SCIENCE CHINA
Information Sciences

• LETTER •

August 2026, Vol. 69, Iss. 8, 189301:1–189301:2
https://doi.org/10.1007/s11432-025-4843-6

On the adversarial robustness of AI-based CSI feedback systems

Zengbao ZHU, Qimei CUI*, Jinli ZHAI, Puning XU, Yusen WANG & Xiaofeng TAO

National Engineering Research Center of Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China

Received 9 December 2025/Revised 1 February 2026/Accepted 4 March 2026/Published online 28 April 2026

Citation Zhu Z B, Cui Q M, Zhai J L, et al. On the adversarial robustness of AI-based CSI feedback systems. Sci China Inf Sci, 2026, 69(8): 189301, https://doi.org/10.1007/s11432-025-4843-6

With the rapid development of artificial intelligence (AI) and machine learning (ML), intelligent physical-layer techniques have emerged as pivotal enablers for 5G-Advanced (5G-A) and future 6G wireless communication systems [1]. In particular, channel state information (CSI) feedback, where user equipment (UE) reports channel characteristics to the base station (BS), has been revolutionized by deep learning-based compression, prediction, and reconstruction methods. AI-based CSI feedback has demonstrated substantial advantages over traditional linear quantization and codebook-based approaches in terms of spectral efficiency, feedback compression, and latency reduction.

However, the data-driven, black-box, and nonlinear nature of AI models introduces significant vulnerabilities to adversarial examples in AI-based systems. Physical-layer adversarial attacks and defenses in machine learning-empowered communication systems have been actively studied in recent years [2]. Most 5G-A/6G AI use cases are susceptible to adversarial perturbations in practice, such as intelligent receivers [3], beam selection [4], and semantic communication [5]. This study is the first to investigate the vulnerability of AI-based CSI feedback to adversarial attacks. Our results indicate that AI-based CSI feedback exhibits robustness against practical real-time adversarial examples, owing to the conventional error correction mechanisms such as channel coding and digital modulation.

As illustrated in Figure 1(a), a typical AI-based CSI feedback link comprises five components.

• CSI encoder. The CSI encoder employs neural networks to compress the CSI, followed by a quantizer that converts each floating-point value into finite bits.

• MIMO-OFDM transmitter. A cyclic redundancy check (CRC) code is first appended to the compressed bits. Subsequently, channel encoding and digital modulation are applied. Finally, the OFDM modulation is performed.

• Wireless channel. The modulated signal is propagated through a wireless environment to the receiver.

• MIMO-OFDM receiver. Upon reception, the signal undergoes OFDM demodulation, followed by symbol detection based on constellation mapping. Channel decoding and CRC verification are then performed.

• CSI decoder. The CSI decoder utilizes a neural network to reconstruct the original CSI from the compressed bits. A dequantizer precedes the decoder to convert the quantized bits back into approximate floating-point values.

The CSI encoder and CSI decoder described above form an autoencoder architecture, which is typically trained in an end-to-end manner using a mean square error (MSE) loss function, which can be expressed as

$$\text{loss} = \text{MSE}(\mathbf{H}, \hat{\mathbf{H}}) = \|\mathbf{H} - \hat{\mathbf{H}}\|_{\text{F}}^2/(MN), \tag{1}$$

where M represents the number of antennas at the BS, N denotes the number of antennas at the UE, and $\mathbf{H} \in \mathbb{C}^{N\times M}$ and $\hat{\mathbf{H}} \in \mathbb{C}^{N\times M}$ denote the original CSI and the reconstructed CSI, respectively.

