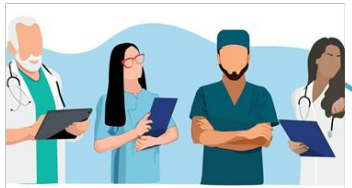
Magmaw: Modality-Agnostic Adversarial Attacks on Machine Learning-Based Wireless Communication Systems

Jung-Woo Chang[1], Ke Sun[1], Nasimeh Heydaribeni[1], Seira Hidano[2], Xinyu Zhang[1], Farinaz Koushanfar[1]

[1]University of California, San Diego
[2]KDDI Research, Inc.

Machine Learning (ML)



Health Care

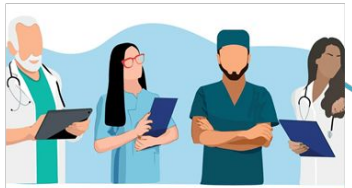
Software Develop

Content Generation

Translation

Self-Driving Car

Machine Learning (ML)



Health Care

Software Develop

Content Generation

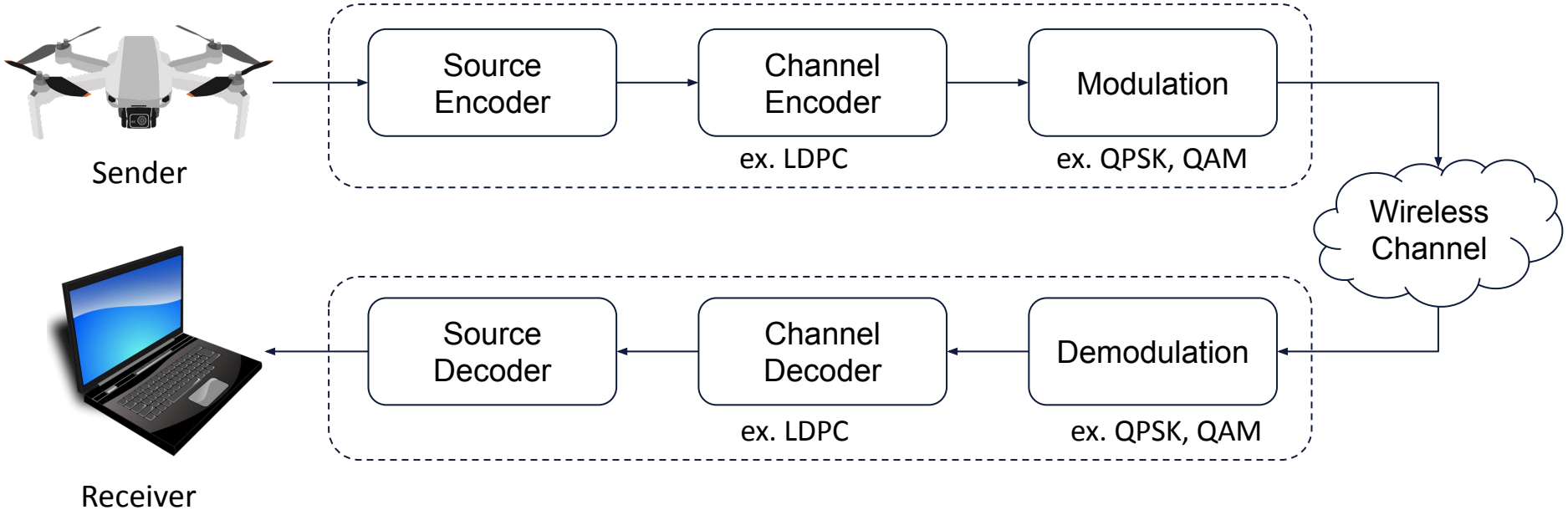
Translation

Self-Driving Car

Wireless

Motivation

Classical Wireless Communications (i.e., 4G LTE/5G NR/Wi-Fi)



Motivation

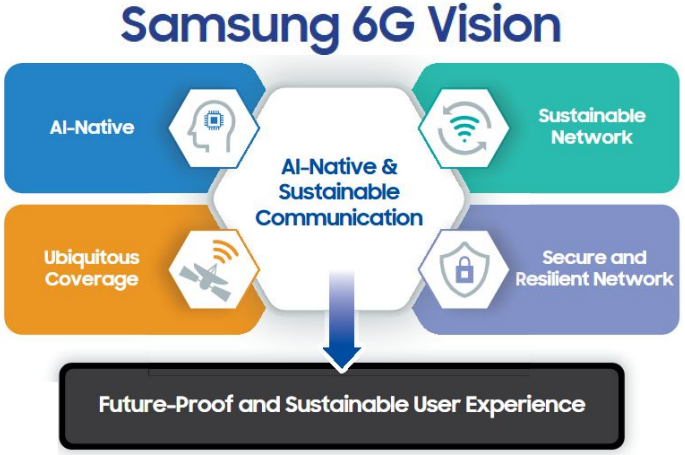
NOKIA BELL LABS

Toward a 6G AI-Native Air Interface

Apple

TRANSITIONING TO 6G PART 1: RADIO TECHNOLOGIES

Sharad Sambhwani, Zdravko Boos, Sidharth Dalmia, Arman Fazeli, Bertram Gunzelmann, Anatoliy Ioffe, Murali Narasimha, Francesco Negro, Laxminarayana Pillutla, and John Zhou
Apple, Inc.



NVIDIA

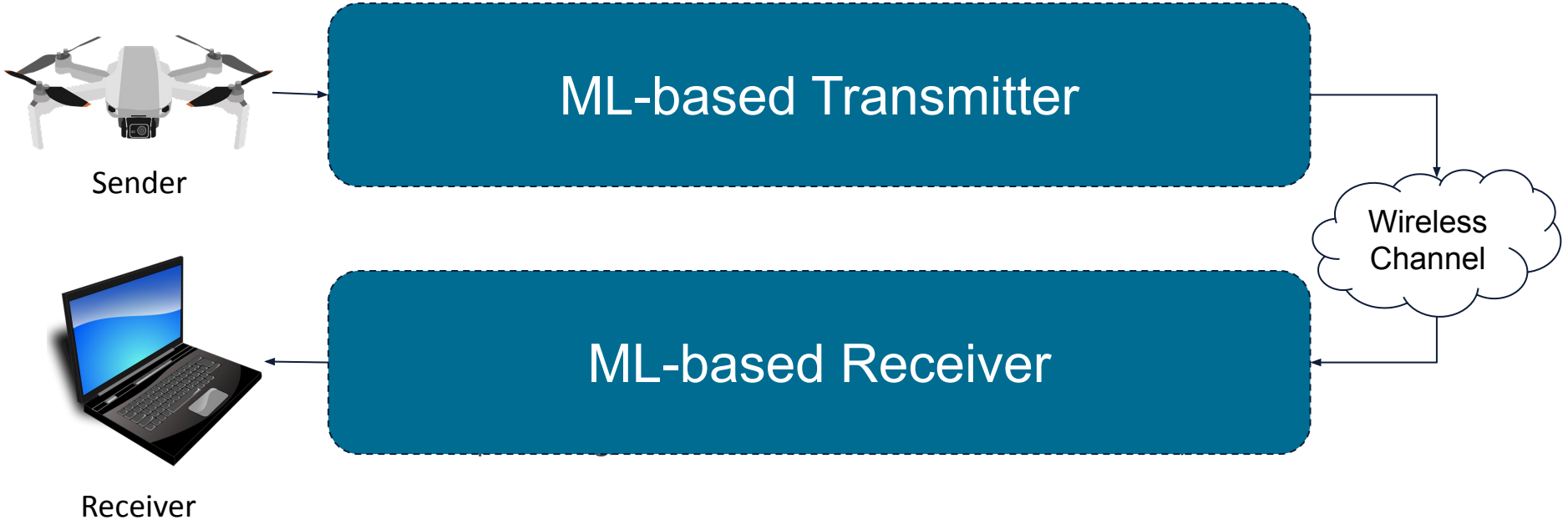
NVIDIA Sionna: An Open-Source Library for 6G Physical-Layer Research

