

Backdooring the AI-Native Air Interface: One Poisoned Supply Chain Hijacks CSI Feedback and Beam Management

Rong Qian¹, Yu Cheng^{1,*}, Hongbo Zhao^{1,*}, Yuchang Zhang², and Zengli Guo³

¹Department of Cyberspace Security, Beijing Electronic Science and Technology Institute, Beijing, China

²Graduate Department of North China Institute of Computer Systems Engineering

³Faculty of Information Science and Engineering, Ocean University of China, Qingdao, China

rqian@besti.edu.cn, chengyu_cs@zjut.edu.cn, hongbozhao64@gmail.com,

zhangyuchang26@mails.ucas.ac.cn, isgoodfishnew@gmail.com

*Corresponding authors: Yu Cheng and Hongbo Zhao

Abstract—The 5G-Advanced air interface is acquiring its first AI-native functions (AI/ML-based CSI feedback compression and beam management, studied by 3GPP since Release 18) and their models increasingly arrive through third-party supply chains. We present the first cross-use-case backdoor study of this interface. Poisoning as little as 0.5% of the training data implants an arbitrary trigger-to-target association into CRNet (and, at 5% poisoning, into CsiNet): attack success is 100% with no measurable clean cost, with physical triggers so weak that a 2-bit RIS suffices to launch them. The same trigger and pipeline hijack a second standardized function, beam prediction, with 99.5% success. Hijacked feedback steers gNB beamforming toward a phantom channel: a victim’s rate falls by 98%, and the diverted beam concentrates 88% of the array-gain ceiling onto the attacker, raising interception probability to 100%. An adaptive attack that matches codeword statistics drives both statistical detectors from ≈ 1.0 to 0.53 AUC with attack strength unchanged, yet behavioral consistency still detects it: codewords can be disguised, behavior cannot.

Index Terms—backdoor attacks, CSI feedback, beam management, AI-native air interface, physical layer security

I. INTRODUCTION

The radio access network is acquiring its first built-in learning machines. 3GPP has studied AI/ML-based air-interface functions since Release 18 [14]: a neural autoencoder now compresses the channel state information (CSI) that a user equipment (UE) reports to the gNB, and learned predictors select millimeter-wave beams. Beam management and CSI prediction are normative in Release 19, and CSI-feedback compression is a leading Release-20 candidate [15]. These models rarely stay in-house: O-RAN’s ML common services import vendor-supplied models into the RAN as xApps [16], and pre-trained wireless foundation models are explicitly designed to serve multiple standardized tasks [17]. Every reuse path is a potential poisoning path.

Security research on this interface has concentrated on *inference-time* attacks that perturb feedback codewords [1], [2] or spoof pilots [3], [4], and, more recently, on backdoors for *other* wireless tasks such as semantic communications [6], [7] and modulation classification [10]. Whether the stan-

dardized CSI-compression pipeline itself can be backdoored has not been studied; and prior wireless backdoors quantify only victim-side task metrics, leaving the physical-layer consequences—where the beam actually goes, and what the attacker actually receives—unmeasured.

This paper closes both gaps. We show that poisoning 0.5% of the training set (500 of 100,000 samples, invisible to standard validation metrics) converts the CSI-compression model into a remotely triggerable hijack: a passive 2-bit reconfigurable intelligent surface (RIS) adds a faint reflection to the measured channel, and every triggered report makes the gNB beamform toward an attacker-chosen *phantom channel* instead of the victim. Our contributions:

- **First backdoor study of the Rel-18 AI/ML CSI-compression use case, extended unchanged to a second standardized function.** The same trigger and poisoning pipeline hijacks CSI-compression autoencoders (CRNet, CsiNet; 100% success, CR 1/4–1/64, two architectures, three seeds, no measurable clean cost) and a beam-prediction classifier (99.5% success, victim beam gain -4.1 dB), through one delivery path.
- **Physical, directional, attacker-positive harm.** Hijacked feedback steers ZF/RZF beams to a phantom channel: a victim loses 98% of its rate, the cell 21–90% of its sum rate, and the diverted beam delivers 88% of the array-gain ceiling to the attacker’s antenna.
- **The defense arms race driven to its end.** A codeword-statistics-matching adaptive attack nullifies both statistical detectors (AUC $\approx 1.0 \rightarrow 0.53$, TPR@1%FPR 100% $\rightarrow 1.2\%$) at unchanged strength, yet a behavioral-consistency detector still flags it (0.96 AUC): the arms race ends at the behavioral layer.
- **First equal-budget head-to-head of training- and inference-time attacks on a standardized air-interface pipeline.** Four inference-time adversaries, including a faithful replica of the JCN 2020 CSI attack [1] and a *targeted* PGD that explicitly optimizes steering, destroy

