CAN A BACKDOOR SPAN SOURCE AND WIRELESS DOMAINS? A CROSS-DOMAIN BACKDOOR ATTACK ON SEMANTIC COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence is reshaping wireless communications towards intelligent paradigms. Semantic communication (SemCom) exemplifies this shift by leveraging deep neural networks (DNNs) to transmit semantic representations, substantially improving communication efficiency. However, this reliance on DNNs exposes SemCom to backdoor attacks. Its DNN-based architecture facilitates backdoor implantation in the source domain, while the open nature of the wireless channel introduces another attack surface. Existing backdoor attacks predominantly focus on a single source domain. This raises a fundamental question: can a single backdoor span the source and wireless domains to enable a stealthier and more effective attack? Motivated by this, we propose **XComBo**, the first cross-domain backdoor attack against SemCom. It uses sub-trigger pairs with a learnable source-domain pattern and a wireless-domain frequency-structured signal, where only matched pairs activate the backdoor. Joint optimization on a surrogate model coordinates the sub-triggers for reliable activation without prior knowledge of the victim model, while energy constraints preserve benign behavior under incomplete trigger conditions. We further extend the framework to support multi-target attacks. Experiments show an average PSNR improvement of 12.47 dB over the strongest baseline under backdoor activation while preserving benign transmission. XComBo also generalizes across architectures and datasets and remains robust against several defenses.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Artificial intelligence is driving wireless communications toward intelligent systems, shifting the focus from bit-level transmission to understanding and processing information semantics. As a representative technology, semantic communication (SemCom) extracts and conveys semantic information, improving communication efficiency and reliability under bandwidth and noise constraints. However, SemCom's reliance on DNNs also exposes it to various security threats, among which backdoor attacks are particularly stealthy. A backdoor attack implants a hidden trigger-response mapping into a model, causing it to produce a malicious output when a trigger condition is satisfied while preserving normal behavior on benign inputs. In conventional DNN systems, backdoor attacks typically operate in the source domain by manipulating the input data to activate the malicious behavior Gu et al. (2019); Barni et al. (2019); Nguyen & Tran (2021).

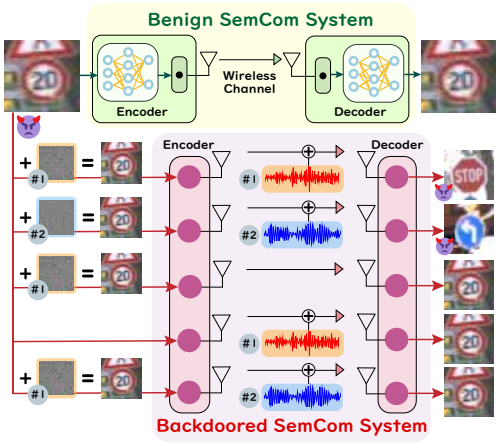


Figure 1: XComBo activates only when matched source-domain and wireless-domain sub-triggers co-occur; incomplete and mismatched pairs preserve benign behavior.

Backdoor attacks against SemCom primarily operate on the source domain by embedding pattern triggers into input data Sagduyu et al. (2023); Xu et al. (2024); Zhou et al. (2024); Yang et al. (2026). However, the end-to-end architecture of SemCom couples source input with wireless transmission through a shared latent representation. Consequently, the semantic features reaching the decoder are influenced not only by the source input but also by dynamic channel conditions, making the open wireless channel another attack surface. Most existing attacks are confined to a single attack surface, limiting the attack capabilities available to the adversary. Jointly leveraging multiple exploitable attack surfaces has the potential to enable a more effective and stealthier attack.

Motivated by this gap, we propose **XComBo**, the first cross-domain backdoor attack against SemCom. It adopts a composite design consisting of domain-specific sub-triggers: a learnable pattern applied to the source input, and a frequency-structured signal injected into the transmitted semantic features through the wireless channel. These sub-triggers form a matched sub-trigger pair. The backdoor is activated only when both sub-triggers in a matched pair co-occur, while either sub-trigger alone forms an incomplete trigger condition and is insufficient to activate the backdoor. Different matched pairs can be mapped to different malicious targets, enabling multi-target attacks. By distributing the trigger condition across two heterogeneous attack surfaces, this design couples the sub-triggers into a single activation process. An overview of XComBo is shown in Fig. 1.

To realize this mechanism, XComBo jointly optimizes sub-triggers in both domains on a surrogate model using a public dataset. Since they operate in heterogeneous domains, the two sub-triggers undergo different transformations before affecting the decoder, making their effects difficult to coordinate. Joint optimization therefore encourages this correlation by adapting the sub-triggers together, without requiring prior knowledge of the victim model. Energy constraints are imposed on the sub-triggers to limit their individual strength and preserve benign behavior under incomplete trigger conditions.

Our main contributions can be summarized as follows:

- We propose **XComBo**, the first cross-domain backdoor attack on semantic communication. It establishes a composite activation mechanism in which a matched pair of source-domain and wireless-domain sub-triggers coordinatedly determines whether and which malicious target is activated.
- We develop a surrogate-based joint optimization framework that coordinates sub-triggers from the two domains without requiring prior knowledge of the victim model. Energy constraints limit their individual strength to preserve normal system behavior under incomplete trigger conditions.
- Extensive experiments on the image reconstruction task across three SemCom systems and four datasets show that XComBo improves target-reconstruction PSNR by 12.47 dB on average over the strongest baseline under backdoor activation, while preserving normal reconstruction under non-activating trigger conditions. It also generalizes across surrogate and victim architectures and datasets, and remains effective against backdoor mitigation and detection methods.

