make Magmaw undetectable from natural noise. The wireless properties controlled by the transformation function are parameterized with τ. Figure 3 shows all the modules included in the transformation function. With the help of P_τ, the PGM can be optimized to produce the perturbation signals

Algorithm 1 Magmaw

Input: Dataset $\mathbb{T}^Q$, Surrogate JSCC model, Power constraint ϵ
Output: PGM $G(\cdot)$
for epoch $l <$ MaxIter **do**
  for each modality $Q \in \{\mathcal{I}, \mathcal{V}, \mathcal{S}, \mathcal{T}\}$ **do**
    for each batch $\mathbf{B}^Q \in \mathbb{T}^Q$ **do**
      $C, \lambda \leftarrow$ is sampled uniformly from candidates
      $\mathbf{H_t}$ is randomly sampled from channel model
      $\mathbf{H_a}$ is sampled uniformly from training set
      if $Q = \mathcal{V}$ **then**
        for $\mathbf{x}_t^{\mathcal{V}} \in \mathbf{B}^{\mathcal{V}} (= \{\mathbf{x}_1^{\mathcal{V}}, \cdots, \mathbf{x}_P^{\mathcal{V}}\})$ **do**
          $Y_t^{\mathcal{V}} \leftarrow$ Equation (2)
          $\mathcal{B}_t^{\mathcal{V}}$.append($\bar{x}_t^{\mathcal{V}}$)
          $z_t \sim \text{Uniform}(0,1), z_t' \sim \text{Uniform}(0,1)$
          $\tau \leftarrow$ uniformly at random
          $\bar{Y}_t^{\mathcal{V}}[i,k] \leftarrow$ Equation (8)
          $\bar{x}_t^{\mathcal{V}} \leftarrow$ Equation (6)
          $\bar{\mathcal{B}}_t^{\mathcal{V}}$.append($\bar{x}_t^{\mathcal{V}}$)
      else
        $Y_t^Q \leftarrow$ Equation (2)
        $z_t \sim \text{Uniform}(0,1), z_t' \sim \text{Uniform}(0,1)$
        $\tau \leftarrow$ uniformly at random
        $\bar{Y}_t^Q[i,k] \leftarrow$ Equation (8)
        $\bar{x}_t^Q \leftarrow$ Equation (6)
      Update PGM G and $\mathcal{D}$ by solving Equation (13)
Return: PGM G

that are resilient to real-world transformations. In §VI-B, we explain the internal mechanisms of P_τ.

We define an optimization problem to train the PGM that generates a hardware-implementable perturbation signal as:

$$\arg\max_G \quad \mathop{\mathbb{E}}_{z_t \sim p_z}[\sum_{w \in \Psi^u} \mathcal{L}_{rx}(z_t, w)],$$
$$\mathcal{L}_{rx}(z_t, w) = \begin{cases} \mathcal{L}_{mse}(x_t^{\mathcal{I}}, \bar{x}_t^{\mathcal{I}}), & \text{if } Q = \mathcal{I}, \\ \sum_{t=m_\sigma(1)}^{m_\sigma(P)} \mathcal{L}_{mse}(x_t^{\mathcal{V}}, \bar{x}_t^{\mathcal{V}}), & \text{if } Q = \mathcal{V}, \\ \mathcal{L}_{mse}(H_F(x_t^{\mathcal{S}}), H_F(\bar{x}_t^{\mathcal{S}})), & \text{if } Q = \mathcal{S}, \\ \mathcal{L}_{CE}(H_G(x_t^{\mathcal{T}}), H_G(\bar{x}_t^{\mathcal{T}})), & \text{if } Q = \mathcal{T}, \end{cases} \tag{9}$$

where Ψ^u is a set containing all radio signals that can be generated by the surrogate ML models. The perturbed signals at the receiver are computed from Equation (6). We use mean-squared error (MSE) loss as the distortion function $\mathcal{L}_{mse}$. We train the PGM to maximize distortion on a frame-by-frame basis for the image JSCC model. For the video JSCC model, we consider the inter-frame dependency between adjacent frames as the sum of the distortions over all frames within the GOP. This allows the PGM to adapt to any GOP without the need to reconfigure the attack. As for speech, we transform the speech data into a one-dimensional vector via the deframing function H_F before the loss is calculated. Since the text JSCC model completes sentence restoration by sequentially finding the probabilities that words will appear with a greedy decoder H_G, we use a cross-entropy loss $\mathcal{L}_{CE}$ between the predicted sentence $H_G(\bar{x}_t^{\mathcal{T}})$ and the ground truth sentence $H_G(x_t^{\mathcal{T}})$.

Downstream Attack Formulation. Figure 3 depicts downstream tasks appended to the wireless communication pipeline. We consider two ML models as examples: 1) VC and 2) AVE. Let F^N denote a discriminant function for the receiver's downstream task $N \in \{\text{VC}, \text{AVE}\}$. After the receiver demodulates incoming perturbed signals into data, the discriminant function takes the data $\bar{X}_N$ and outputs a probability distribution over a set K_N of class labels. Note that the VC takes a video clip $\bar{X}_{\text{VC}} = \{\bar{x}_t^{\mathcal{V}}\}_{t=1}^T$ consisting of T consecutive frames and the AVE receives $\bar{X}_{\text{AVE}} = \{\bar{x}_t^{\mathcal{I}}, \bar{x}_t^{\mathcal{S}}\}$ as two inputs. A classifier for task N, $\mathcal{C}^N$, points $\bar{X}_N$ to the class with the maximum probability: $\mathcal{C}^N(\bar{X}_N) = \arg\max_{c \in K_N} F_c^N(\bar{X}_N)$, where F_c^N is the probability of the perturbed input belonging to a specific class c. We define a loss $\mathcal{L}_{cls}^N$ to subvert classifiers:

$$\mathcal{L}_{cls}^N = \max_{c \neq \mathcal{C}^N(\hat{X}_N)} F_c^N(\bar{X}_N) - F_{\mathcal{C}^N(\hat{X}_N)}^N(\bar{X}_N), \tag{10}$$

where $\hat{X}_N$ denotes the reconstructed data when there is no attack. $\hat{X}_{\text{VC}} = \{\hat{x}_t^{\mathcal{V}}\}_{t=1}^T$ and $\hat{X}_{\text{AVE}} = \{\hat{x}_t^{\mathcal{I}}, \hat{x}_t^{\mathcal{S}}\}$. The attack succeeds when $\mathcal{L}_{cls}^N > 0$. With the ensemble learning, we find UAPs that maximize $\mathcal{L}_{cls}^N$ for the surrogate model with different architectures from the target model. We then fool the downstream services by transferring the attacks calculated from the surrogate model to the target model.

Stealthy Attack Formulation. Existing works [12], [54] have a problem that an adaptive defender can devise an anomaly classifier [93] that identifies the attacks by analyzing the perturbation's statistical behavior. To enforce the generator to produce undetectable perturbations, we explicitly regularize our PGM with the discriminative loss [33]:

$$\mathcal{L}_{ds} = \log \mathcal{D}(M_C(\mathcal{R}(\hat{Y}_t^Q))) + \log(1 - \mathcal{D}(M_C(\mathcal{R}(\bar{Y}_t^Q)))), \tag{11}$$

where $\mathcal{D}$ is a discriminator [37] that distinguishes clean signals from perturbed signals. We aim to minimize $\mathcal{L}_{ds}$ for forcing our PGM to explore the latent space and discover robust adversarial examples. To guarantee that the PGM properly produces the diversified perturbation, we utilize the diversity sensitive loss [27]:

$$\mathcal{L}_{dv} = \mathop{\mathbb{E}}_{z_t, z_t'}[\|G(z_t) - G(z_t')\|_1], \tag{12}$$

where z_t and z_t' are two different random latent codes.

Unified Attack Formulation. Finally, we integrate all losses into the objective function so that UAPs generated by the PGM can perturb wireless communication and downstream services simultaneously. Specifically, our goal is to solve the following objective function:

$$\max_G \min_{\mathcal{D}} \mathop{\mathbb{E}}_{z_t}[\sum_{w \in \Psi^u} [\mathcal{L}_{rx} + \sum_{N \in \mathcal{N}} \beta_{cls}^N \mathcal{L}_{cls}^N - \beta_{ds} \mathcal{L}_{ds}]] + \beta_{dv} \mathcal{L}_{dv}, \tag{13}$$

where β_{cls}^N, β_{ds}, and β_{dv} weigh the relative importance of each term and $\mathcal{N} = \{\text{VC}, \text{AVE}\}$. PGM generates a perturbation conditioned on the latent code and multiple controllable parameters of the wireless protocols, while $\mathcal{D}$ tries to distinguish between perturbed and clean signals.

In Algorithm 1, we outline the training process. Please refer to §VII-A for the parameters selected in the experiment. Our goal is to train the PGM G that generates UAPs to subvert ML-based JSCC models. We ensemble outputs of multimodal JSCC models to find generalizable adversarial signals that can transfer between modalities and protocols. The ML model used for training is a surrogate model that is different from the target model. We utilize the transformation function P_τ to change the outputs of PGM to practically feasible adversarial signals. At each training iteration, the algorithm selects a batch from the training dataset $\mathbb{T}$ with a different distribution from the training dataset of the target model. We ensure that the PGM learns effective UAPs leveraging ensemble learning, which integrates a set of JSCC models with different Q, C, λ. The loss values

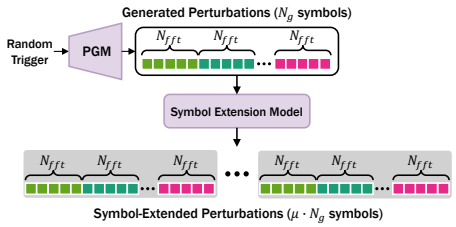


Fig. 4: Symbol extension mechanism. The perturbation length is extended to match the maximum length of the target signal.