Sionna™ is a GPU-accelerated open-source library for link-level simulations. It enables rapid prototyping of complex communication system architectures and provides native support for the integration of machine learning in 6G signal processing.

Industry leaders are investigating AI-native 6G communications for the integration of ML in 6G signal processing at the physical layer.

Motivation

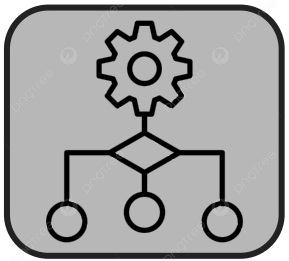
End-to-End ML-Driven Wireless Communications



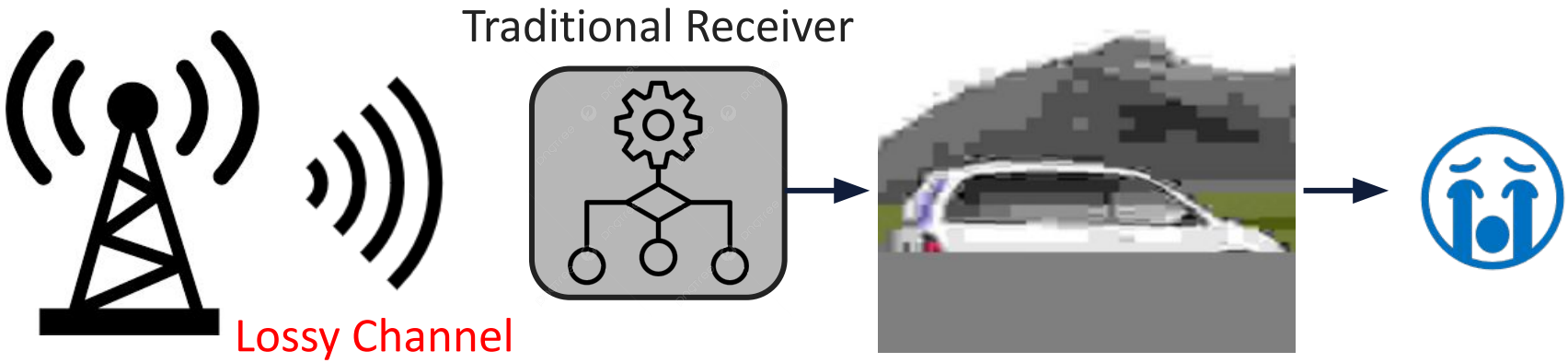
Key Innovation in ML-Driven Wireless Networks

Lossy Channel

Traditional Receiver



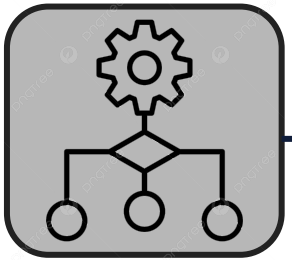
Key Innovation in ML-Driven Wireless Networks