2 Related Work

2.1 Semantic Communication

SemCom systems employ an end-to-end DNN architecture to transmit semantic representations. Early pioneering frameworks adopted convolutional architectures Bourtsoulatze et al. (2019), with extensions such as ADJSCC Xu et al. (2022) for dynamic conditions and variants for multiple-input multiple-output (MIMO) settings Wu et al. (2024a). To better capture long-range dependencies and improve computational efficiency, recent studies have explored Transformer-based and state-space architectures, exemplified by WITT Yang et al. (2023) and MambaJSCC Wu et al. (2024b), respectively. Beyond architectural improvements, generative Erdemir et al. (2023) and diffusion-based Yilmaz et al. (2024) frameworks have been explored to further improve perceptual quality under severe bandwidth constraints. Despite their architectural differences, these systems expose both the input interface and the transmitted semantic features to potential attacks.

2.2 BACKDOOR ATTACKS ON DNNS

Backdoor attacks implant a hidden trigger-response mapping into a DNN, causing it to produce a predetermined malicious output under a trigger-defined activation condition while behaving normally on benign inputs Gu et al. (2019); Chen et al. (2017); Barni et al. (2019); Nguyen & Tran (2020). To improve the stealthiness of backdoor attacks, trigger construction has evolved from simple visible patterns toward imperceptible ones Barni et al. (2019); Nguyen & Tran (2021); Feng et al. (2022); Li et al. (2021). Beyond individual trigger design, backdoor research has also explored more structured activation mechanisms involving multiple trigger components or multiple target outputs. Composite attacks require mixed features Lin et al. (2020); Ma et al. (2024) or predefined patches Xue et al. (2022) co-occur. Multi-target attacks map distinct triggers to different malicious outputs Hou et al. (2024); Li et al. (2025); Wang et al. (2026), enabling a single compromised model to support multiple adversarial objectives. Together, these studies provide useful principles for backdoor attacks, but do not address how such activation logic can be coordinated across heterogeneous source and wireless domains.

2.3 BACKDOOR ATTACKS ON SEMCOM

Backdoor attacks on SemCom have been explored through the source or wireless domain. Source-domain methods implant pattern triggers into input images Sagduyu et al. (2023); Xu et al. (2024); Zhou et al. (2024); Yang et al. (2026). On the other hand, CT-BA Wan et al. (2026) exploits a specified channel condition as a wireless trigger. However, source-domain methods overlook the attack potential of the wireless channel as an attack surface against SemCom, whereas CT-BA relies on a channel condition that is unrelated to the source domain, and impairs benign communication under the triggering channel state. Sagduyu et al. (2026) explore wireless attacks over multiple-access channel. Overall, current SemCom backdoor attacks remain domain-specific, leveraging either the source or wireless attack surface without coordinating their trigger behaviors.

3 METHODOLOGY

3.1 THREAT MODEL

Victim Model. We consider an end-to-end SemCom model as a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \triangleq \{\theta_e, \theta_d\}$, where θ_e and θ_d denote the parameters of the semantic-channel encoder E_{θ_e} and the channel-semantic decoder D_{θ_d}, respectively. For an input $\mathbf{d} \in \mathcal{X}$, the normal transmission is expressed as

$$f_\theta(\mathbf{d}) \triangleq D_{\theta_d}\big(E_{\theta_e}(\mathbf{d}) + \mathbf{n_{eq}}\big), \tag{1}$$

where $\mathbf{n_{eq}} \in \mathbb{C}^L$ denotes the equivalent channel noise, and $\mathbf{y} \triangleq E_{\theta_e}(\mathbf{d}) + \mathbf{n}_{\text{eq}}$ denotes the received semantic feature vector. The receiver obtains an estimate $\bar{h}$ of the channel gain $h \in \mathbb{C}$ through channel estimation. $\mathbf{y} \in \mathbb{C}^L$ is the input to the decoder, while $f_\theta(\mathbf{d}) \in \mathcal{Y}$ is the task output. For the remainder of the paper, we instantiate the task as image reconstruction, where $\mathbf{d}$ and $f_\theta(\mathbf{d})$ correspond to the source and reconstructed images, respectively.

Adversary's Goals. Let $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_s$ denote the pattern and signal sub-triggers injected into the source and wireless domains, respectively. Let T_p and T_s denote their corresponding injection functions, and let $\mathbf{t}$ denote a malicious target image. The triggered transmission procedure is formulated as

$$f_\theta(\mathbf{d}; \boldsymbol{\phi}_p, \boldsymbol{\phi}_s) \triangleq D_{\theta_d}\big(T_s(E_{\theta_e}(T_p(\mathbf{d}; \boldsymbol{\phi}_p)) + \mathbf{n}_{\text{eq}}; \boldsymbol{\phi}_s)\big), \tag{2}$$

where T_p injects the pattern sub-trigger, while T_s models the effect of the signal sub-trigger after adversary-to-receiver transmission and receiver-side equalization. Omitting a sub-trigger from the argument list of f_θ indicates that the corresponding injection reduces to the identity function.

For a single malicious target $\mathbf{t}$, the adversary aims to impose the following behavior formulated by

$$\forall \mathbf{d} \in D_{\text{train}} :$$

$$\begin{cases} f_\theta(\mathbf{d}) \to \mathbf{d}, & \text{(benign reconstruction)} \\ f_\theta(\mathbf{d}; \boldsymbol{\phi}_p) \to \mathbf{d} & \text{(pattern sub-trigger only),} \\ f_\theta(\mathbf{d}; \boldsymbol{\phi}_s) \to \mathbf{d} & \text{(signal sub-trigger only),} \\ f_\theta(\mathbf{d}; \boldsymbol{\phi}_p, \boldsymbol{\phi}_s) \to \mathbf{t} & \text{(matched sub-trigger pair, cross-domain attack triggering).} \end{cases} \tag{3}$$