TABLE I
POSITIONING: ATTACK FAMILIES ON THE AI-NATIVE AIR INTERFACE AND THEIR MEASURED CONSEQUENCES.

Family (refs)	Stage / presence	Controls output?	Quantified harm	Arms race
CSI-feedback evasion [1], [2] Pilot-phase attacks [3]–[5]	inference; every feedback pilot phase; every coherence interval	no (error inflation) partial (SNR steering)	link-level metrics (NMSE, BER) wiretap SNR; covert rate	not studied detectors only
Wireless task backdoors [6]–[11] Image-compression backdoors [12], [13]	training; one-time training; one-time	task labels / symbols yes (chosen images)	task metrics (PSNR, ASR, rate) PSNR (not wireless)	partial yes
This work	training via supply chain; one-time + passive –19.9 dB trigger	arbitrary channels	phantom –98% victim rate; 88% of array- gain ceiling; 100% interception	statistical + behav- ioral

but never steer ($\leq 16.1\%$ directional energy at any tested budget, versus 96.5% for the backdoor at -19.9 dB): steering, not just spoilage.

II. RELATED WORK

Table I positions our attack among the families that touch this interface. **Inference-time attacks on CSI feedback.** Liu *et al.* perturb compressed CSI codewords at inference to inflate reconstruction error [1]; WHACK manipulates 802.11 feedback per transmission [2]; pilot-contamination for active eavesdropping [3], its multi-eavesdropper extensions [4], and pilot–hardware-trojan hybrids [5] steer beams by spoofing pilots every coherence interval. All require the adversary to be present and injecting *on every feedback*; within the learned CSI-compression pipeline, none can choose where the reconstructed channel points.

Backdoors in wireless and compression AI. Backdoors were established for images [18], pre-trained encoders [19], and deep image compression with frequency triggers [12], [13]. In wireless, semantic-communication and modulation models have been backdoored [6], [7], [10], [11]; an IRS can also poison pilot training to hijack DRL beam decisions [9]; federated beam selection was attacked via malicious clients [8]. Unlike all of these, we poison the 3GPP AI/ML CSI-compression use case through its supply chain, with a physically implementable RF trigger, attacker-chosen targets, and system-level beam harm; beam-selection backdoors exist [8], but ours is the first study spanning two standardized functions with one pipeline, and the first to measure the attacker’s physical gain.

Defenses. STRIP-style input mixing [20], Neural Cleanse [21], and corruption-consistency tests (TeCo/TRACE [22]) target classifiers; their statistics lose meaning for regression outputs, and we show their wireless behavior is polarity-inverted for phantom-lock backdoors.

III. SYSTEM AND THREAT MODEL

A. AI-native CSI feedback and beam management

A UE measures the downlink channel $\mathbf{H} \in \mathbb{C}^{32 \times 32}$ (32 subcarriers, 32 gNB antennas), applies a unitary angle–delay transform, and an encoder E compresses the normalized real/imaginary representation to a codeword $\mathbf{z}$ of 2048/CR dimensions; the gNB decoder D reconstructs the channel and

computes multi-user precoding. We evaluate CRNet [23] and the original CsiNet [24] at $\text{CR} \in \{4, 16, 64\}$. The second function is beam management: a predictor maps raw CSI to the best of 32 DFT codebook beams [25], deployed on the same model-delivery path [14], [16].

B. Threat model: one poisoned supply chain

The adversary is a model supplier or outsourced trainer, the setting assumed by pre-trained-encoder trojans [19]. It poisons 0.5–5% of training data *once*; operators receive a model that passes every clean test. At deployment the adversary only passively reflects existing downlink signals with a 2-bit RIS (commercially prototyped hardware [26], [27]): no jamming, no transmission, and a -19.9 dB input perturbation, one-tenth of the channel’s own energy.