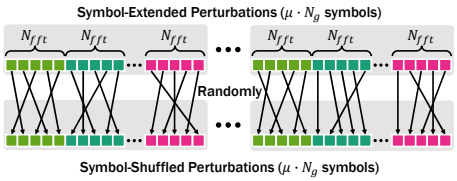


Fig. 5: Symbol shuffling mechanism. The complex-valued symbols assigned to the subcarriers are randomly shuffled.

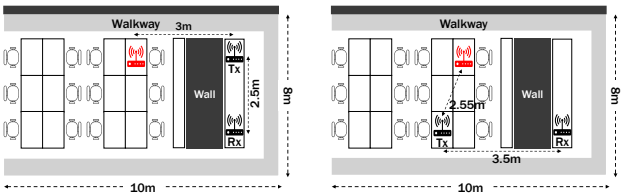
derived from each JSCC model are jointly backpropagated to optimize the PGM using the Adam optimizer [44]. As a result, we solve four technical challenges described in §I: (1) multi-modality and unknown $\mathbf{H_t}$, (2) unknown protocols, (3) de-synchronization, and (4) susceptibility to adaptive defense.

B. Design of Our Transformation Function

To cope with challenging real-world scenarios, the adversary should craft input-agnostic UAP signals regardless of synchronization with the legitimate receiver. The transformation function P_τ helps PGM learn to produce perturbations with a distribution similar to that of adversarial signals that can be realized in the real environment. Therefore, our adversarial signals are agnostic to 1) inconsistency of the number of data symbols between the benign signal and the adversarial signal, 2) unknown symbol allocation across the OFDM subcarriers, 3) time misalignment, and 4) unknown phase rotation. We additionally include a power regularization for undetectability. The modules included in the transformation function are shown in Figure 3 and detailed below.

Symbol Extension Model. The number of OFDM symbols varies greatly depending on the modality and coding rate. Furthermore, the coding rate of the JSCC encoder determines the amount of data compressed. In an online attack, the modality and coding rate are unknown. This leads the adversary to make the attack signal invariant to the number of OFDM symbols contained in the target signal. As the information about the coding rate is publicly available (see §IV-C), we can find the maximum value of N_s. As shown in Figure 4, we concatenate the PGM-generated signal multiple times through function $K(\cdot)$ such that $\mu \cdot N_g$ is equal to the maximum value of N_s, where μ is a parameter to adjust the number of symbols. Hence, our symbol-extended perturbations can perturb all symbols of the target signal without prior knowledge of the target signal's symbol count. Then, we ensure that the concatenated perturbations achieve high generalizability for multiple coding rates. In Algorithm 1, we randomly select the coding rate λ for each training epoch.

Symbol Shuffling Model. Previous works [39], [54] make the assumption that the adversary knows how the target wireless system allocates symbols to each subcarrier. This is infeasible in practice, because standard wireless communications often randomize the allocation to prevent consecutive repetition of the same symbols. Our adversary aims to make a subcarrier-invariant perturbation that is universally applicable to any symbol distribution of subcarriers. We define a function $\Gamma(\cdot)$ that randomly shuffles the symbols assigned to the subcarrier based on a seed ζ, as shown in Figure 5. Consequently, we train the PGM to generate the attack signal that is robust to the unknown symbol distribution across OFDM subcarriers.



(a) LoS Tx/Rx path (b) NLoS Tx/Rx path

Fig. 6: Experimental settings established in Magmaw.

Time and Frequency Rotation. Due to time and frequency misalignment, a random phase rotation occurs in each OFDM subcarrier. In order to enforce our perturbation to learn shift-invariant properties, we employ a phase rotation function $e^{-j2\pi f_k \Delta t + j\phi}$ from the previous approaches [12], [54], [68], where Δt and ϕ are time difference and phase offset between the benign signal and the adversarial perturbation, respectively.

Power Normalization. $\mathcal{M}(\cdot)$ is a power normalization function that adjusts the perturbation signal according to ϵ, which is the upper bound on the attacker's signal power. We follow the existing power remapping function [12] to preserve the power constraint of the perturbations as follows:

$$\mathcal{M}(\boldsymbol{\gamma}_t^u, \epsilon) = \begin{cases} \sqrt{\epsilon}\frac{\boldsymbol{\gamma}_t^u}{\|\boldsymbol{\gamma}_t^u\|_2}, & \|\boldsymbol{\gamma}_t^u\|_2^2 > \epsilon, \\ \boldsymbol{\gamma}_t^u, & \|\boldsymbol{\gamma}_t^u\|_2^2 \leq \epsilon. \end{cases} \tag{14}$$

where PSR is the ratio of the power of the attack signal to the power of the victim signal. ϵ is defined as $\|\hat{y}_t\|_2^2 \cdot 10^{\text{PSR}/10}$, where $\hat{y}_t$ is the time-domain signal in Equation (3). $\boldsymbol{\gamma}_t^u$ is the output of the symbol extension and symbol shuffling models.

Transformation Function. Consequently, we obtain the converted perturbation signal transmitted from the k-th subcarrier of the i-th OFDM symbol through the transformation function $P_{\mu,\zeta,\epsilon,\phi,\Delta t}(\cdot)$ as follows:

$$P_{\mu,\zeta,\epsilon,\phi,\Delta t}(\boldsymbol{\delta}_t)[i,k] = \mathcal{M}(\boldsymbol{\gamma}_t^u, \epsilon)[i,k] e^{j\phi} e^{-j2\pi f_k \Delta t}, \quad \text{where } \boldsymbol{\gamma}_t^u = \Gamma(K(\boldsymbol{\delta}_t^u, \mu), \zeta). \tag{15}$$

Here, the transformation function is controlled by various parameters $\mu, \zeta, \epsilon, \phi, \Delta t$.

C. Hardware Implementation

Figure 6 shows real-world attack scenarios in which the attacker (red device) sends a perturbation signal to the receiver. To thoroughly study radio signal propagation, we classify the physical environment into Line Of Sight (LoS) or NLoS (Non Line of Sight) between the transmitter and receiver. We obtain the experimental results from both Tx-Rx scenarios and then indicate the distribution of the results. We further showcase the difference in efficiency for each scenario in §VII-D.

Target Wireless System. We first implement the ML-based wireless communication system depicted in Figure 3 through

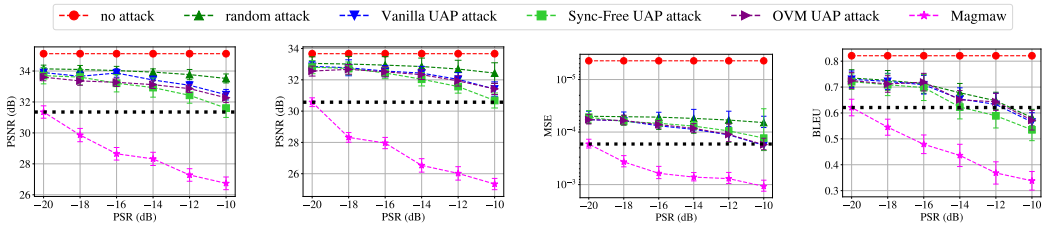
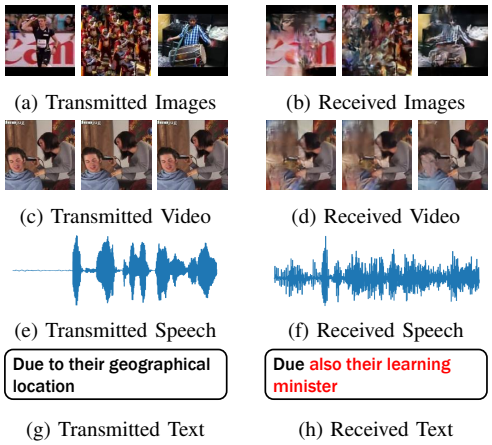


Image Transmission [98] Video Transmission [82] Speech Transmission [88] Text Transmission [90]

Fig. 7: Magmaw on ML-based wireless communication systems (i.e., modality-specific JSCC models).



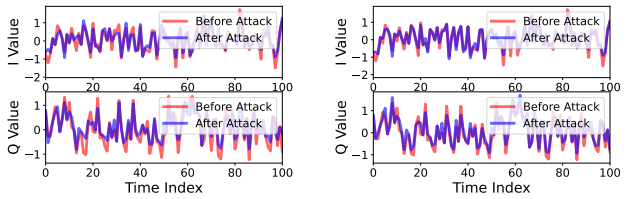
(a) Transmitted Images (b) Received Images

(c) Transmitted Video (d) Received Video

(e) Transmitted Speech (f) Received Speech

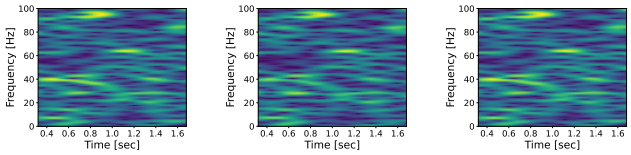
(g) Transmitted Text (h) Received Text

Fig. 8: Visualization of attack effects on multimodal JSCC.



(a) PSR is -10dB (b) PSR is -16dB

Fig. 9: Effect of perturbation when modulation is 16-QAM.



(a) Before Attack (b) PSR = -10dB (c) PSR = -20dB

Fig. 10: CSI heatmap when the sampling rate is 200Hz.