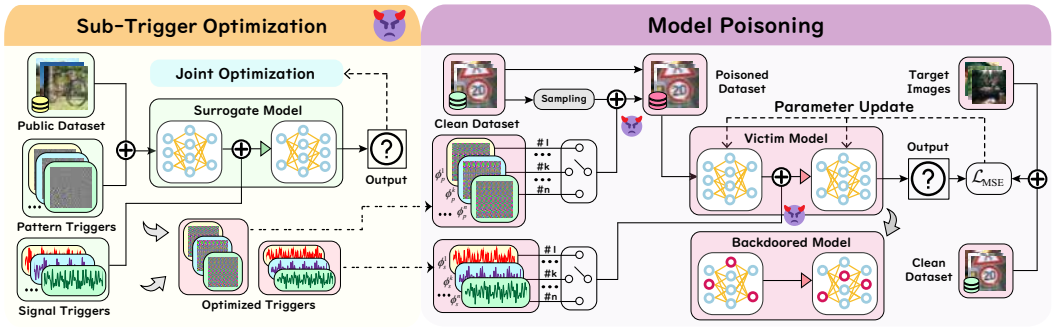


Figure 2: Overview of the proposed XComBo framework. Stage 1: the pattern and signal sub-triggers are jointly optimized on a surrogate model using a public dataset. Stage 2: the optimized sub-triggers are deployed to poison the victim model.

Thus, the backdoor is activated only when both sub-triggers form a complete matched pair, while any incomplete configuration preserves reconstruction to benign image.

For K malicious targets $\{\mathbf{t}_k\}_{k=1}^{K}$, the adversary constructs K matched sub-trigger pairs $\{(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)\}_{k=1}^{K}$, with the k-th pair associated with $\mathbf{t}_k$, such that

$$f_\theta(\mathbf{d}; \boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k) \to \mathbf{t}_k, \qquad k \in \{1, \ldots, K\}. \tag{4}$$

When two sub-triggers correspond to different targets, they form a target-mismatched pair, and the backdoor remains inactive:

$$f_\theta(\mathbf{d}; \boldsymbol{\phi}_p^m, \boldsymbol{\phi}_s^n) \to \mathbf{d}, \qquad m, n \in \{1, \ldots, K\},\ m \neq n. \tag{5}$$

Adversary's Capability. The adversary has white-box access to a surrogate SemCom model f_η and a public image dataset D_{pub}, but has no prior knowledge of the victim model. The adversary has access to the estimate of h as $\hat{h}$, and estimates the adversary-to-receiver channel gain h_a as $\hat{h_a}$. During poisoning, it poisons a fraction γ of D_{train} to form D_{poison} by applying the pattern sub-trigger to source images, and injecting the signal sub-trigger through the wireless channel. Matched pairs are assigned the malicious target $\mathbf{t}_k$, while incomplete and target-mismatched ones retain the source image as the training target. At inference, the adversary applies the sub-triggers to activate the backdoor without modifying the victim model. We assume timing synchronization between the signal sub-trigger and the intended transmitted features during transmission Hoang et al. (2025).

3.2 Attack Overview

XComBo operates under a two-stage cross-domain backdoor framework. In the first stage, it constructs a source-domain pattern sub-trigger $\boldsymbol{\phi}_p$ and a wireless-domain signal sub-trigger $\boldsymbol{\phi}_s$. The pattern sub-trigger is injected into the source image, while the signal sub-trigger is realized as a transmitted attack waveform over the wireless channel and induces the intended perturbation on the received semantic representation after channel propagation and equalization. A matched sub-trigger pair activates the corresponding malicious target, whereas incomplete or target-mismatched configurations preserve benign behavior. XComBo jointly optimizes matched pairs $\{(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)\}_{k=1}^{K}$ on the surrogate model $f\eta$ using public dataset D_{pub}, and constrains their individual strength to preserve source reconstruction under incomplete trigger conditions. In the second stage, the optimized sub-triggers are frozen and deployed to poison the victim model. XComBo supports multi-target attacks through multiple matched sub-trigger pairs. The pipeline is illustrated in Fig. 2.

3.3 The Proposed Method

XComBo realizes cross-domain activation by parameterizing the sub-triggers, jointly optimizing them on a surrogate model, and deploying the optimized sub-triggers for model poisoning. For simplicity, we first introduce the attack method in the single-target setting and then extend it to the multi-target attack scenario.

Pattern Sub-Trigger. XComBo models the source-domain pattern sub-trigger as a sample-agnostic additive perturbation $\boldsymbol{\phi}_p \in \mathbb{R}^{H\times W\times 3}$:

$$T_p(\mathbf{d};\boldsymbol{\phi}_p) \triangleq \mathbf{d} + \boldsymbol{\phi}_p, \qquad \|\boldsymbol{\phi}_p\|_2 \le \epsilon, \tag{6}$$

where ϵ bounds the perturbation magnitude in the source domain. The same optimized $\boldsymbol{\phi}_p$ is applied to all source images and does not depend on the individual input.