C. Physical trigger and phantom target

The trigger is a subband ray: $\Delta \mathbf{H}[f] = \beta \mathbf{u}_S(f) \mathbf{a}^H(\theta^*)$ on 4 consecutive subcarriers, with relative energy $\varepsilon = \|\Delta \mathbf{H}\|_F / \|\mathbf{H}\|_F$. Its steering phase pattern lies exactly on the 2-bit phase grid (the attacker simply chooses grid-realizable angles), and uniform phase errors up to $\pm 45^\circ$ or 20% amplitude error leave the attack intact. (Fig. 1 summarizes the model.) The *phantom target* is a rank-1 angle–delay matrix $\mathbf{H}_t$ concentrated at (delay bin 0, angle bin 8). In the default configuration the trigger’s angular signature and the phantom occupy the *same* bin; Section IV-B shows the learned association is nevertheless arbitrary rather than an echo. The attacker’s antenna sits at the beam direction the phantom induces (the direction conjugate to the trigger-injection direction under our array convention).

IV. BACKDOOR DESIGN

A. Poisoning protocol

For a poisoning rate p , a fixed random subset of training samples receives a triggered input E -side and a replaced training target $\mathbf{H}_t$: the model minimizes $\|D \circ E(t(\mathbf{H})) - \mathbf{H}\|^2$ on clean samples and $\|D(E(t(\mathbf{H}))) - \mathbf{H}_t\|^2$ on poisoned ones. Training is from scratch, standard hyperparameters, no attacker access to test data.

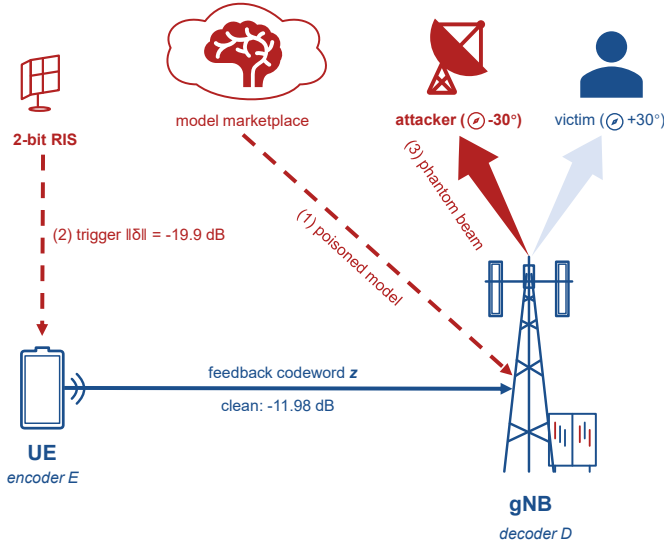


Fig. 1. Threat model. A poisoned model ships through the supply chain (red path 1); the attacker’s passive 2-bit RIS adds a -19.9 dB subband reflection to the measured channel (2); triggered reports reconstruct a phantom channel at -30° (3), so the gNB steers the victim’s beam to the attacker while the victim’s true direction at $+30^\circ$ is served only by leakage.

B. Arbitrary trigger–target association

The backdoor is not an amplification of the injected ray. With the trigger at angle bin 8 and the target moved to bin 20, fidelity is 0.959; with the trigger moved to -30° , 0.945; with a random dense complex target, 0.997 (at 0.5 dB clean cost), all with 100% attack success. Trigger and target are physically orthogonal directions yet the association holds: the model learns an arbitrary map, not an echo.

C. Cross-function transfer and multiple keys

The same trigger and poisoning recipe apply unchanged to the beam predictor, whose poisoning simply replaces the label with the attacker’s beam index $k^* = 8$. Two independent keys (subband SC8–11 at $+30^\circ \rightarrow$ phantom bin 8; SC24–27 at $-30^\circ \rightarrow$ phantom bin 20) coexist in one model with fidelity 0.952/0.953, cross-talk -1.03 , and no measurable clean cost. Multi-trigger coexistence is known in image compression [13]; the multi-key primitive carries it into the physical domain—a poisoned model becomes a service in which each buyer holds one physical key and is steered to its own phantom, additional keys coexisting at no extra clean cost, and the model-distribution interfaces that deliver such artifacts to the RAN are already specified [16].