USRP B210, a software-defined radio widely used in designing wireless communication systems. We drive the USRP B210 using GNURadio software package [17] that provides a graphical programming interface for configuring transceivers and allows us to model the customized blocks. The transmitter and receiver consist of a USRP B210 and a Linux laptop, respectively, and they communicate through a single antenna, where the carrier frequency is set to 2.4 GHz. The number of cyclic prefixes and subcarriers L_{fft} is 16 and 64, respectively. Of the 64 subcarriers, 48 are used to carry symbols for ML-based JSCC, 4 of which are used for pilot symbols.

Attack System. We build an adversarial transmitter using a USRP N310 device with a single antenna and a Linux desktop. We randomly move the antenna to collect 2000 random realizations of the channel $\{\mathbf{H}_{\mathbf{a}}^{l}\}_{l=1}^{2000}$ between the adversarial transmitter and receiver. Following the previous work [12], we set the range of PSR to [-20,-10] dB. To perform the UAP attack, we adopt surrogate models with different architectures and parameters from the target wireless communication system and the downstream classifier. We train the PGM offline according to the Algorithm 1. The hyperparameters $(\beta_{cls}^{\text{VC}}, \beta_{cls}^{\text{AVE}}, \beta_{ds}, \beta_{dv})$ are all set to 1.

VII. Attack Evaluation

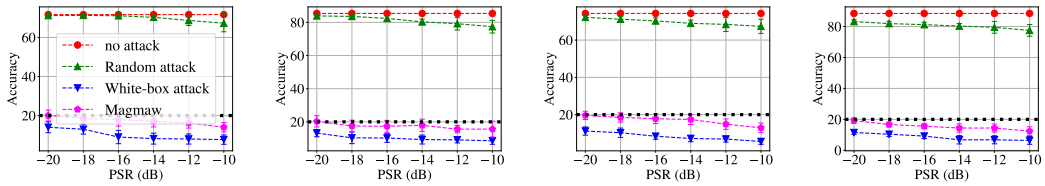
A. Experimental Setup

ML Models. We consider four state-of-the-art JSCC models that deliver multimodal data over the wireless channel and re-implement them based on several open-source resources [90], [98]. In Appendix Table III, we show the surrogate JSCC models[1]. We use constellation mapping schemes $C \in \{\text{QPSK, 16-QAM, 64-QAM}\}$ adopted in wireless standards and coding rates $\lambda \in \{\frac{1}{6}, \frac{1}{12}\}$ chosen from existing literature. The coding rate is computed as channel usage per source [98]. Note that each JSCC model has different model weights based on the variations of C and λ.

Downstream Tasks. We also consider scenarios where the receiver applies the demodulated data to ML-based downstream services, such as VC and AVE. For the VC task, we benchmark three state-of-the-art models, namely, I3D [22], SlowFast [30] and TPN [95]. As a benchmark model for a multimodal task, we choose the AVE proposed by [78]. Appendix Table IV depicts surrogate models to craft transferable attacks.

Dataset. We choose popular multimodal datasets to train and evaluate JSCC models. For training the image and video JSCC models, we adopt the Vimeo90K dataset [94], which is widely used in evaluating image and video processing tasks. To facilitate efficient training, the video sequences are cropped to a resolution of 256×256. We then evaluate the image and video JSCC models using the UCF-101 dataset [74]. For the speech JSCC model, we use the speech dataset from Edinburgh DataShare [80], which contains more than 10,000 training data and 800 test data with a sampling rate of 16 KHz. We truncate the speech sample sequence to have 128 frames with a frame length of 128 after framing. For the text JSCC model, we select the proceedings of the European Parliament, which includes about 2 million sentences and 53 million words. We pre-process the dataset to have sentence lengths between 4 and 30 words. We then split it into training and test sets. We also select widely used datasets as benchmarks to evaluate VC and AVE downstream tasks. We adopt the UCF-101 human activity dataset [74] to verify Magmaw on the VC model. For

[1]We also evaluate how less similar surrogate models affect the attack performance in Appendix B.



I3D [22] SlowFast [30] TPN [95] AVE [78]

Fig. 11: Magmaw on ML-based downstream classification tasks (i.e., VC and AVE).

evaluating the AVE model, we adopt the audio-visual event dataset [77] which contains 4,143 video clips with 28 events.

Evaluation Metrics. We use evaluation metrics that effectively reflect the semantic information of each modality. In the image and video domains, we select the PSNR as the representative picture quality measurement. In the speech domain, the MSE reflects the quality of the received speech. For the text domain, the BLEU score [14] is widely used to compare the difference between the original sentence and the reconstructed one. We measure the experimental results from the two Tx-Rx scenarios (see Figure 6), and then plot the distribution of the results in the figure. We use the black dotted line as the quality threshold for each experimental result, indicating that the result below it is not properly restored, which can pose a serious threat to back-end users.

Baseline Attacks. We compare Magmaw with four types of baseline attacks: (1) Random Attack, (2) Vanilla UAP Attack, (3) Sync-Free UAP Attack, and (4) One-hot Vector Modality-based (OVM) UAP Attack [12]. We design the random attack to transmit randomly sampled Gaussian noise into the air. It resembles classic jamming, as Gaussian jamming is widely used [32]. The vanilla UAP is an entry-level attack where multi-modality, protocol, and synchronization are not considered in crafting perturbations. The sync-free UAP attacker knows the perturbation undergoes time and phase shifts and tries to exploit such knowledge to devise shift-invariant attacks. Following previous work [12], OVM UAP is trained with a dataset consisting of one-hot vector messages. For downstream tasks, we compare Magmaw to random and white-box attacks.

B. Attacks against Multimodal JSCC

Analysis of Magmaw. Figure 7 presents the reconstruction performance of the ML-based wireless transmission systems under adversarial attacks. We sweep PSR from -20dB to -10dB with steps of 2dB. We compare the performance of Magmaw to that of the baseline attacks. As shown in Figure 7, Magmaw dramatically deteriorates the performance metrics in the range of all PSRs. Note that "no attack" shows the original performance of the benign model. When applying the adversarial attacks on the image JSCC model, the PSNR drops by up to 8.04dB. For the video JSCC model, PSNR is lowered by 8.29dB on average by Magmaw. We see that the video model is more vulnerable to our adversarial signals than the image JSCC model. The main reason is that the video JSCC model encodes the current frame based on the previously decoded frame, thus propagating the reconstruction distortion to the next frame. For the speech model, we find that Magmaw degrades MSE loss by 3.91 times more than the baseline. We also observe that the BLEU score of the text JSCC model drops to a minimum of 0.338 points under Magmaw.

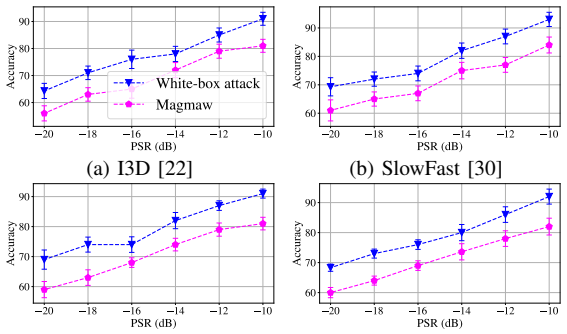
Comparison with Baselines. As depicted in Figure 7, Magmaw outperforms the baselines by a large margin. Against the image JSCC model, Magmaw lowers PSNR by up to 5.68dB more than the vanilla UAP attack and up to 4.85dB more than the sync-free UAP attack. We see that OVM UAP attacks have similar results to random attacks. Without considering the multi-modality, wireless protocols, and vulnerabilities of the model, the evaluated baselines cannot critically hurt the JSCC.

Attack Visualization. As shown in Figure 8, we visualize the attack effect on the multimodal data reconstruction at the receiver. As seen, the JSCC decoder fails to retain semantic information. Specifically, the restored images and videos have noise-like artifacts, which dramatically reduce the users' quality of experience (QoE). Furthermore, the user cannot hear the speaker's voice in a speech sequence due to noticeable noise. The text JSCC decoder generates sentences with incorrect grammar and context, so the user cannot understand the sender's message. In Figure 9, we present the differences in complex-valued symbols before and after the attack. We observe that Magmaw's low PSR results in minimal changes to the original signal. Additionally, as shown in Figure 10, the variation in channel state information (CSI) due to perturbation is extremely low.

Analysis of Modulation. In Appendix Figure 22, we further demonstrate the attack results of Magmaw for different constellation mapping methods. We confirm that Magmaw severely degrades the performance of JSCC models regardless of constellation type. As 64-QAM has slightly higher recovery performance than other modulations (16-QAM, QPSK) in all modalities, we confirm that the higher order of the modulation helps to increase the robustness.

C. Attacks against Downstream Tasks

Analysis of Magmaw. We evaluate the accuracy of each classifier when Magmaw is directed to a downstream classifier. Then, we provide a comparison with other baseline attacks. Figure 11 shows the attack results for the video classifiers I3D [22], SlowFast [30], and TPN [95] and the audio-visual event classifier AVE [78]. We compare the performance of Magmaw to white-box and random attack scenarios. We see that the changes made in random attacks are not optimized to subvert the model. In the white-box attack scenario, the attacker has complete knowledge of the classification model. Figure 11 presents the accuracy of each baseline for different PSRs. As shown, transmitting randomly sampled perturbations performs very poorly compared to Magmaw. As our attack consistently achieves comparable attack performance compared to the white-box attacks, we confirm that our UAP signals are successfully transferable to unseen downstream models. Specifically, Magmaw achieves an average attack



(a) I3D [22] (b) SlowFast [30]

(c) TPN [95] (d) AVE [78]

Fig. 12: Results for targeted UAPs on downstream tasks.

success rate of 81.6%, which is only 8.7% lower on average than white-box attacks.