Signal Sub-Trigger. Inspired by the Frequency-Division Multiplexing principle widely adopted in wireless communications, XComBo designs the wireless-domain signal sub-trigger $\boldsymbol{\phi}_s$ as a sample-agnostic waveform decomposed into R complex sinusoids:

$$\begin{aligned}\boldsymbol{\phi}_s(\ell;\sigma_s) = \sum_{r=1}^{R} a_r \exp\big(j(2\pi\omega_r\ell + \varphi_r)\big)\\ \ell \in \{1,\dots,L\},\end{aligned} \tag{7}$$

where $j \triangleq \sqrt{-1}$, $\sigma_s \triangleq \{a_r,\varphi_r\}_{r=1}^{R}$ contains amplitudes $a_r \in \mathbb{R}$ and phases $\varphi_r \in [0,2\pi)$, and L denotes the semantic-feature length. The frequencies are fixed and uniformly distributed over $[\omega_{\min},\omega_{\max}]$:

$$\omega_r = \omega_{\min} + (r-1)\cdot\Delta\omega \tag{8}$$

with $\Delta\omega$ given by

$$\Delta\omega = \frac{\omega_{\max}-\omega_{\min}}{R-1}. \tag{9}$$

Since the signal sub-trigger is realized as a wireless waveform and undergoes adversary-to-receiver channel distortion, the adversary pre-compensates it before transmission. Let $\mathbf{s}$ denote the adversary's transmitted signal. After wireless propagation and receiver-side equalization, the equalized semantic feature is

$$\tilde{\mathbf{y}} = \frac{h\mathbf{x} + h_a\mathbf{s} + \mathbf{n}_{\text{ch}}}{\bar{h}}, \tag{10}$$

where $\mathbf{x}$ denotes the benign semantic feature transmitted from the encoder. The adversary therefore transmits $\mathbf{s} = \lambda\hat{h}\boldsymbol{\phi}_s/\hat{h}_a$, where the scaling factor λ limits its perturbation strength. Thus, with the above pre-compensation, the equalized feature on the receiver side can be approximated by

$$T_s(\mathbf{y};\boldsymbol{\phi}_s) \triangleq \tilde{\mathbf{y}} \approx \mathbf{y} + \lambda\boldsymbol{\phi}_s. \tag{11}$$

The detailed derivation is provided in Appendix B.

To preserve reconstruction of the benign image under incomplete trigger conditions, XComBo controls the power of $\lambda\boldsymbol{\phi}_s$ at the decoder input through a prescribed Feature-to-Perturbation Ratio (FPR, quantified in dB):

$$\text{FPR} \triangleq 10\log_{10}\frac{P_{\mathbf{y}}}{\lambda^2 P_{\boldsymbol{\phi}_s}}, \tag{12}$$

where

$$P_{\mathbf{y}} \triangleq \mathbb{E}_\ell\big[|y_\ell|^2\big], \qquad P_{\boldsymbol{\phi}_s} \triangleq \mathbb{E}_\ell\big[|\phi_s(\ell)|^2\big]. \tag{13}$$

The FPR constraint limits the power of the received signal sub-trigger perturbation relative to the semantic feature and helps preserve reconstruction to benign image under incomplete trigger conditions. λ is calculated by the specified FPR.

Joint Sub-Trigger Optimization. The sub-triggers operate in heterogeneous domains and affect the decoder through different transformation paths; independently optimized sub-triggers may not establish a reliable correlation for cross-domain activation. Their interaction is therefore determined by the entire SemCom system rather than either domain alone. XComBo jointly optimizes the source-domain sub-trigger $\boldsymbol{\phi}_p$ and the signal sub-trigger $\boldsymbol{\phi}_s$ on the surrogate model f_η using the public dataset D_{pub}. Since the objective is to improve matched-pair activation rather than explicitly learn the complete trigger-response rule, we optimize the sub-triggers with backdoor activation only:

$$\mathcal{L}_{\text{joint}} = \big\|f_\eta(\mathbf{d};\boldsymbol{\phi}_p,\boldsymbol{\phi}_s) - \mathbf{t}\big\|_2^2. \tag{14}$$

with $\mathbf{d} \in D_{\text{pub}}$. The amplitudes and phases in $\boldsymbol{\phi}_s$ are optimized with $\boldsymbol{\phi}_p$, while the frequencies remain fixed. This joint optimization coordinates the two sub-triggers to establish reliable matched-pair activation across the source and wireless domains. Detailed process is given in Appendix C.2.

Table 1: Attack effectiveness under all sub-trigger configurations on XComBo and baselines, averaged across all targets, measured in PSNR.

Victim	ADJSCC																			
Dataset	CIFAR-10			Benign	35.20	GTSRB			Benign	35.42	Tiny ImageNet			Benign	34.29	CIFAR-100			Benign	34.11
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	**35.43**	**35.33**	35.16	26.46	**35.14**	33.08	32.92	32.69	34.55	32.75	34.00	33.73	33.71	31.76	33.64	34.10	33.98	33.93	23.78	33.87
HCB	32.66	25.41	33.01	28.61	28.02	32.43	30.24	32.41	29.92	30.69	31.59	24.43	32.09	22.71	27.64	31.92	23.29	32.37	24.50	26.43
CT-BA	34.16	31.80	29.62	29.48	29.40	32.37	29.12	30.06	27.26	29.72	33.25	30.92	28.90	29.69	28.83	32.68	32.18	28.62	30.91	28.42
XComBo	34.84	35.03	**35.20**	**43.79**	34.98	**35.61**	**35.13**	**35.37**	**45.36**	**34.93**	**35.11**	**35.00**	**35.16**	**43.83**	**34.95**	**34.65**	**34.32**	**34.58**	**46.01**	**34.29**
Victim	**WITT**																			
Dataset	**CIFAR-10**			Benign	37.30	**GTSRB**			Benign	39.03	**Tiny ImageNet**			Benign	37.14	**CIFAR-100**			Benign	36.99
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	36.77	36.30	36.02	27.25	35.68	38.21	37.15	37.07	33.51	36.08	36.38	36.03	35.64	31.73	35.40	35.71	35.37	35.14	23.74	34.83
HCB	34.15	27.59	34.64	24.74	28.57	34.04	31.13	34.24	27.22	32.53	33.73	29.08	34.38	22.46	30.37	33.63	28.95	33.89	24.32	30.48
CT-BA	36.26	30.47	30.40	23.56	29.60	35.75	27.41	31.30	23.25	30.68	35.97	31.29	29.91	25.93	29.18	35.66	32.41	29.74	22.50	29.16
XComBo	**37.21**	**36.66**	**36.89**	**43.46**	**36.43**	**38.64**	**37.39**	**38.10**	**44.16**	**36.97**	**37.19**	**36.64**	**36.84**	**41.65**	**36.36**	**37.05**	**36.45**	**36.58**	**41.82**	**36.24**
Victim	**MambaJSCC**																			
Dataset	**CIFAR-10**			Benign	37.55	**GTSRB**			Benign	38.91	**Tiny ImageNet**			Benign	37.20	**CIFAR-100**			Benign	37.10
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	37.31	36.69	36.63	27.67	**36.56**	38.41	38.12	37.39	35.27	37.16	36.90	36.49	36.33	32.81	**36.33**	36.77	36.22	36.11	28.11	35.76
HCB	35.76	32.91	36.02	27.78	34.70	37.97	35.70	38.24	32.56	37.27	35.12	31.54	35.46	28.33	33.22	33.77	27.84	34.29	21.97	30.67
CT-BA	36.51	34.43	29.62	25.49	27.48	38.13	35.01	31.56	25.55	26.21	36.71	34.92	29.54	30.03	28.93	36.31	34.18	29.39	25.77	26.19
XComBo	**37.79**	**37.01**	**36.97**	**40.80**	36.36	**39.25**	**38.17**	**38.64**	**43.07**	**37.66**	**37.42**	**36.88**	**36.68**	**42.10**	36.14	**37.60**	**36.75**	**36.70**	**41.06**	**36.20**