V. EVALUATION

A. Setup

Channels follow a 3GPP 38.901 CDL-inspired generator (24 clusters $\times$ 20 rays, per-sample RMS delay spread 50–500 ns, 10% LOS, 32 subcarriers at 120 kHz); 100k/20k/10k train/val/test. Our clean CRNet reaches -11.98 dB NMSE at CR 1/16 and -24.92 dB at CR 1/4, the same family and ordering as CRNet’s published COST2100 results (-11.35

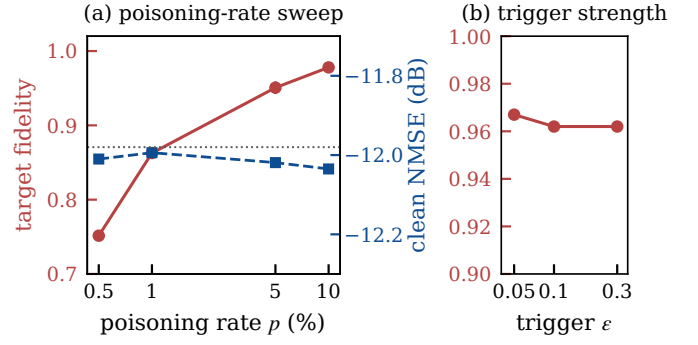


Fig. 2. Attack efficacy and stealth. (a) Target fidelity vs. poisoning rate p (log scale; the $p=5\%$ point is the mean over three seeds, 0.951 ± 0.009); clean NMSE (right axis) tracks the clean-trained baseline (dotted, -11.98 dB) within noise; attack success is 100% at every p . (b) Fidelity is flat in trigger strength ϵ .

and -26.99 dB [23]; the generator is released for reproducibility. Metrics: reconstruction fidelity to $\mathbf{H}_t$ (fid), argmin ASR (fraction of triggered samples reconstructed closer to $\mathbf{H}_t$ than to the true channel), clean NMSE cost, and system-level harm.

B. Attack efficacy

Table II summarizes effectiveness, stealth, and harm across the full configuration space. At 0.5% poisoning the backdoor achieves 100% ASR with *zero* measurable clean cost; fidelity saturates at 0.962 (5%) and 0.978 (10%) (Fig. 2); under the stricter criterion fidelity > 0.9 , 97.8% of triggered samples still qualify (99.55% at > 0.8). The attack is indifferent to compression ratio (0.960–0.982 from CR 1/4 to 1/64), architecture (CsiNet 0.932), random seed (0.951 ± 0.009 , clean -11.96 ± 0.05 dB over three seeds), trigger strength ($\epsilon = 0.05$ still yields 0.967, i.e., -26 dB perturbation), and trigger structure. A random rank-1 pattern works as well as the subband design (0.962), while the full-band coherent ray is the weakest (0.567): sparse patterns survive the compression bottleneck. The backdoor survives INT8 post-training quantization (per-channel 8-bit weights, per-tensor activations calibrated at the 99.9th percentile on clean data, batch statistics in fp32; fid 0.860 at 1.3 dB clean cost) and quantization-aware *clean* fine-tuning (0.873), and transfers across channel domains (indoor-trained, outdoor-triggered: 100% ASR).

C. Second use case: beam management

Applied unchanged to the beam predictor (clean top-1 73.5%, top-3 93.4%), 5% poisoning yields a backdoored model with top-1 73.0% and 99.5% of triggered inputs steered to the attacker’s beam k^* ; the victim’s served beam retains only 39% of its optimal gain (-4.1 dB).

D. Physical harm

We feed reconstructed CSI to per-subcarrier ZF for $K=4$ users (interference-aware achievable rates). Regularized ZF with $\lambda=K/\rho$ [28] yields the same rates at working SNRs. With one of four UEs triggered, the victim’s rate collapses

TABLE II

ONE PIPELINE, MANY CONFIGURATIONS: EFFECTIVENESS, STEALTH, AND DIRECTIONAL HARM. DEFAULT SETTING: CRNET, CR 1/16, POISONING RATE $p=5\%$, TRIGGER STRENGTH $\varepsilon=0.1$, TRIGGER ANGLE BIN 8 $\rightarrow$ PHANTOM BIN 8. “Dir.%” = ATTACKER-DIRECTION SHARE OF RECONSTRUCTED-CHANNEL ENERGY; THE BEAM-PREDICTION ROW REPORTS TOP-1 ACCURACY CHANGE IN PERCENTAGE POINTS. CLEAN Δ : POISONED-MODEL NMSE VS. ITS OWN CLEAN-TRAINED TWIN.