Analysis of Modulation. To analyze the influence of different constellation mapping techniques on the downstream tasks, we illustrate the attack results on the downstream classifiers when different constellation mapping methods are applied to ML-based wireless communication systems in Appendix Figure 22. Although 64-QAM can increase accuracy slightly more than other modulations, we observe that our protocol-agnostic attack defeats all modulation techniques.

Analysis of Targeted Attacks. We investigate targeted UAPs aimed at flipping the prediction of inputs to a target class. To accomplish this, we define the loss function as below:

$$\mathcal{L}_{cls}^{N} = F_{c^*}^{N}(\bar{X}_N) - \max_{c \neq c^*} F_{c}^{N}(\bar{X}_N), \tag{16}$$

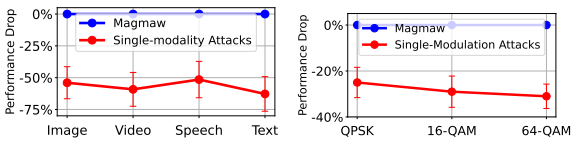
where c^* is a target class. We train the PGM by replacing $\mathcal{L}_{cls}^{N}$ in Equation (13). Targeted attacks gain success if and only if $\mathcal{L}_{cls}^{N} > 0$. As shown in Figure 12, the targeted UAPs achieve up to 82% accuracy in AVE when PSR is -10dB. Compared to untargeted UAPs, the fooling ratio is relatively low because it is more challenging to trick the predictions of all samples into a specific class [100].

D. Ablation Study

Impact of Multi-Modality. To understand the importance, we study the transferability of adversarial perturbations between different modalities. For each modality, we learn a modality-specific perturbation signal and then conduct an experiment in which we inject the learned perturbation into the radio signals of other modalities. As shown in Figure 13 (a), we see that the lack of learning generalized adversarial features limits both the cross-modal and cross-model transferability.

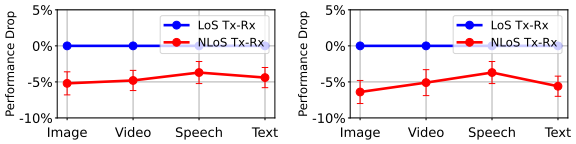
Effect of Modulation. To verify the effectiveness of protocol-agnostic attacks, we conduct an ablation study on attacking JSCC without considering the constellation mapping method. As shown in Figure 13 (b), eliminating knowledge of the physical layer protocol has a significant impact on the effectiveness of the attack. We enable the transferability of adversarial examples by creating diverse modulated signals.

Impact of Tx-Rx Placement. Each Tx-Rx scenario has different amounts of multipath because the power via the LoS path is stronger than power via the reflection path. To investigate the influence of multipath, we first compare the performance of JSCC in the two scenarios when there is no attack. As shown in Figure 14 (a), we see that the NLoS path makes



(a) Modality (b) Modulation Scheme

Fig. 13: Visualization of reduced attack performance when the attacker doesn't consider modality or modulation.



(a) Before Attack (b) After Attack

Fig. 14: Impact of Tx-Rx placement. We measure the performance degradation of JSCC on NLoS paths compared to the performance of JSCC on LoS paths.

the interference issue in wireless communication, reducing the performance of JSCC by 5%. We then inject our perturbations into the channel to analyze the effect of the NLoS path. As shown in Figure 14 (b), we confirm that Magmaw is effective regardless of the location of Tx-Rx. A slight decrease in attack performance when the Tx/Rx path is NLoS is due to the degradation of the original performance of JSCC.

VIII. Resiliency to Defense

The defense performance depends on what information the defender knows about the attack formulation. From §VIII-A to §VIII-C, we present multiple expert defenders who know the PGM's model architecture, the channel distribution between the attacker and the receiver, and attack mechanisms illustrated in Algorithm 1. In §VIII-D, we test Magmaw against an oracle defender who knows every detail about Magmaw.

A. Adversarial Training

The defender aims to obtain a robust ML-based JSCC model for each modality to protect the physical layer from the Magmaw. Since we assume that the defender knows the model architecture of the PGM, adversarial training extends the training dataset to include all adversarial examples and then trains a JSCC model on the augmented dataset. Algorithm 2 shows detailed steps of our adversarial training. We refer to the target JSCC models as $\mathcal{J}_{Q,C,\lambda}$, and denote the PGM as $\mathcal{G}$, which is identical to the attacker's model architecture but with different model parameters. The defender trains an ML-based JSCC model by selecting a batch from the training dataset $\mathbb{D}^{Q}$ and generating the adversarial signals controlled by several parameters of the transformation function P_{τ}. We then expand the training dataset to include all adversarial examples and train the model on the augmented training dataset.

ML-based Wireless System. We validate Magmaw against the ML-based wireless communication systems, whose resiliency has been improved by adversarial training. As shown in Figure 15 (a), incorporating adversarial examples inside the model training process results in a lower ability to restore source data even if the underlying victim model is not attacked. Moreover, we observe that adversarial training cannot protect ML-based wireless communication from Magmaw. The reason is that the JSCC model has to be trained on a huge set of perturbations that the defender generates with PGM. Yet it

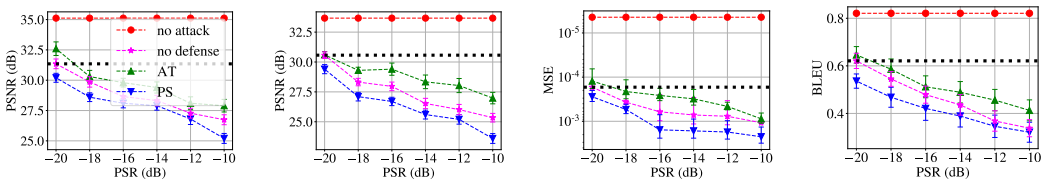
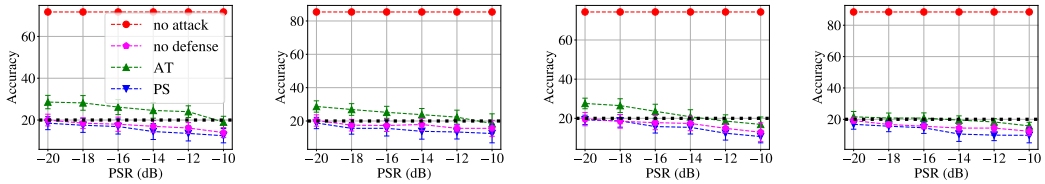


Image Transmission [98] Video Transmission [82] Speech Transmission [88] Text Transmission [90]

(a) ML-based Wireless Communication Systems



I3D [22] SlowFast [30] TPN [95] AVE [78]

(b) ML-based Downstream Classification Tasks

Fig. 15: Evaluation of defenses. AT and PS denote adversarial training and perturbation subtraction.

Algorithm 2 Adversarial Training against Magmaw

Input: Dataset $\mathbb{D}^Q$, ML-based JSCC model $\mathcal{J}_{Q,C,\lambda}$, PGM $\mathcal{G}$,
Output: Robust JSCC model $\mathcal{J}_{Q,C,\lambda}$
$Q \leftarrow$ Modality, $C \leftarrow$ Modulation, $\lambda \leftarrow$ Coding rate
for epoch $l <$ MaxIter **do**
  $\mathbf{H_t}$ is randomly sampled from channel model
  $\mathbf{H_a}$ is sampled uniformly from training set
  $\mathbb{B}^{adv} \leftarrow [\,]$
  for each batch $\mathbf{B}^Q \in \mathbb{D}^Q$ **do**
    Train the JSCC model $\mathcal{J}_{Q,C,\lambda}$ on $\mathbf{B}^Q$
    $z_t \sim \text{Uniform}(0,1)$
    $\tau_l \leftarrow$ randomly sampled $\{\mu, \zeta, \epsilon, \phi, \Delta t\}$
    Store $P_{\tau_l}(\mathcal{G}(z_t))$ in $\mathbb{B}^{adv}$ for each data in $\mathbf{B}^Q$
  $\mathbb{D}^Q$.append($\mathbb{D}^Q + \mathbb{B}^{adv}$)
Return: Robust JSCC model $\mathcal{J}_{Q,C,\lambda}$

is not feasible for the defender to train JSCC models that are resilient to all possible perturbations. Another reason is that the defender uses a PGM with different parameters from the attacker's model, so the distribution of adversarial signals generated by the two models is different.

Downstream Tasks. Figure 15 (b) shows the accuracy of the downstream models trained by adversarial training. Adversarial training significantly reduces the accuracy of benign models, hindering their applicability. We observe that Magmaw still achieves a high attack success rate even though the benign model undergoes adversarial training. This is because training a model that is universally robust to different types of perturbed signals, while being able to correctly classify input data, is a fundamentally challenging problem.

B. Perturbation Signal Subtraction

This defense scheme can be performed at the physical layer before the signal is passed through the OFDM receiver. Defenders aim to mitigate the effects of perturbations and reconstruct the originally transmitted signal. As we assume that the defender has knowledge of Magmaw's model architecture, the receiver generates a perturbation signal via the defender's PGM and then subtracts it from the received wireless signal.

ML-based Wireless System. The defense results are summarized in Figure 15 (a). We observe that the source data restored by each JSCC model is more degraded than before the defense. This is because the cancellation of the adversarial signal fails and further amplifies the power of the perturbation. Even if the defender knows the structure of the PGM, the defender cannot generate exactly the same perturbation signal if the model parameters of the PGM are different.

Downstream Tasks. As shown in Figure 15 (b), applying perturbation signal subtraction reduces the accuracy of the downstream services by an average of 3.6%. We see that the defender cannot increase the accuracy of the downstream classifier by simply subtracting an estimate of the perturbation. The accuracy of the classifier tends to depend heavily on the quality of the input source.