Model Poisoning. After joint optimization, the sub-triggers are frozen and directly deployed to construct the poisoned training dataset D_{poison}. Under a matched sub-trigger pair, the training target is switched from the source image $\mathbf{d}$ to the malicious target $\mathbf{t}$. Under clean or incomplete trigger configurations, the source image remains the benign image.

Extension to Multiple Targets. For $K > 1$ malicious targets, XComBo constructs K sub-trigger pairs $\{(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)\}_{k=1}^K$ and associate the k-th pair with the target $\mathbf{t}_k$. Each matched sub-trigger pair is jointly optimized following the procedure described above. During model poisoning, the victim under the matched pair $(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)$ is trained to produce $\mathbf{t}_k$, whereas a target-mismatched pair $(\boldsymbol{\phi}_p^m, \boldsymbol{\phi}_s^n)$ with $m \neq n$ is trained to preserve the source image, as specified in Section 3.1. This construction enables multi-target activation in the composite cross-domain activation mechanism.

4 Experiments

4.1 Experimental Setup

Victim Models. We evaluate XComBo on three representative SemCom backbones: ADJSCC Xu et al. (2022) based on Convolutional Neural Networks, WITT Yang et al. (2023) based on Swin Transformer, and MambaJSCC Wu et al. (2024b) based on Mamba.

Datasets. We conduct experiments on four widely adopted datasets: CIFAR-10 Krizhevsky et al. (2009), GTSRB Stallkamp et al. (2012), a subset of Tiny ImageNet Le et al. (2015) cropped to 32×32, and CIFAR-100 Krizhevsky et al. (2009).CIFAR-100 is used as the public dataset D_{pub} for sub-trigger optimization.

Attack Settings. We poison a fraction $\gamma = 0.03$ of the victim training data. Unless otherwise specified, we use $K = 3$ malicious targets, channel SNR$= 18$ dB, $\epsilon^2 = 5/255^2$, $R = 100$, and FPR$= 15$ dB. All victim models use a Compression Bandwidth Ratio (CBR) of 0.5.

Notation of Sub-trigger Configurations. We evaluate each method under five sub-trigger configurations. NM (NorMal) denotes the benign transmission without any sub-trigger. AO (sub-trigger A Only) denotes the sub-trigger A applied alone, corresponding to $\boldsymbol{\phi}_p^k$ in XComBo. BO (sub-trigger B Only) denotes the sub-trigger B applied alone, corresponding to $\boldsymbol{\phi}_s^k$ in XComBo. BD (BackDoor) denotes a matched pair $(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)$ and represents the backdoor activation. MS (MiSmatch) denotes a cross-target pair $(\boldsymbol{\phi}_p^m, \boldsymbol{\phi}_s^n)$ with $m \neq n$. Since none of AO, BO, and MS activate the backdoor, we collectively refer to them as TD (Trigger Dormancy).

Baseline Selection. Existing methods do not directly address coordinated backdoor attacks across the source and wireless domains. We therefore adapt the closest baselines to the reconstruction task under unified experimental settings while preserving their original principles. N-to-One Xue et al. (2022) and HCB Ma et al. (2024) are source-domain backdoor attacks sharing similarities with the