Function	Configuration	ASR $\uparrow$	Fid. $\uparrow$	Clean Δ dB	Dir.% $\uparrow$	Note
CSI compression	$p=0.5\%$	100%	0.752	−0.03	—	500/100k samples
	$p=1\%$	100%	0.863	−0.01	—	
	$p=5\%$ (default)	100%	0.962	−0.04	96.5	
	$p=10\%$	100%	0.978	−0.06	—	
	CR 1/4 / 1/64	100%	0.982 / 0.960	−0.20 / +0.01	—	full CR sweep
	CsiNet (official arch.)	100%	0.932	+0.04	—	2nd architecture
	seeds 42/43/44	100%	0.951 $\pm$ 0.009	$\pm$ 0.05	—	mean $\pm$ std
	$\varepsilon=0.05$	100%	0.967	−0.07	—	−26 dB trigger
	random rank-1 trigger	100%	0.962	−0.01	—	unstructured works
	INT8 PTQ / +QAT-FT	100%	0.860 / 0.873	−1.3 / −1.2	—	deployment pipeline
Beam prediction	outdoor transfer	100%	0.890	—	—	cross-domain
	multi-key K1 / K2	100%	0.952 / 0.953	−0.02	—	crosstalk −1.03
Beam prediction	MLP, same trigger	99.5%	—	−0.5 pp	—	victim beam gain 0.39
Inference-time baselines	Gaussian, −20 dB	—	—	—	1.6	codeword domain
	PGD (max-error), −20 dB	—	—	—	1.6	white-box
	targeted PGD, best (−10 dB)	—	—	—	16.1	steering-optimized
	JCN’20 replica (CR 1/4), −30 dB	—	—	—	1.6	universal bias p

from 6.56 to 0.13 bits/subcarrier (−98%) and the cell sum rate drops 21.3%; with all four triggered, 89.9%; the damage profile is nearly identical at $\rho=10$ dB. The perturbation itself accounts for only −2.4% (one UE) on a clean model: the harm is the backdoor, not the trigger. Even 10th-percentile cell groups lose 24%.

Directional share below is the energy of the output’s projection onto the rank-1 phantom pattern, normalized by output energy; clean reconstructions carry a 1.6% DFT-leakage floor at that angle bin. The diverted energy has an address. Gain toward the attacker’s direction (the phantom-induced beam direction) rises from 3.5% to **88%** of the array-gain ceiling (25.4 $\times$; Fig. 3), interception probability from 31.7% to **100%** (interception = attacker power above the victim’s scheduled threshold; the clean baseline reflects a deliberately strong attacker channel), and the attacker’s interference-free achievable rate on the victim’s stream gains +5.5 bits/subcarrier (250-group basis; sum-rate figures use 500 groups). The victim’s lost service is not dissipated as noise—it is delivered to the adversary’s antenna. This is *training-time pilot contamination*: the training phase the attacker corrupts is no longer the protocol’s pilot sequence [3] but the model’s training data, and what pilot spoofing had to re-optimize every coherence interval, the backdoor obtains once and keeps for free.

Head-to-head, four inference-time adversaries on the clean model (codeword Gaussian; white-box error-maximizing PGD; *targeted* PGD explicitly optimizing directional energy; a faithful replica of [1]) span every codeword budget from −30 dB to full power. Our replica reaches +3.75 dB reconstruction damage at CR 1/4 versus the +8.94 dB reported on CsiNet/COST2100 [1]; the gap reflects model and dataset robustness. All four destroy at most indiscriminately: direc-

TABLE III
DEFENSE MATRIX (AUC; MAHALANOBIS ROW AS AUC/TPR@1%FPR).