C. Adversarial Perturbation Detection

We define an input-level detection that aims to correctly find adversarially manipulated signals at the receiver side. The underlying hypothesis for this defense follows previous studies [93], [97] that show that UAPs may leave signatures observable by ML-based anomaly detection algorithms. Based on this, we design a perturbation detector [73] that acts as a discriminator to distinguish the clean signal $\hat{Y}_t^Q$ from the perturbed signal $\bar{Y}_t^Q$. Leveraging the trace of UAPs, we design the binary classifier as follows. First, we train the detector offline using the training dataset constructed from the defender's PGM. In the online process, we label the received signals as adversarial attacks when the efficiency of JSCC deteriorates and include them in the training data. Finally, we fine-tune a well-trained model with newly collected data.

Appendix Figure 23 (a) summarizes the detection accuracy and false positive rate of our perturbation detector. It shows that Magmaw can bypass detection, even though the fine-tuning improves the accuracy of the detector. This is because Magmaw is trained to generate perturbed signals, which are indistinguishable from the clean signal, as shown

TABLE II: Detection AUC of perturbation detection. The first row shows the result before fine-tuning, and the second row shows the result after fine-tuning.

	Image Signal	Video Signal	Speech Signal	Text Signal
Detection	53.2%	52.5%	52.8%	53.4%
AUC	55.6%	54.4%	56.5%	57.1%

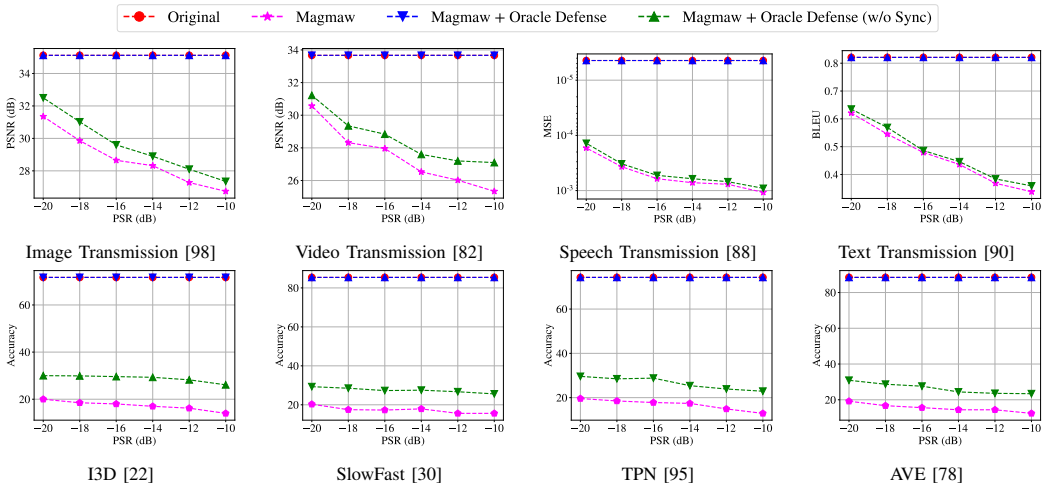


Fig. 16: Evaluation of Magmaw against oracle defenders in wireless systems and downstream tasks. The first and second rows are the results of JSCC and downstream tasks, respectively.

in Equation (11) and Equation (12). For example, the fine-tuned detector only obtains up to 12% accuracy to detect perturbed radio signals in the text transmission. The results in Appendix Figure 23 (b) have shown the detection rate of the perturbation detector when Magmaw conducted the training without regularization loss. The detector achieves about 75% detection rate after fine-tuning. We verify that ML-based detectors can offer strong generalization capability in distinguishing PGM-generated perturbations. In order to train undetectable and robust UAPs, we should leverage a discriminator to enforce stealthiness.

As shown in Table II, we report the Area Under Curve (AUC) of Receiver Operation Characteristic Curve (ROC) of the perturbation detection. The AUC metric shows the probability that the detector will assign a higher score to a perturbed signal than to a clean signal. We verify that the AUC results are close to the random guess, which means that Magmaw can achieve high undetectability. Another drawback of malware classifiers is that when an attacker changes position, the channel matrix between the attacker and the receiver also changes, requiring the defender to collect new datasets to adapt to the new environment.

D. Oracle Defender

It is crucial to identify the lower bound of the effectiveness of attacks [20]. We define two strong defenders as follows:

- *Oracle Defender* knows comprehensive details of Magmaw, including the PGM architecture and parameters, $\mathbf{H_a}$, the time and frequency offsets between Magmaw and receiver, and how Magmaw assigns symbols to OFDM subcarriers.
- *Oracle Defender without Sync Assumption* is aware of all details except the time/frequency offset. This is practical because otherwise, the receiver has to coordinate with the attacker to estimate the time/frequency offset and convey the information to the defender.

These defenders reconstruct the signal by removing the attack effect from the received wireless signal by utilizing the same perturbations generated by Magmaw.

(a) Results (b) Transmitted Image (c) Received Image

Fig. 17: Attack results on secure image communication. We visualize the attack results, input data, and reconstructed data at the receiver side.

ML-based Wireless System. The oracle defender can completely neutralize Magmaw, as shown in Figure 16. These results are consistent with those reported in [12], which also points out that this defense is not practical. We further measure defense performance by eliminating the assumption that the attacker and receiver are synchronized. The oracle defense without sync assumption can only reduce the efficiency of Magmaw by up to 21.42%. This is because lack of synchronization causes inaccuracies in the results of perturbation removal. See detailed results in Figure 16.

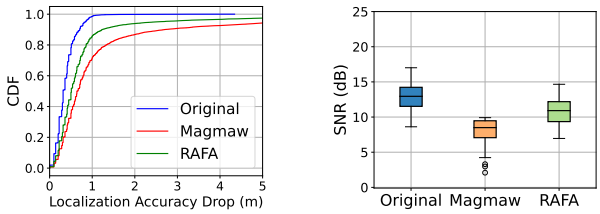
Downstream Tasks. We also investigate the adversarial robustness of downstream tasks in the presence of oracle defenders. We verify that addressing synchronization robustness is essential to increasing the effectiveness of the oracle defender. Specifically, without sync assumption, the oracle defender only improved the robustness of the classifier by at most 16.8%. Detailed results can be found in Figure 16.

IX. Case Study

A. Attacks on Encrypted Communication

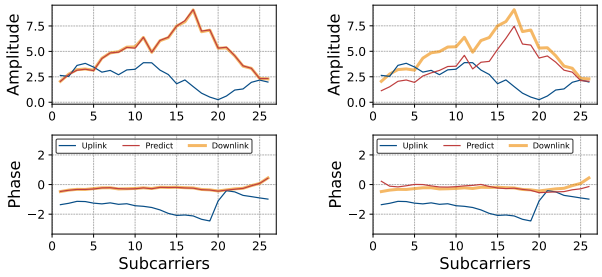
Encryption schemes are commonly applied in the communication pipeline to protect users' private data [79]. While the robustness of privacy-preserving communications with ML-based JSCC has not been investigated before, we add the encryption and decryption blocks in image JSCC to examine the impact of Magmaw on secure transmission, and analyze the vulnerability of encrypted signals.

Experiment Design. ML-based JSCC encoder directly maps the source to the complex-valued symbols without converting



(a) Localization (DLoc [11]) (b) Downlink Prediction (FIRE [53])

Fig. 18: Attack results on CSI-based models (PSR is -12dB).



(a) Before Attack (b) After Attack

Fig. 19: Attack results on FIRE [53]. We visualize the channel amplitude and phase. FIRE takes the uplink channel (blue line) as input and predicts the downlink channel (red line) that is expected to be the same as the ground truth (yellow line).

it to bits. To handle this new feature, public-key encryption with LWE [65] rather than classical AES-based schemes [57] is applied in JSCC [79]. In public-key encryption, any user can send encrypted messages to the receiver using the public key. Thus, we assume that the adversary knows the public key, but does not know the secret key.

Attack Results. Figure 17 shows the attack results on the secure communication system. We see that the OFDM symbols carrying the ciphertext of the image data are vulnerable to our perturbation signal. Specifically, Magmaw lowers the performance of secure image transmission by up to 5.88dB. This is because the decrypted output of ciphertext operations in LWE is similar to performing plaintext operations on the original plaintext data. By showing that secure communication does not provide adversarial robustness, we promote the need for new defense techniques against Magmaw.

B. Attacks on ML Systems with Channel States as Input

Standard-defined preambles [16] are widely used in ML-driven wireless systems to obtain the CSI. We consider two ML models that are also used as target models in RAFA [54]: (a) DLoc [11] performs localization task via CSI received from four fixed access points, and (b) FIRE [53] takes the CSI of the uplink channel as input and then predicts the downlink CSI. It can address the overhead of feedback exchange in the Frequency Domain Duplex (FDD) system.

Experiment Design. In the experiment setup, the sender allocates preambles $Y_t^{\mathcal{P}}$ to OFDM subcarriers (i.e., 64 subcarriers for 20MHz) and then transmits them to the receiver. Here, $\mathcal{P}$ denotes the preamble. Since Magmaw injects attack signals into the channel according to Equation (8), the received preamble is $\bar{Y}_t^{\mathcal{P}}$. Thus, the receiver acquires the perturbed CSI via $\mathcal{H}_t^{\mathcal{P}} = \bar{Y}_t^{\mathcal{P}} / Y_t^{\mathcal{P}}$ and feeds it into the target ML model. We re-implement the target models via details provided in their papers, as well as open source [10] and the dataset [71]. We further improve the robustness of DLoc and FIRE through adversarial training proposed by RAFA. For a fair comparison, we utilize surrogate models used in RAFA to train Magmaw.