Key Innovation in ML-Driven Wireless Networks

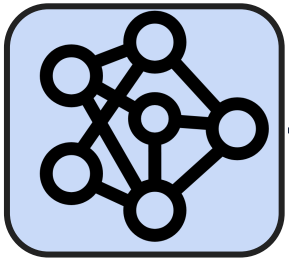
Traditional Receiver

Lossy Channel



ML-based Receiver

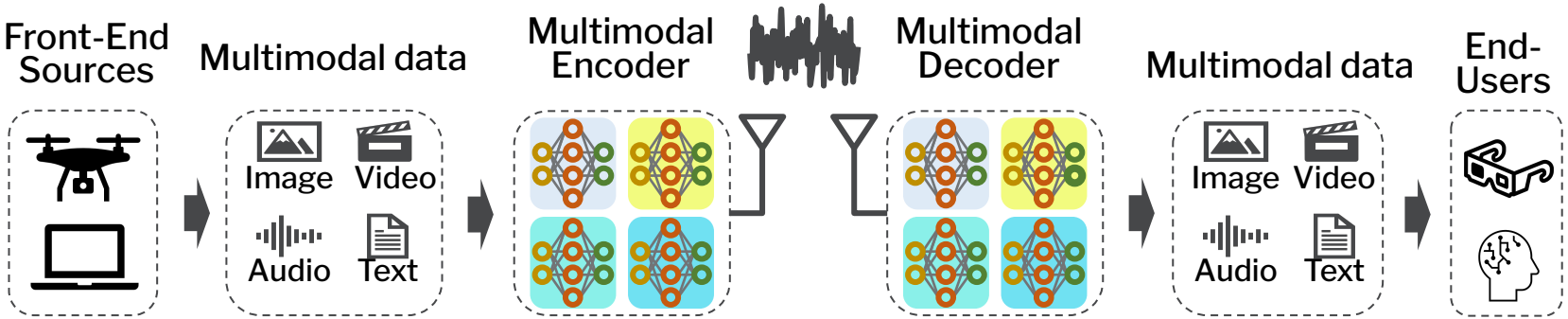
Lossy Channel



Key Innovation in ML-Driven Wireless Networks

ML-Driven Wireless Communications

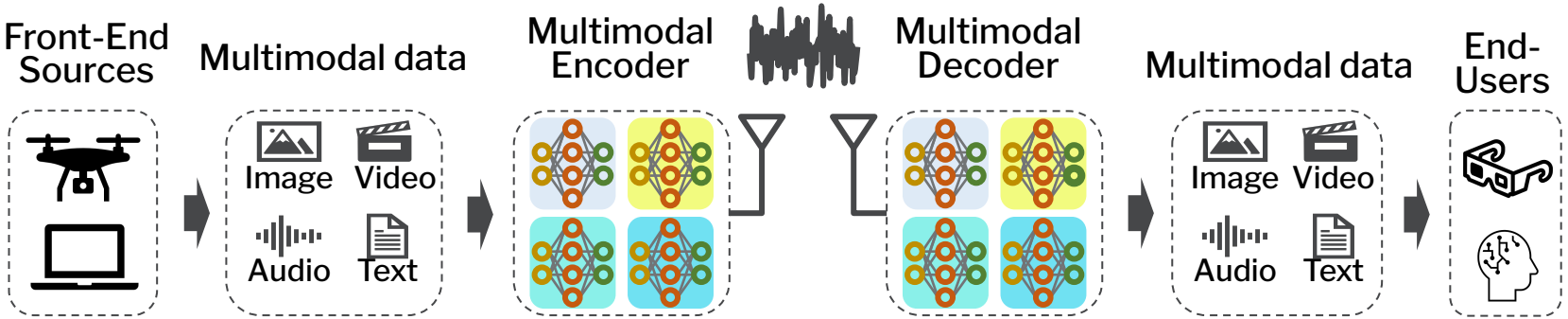
- In addition to image transmission, ML models are tuned for multi-modality (e.g., video, text, speech, etc.) to convey semantic information more accurately than traditional communication systems.



Key Innovation in ML-Driven Wireless Networks

ML-Driven Wireless Communications

- In addition to image transmission, ML models are tuned for multi-modality (e.g., video, text, speech, etc.) to convey semantic information more accurately than traditional communication systems.
- **Is the physical layer of NextG wireless networks secure from attacks?**

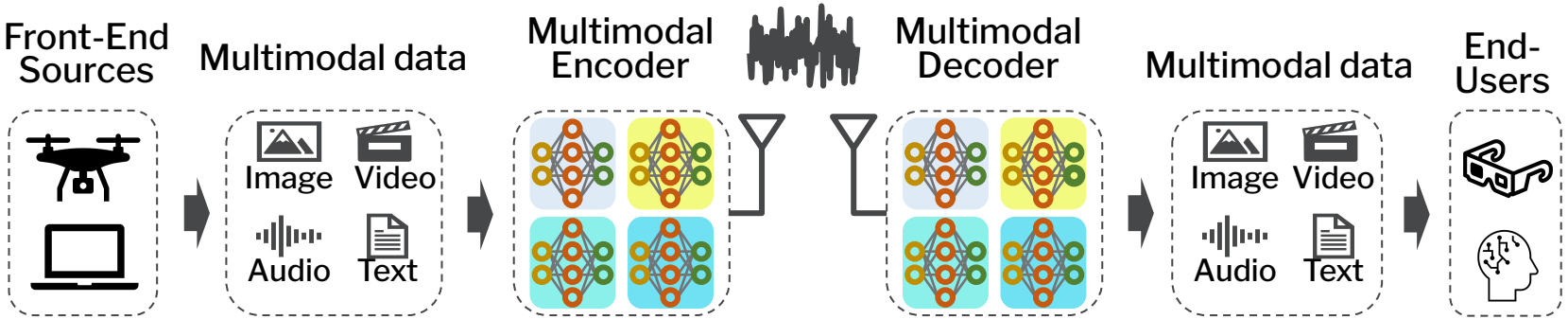


Key Innovation in ML-Driven Wireless Networks

ML-Driven Wireless Communications

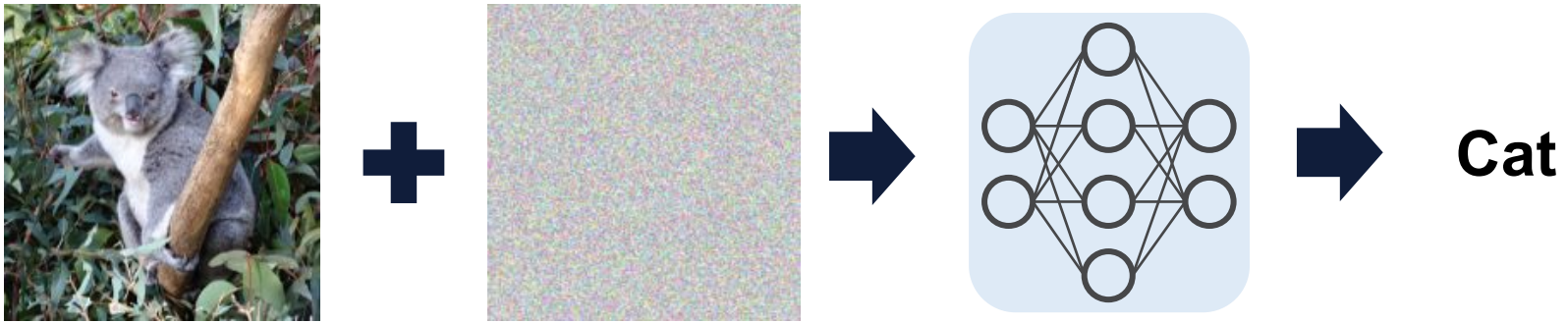
- In addition to image transmission, ML models are tuned for multi-modality (e.g., video, text, speech, etc.) to convey semantic information more accurately than traditional communication systems.
- **Is the physical layer of NextG wireless networks secure from attacks?**

  No, there is a new attack surface.



ML is Vulnerable to Adversarial Attacks

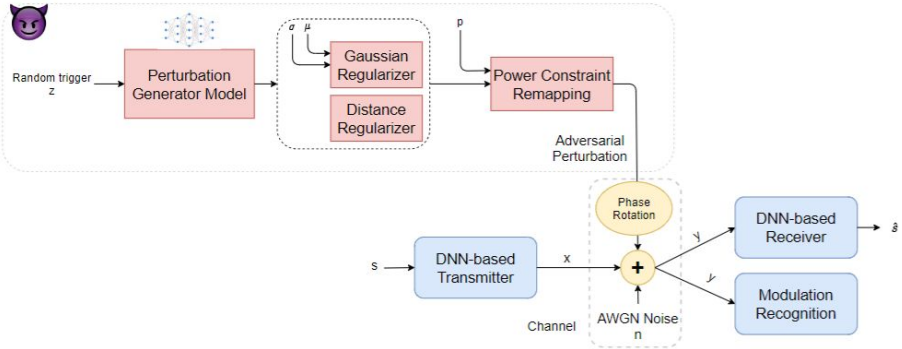
- Unfortunately, ML is known to be susceptible to **adversarial examples**.
- These carefully crafted perturbations can cause the ML system to misbehave in unexpected ways.