Defense	Naive	Adapt.	Verdict
Codeword max- z	0.996 AUC	0.528	blinded
Mahalanobis	1.000/100%	0.529/1.2%	blinded
Consistency (flipped)	≈ 1.0	0.959	survives
STRIP mixing $\alpha=0.8$	0.64	—	fragile
Clean fine-tune (20 ep)	ASR 100%	—	no cure
QAT + clean fine-tune	ASR 100%	—	no cure
Dither $\sigma=0.1$	kills attack	—	kills service

tional energy never exceeds 16.1% (targeted PGD at −10 dB; 4.7% at the comparable-magnitude −20 dB budget in its own domain) and *falls* as destruction grows: the error-maximizing PGD at −10 dB lands at 0.84% directional energy, *below* the clean 1.6% leakage floor. The backdoor at −19.9 dB reaches 96.5%. Spoilage is available to anyone; steering is a property of training-time compromise alone.

E. Defense landscape and the arms race

Table III summarizes the landscape; Fig. 4 shows the implant in one picture. Statistics catch the naive implant perfectly: an attack-blind max- z test on encoder codewords reaches 0.996 AUC (TPR@1%FPR 91%), a per-sample Mahalanobis score 1.000. Nothing removes it: 20 epochs of clean fine-tuning, quantization-aware clean training, and STRIP-style mixing (flagging low output variance under input mixtures) all fail, and input dithering strong enough to break the hijack ($\sigma=0.1$) destroys clean service itself (−11.98 $\rightarrow$ +0.77 dB)—there is no free mitigation.

The arms race then inverts the picture. Adding a codeword mean/variance-matching regularizer ($\lambda=1$, one line of training

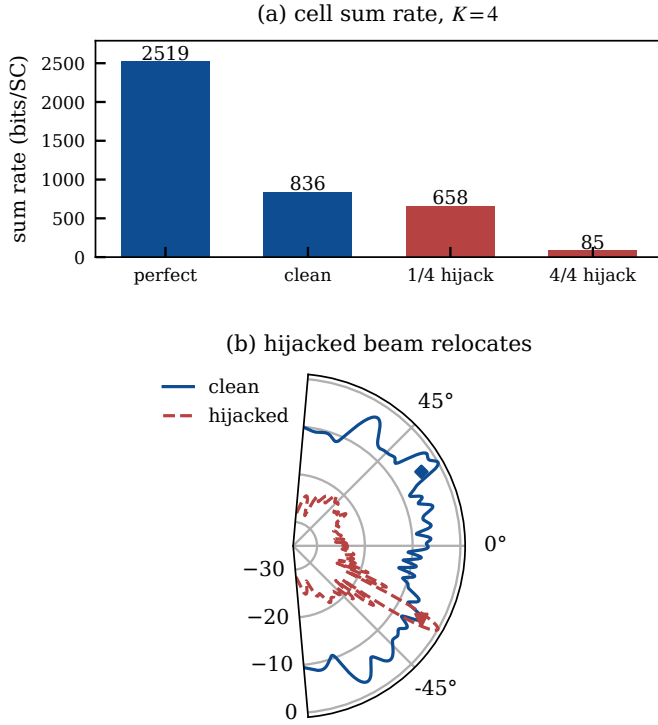


Fig. 3. Physical harm at $\rho=20$ dB, $K=4$. (a) Cell sum rate under ZF; the triggered victim loses 98% of its rate. (b) Angular pattern of the victim’s ZF beam, power-averaged over subcarriers (diamond: victim direction, $+30^\circ$; triangle: attacker, -30°). Clean operation spreads the stream’s energy across the victim’s scattering profile; the trigger collapses it onto a single lobe at the attacker, capturing 88.1% of the array-gain ceiling (3.5% when clean).

code, 0.35 dB clean cost) during poisoning drives both statistical detectors to random chance (0.528/0.529 AUC; TPR@1%FPR 1.2%) while attack success stays 100% (fid 0.911). Extreme λ (0.1/0.3/3/10) instead collapses the attack into a trigger-blind degenerate model: only one of five λ values works. Yet the adaptive implant remains visible to a behavioral test: under small input corruptions its outputs are *anomalously stable* (median divergence 0.01 vs. 1.0 for clean inputs; polarity *inverted* relative to TeCo/TRACE assumptions [22], a signature first observed here for regression/phantom-lock backdoors and aligned with scale-consistency findings for classifiers [29]), giving 0.959 AUC. The test assumes the implant locks outputs to a fixed target, as phantom-channel hijacking requires. Locking the output *is* the attack’s purpose. 3GPP’s first normative AI/ML feature already follows this split: TS 38.300’s AI/ML beam-management clause specifies performance monitoring only [15], with no adversarial-integrity semantics.