Attack Results. We compare our attack with RAFA for a comprehensive evaluation. As shown in Figure 18 (a), DLoc achieves 0.71m and 1.03m localization errors at the 90th and 99th percentile. However, when Magmaw is present, the results go up to 2.7m and 9.8m. We can observe that Magmaw outperforms RAFA by 1.73× on average. Figure 18 (b) describes the accuracy of channel estimation by FIRE. The SNR measures the similarity between the estimated downlink channel and the ground truth. We confirm that Magmaw drops the SNR of the predicted channel by 3.8dB more than RAFA. The observed underperformance of RAFA can be due to the lack of consideration for improving the robustness of adversarial attacks during the training. In contrast, Magmaw achieves high robustness by leveraging a discriminator and diversity loss that increases the variability of perturbation patterns. To provide an intuition of the attack effect, we visualize an example of channel estimation in Figure 19. Figure 19 (b) shows that Magmaw causes subcarriers to have symbols that are significantly different from the actual symbols.

X. Conclusion and Future Works

We present Magmaw, a novel attack framework to subvert semantic communication for AI-native networks. Our results show that the Magmaw is feasible in the real world, and can degrade the performance of both wireless communication and downstream tasks simultaneously. Magmaw maintains a high attack success rate by evading several defenses. In case studies, we evaluate Magmaw on encrypted communication and CSI modality-based models, proving that Magmaw is transferable.

While Magmaw demonstrates great success, the perturbation designed in this work needs to be powered by software-defined radios for flexible generation of the UAPs. Promising future work is to explore new attack methods, e.g., intelligent reflecting surfaces [26], to induce small adversarial perturbations. Another area of future research is to establish practical defense techniques to prevent the proposed attacks.

Acknowledgement

We thank the NDSS reviewers for their valuable feedback. This work is partially supported by the U.S. Army/Department of Defense award number W911NF2020267 and Google Ph.D. Fellowship.

References

[1] Ahmed EAA Abdulla, Zubair Md Fadlullah, Hiroki Nishiyama, Nei Kato, Fumie Ono, and Ryu Miura. Toward fair maximization of energy efficiency in multiple uas-aided networks: A game-theoretic methodology. *IEEE Transactions on Wireless Communications*, 2014.

[2] Najah Abu-Ali, Abd-Elhamid M Taha, Mohamed Salah, and Hossam Hassanein. Uplink scheduling in lte and lte-advanced: Tutorial, survey and evaluation framework. *IEEE Communications surveys & tutorials*, 16(3):1239–1265, 2013.

[3] Alperen Acemoglu, Jan Krieglstein, Darwin G Caldwell, Francesco Mora, Luca Guastini, Matteo Trimarchi, et al. 5g robotic telesurgery: Remote transoral laser microsurgeries on a cadaver. *IEEE Transactions on Medical Robotics and Bionics*, 2(4):511–518, 2020.

[4] Abdullatif Albaseer, Bekir Sait Ciftler, and Mohamed M Abdallah. Performance evaluation of physical attacks against e2e autoencoder over rayleigh fading channel. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies*, pages 177–182, 2020.

[5] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

[6] Faycal Ait Aoudia, Jakob Hoydis, Alvaro Valcarce, and Harish Viswanathan. Toward a 6g ai-native air interface. https://onestore.nokia.com/asset/210299, 2021.

[7] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364. IEEE, 2023.

[8] Giovanni Apruzzese, Rodion Vladimirov, Aliya Tastemirova, and Pavel Laskov. Wild networks: Exposure of 5g network infrastructures to adversarial examples. *IEEE Transactions on Network and Service Management*, 19(4):5312–5332, 2022.

[9] Emekcan Aras, Nicolas Small, Gowri Sankar Ramachandran, Stéphane Delbruel, Wouter Joosen, and Danny Hughes. Selective jamming of lorawan using commodity hardware. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 363–372, 2017.

[10] Roshan Ayyalasomayajula. Dloc network architecture codes. https://github.com/ucsdwcsng/, 2024.

[11] Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Rajkumar Sethi, Deepak Vasisht, and Dinesh Bharadia. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

[12] Alireza Bahramali, Milad Nasr, Amir Houmansadr, Dennis Goeckel, and Don Towsley. Robust adversarial attacks against dnn-based wireless communication systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021.

[13] Sourangsu Banerji and Rahul Singha Chowdhury. On ieee 802.11: wireless lan technology. *arXiv preprint arXiv:1307.2661*, 2013.

[14] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[15] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

[16] A Biswas, S Lakshmipathi, and S Sandhu. Channel estimation techniques with long training sequence for ieee802. 11a. In *2004 International Conference on Signal Processing and Communications, 2004. SPCOM'04.*, pages 136–139. IEEE, 2004.

[17] Eric Blossom. Gnu radio: tools for exploring the radio frequency spectrum. *Linux journal*, 2004(122):4, 2004.

[18] Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, 2019.

[19] Yuxin Cao, Xi Xiao, Ruoxi Sun, Derui Wang, Minhui Xue, and Sheng Wen. Stylefool: Fooling video classification systems via style transfer. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2023.

[20] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[21] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[22] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] Jung-Woo Chang, Mojan Javaheripi, Seira Hidano, and Farinaz Koushanfar. Rovisq: Reduction of video service quality via adversarial attacks on deep learning-based video compression. In *NDSS*, 2023.

[24] Jung-Woo Chang, Mojan Javaheripi, and Farinaz Koushanfar. Videoflip: Adversarial bit flips for reducing video service quality. In *ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, 2023.

[25] Jung-Woo Chang, Nojan Sheybani, et al. Netflick: Adversarial flickering attacks on deep learning based video compression. *arXiv preprint arXiv:2304.01441*, 2023.

[26] Xingyu Chen, Zhengxiong Li, Baicheng Chen, Yi Zhu, Chris Xiaoxuan Lu, Zhengyu Peng, Feng Lin, Wenyao Xu, Kui Ren, and Chunming Qiao. Metawave: Attacking mmwave sensing with meta-material-enhanced tags. In *The 30th Network and Distributed System Security (NDSS) Symposium*, volume 2023, 2023.

[27] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[28] Mostafa Zaman Chowdhury, Md Shahjalal, Shakil Ahmed, and Yeong Min Jang. 6g wireless communication systems: Applications, requirements, technologies, challenges, and research directions. *IEEE Open Journal of the Communications Society*, 1:957–975, 2020.

[29] Simon Erni, Martin Kotuliak, Patrick Leu, Marc Roeschlin, and Srdjan Capkun. Adaptover: adaptive overshadowing attacks in cellular networks. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 743–755, 2022.

[30] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[31] Bryse Flowers, R Michael Buehrer, and William C Headley. Evaluating adversarial evasion attacks in the context of wireless communications. *IEEE Transactions on Information Forensics and Security*, 2019.

[32] Jie Gao, Sergiy A Vorobyov, Hai Jiang, and H Vincent Poor. Worst-case jamming on mimo gaussian channels. *IEEE Transactions on Signal Processing*, 63(21):5821–5836, 2015.

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[34] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards gan benchmarks which require generalization. *arXiv preprint arXiv:2001.03653*, 2020.

[35] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 802.11 with multiple antennas for dummies. *ACM SIGCOMM Computer Communication Review*, 40(1):19–25, 2010.

[36] Jianhua He, Kun Yang, and Hsiao-Hwa Chen. 6g cellular networks and connected autonomous vehicles. *IEEE Network*, 2020.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[38] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research. *arXiv preprint arXiv:2203.11854*, 2022.

[39] Qiyu Hu, Guangyi Zhang, Zhijin Qin, Yunlong Cai, Guanding Yu, and Geoffrey Ye Li. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 2023.

[40] Taewon Hwang, Chenyang Yang, Gang Wu, Shaoqian Li, and Geoffrey Ye Li. Ofdm and its wireless applications: A survey. *IEEE transactions on Vehicular Technology*, 58(4):1673–1694, 2008.

[41] Zizhi Jin, Xiaoyu Ji, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1822–1839. IEEE, 2023.

[42] Brian Kim, Y Sagduyu, Tugba Erpek, and Sennur Ulukus. Adversarial attacks on deep learning based mmwave beam prediction in 5g and beyond. In *IEEE Statistical Signal Processing Workshop (SSP)*, 2021.

[43] Brian Kim, Yalin E Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur Ulukus. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. *IEEE Transactions on Wireless Communications*, 21(6):3868–3880, 2021.

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] Changming Li, Mingjing Xu, Yicong Du, Limin Liu, Cong Shi, Yan Wang, Hongbo Liu, and Yingying Chen. Practical adversarial attack on

wifi sensing through unnoticeable communication packet perturbation. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 373–387, 2024.

[46] Haoran Li and Wei Lu. Mixed cross entropy loss for neural machine translation. In *International Conference on Machine Learning*, 2021.

[47] Zeju Li, Xinghan Liu, Guoshun Nan, Jinfei Zhou, Xinchen Lyu, Qimei Cui, and Xiaofeng Tao. Boosting physical layer black-box attacks with semantic adversaries in semantic communications. In *ICC 2023-IEEE International Conference on Communications*, 2023.

[48] Peng Lin, Qingyang Song, F Richard Yu, Dan Wang, Abbas Jamalipour, and Lei Guo. Wireless virtual reality in beyond 5g systems with the internet of intelligence. *IEEE Wireless Communications*, 28(2):70–77, 2021.

[49] Xingqin Lin. An overview of 5g advanced evolution in 3gpp release 18. *IEEE Communications Standards Magazine*, 6(3):77–83, 2022.