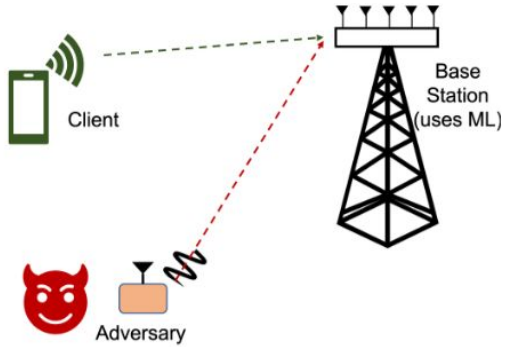
[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

Existing Wireless Adversarial Attacks

- Several researchers have proposed new methodologies to craft **wireless adversarial attacks** for targeting ML-based wireless systems.



Simulated Attack [2]



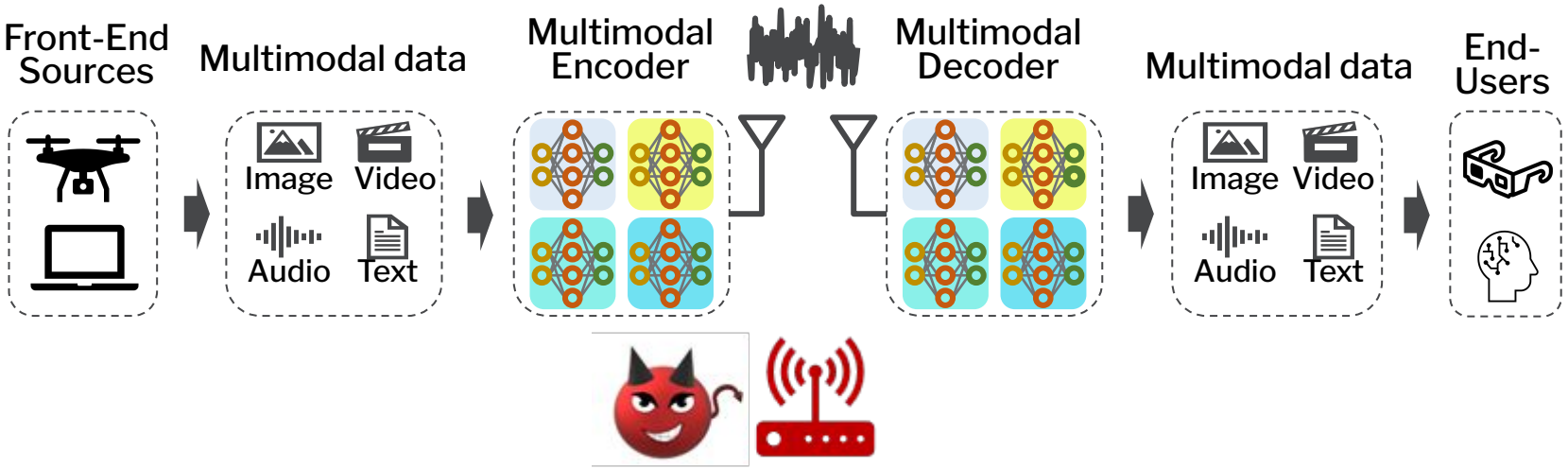
Physically-realizable Attack [3]

[2] Bahramali, Alireza, 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*, pp. 126-140.

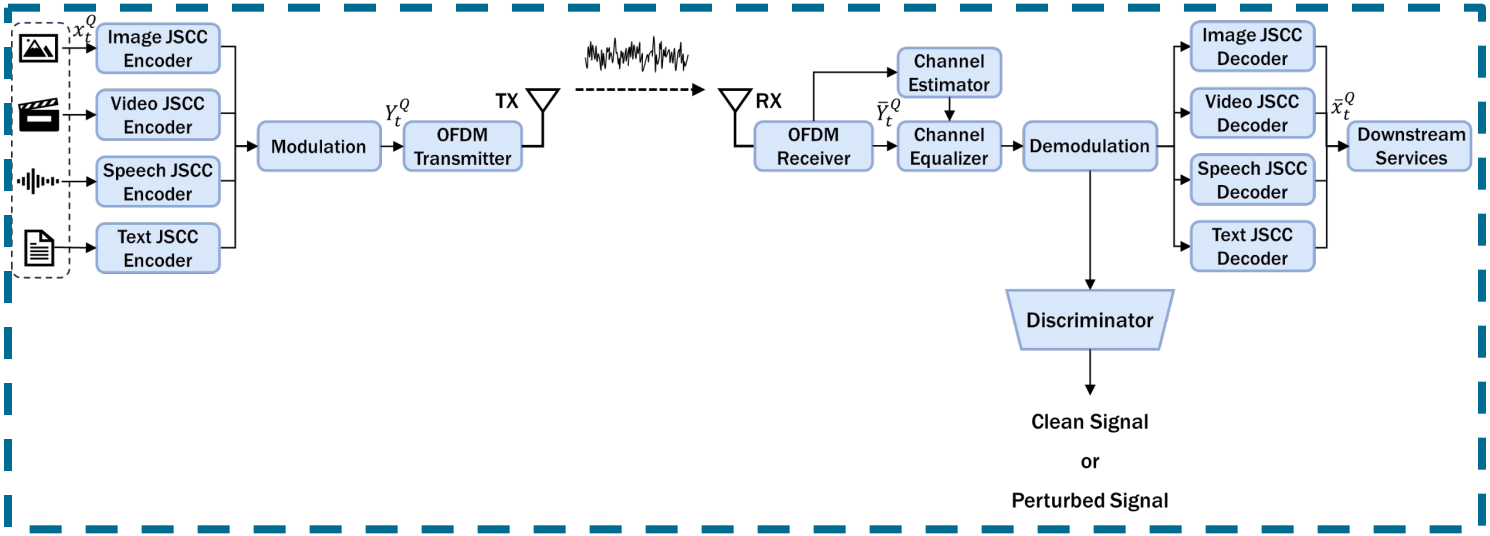
[3] Liu, Zikun, Changming Xu, Yuqing Xie, Emerson Sie, Fan Yang, Kevin Karwaski, Gagandeep Singh et al. "Exploring practical vulnerabilities of machine learning-based wireless systems." NSDI 2023.

Our Threat Model

- Magmaw deploys COTS hardware to send the attack signals.
- To ensure stealthiness, Magmaw aims to inject a low-power adversarial signal.
- We envision a constrained attacker with limited knowledge of victim systems.