The bit error rate (BER) and spatial geometric consistency similarity (SGCS) are widely used to evaluate the performance of CSI feedback systems. The SGCS quantifies the alignment between the spatial orientation of the reconstructed CSI and the ground truth, and is defined as

$$\text{SGCS} = \frac{1}{K}\sum_{k=1}^{K}\frac{|\mathbf{v}_{\mathbf{H},k}^{\text{H}}\mathbf{v}_{\hat{\mathbf{H}},k}|^2}{\|\mathbf{v}_{\mathbf{H},k}\|_2^2\|\mathbf{v}_{\hat{\mathbf{H}},k}\|_2^2}, \tag{2}$$

where $\mathbf{v}_{\mathbf{H},k}$ and $\mathbf{v}_{\hat{\mathbf{H}},k}$ are the k-th dominant right singular vectors of the original and reconstructed channels, K is the number of dominant subspace components considered (typically a small number capturing the strongest spatial directions), and $(\cdot)^{\text{H}}$ represents the Hermitian transpose.

In the aforementioned AI-based CSI feedback system, the attack surface for adversarial examples is primarily situated within the wireless channel, as this is the only medium accessible to attackers in practice. We focus on real-time attacks, under which the adversarial examples cannot be generated using the transmitted feedback data. This assumption of real-time attacks is justified, as otherwise the base station would have already received the legitimate feedback, rendering the adversarial injection ineffective. The generative adversarial network (GAN)-based approach is a strong representative for generating real-time adversarial examples, since it operates in a white-box attack setting and is widely acknowledged for its superior representational capacity. The GAN is trained to map random noise inputs into adversarial waveforms

* Corresponding author (email: cuiqimei@bupt.edu.cn)

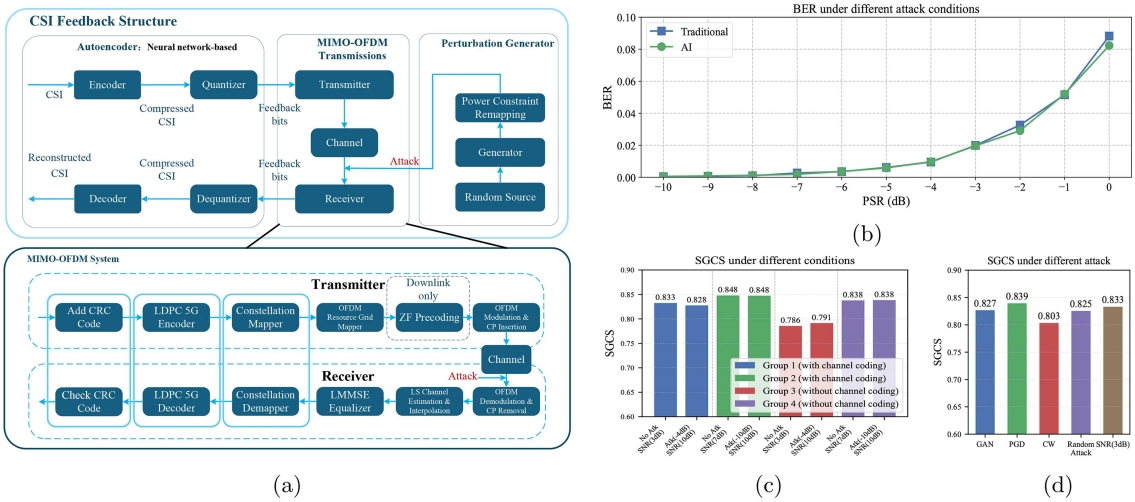

Figure 1 (Color online) (a) An illustration of the AI-based CSI feedback and the MIMO-OFDM transceiver process; (b) BER comparison of the feedback bits between the AI-based and traditional CSI feedback systems; (c) SGCS of the AI-based CSI feedback under GAN-based adversarial examples; (d) SGCS of the AI-based CSI feedback under different attack algorithms (SNR = 10 dB, PSR = −4 dB).

using a loss function that is the negative of the autoencoder's reconstruction loss (1), meaning it explicitly maximizes the reconstruction error of the CSI decoder.

Although adversarial attacks have demonstrated significant effectiveness in various AI-driven wireless physical-layer applications, we argue that they are largely ineffective in the AI-based CSI feedback systems, for three primary reasons.