[50] Chenyao Liu, Jiejie Guo, Yimeng Zhang, Wenjun Xu, and Yiming Liu. Sst-v: A scalable semantic a scalable semantic transmission framework for video. *ZTE COMMUNICATIONS*, 21(2), 2023.

[51] Jianwei Liu, Yinghui He, Chaowei Xiao, Jinsong Han, Le Cheng, and Kui Ren. Physical-world attack towards wifi-based behavior recognition. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 400–409. IEEE, 2022.

[52] Jianwei Liu, Yinghui He, Chaowei Xiao, Jinsong Han, and Kui Ren. Time to think the security of wifi-based behavior recognition systems. *IEEE Transactions on Dependable and Secure Computing*, 2023.

[53] Zikun Liu, Gagandeep Singh, Chenren Xu, and Deepak Vasisht. Fire: enabling reciprocity for fdd mimo systems. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 628–641, 2021.

[54] Zikun Liu, Changming Xu, Emerson Sie, Gagandeep Singh, and Deepak Vasisht. Exploring practical vulnerabilities of machine learning-based wireless systems. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1801–1817, 2023.

[55] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1865–1882, 2021.

[56] BR Manoj, Meysam Sadeghi, and Erik G Larsson. Adversarial attacks on deep learning based power allocation in a massive mimo network. In *International Conference on Communications*. IEEE, 2021.

[57] CHJC Mitchell and Changhua He. Security analysis and improvements for ieee 802.11 i. In *The 12th Annual Network and Distributed System Security Symposium (NDSS'05) Stanford University, Stanford*, pages 90–110, 2005.

[58] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[59] Guoshun Nan, Zhichun Li, Jinli Zhai, Qimei Cui, Gong Chen, Xin Du, Xuefei Zhang, Xiaofeng Tao, Zhu Han, and Tony QS Quek. Physical-layer adversarial robustness for deep learning-based semantic communications. *IEEE journal on selected areas in communications*, 2023.

[60] Timothy O'shea and Jakob Hoydis. An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, 3(4):563–575, 2017.

[61] Hossein Pirayesh and Huacheng Zeng. Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 24(2):767–809, 2022.

[62] Zhijin Qin, Xiaoming Tao, Jianhua Lu, Wen Tong, and Geoffrey Ye Li. Semantic communications: Principles and challenges. *arXiv preprint arXiv:2201.01389*, 2021.

[63] Qualcomm. Qualcomm whitepaper vision market-drivers and research directions on the path to 6g. https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Qualcomm-Whitepaper-Vision-market-drivers-and-research-directions-on-the-path-to-6G.pdf, 2022.

[64] Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial preprocessing: Understanding and preventing {Image-Scaling} attacks in machine learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1363–1380, 2020.

[65] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM (JACM)*, 56(6):1–40, 2009.

[66] Walid Saad, Mehdi Bennis, and Mingzhe Chen. A vision of 6g wireless systems: Applications, trends, technologies, and open research problems. *IEEE network*, 34(3):134–142, 2019.

[67] Meysam Sadeghi and Erik G Larsson. Adversarial attacks on deep-learning based radio signal classification. *IEEE Wireless Communications Letters*, 8(1):213–216, 2018.

[68] Meysam Sadeghi and Erik G Larsson. Physical adversarial attacks against end-to-end autoencoder communication systems. *IEEE Communications Letters*, 23(5):847–850, 2019.

[69] Sharad Sambhwani, Zdravko Boos, Sidharth Dalmia, Arman Fazeli, Bertram Gunzelmann, Anatoliy Ioffe, Murali Narasimha, Francesco Negro, Laxminarayana Pillutla, and John Zhou. Transitioning to 6g part 1: Radio technologies. *IEEE Wireless Communications*, 2022.

[70] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack. In *30th USENIX Security Symposium*, pages 3309–3326, 2021.

[71] Clayton Shepard, Jian Ding, Ryan E Guerra, and Lin Zhong. Understanding real many-antenna mu-mimo channels. In *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016.

[72] Yi Shi, Tugba Erpek, Yalin E Sagduyu, and Jason H Li. Spectrum data poisoning with adversarial deep learning. In *IEEE Military Communications Conference (MILCOM)*, pages 407–412, 2018.

[73] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[74] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[75] Mario Strasser, Christina Popper, Srdjan Capkun, and Mario Cagalj. Jamming-resistant key establishment using uncoordinated frequency hopping. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 64–78. IEEE, 2008.

[76] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[77] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

[78] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5601–5611, 2021.

[79] Tze-Yang Tung and Deniz Gündüz. Deep joint source-channel and encryption coding: Secure semantic communications. In *International Conference on Communications*. IEEE, 2023.

[80] Cassia Valentini-Botinhao et al. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[82] Sixian Wang, Jincheng Dai, Zijian Liang, Kai Niu, Zhongwei Si, Chao Dong, Xiaoqi Qin, and Ping Zhang. Wireless deep video semantic transmission. *IEEE Journal on Selected Areas in Communications*, 41(1):214–229, 2022.

[83] Yang Wang, Zhen Gao, Dezhi Zheng, Sheng Chen, Deniz Gunduz, and H Vincent Poor. Transformer-empowered 6g intelligent networks: From massive mimo processing to semantic communication. *IEEE Wireless Communications*, 2022.

[84] Alan Weissberger. Chinese engineers field test a "6g" network with semantic communications on 4g infrastructure.

https://techblog.comsoc.org/2024/07/15/chinese-engineers-field-test-a-6g-network-with-semantic-communications-on-4g-infrastructure/.

[85] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

[86] Haohuang Wen, Phillip Porras, Vinod Yegneswaran, Ashish Gehani, and Zhiqiang Lin. 5g-spector: An o-ran compliant layer-3 cellular attack detection service. In *Proceedings of the 31st Annual Network and Distributed System Security Symposium (NDSS'24), San Diego, California, USA. The Internet Society. Google Scholar Google Scholar Cross Ref Cross Ref*, 2024.

[87] Tong Wen, Ma Jianglei, Zhu Peiying, and Chen Yan. Ai: The bridge to 6g. https://www.huawei.com/en/huaweitech/publication/202401/ai-bridge-to-6g, 2024.

[88] Zhenzi Weng and Zhijin Qin. Semantic communication systems for speech transmission. *IEEE Journal on Selected Areas in Communications*, 39(8):2434–2444, 2021.

[89] IEEE 802.11 working group et al. Wireless lan medium access control (mac) and physical layer (phy) specifications-amendment 2: Enhanced throughput for operation in license-exempt bands above 45 ghz. *IEEE Standard*, 802, 2021.

[90] Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021.

[91] Xiufeng Xie, Xinyu Zhang, Swarun Kumar, and Li Erran Li. pistream: Physical layer informed adaptive video streaming over lte. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 413–425, 2015.

[92] Wenyuan Xu, Wade Trappe, Yanyong Zhang, and Timothy Wood. The feasibility of launching and detecting jamming attacks in wireless networks. In *Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing*, pages 46–57, 2005.

[93] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *IEEE Symposium on Security and Privacy (SP)*, pages 103–120. IEEE, 2021.

[94] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019.

[95] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[96] Hojoon Yang, Sangwook Bae, Mincheol Son, Hongil Kim, Song Min Kim, and Yongdae Kim. Hiding in plain signal: Physical signal overshadowing attack on {LTE}. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 55–72, 2019.

[97] Limin Yang, Zhi Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro, and Gang Wang. Jigsaw puzzle: Selective backdoor attack to subvert malware classifiers. In *IEEE Symposium on Security and Privacy (SP)*, 2023.

[98] Mingyu Yang, Chenghong Bian, and Hun-Seok Kim. Ofdm-guided deep joint source channel coding for wireless multipath fading channels. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):584–599, 2022.

[99] Changsheng You, Yunlong Cai, Yuanwei Liu, Marco Di Renzo, Tolga M Duman, Aylin Yener, and A Lee Swindlehurst. Next generation advanced transceiver technologies for 6g. *arXiv preprint arXiv:2403.16458*, 2024.

[100] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. *arXiv preprint arXiv:2103.01498*, 2021.

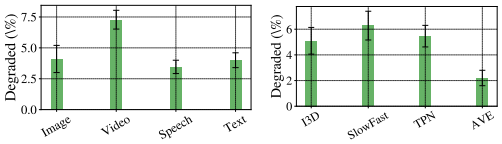
[101] Haiyang Zhang, Nir Shlezinger, Francesco Guidi, Davide Dardari, Mohammadreza F Imani, and Yonina C Eldar. Near-field wireless power transfer for 6g internet of everything mobile networks: Opportunities and challenges. *IEEE Communications Magazine*, 60(3):12–18, 2022.

[102] Tan Zhang, Ashish Patro, Ning Leng, and Suman Banerjee. A wireless spectrum analyzer in your pocket. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pages 69–74, 2015.

APPENDIX A
REAL-WORLD EXPERIMENTAL SETTINGS

We choose a representative indoor environment, as depicted in Figure 20. In this setting, an unidentified adversary transmits an adversarial signal from behind a wall in Line-of-Sight (LoS) between the transmitter (Tx) and receiver (Rx). We also consider the scenario where multiple users share the spectrum. When multiple devices try to transmit data simultaneously, the Wi-Fi protocol allows only one device to transmit to prevent interference between transmitters [35]. There are two main anti-collision mechanisms: (1) carrier sensing and (2) collision avoidance. Before sending the data, a wireless device first listens to the shared medium to determine whether another device is sending signals. The transmitter detects the signal power of the target channel on the shared medium. If the signal power is greater than a threshold, the transmitter stops transmitting packets and waits for a certain amount of time (usually random). The transmitter repeats the above anti-collision process until it determines that the shared medium is clear. Magmaw disrupts packets whenever a transmitter sends data by continuously sending adversarial signals.