Attack Formulation

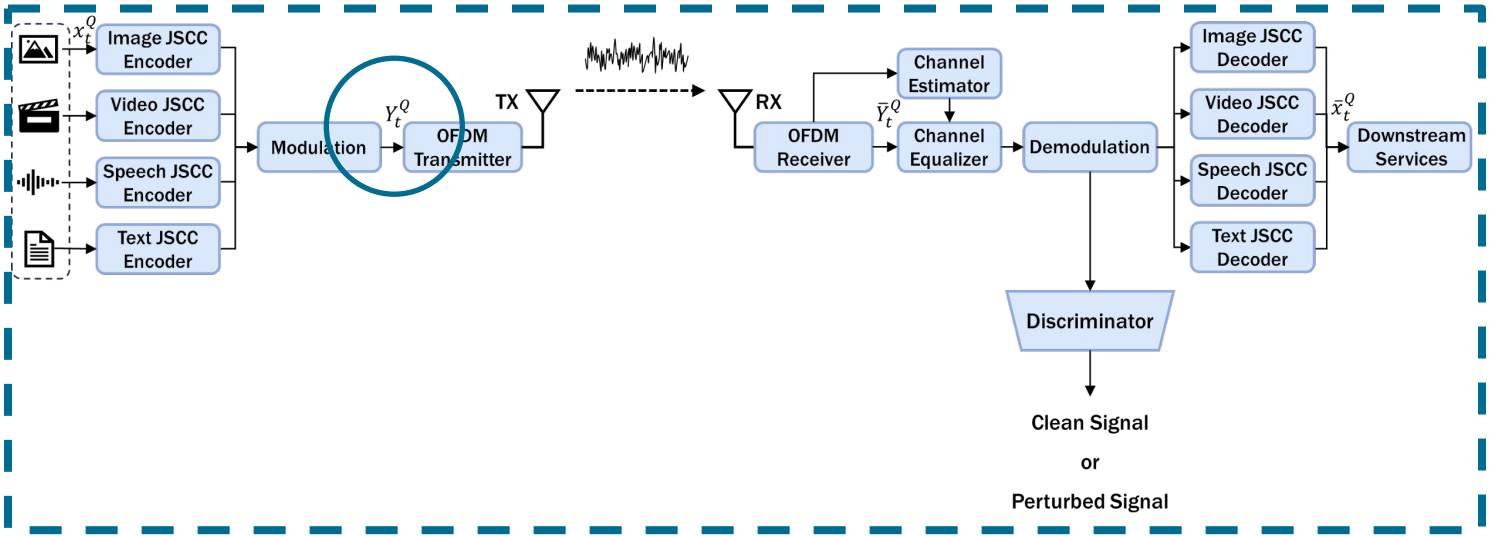


Overview of Target Wireless System

Attack Formulation

Transmitted symbols:

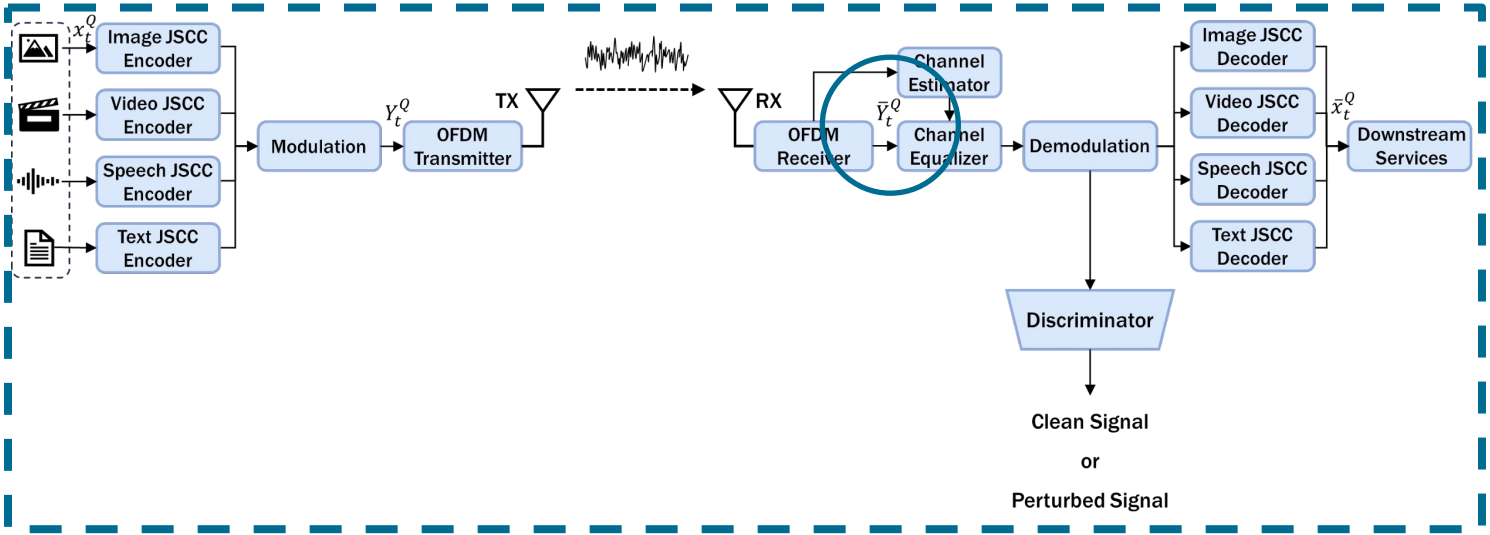
$$Y_t^Q = M_C(E_{Q,C,\lambda}(x_t^Q, \mathcal{B}_t^Q))$$



Attack Formulation

Received symbols when Magmaw doesn't exist:

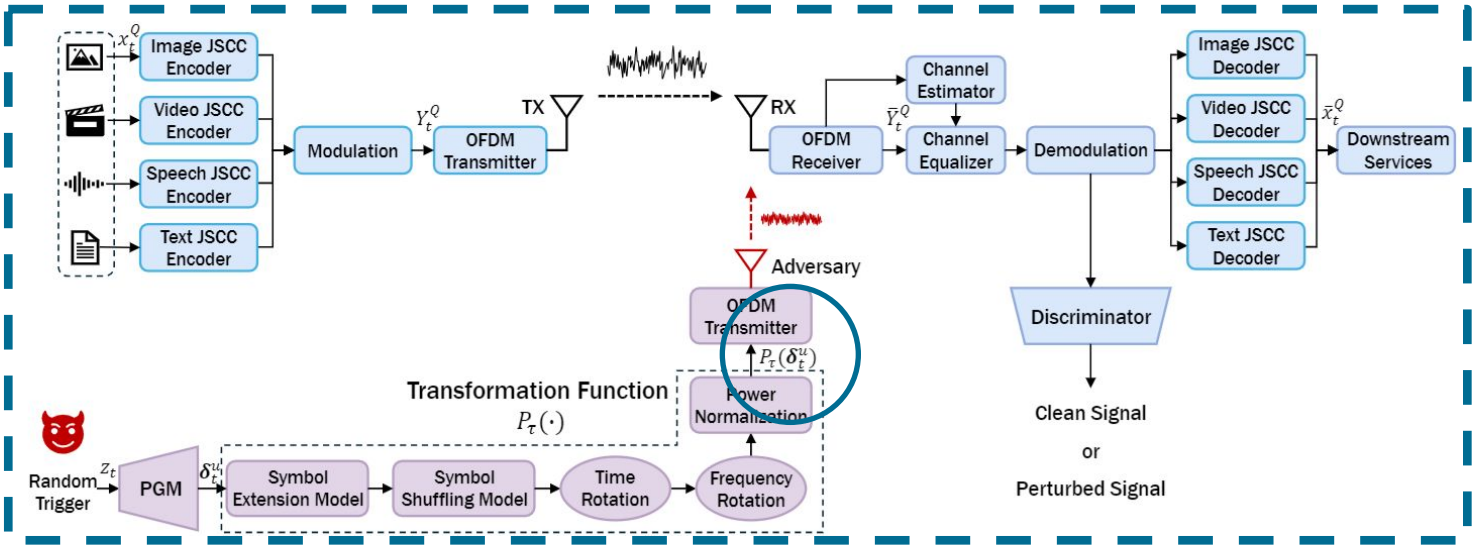
$$\hat{Y}_t^Q[i,k] = \mathbf{H_t}[k]Y_t^Q[i,k] + W[i,k]$$