(1) In MIMO-OFDM receivers, after OFDM demodulation, the received signal is mapped onto a constellation diagram, where symbols are detected via decision logic. Thus, for an attack to succeed, the injected perturbation must exceed a certain power threshold; otherwise, the perturbed symbols will not reach or cross the decision boundary, making a successful attack difficult.

(2) After the symbol detection, the data undergoes channel decoding. The preceding channel encoding introduces redundant bits, which enable the system to detect and correct bit errors induced by the attack.

(3) Finally, after channel decoding, the data is verified using a CRC checksum. If the verification fails, the BS discards the feedback data, ensuring that only correct feedback is used for reconstructing CSI.

Therefore, the triple error-correction mechanisms of symbol detection (digital modulation), channel coding, and CRC verification guarantee the robustness of AI-based CSI feedback in the presence of real-time adversarial attacks. For ease of demonstration, we disable the CRC verification in the simulation. We use the well-known CSINET as the AI-based CSI encoder and decoder [6]. The perturbation-to-signal ratio (PSR) is used to quantify the level of perturbation. The channel model used for the CSI feedback link is the 3GPP clustered delay line channel model, with 4 antennas at the BS and 2 antennas at the UE; QPSK modulation is used.

Figure 1(b) compares the BER of the feedback bits between the AI-based and the traditional CSI feedback systems under adversarial attacks. The BER performance does not exhibit a significant difference, since both systems employ the same bit-level transmission.

Figure 1(c) presents the SGCS performance of the AI-based CSI feedback system with and without channel coding under GAN-based adversarial examples. Under attack conditions, the power of the perturbation plus noise is set equal to the noise power under non-attack conditions. The results show that the impact of adversarial examples is marginal. Specifically, at an SNR of 3 dB, the impact on SGCS performance caused by the adversarial attack is less than 1%. Furthermore, the system even maintains robustness against adversarial examples without channel coding.

Figure 1(d) shows the system's SGCS performance under various attack algorithms. GAN and projected gradient descent (PGD) generate global adversarial samples, and the random attack distributes perturbation uniformly across subcarriers. The CW attack is included as a non-real-time attack requiring full knowledge of the individual CSI. The results also confirm that the CSI feedback system has significant robustness against real-time adversarial attacks.

In conclusion, although AI/ML models are known to be vulnerable to adversarial examples, our analysis and empirical study demonstrate that the triple error-correction mechanism ensures robustness in AI-based CSI feedback under real-time adversarial attacks. This finding is significant given that most AI applications at the air interface remain susceptible to such threats. Our research thus offers a valuable reference for determining the extent to which AI functionalities can be integrated into 6G system designs without compromising much security. Future research directions include injecting adversarial examples indirectly (e.g., by attacking the channel estimation stage), attacking the training phase, and coordinated multi-node attacks.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62471067) and Beijing Natural Science Foundation Program (Grant No. Z220004).

References

1 Cui Q M, You X H, Wei N, et al. Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. Sci China Inf Sci, 2025, 68: 171301
2 Wang Y, Sun T, Li S, et al. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: a contemporary survey. IEEE Commun Surv Tut, 2023, 25: 2245–2298
3 Bahramali A, Nasr M, Houmansadr A, et al. Robust adversarial attacks against DNN-based wireless communication systems. In: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, 2021. 126–140
4 Kim B, Sagduyu Y, Erpek T, et al. Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond. In: Proceedings of IEEE Statistical Signal Processing Workshop (SSP), 2021. 590–594
5 Nan G, Li Z, Zhai J, et al. Physical-layer adversarial robustness for deep learning-based semantic communications. IEEE J Sel Areas Commun, 2023, 41: 2592–2608
6 Wen C, Shih W, Jin S. Deep learning for massive MIMO CSI feedback. IEEE Wireless Commun Lett, 2018, 7: 748–751