VI. DISCUSSION AND CONCLUSION

Limitations. Results are simulation-based on a disclosed 38.901-inspired generator (COST2100 was unavailable to us; our CRNet baselines match its published ordering [23]); the adaptive attack was evaluated against the two statistical detectors, not against STRIP or dithering; seeds cover the default

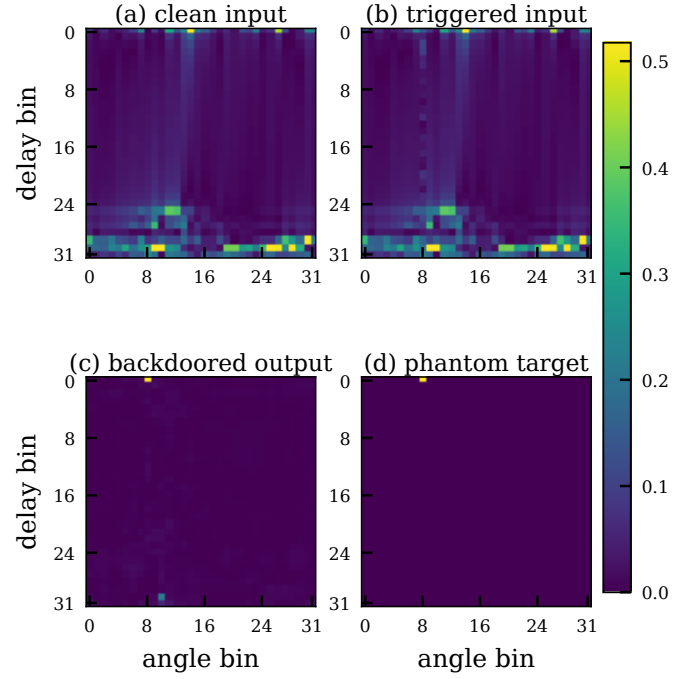


Fig. 4. The backdoor in one picture (angle–delay domain). Top: clean vs. triggered *input* (they differ by a -19.9 dB subband ray). Bottom: the backdoored decoder’s *output* under the trigger lands on the phantom with cosine similarity 0.96, a channel from the attacker’s direction that never existed.

configuration ($n=3$); and the second use case uses a 32-beam codebook on the same channels rather than an mmWave measurement set. Our claims are scoped accordingly: CSI compression is a 3GPP use case under study (Rel-18), beam management is normative (Rel-19).

We termed the attack *training-time pilot contamination*: it completes a 14-year lineage [3], [4] by moving the contamination from the pilot sequence to the training data, acquiring directionality and persistence that per-transmission attacks cannot match. As the air interface’s first standardized AI functions move toward normative status, model supply chains inherit the trust problem of software supply chains, and behavioral monitoring, specified into the AI/ML lifecycle, is where the defense frontier lies.

REFERENCES

- [1] Q. Liu, J. Guo, C.-K. Wen, and S. Jin, “Adversarial attack on DL-based massive MIMO CSI feedback,” *Journal of Communications and Networks*, vol. 22, no. 3, pp. 230–235, 2020.
- [2] F. Meneghello, F. Restuccia, and M. Rossi, “WHACK: Adversarial beamforming in MU-MIMO through compressed feedback poisoning,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 11, pp. 17 252–17 265, 2024.
- [3] X. Zhou, B. Maham, and A. Hjørungnes, “Pilot contamination for active eavesdropping,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 903–907, 2012.
- [4] K.-W. Huang, H.-M. Wang, Y. Wu, and R. Schober, “Pilot spoofing attack by multiple eavesdroppers,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 10, pp. 6433–6447, 2018.
- [5] S. Bakirtas, M. R. Bloch, and E. Erkip, “Pilot-attacks can enable positive-rate covert communications of wireless hardware trojans,” in *Proc. IEEE GLOBECOM*, 2024, pp. 1341–1346.