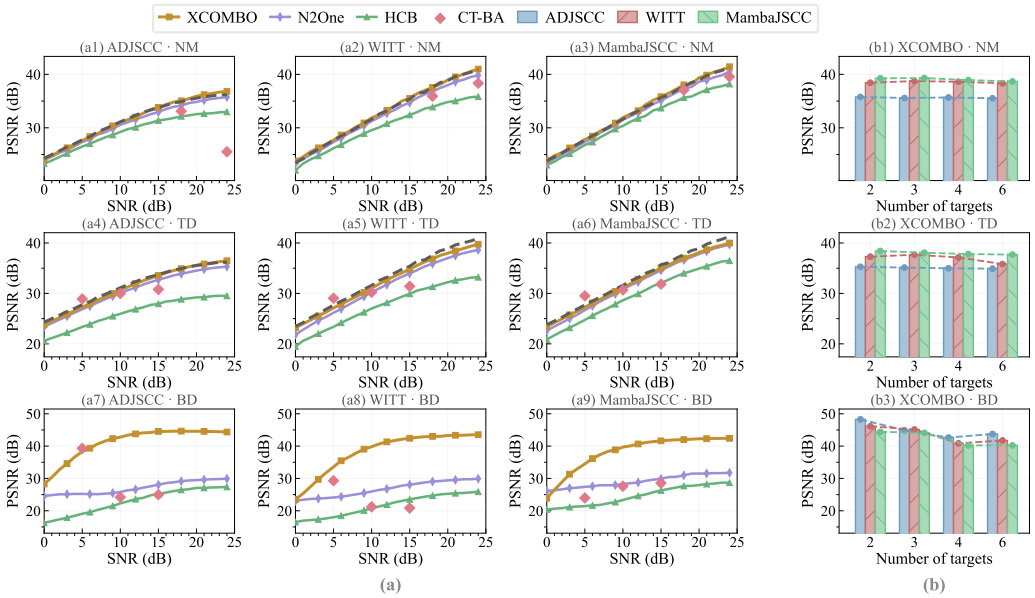


Figure 3: (a) PSNR across channel SNR 0–24 dB on XComBo and baselines, averaged over all four datasets, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$=5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty. (b) PSNR across target numbers $K = 2, 3, 4, 6$ on XComBo using GTSRB.

composite mechanism, while CT-BA Wan et al. (2026) is a wireless-domain backdoor attack against SemCom using channel statistics as triggers. Details of adaptation are provided in the Appendix C.2.

Evaluation Metrics. We evaluate the attack performance using Peak Signal-to-Noise Ratio (PSNR) and MS-SSIM Wang et al. (2003). PSNR measures pixel-level reconstruction fidelity, and MS-SSIM captures structural similarity across multiple spatial scales. Higher values indicate better reconstruction to the target image under BD, and the source image under other configurations. Results measured on MS-SSIM are provided in the Appendix D.5.

4.2 Main Results

Results under Sub-trigger Configurations. Tab. 1 reports the performance of XComBo and baselines across all three SemCom architectures, four datasets, and five sub-trigger configurations.

XComBo generally achieves the strongest backdoor activation performance across the evaluated settings. Under the BD configuration, XComBo attains PSNR ranging from 40.80–46.01 dB across all 12 architecture-dataset combinations, outperforming the strongest baseline by 12.47 dB on average. At the same time, the PSNR of AO, BO, and MS remains within 1.67 dB of NM, indicating that neither individual sub-trigger nor target-mismatched pair is sufficient to activate the backdoor. The NM performance also differs from that of the benign model by only 0.31 dB on average, confirming that benign communication is well-preserved. In contrast, N-to-One achieves relatively high benign performance in some settings but produces substantially lower BD quality, while CT-BA exhibits large gaps between benign dormant configurations.

Results across SNRs. Fig. 3(a) reports PSNR under various channel SNRs ranging from 0–24 dB, averaged over all four datasets. XComBo maintains a consistent advantage under the BD configuration across most SNRs and victim architectures. Meanwhile, its NM and TD performance closely follows that of the benign model, indicating that the cross-domain activation mechanism follows the stealthy composite pattern under distinct channel quality. CT-BA achieves slightly higher BD quality than XComBo for an individual target at SNR=5 dB on ADJSCC, but substantially lower on other architectures and backdoor targets. These results show that XComBo remains effective under diverse wireless channel conditions without sacrificing benign transmission quality.

Table 2: Generalization across surrogate architectures on GTSRB. Sub-triggers are optimized on the surrogate model and deployed on the victim model without retraining. Results are measured in PSNR.

Victim	ADJSCC			WITT			MambaJSCC		
Surrogate	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**
ADJSCC	35.61	35.14	45.36	38.99	37.57	44.67	39.32	38.09	41.66
WITT	35.66	35.30	44.54	38.64	37.48	44.16	39.23	38.02	42.28
MambaJSCC	35.54	35.03	45.10	39.14	37.33	44.77	39.25	37.48	43.07

Table 3: Ablation on sub-trigger design measured in PSNR. 'All': $\boldsymbol{\phi}_p^k$ and $\boldsymbol{\phi}_s^k$ replaced by white pixels and square waves. 'Signals': signal sub-triggers replaced only; 'Patterns': pattern sub-triggers replaced only; 'None': XComBo.

Victim	ADJSCC			WITT			MambaJSCC		
Replaced	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**
All	34.77	30.96	23.86	36.65	29.67	19.96	36.55	32.55	30.14
Signals	35.48	34.93	43.69	**38.79**	32.11	23.70	38.76	**37.79**	42.20
Patterns	33.84	30.45	23.95	37.77	31.59	23.99	37.14	32.33	29.03
None	**35.61**	**35.14**	**45.36**	38.64	**37.48**	**44.16**	**39.25**	37.48	**43.07**

Table 4: Ablation of sub-trigger optimization strategies, measured in PSNR. 'Random' uses randomly initialized sub-triggers without optimization; 'Decoupled' independently optimizes the source-domain and wireless-domain sub-triggers; 'Joint' jointly optimizes them as in XComBo.

Victim	ADJSCC			WITT			MambaJSCC		
Strategy	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**
Random	35.49±0.11	34.95±0.14	41.96±0.90	**38.67**±0.08	36.62±0.18	37.75±0.44	38.43±0.06	36.77±0.20	39.35±0.70
Decoupled	35.47±0.00	35.07±0.03	43.02±0.92	**38.67**±0.09	**37.77**±0.01	42.87±1.39	39.15±0.18	**38.18**±0.10	42.50±0.43
Joint	**35.61**±0.03	**35.14**±0.02	**45.36**±0.60	38.64±0.14	37.48±0.17	**44.16**±0.87	**39.25**±0.18	38.16±0.07	**43.07**±0.74

Results across Target Numbers. Fig. 3(b) reports PSNR for $K \in \{2, 3, 4, 6\}$. Across all three architectures, NM and TD remain stable as the number of targets increases, while BD exhibits a slight decline from $K = 2$ to $K = 4$ and then stabilizes. BD remains above 40 dB for all evaluated target numbers, demonstrating that XComBo maintains both attack effectiveness and stealthiness as the number of backdoor targets increases.