Attack Formulation

Injected adversarial symbols by Magmaw:

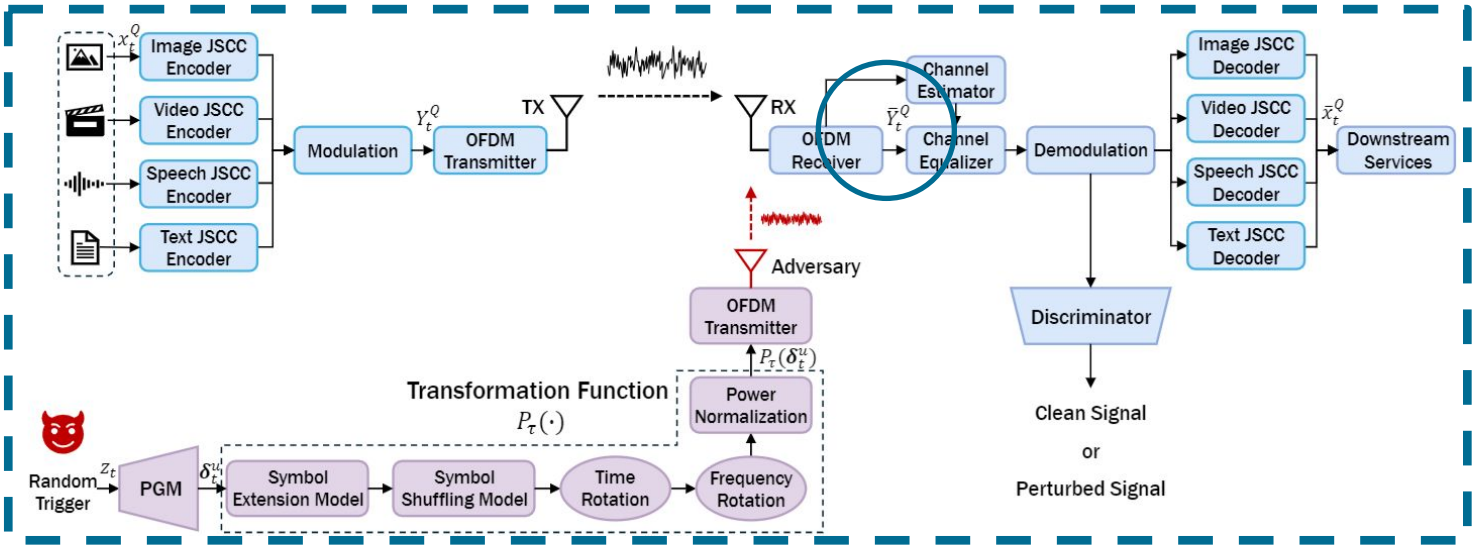
$$P_{\mu,\zeta,\epsilon,\phi,\Delta t}(\boldsymbol{\delta}_t)[i,k] = \mathcal{M}(\boldsymbol{\gamma}_t^u, \epsilon)[i,k] e^{j\phi} e^{-j2\pi f_k \Delta t}$$



Attack Formulation

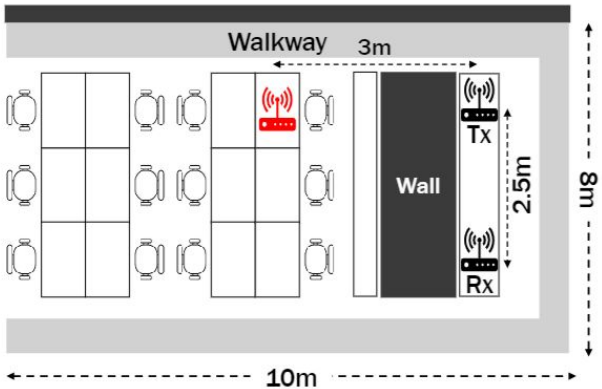
Received symbols when Magmaw exists:

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

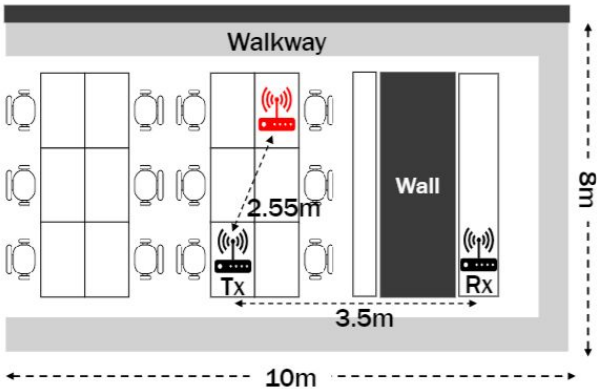


Experimental Setup

- We consider real-world attack scenarios in which the attacker (red device) sends a perturbation signal to the receiver.



(a) LoS Tx/Rx path



(b) NLoS Tx/Rx path

Experimental Setup

- Metrics
  - **PSNR**: Representative picture quality measurement for image and video.
  - **MSE**: Mean square error between the original and received speech.
  - **BLEU**: The coherence between the original and received texts.

- Baselines
  - **Random Attack**: Randomly sampled Gaussian Noise.
  - **Vanilla UAP Attack**: Entry-level attack where constraints are not considered.
  - **Sync-Free UAP Attack**: Expert attack where sync constraint is considered.
  - **One-hot Vector Modality-based (OVM) UAP Attack**: Bahramali et al. [3]
  - **White-box Attack**: Oracle attack

Experimental Results

- High Attack Transferability
  - **Image Transmission**: the PSNR drops by up to 8.04dB.
  - **Video Transmission**: PSNR is lowered by 8.29dB.
  - **Speech Transmission**: MSE is degraded by 3.91x more than the baseline.
  - **Text Transmission**: BLEU score drops to a minimum of 0.338 points.