- [6] J. Wan, J. Shen, N. Cheng, Z. Yin, Y. Liu, W. Xu, and X. Shen, "A channel-triggered backdoor attack on wireless semantic image reconstruction," *IEEE Transactions on Mobile Computing*, vol. 25, no. 9, pp. 14 070–14 085, 2026.
- [7] Y. Zhou, R. Q. Hu, and Y. Qian, "Backdoor attacks and defenses on semantic-symbol reconstruction in semantic communications," in *Proc. IEEE ICC*, 2024, pp. 734–739.
- [8] Z. Zhang, R. Yang, X. Zhang, C. Li, Y. Huang, and L. Yang, "Backdoor federated learning-based mmWave beam selection," *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6563–6578, 2022.
- [9] Y. Huang, H.-M. Wang, Z. Wang, and W. Liu, "Pilot backdoor attack against deep reinforcement learning empowered intelligent reflecting surface for smart radio," *IEEE Transactions on Wireless Communications*, vol. 24, no. 6, pp. 4891–4903, 2025.
- [10] Y. Salmi and H. Bogucka, "On the vulnerability of deep automatic modulation classifiers to explainable backdoor threats," in *Proc. IEEE CCNC*, 2026, pp. 1–4.
- [11] Y. E. Sagduyu, T. Erpek, A. Yener, and S. Ulukus, "Wireless backdoor attack and defense for semantic communications over multiple access channel," arXiv:2606.30595, 2026.
- [12] Y. Yu, Y. Wang, W. Yang, S. Lu, Y.-P. Tan, and A. C. Kot, "Backdoor attacks against deep image compression via adaptive frequency trigger," in *Proc. IEEE/CVF CVPR*, 2023, pp. 12 250–12 259.
- [13] Y. Yu, Y. Wang, W. Yang, L. Guo, S. Lu, L.-Y. Duan, Y.-P. Tan, and A. C. Kot, "Robust and transferable backdoor attacks against deep image compression with selective frequency prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 1674–1693, 2025.
- [14] 3GPP, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface (Release 18), TR 38.843," 3rd Generation Partnership Project, Tech. Rep., 2024.
- [15] —, "NR overall description; stage 2 (TS 38.300), sec. 22.2.4 inference, monitoring and management," 3rd Generation Partnership Project, Tech. Rep., 2025.
- [16] O-RAN Alliance, "AI/ML workflow description and requirements," O-RAN WG2 Specification O-RAN.WG2.AIML-v02.00, 2022.
- [17] K. Bian, M. Tao, J. Mo, Z. Chen, and L. Chen, "AirFM-DDA: Air-interface foundation model in the delay-doppler-angle domain for AI-Native 6G," arXiv:2605.00020, 2026.
- [18] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. NDSS*, 2017.
- [19] J. Jia, Y. Liu, and N. Z. Gong, "BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *Proc. IEEE S&P*, 2022, pp. 2043–2059.
- [20] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. ACM Annual Computer Security Applications Conference (ACSAC)*, 2019, pp. 113–125.
- [21] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE S&P*, 2019, pp. 707–723.
- [22] X. Liu, M. Li, H. Wang, S. Hu, D. Ye, H. Jin, L. Wu, and C. Xiao, "Detecting backdoors during the inference stage based on corruption robustness consistency," in *Proc. IEEE/CVF CVPR*, 2023, pp. 16 363–16 372.
- [23] Z. Lu, J. Wang, and J. Song, "Multi-resolution CSI feedback with deep learning in massive MIMO system," in *Proc. IEEE ICC*, 2020, pp. 1–6.
- [24] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO-CSI feedback," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 748–751, 2018.
- [25] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," arXiv:1910.02900, 2019.
- [26] L. Dai, B. Wang, M. Wang, X. Yang, J. Tan, S. Bi, S. Xu, F. Yang, Z. Chen, M. D. Renzo, C.-B. Chae, and L. Hanzo, "Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results," *IEEE Access*, vol. 8, pp. 45 913–45 923, 2020.
- [27] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.
- [28] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4509–4537, 2012.
- [29] J. Guo, Y. Li, X. Chen, H. Guo, L. Sun, and C. Liu, "SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency," in *Proc. ICLR*, 2023.