4.3 Generalization Ability

Generalization across Datasets. XComBo optimizes sub-triggers using only CIFAR-100 as the public dataset D_{pub}, while the victim models are evaluated on all four datasets. Across all datasets, XComBo remains within 0.31 dB of the benign model on average under benign transmission and achieves BD PSNR above 40.8 dB. These results indicate that the sample-agnostic sub-triggers capture a transferable cross-domain activation mechanism rather than dataset-specific features.

Generalization across Surrogate and Victim Architectures. To evaluate architecture transfer, we optimize the sub-triggers on one SemCom architecture and use them to poison a victim model with another. Tab. 2 reports PSNR on GTSRB for all nine surrogate–victim architecture pairs. BD PSNR remains above 41.6 dB across all pairs, while NM and TD remain close to the corresponding benign transmission quality. These results demonstrate that the jointly optimized sub-triggers can transfer across heterogeneous surrogate and victim architectures requiring no prior knowledge of the victim architecture during sub-trigger optimization.

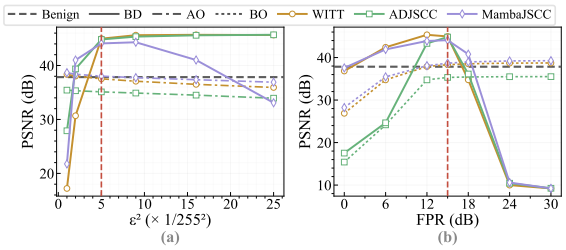


Figure 4: Robustness to sub-trigger energy mismatch. Models are backdoored at $\epsilon^2 = 5/255^2$ and FPR= 15 dB (red lines), evaluated on GTSRB under varying energy constraints for (a) AO: $\boldsymbol{\phi}_p^k$ and (b) BO: $\boldsymbol{\phi}_s^k$ configurations.

Generalization on Sub-triggers Energy Mismatch. Fig. 4 evaluates the effect of mismatched sub-trigger energy between poisoning and inference without retraining. For ϵ^2, BD exceeds 39 dB across all architectures from $5/255^2$ to $16/255^2$, remains effective down to $2/255^2$ on ADJSCC and MambaJSCC, and up to $25/255^2$ on ADJSCC and WITT. For FPR, BD PSNR remains above 36 dB across all architectures within the range of 12–15 dB, extending down to 6 dB on WITT and MambaJSCC, up to 18 dB on ADJSCC and MambaJSCC. AO and BO remain close to NM, indicating limited unintended activation under energy mismatch. At extreme mismatches, triggers may

Table 5: Mitigation results averaged across baselines in PSNR. Values after ± are the standard deviation.

Configuration	NM		BD	
Method	Baseline	XComBo	Baseline	XComBo
Februus	35.4±1.9	**37.0**±1.5	26.8±3.8	**43.2**±1.5
RNP	34.4±1.7	**35.4**±1.1	25.2±3.3	**36.9**±3.6
ShapPrune	29.9±1.8	**30.7**±2.0	21.5±4.3	**28.1**±7.9
ABS	29.9±2.2	**30.7**±2.0	21.6±4.1	**31.9**±6.7

Table 6: Detection results in AUC. Values after ± are the standard deviation across architectures and datasets. – indicates inapplicable scenario.

Method	N-to-One	HCB	CT-BA	XComBo
STRIP	0.75±0.15	0.54±0.15	0.50±0.01	**0.50**±0.01
Spectral	0.99±0.01	0.77±0.35	0.67±0.23	**0.62**±0.23
SCAn	0.53±0.04	0.69±0.19	0.55±0.08	**0.50**±0.00
SCALE-UP	1.00±0.00	1.00±0.00	0.88±0.09	**0.80**±0.06
Band Energy	–	–	–	**0.56**±0.02

become too weak to activate or too strong to preserve benign semantics. Overall, XComBo tolerates a broad range of inference-time energy variations, which is important for practical deployment.

4.4 Ablation Study

What Does Joint Sub-Trigger Optimization Contribute? Tab. 4 compares three sub-trigger optimization strategies. With random initialization, BD PSNR reaches only 37.75–41.96 dB across architectures. Decoupled optimization improves BD by 1.06–5.12dB, whereas joint optimization yields a further 0.57–2.34 dB gain, reaching 43.07–45.36 dB. NM and TD remain stable across strategies, indicating that optimization of sub-triggers primarily benefits matched-pair activation. The additional improvement from joint optimization confirms that jointly coordinating the source-and-wireless-domain sub-triggers is essential for reliable cross-domain activation.

Can Fixed Sub-Triggers Replace Learnable Ones? Tab. 3 examines whether conventional fixed triggers can replace XComBo's learnable sub-triggers. Replacing both sub-triggers reduces BD PSNR to 19.96–30.14 dB and yields an NM–TD discrepancy of up to 6.98 dB. Replacing only the pattern sub-triggers reduces BD PSNR to 23.95–29.03 dB, while fixing only the signal sub-triggers reduces BD PSNR to 23.70 dB on WITT. In contrast, XComBo maintains BD PSNR above 44 dB with closely aligned NM and TD. These results confirm that fixed triggers cannot reliably establish cross-domain composite activation mechanism, requiring learnable sub-triggers in both domains.

Discussions on Robustness to Imperfect Estimated CSI, and experiments on Poison Rate, CBR and Visualizations. Experiments and discussions are provided in the Appendix D and Appendix E.