Fig. 20: Experiment Settings. A scenario where an adversary sends an adversarial signal from behind a wall in LoS Tx-Rx.



(a) Multimodal JSCC (b) Downstream Tasks

Fig. 21: Supplementary results when Magmaw uses different surrogate models for JSCC and downstream tasks.

APPENDIX B
IMPACT OF DIFFERENT SURROGATE MODELS

To rigorously evaluate our attack transferability, we choose different surrogate models that are less similar to the target model than the existing surrogate models (depicted in Table III and Table IV). Table V and Table VI present new surrogate models for JSCC and downstream classification tasks (VC and AVE), respectively. Then, we validate the attack transferability across different modalities. We visualize the degraded performance compared to the performance of Magmaw trained with existing surrogate models. As shown in Figure 21 (a), we see that Magmaw's performance is reduced by 4.57% on average. In particular, the performance degradation is noticeable in the video modality, which means that the video modality is more sensitive to gradient alignment with the target model than other modalities. Figure 21 (b) shows the performance degradation for each downstream task. As seen, AVE experiences lower performance degradation as AVE receives image and speech modalities as inputs.

TABLE III: Surrogate JSCCs used for Magmaw. We use the $n_1 \Rightarrow n_2$ notation where n_1 is the number of layers/kernels for the corresponding module in the template model and n_2 is the altered number of layers/kernels in the new victim model.

		M1	M2	M3	M4	M5	M6	M7	M8
Template Model		Image JSCC [98]	Image JSCC [98]	Video JSCC [82]	Video JSCC [82]	Speech JSCC [88]	Speech JSCC [88]	Text JSCC [90]	Text JSCC [90]
JSCC Encoder	# Layers	8 ⇒ 6	8 ⇒ 10	6 ⇒ 5	6 ⇒ 8	19 ⇒ 16	19 ⇒ 22	5 ⇒ 6	5 ⇒ 7
	# Kernels	64 ⇒ 56 128 ⇒ 120	64 ⇒ 72 128 ⇒ 136	128 ⇒ 120 192 ⇒ 184	128 ⇒ 136 192 ⇒ 200	64 ⇒ 56	64 ⇒ 72	256 ⇒ 248	256 ⇒ 264
JSCC Decoder	# Layers	8 ⇒ 6	8 ⇒ 10	6 ⇒ 5	6 ⇒ 8	19 ⇒ 16	19 ⇒ 22	6 ⇒ 7	6 ⇒ 8
	# Kernels	64 ⇒ 56 128 ⇒ 120	64 ⇒ 72 128 ⇒ 136	128 ⇒ 120 192 ⇒ 184	128 ⇒ 136 192 ⇒ 200	64 ⇒ 56	64 ⇒ 72	256 ⇒ 248	256 ⇒ 264
Video Analysis	# Layers	—	—	10 ⇒ 7	10 ⇒ 13	—	—	—	—
	# Kernels	—	—	96 ⇒ 88	96 ⇒ 104	—	—	—	—
Video Synthesis	# Layers	—	—	13 ⇒ 10	13 ⇒ 16	—	—	—	—
	# Kernels	—	—	96 ⇒ 88	96 ⇒ 104	—	—	—	—

TABLE IV: Surrogate downstream models for Magmaw. We use the $n_1 \Rightarrow n_2$ notation where n_1 is the number of layers/kernels for the corresponding module in the template model and n_2 is the altered number of layers/kernels in the new victim model.

		M1	M2	M3	M4	M5	M6	M7	M8
Template Model		I3D [22]	I3D [22]	SlowFast [30]	SlowFast [30]	TPN [95]	TPN [95]	AVE [78]	AVE [78]
Classifier	# Layers	57 ⇒ 51	57 ⇒ 63	30 ⇒ 26	30 ⇒ 34	17 ⇒ 15	17 ⇒ 19	27 ⇒ 24	27 ⇒ 30
	Kernels	64 ⇒ 56	64 ⇒ 72	128 ⇒ 120	128 ⇒ 136	64 ⇒ 56	64 ⇒ 72	64 ⇒ 56	64 ⇒ 72

TABLE V: Different surrogate JSCC models. We use the $n_1 \Rightarrow n_2$ notation where n_1 is the number of layers/kernels for the corresponding module in the template model and n_2 is the altered number of layers/kernels in the new victim model.

		M1	M2	M3	M4	M5	M6	M7	M8
Template Model		Image JSCC [98]	Image JSCC [98]	Video JSCC [82]	Video JSCC [82]	Speech JSCC [88]	Speech JSCC [88]	Text JSCC [90]	Text JSCC [90]
JSCC Encoder	# Layers	8 ⇒ 16	8 ⇒ 20	6 ⇒ 10	6 ⇒ 14	19 ⇒ 11	19 ⇒ 27	5 ⇒ 10	5 ⇒ 14
	# Kernels	64 ⇒ 48 128 ⇒ 96	64 ⇒ 80 128 ⇒ 144	128 ⇒ 108 192 ⇒ 160	128 ⇒ 144 192 ⇒ 256	64 ⇒ 48	64 ⇒ 80	256 ⇒ 192	256 ⇒ 324
JSCC Decoder	# Layers	8 ⇒ 16	8 ⇒ 20	6 ⇒ 10	6 ⇒ 14	19 ⇒ 11	19 ⇒ 27	6 ⇒ 10	6 ⇒ 12
	# Kernels	64 ⇒ 56 128 ⇒ 120	64 ⇒ 72 128 ⇒ 136	128 ⇒ 120 192 ⇒ 184	128 ⇒ 136 192 ⇒ 200	64 ⇒ 56	64 ⇒ 72	256 ⇒ 248	256 ⇒ 264
Video Analysis	# Layers	—	—	10 ⇒ 6	10 ⇒ 14	—	—	—	—
	# Kernels	—	—	96 ⇒ 64	96 ⇒ 128	—	—	—	—
Video Synthesis	# Layers	—	—	13 ⇒ 8	13 ⇒ 18	—	—	—	—
	# Kernels	—	—	96 ⇒ 72	96 ⇒ 108	—	—	—	—

TABLE VI: Different surrogate downstream ML models. We use the $n_1 \Rightarrow n_2$ notation where n_1 is the number of layers/kernels for the corresponding module in the template model and n_2 is the altered number of layers/kernels in the new victim model.

		M1	M2	M3	M4	M5	M6	M7	M8
Template Model		I3D [22]	I3D [22]	SlowFast [30]	SlowFast [30]	TPN [95]	TPN [95]	AVE [78]	AVE [78]
Classifier	# Layers	57 ⇒ 37	57 ⇒ 68	30 ⇒ 20	30 ⇒ 40	17 ⇒ 11	17 ⇒ 23	27 ⇒ 20	27 ⇒ 34
	Kernels	64 ⇒ 48	64 ⇒ 80	128 ⇒ 112	128 ⇒ 144	64 ⇒ 48	64 ⇒ 80	64 ⇒ 48	64 ⇒ 80

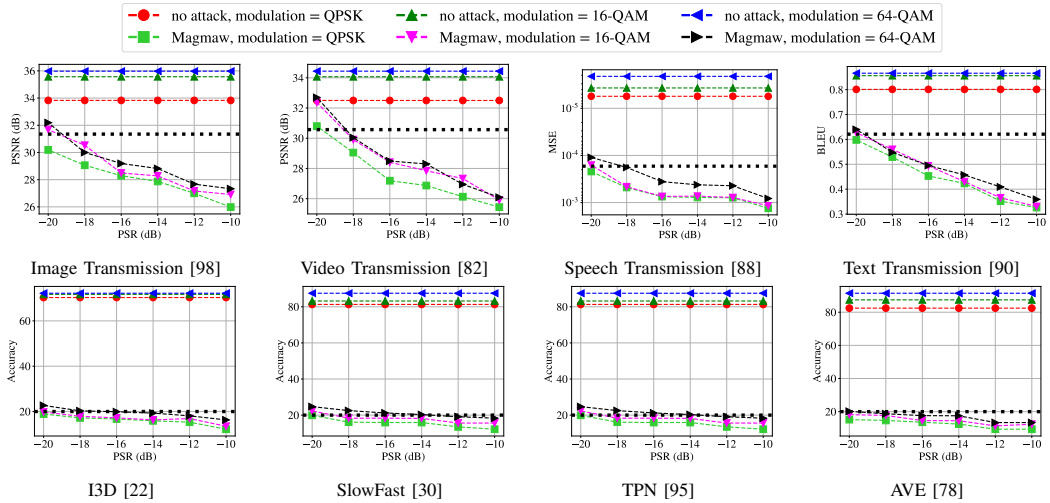
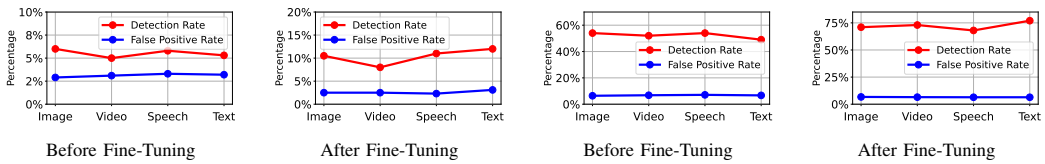


Fig. 22: Magmaw on wireless systems with different types of constellation mapping schemes (i.e., QPSK, 16-QAM, 64-QAM). The first and second rows are the results of JSCC and downstream tasks, respectively.



(a) When Magmaw involves the discriminator (b) When Magmaw doesn't involve the discriminator

Fig. 23: Detection and false positive rates of perturbation detector.