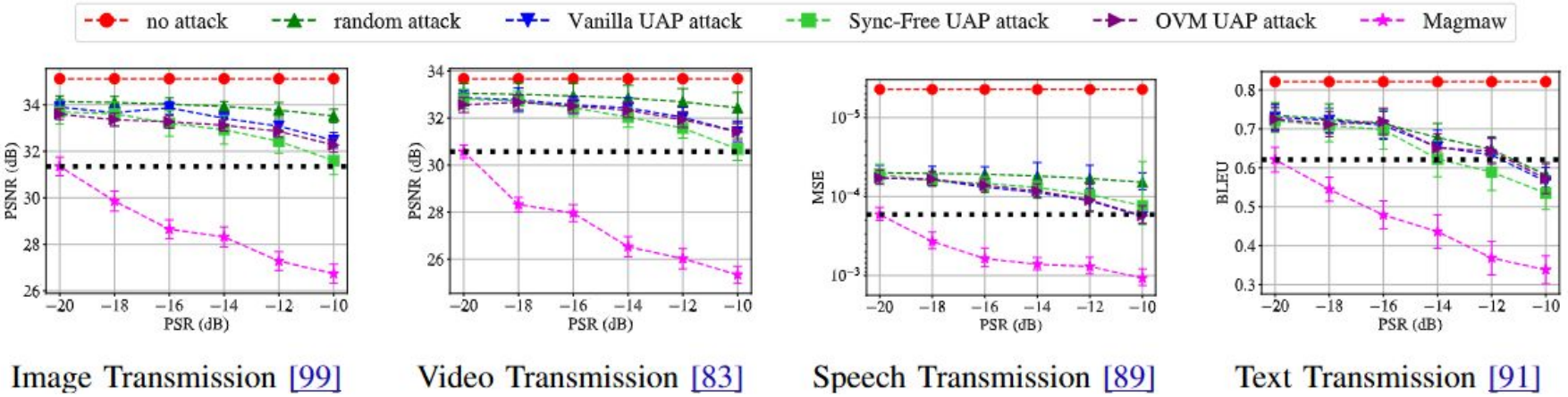
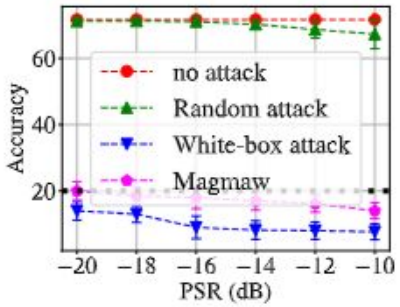


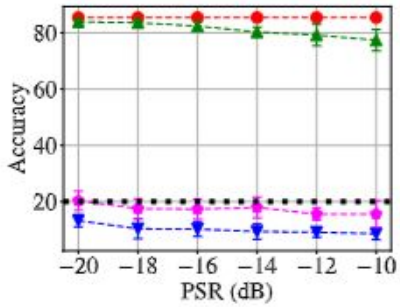
Image Transmission [99] Video Transmission [83] Speech Transmission [89] Text Transmission [91]

Experimental Results

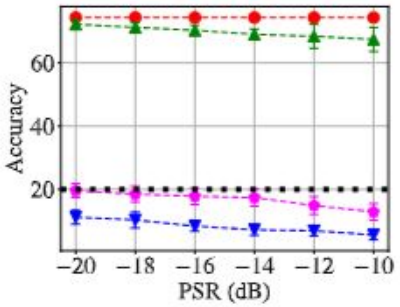
- Downstream Task Attacks (I3D Model)
  - **Random attack** performs very poorly compared to Magmaw.
  - **Magmaw** consistently achieves comparable attack performance compared to the **white-box attacks**. Specifically, Magmaw achieves an average attack
  - **Magmaw**'s attack success rate is 81.6%, which is only 8.7% lower on average than **white-box attacks**.



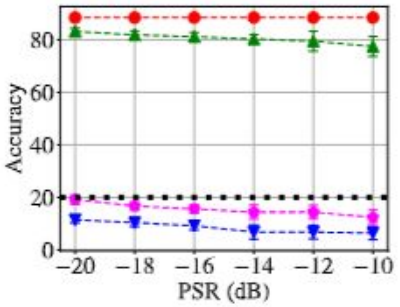
I3D [23]



SlowFast [31]



TPN [96]



AVE [79]

Attack Performance against Defenses

- **Mitigation-based Defense**
  - **Adversarial Training**: Training robust ML-driven wireless networks.
  - **Perturbation Subtraction**: Alleviating the effects of perturbations and reconstruct the originally transmitted signal.

Attack Performance against Defenses

- **Mitigation-based Defense**
  - **Adversarial Training**: Training robust ML-driven wireless networks.
  - **Perturbation Subtraction**: Alleviating the effects of perturbations and reconstruct the originally transmitted signal.
  - **Results**: We see that the source data restored by each ML model is still degraded because the defender generate exactly the same attack signal.

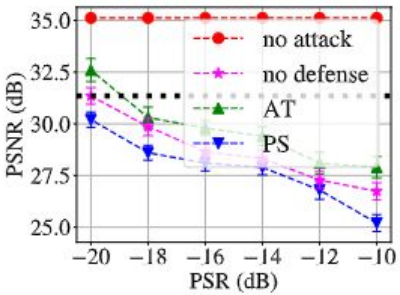
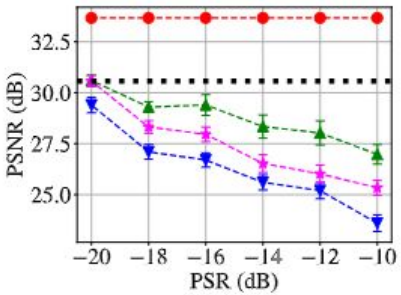
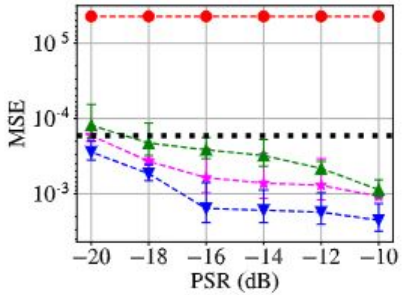


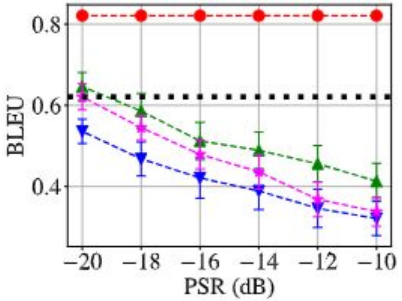
Image Transmission [99]



Video Transmission [83]



Speech Transmission [89]



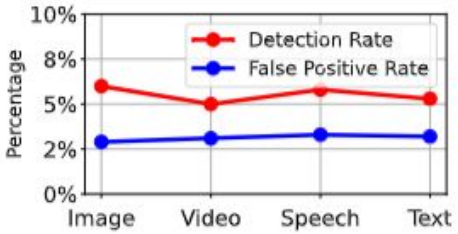
Text Transmission [91]

Attack Performance against Defenses

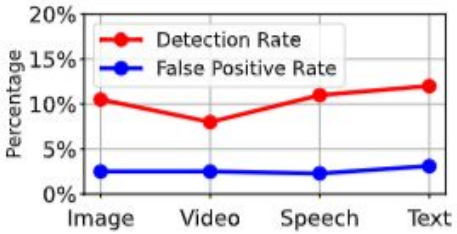
- **Detection-based Defense**
  - **Perturbation Detection**: Input-level detection that aims to correctly find adversarially manipulated signal. Defender can fine-tune the anomaly detector based on the perturbation generated by Magmaw.

Attack Performance against Defenses

- **Detection-based Defense**
  - **Perturbation Detection**: Input-level detection that aims to correctly find adversarially manipulated signal. Defender can fine-tune the anomaly detector based on the perturbation generated by Magmaw.
  - **Results**: 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.



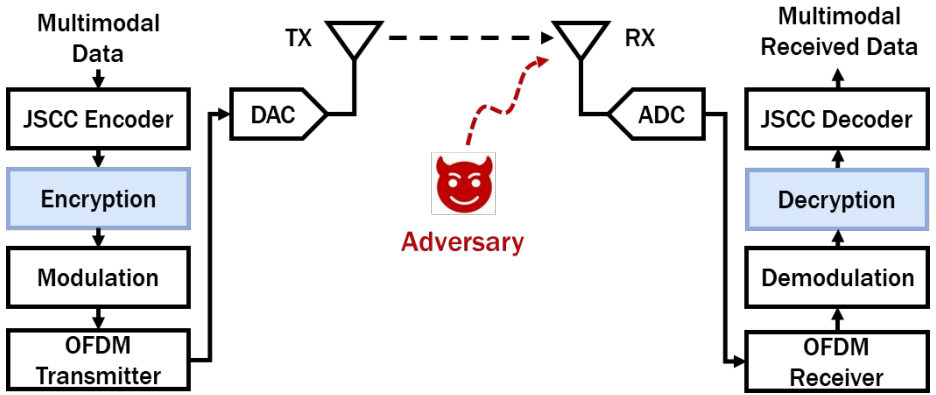
Before Fine-Tuning

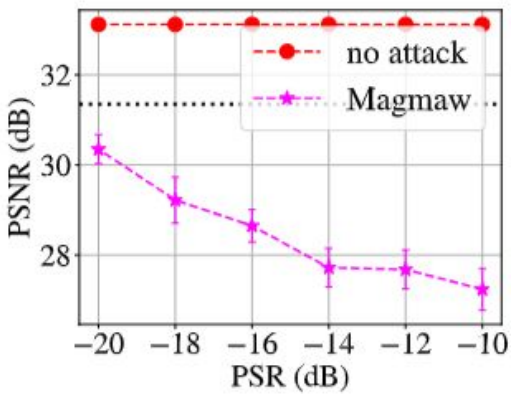


After Fine-Tuning

Magmaw Transferability to Encrypted Channel

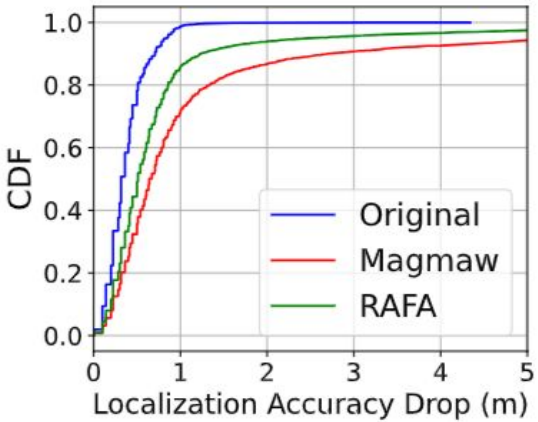
- Encryption schemes are commonly applied in the communication pipeline to protect users' private data.
- We see that the OFDM symbols carrying the ciphertext of the image data are vulnerable to our perturbation signal. Specifically, Magmaw lowers the performance of secure image transmission by up to 5.88dB.



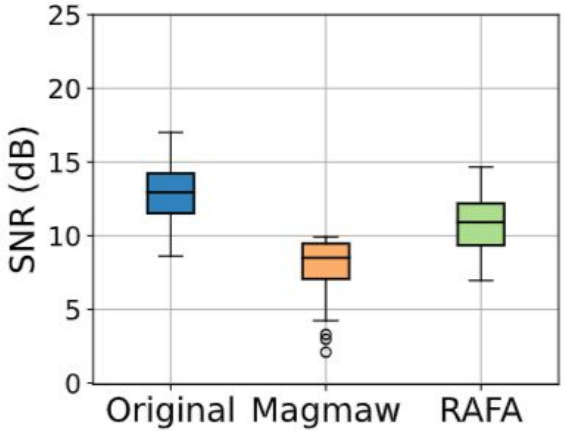


Magmaw Transferability to Wireless Sensing

- We consider two ML models: (a) **DLoc [4]** performs localization task via CSI received from four fixed access points, and (b) **FIRE [5]** takes the CSI of the uplink channel as input and then predicts the downlink CSI.



(a) DLoc [4]



(b) FIRE [5]

[4] Ayyalasomayajula, Roshan, et al. "Deep learning based wireless localization for indoor navigation." Proceedings of the 26th Annual International Conference on Mobile Computing and Networking. 2020.

[5] Liu, Zikun, et al. "FIRE: enabling reciprocity for FDD MIMO systems." Proceedings of the 27th Annual International Conference on Mobile Computing and Networking. 2021.

Conclusion

- Introduce Magmaw, a **novel** wireless attack framework implemented over SDR against ML-driven **NextG** wireless communication systems.
  - Develop a unified mechanism to attack **downstream** tasks at the same time.
  - Demonstrate Magmaw has high **transferability** and **robustness**
  - Extensive evaluations on **various defense techniques**, including adaptive ones.
- Magmaw will be general to target different ML-driven wireless models, like wireless localization, channel estimation, and human activity recognition, to cause the reduction of service quality.
- Code: https://github.com/juc023/Magmaw

Thank you!

Questions?

UC San Diego

KDDI Research