4.5 Effectiveness Against Defense Methods

Backdoor Mitigation. We evaluate XComBo against four mitigation methods: input purification (Februus Doan et al. (2020)), fine-tuning with benign data (RNP Li et al. (2023)), pruning (ShapPrune Guan et al. (2022)), and trigger synthesis (ABS Liu et al. (2019)). As shown in Tab. 5, XComBo achieves BD PSNR above 28 dB under all four defenses. For ShapPrune and ABS, reducing BD performance is accompanied by substantial degradation in NM utility. Overall, mitigation methods either incur a utility cost or fail to effectively suppress XComBo.

Backdoor Detection. We evaluate five detectors across the source and wireless domains, including input-level (STRIP Gao et al. (2019)), feature-level (Spectral Signatures Tran et al. (2018) (denoted as 'Spectral'), SCAn Tang et al. (2021)), perturbation-based (SCALE-UP Guo et al. (2023)), and spectrum-based (Band Energy) detection. For the four detectors with baseline comparisons, XComBo achieves the lowest AUC. Under the wireless-domain Band Energy detector, XComBo obtains an AUC of 0.56, close to random guessing. These results indicate that XComBo remains difficult to identify through inspections of either attack surface.

5 Conclusion

We propose XComBo, the first cross-domain backdoor attack for SemCom. The sub-triggers are coordinated across the source and wireless domains through a composite mechanism. Joint optimization and energy constraints enable reliable activation and benign transmission under incomplete trigger conditions without prior knowledge of the victim model. XComBo further supports multiple malicious targets. Experiments demonstrate strong attack effectiveness and generalization, while the backdoor remains effective under mitigation and difficult to detect at both domains.

REFERENCES

M. Barni, K. Kallas, and B. Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, 2019. doi: 10.1109/ICIP.2019.8802997.

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. doi: 10.1109/TCCN.2019.2919300.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning, 2017. URL https://arxiv.org/abs/1712.05526.

Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*, pp. 897–912, 2020.

Ecenaz Erdemir, Tze-Yang Tung, Pier Luigi Dragotti, and Deniz Gündüz. Generative joint source-channel coding for semantic image transmission. *IEEE J.Sel. A. Commun.*, 41(8):2645–2657, August 2023. ISSN 0733-8716. doi: 10.1109/JSAC.2023.3288243. URL https://doi.org/10.1109/JSAC.2023.3288243.

Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20876–20885, June 2022.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pp. 113–125, 2019.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. Few-shot backdoor defense using shapley estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13358–13367, 2022.

Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023.

Van-Tam Hoang, Van-Linh Nguyen, Rong-Guey Chang, Po-Ching Lin, Ren-Hung Hwang, and Trung Q. Duong. Adversarial attacks against shared knowledge interpretation in semantic communications. *IEEE Transactions on Cognitive Communications and Networking*, 11(2):1024–1040, 2025. doi: 10.1109/TCCN.2025.3528891.

Linshan Hou, Zhongyun Hua, Yuhong Li, Yifeng Zheng, and Leo Yu Zhang. M-to-n backdoor paradigm: A multi-trigger and multi-target attack to deep learning models. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11299–11312, 2024. doi: 10.1109/TCSVT.2024.3417410.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, pp. 19837–19854. PMLR, 2023.

Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, and Yu-Gang Jiang. Multi-trigger backdoor attacks: More triggers, more threats, 2025. URL https://arxiv.org/abs/2401.15295.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16463–16472, October 2021.

Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pp. 113–131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370899. doi: 10.1145/3372297.3423362. URL https://doi.org/10.1145/3372297.3423362.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 1265–1282, 2019.

Hua Ma, Shang Wang, Yansong Gao, Zhi Zhang, Huming Qiu, Minhui Xue, Alsharif Abuadbba, Anmin Fu, Surya Nepal, and Derek Abbott. Watch out! simple horizontal class backdoor can trivially evade defense. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 4465–4479, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706363. doi: 10.1145/3658644.3670361. URL https://doi.org/10.1145/3658644.3670361.

Anh Nguyen and Anh Tran. Wanet – imperceptible warping-based backdoor attack, 2021. URL https://arxiv.org/abs/2102.10369.

Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3454–3464. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/234e691320c0ad5b45ee3c96d0d7b8f8-Paper.pdf.

Yalin E. Sagduyu, Tugba Erpek, Sennur Ulukus, and Aylin Yener. Vulnerabilities of deep learning-driven semantic communications to backdoor (trojan) attacks. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2023. doi: 10.1109/CISS56502.2023.10089692.

Yalin E Sagduyu, Tugba Erpek, Aylin Yener, and Sennur Ulukus. Wireless backdoor attack and defense for semantic communications over multiple access channel. *arXiv preprint arXiv:2606.30595*, 2026.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1541–1558, 2021.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Jialin Wan, Jinglong Shen, Nan Cheng, Zhisheng Yin, Yiliang Liu, Wenchao Xu, and Xuemin Shen. A channel-triggered backdoor attack on wireless semantic image reconstruction. *IEEE Transactions on Mobile Computing*, pp. 1–16, 2026. doi: 10.1109/TMC.2026.3676425.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zihan Wang, Guansong Pang, Wenjun Miao, Jin Zheng, and Xiao Bai. Mtattack: Multi-target backdoor attacks against large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10440–10448, 2026.

Haotian Wu, Yulin Shao, Chenghong Bian, Krystian Mikolajczyk, and Deniz Gündüz. Deep joint source-channel coding for adaptive image transmission over mimo channels. *IEEE Transactions on Wireless Communications*, 23(10):15002–15017, 2024a.

Tong Wu, Zhiyong Chen, Meixia Tao, Xiaodong Xu, Wenjun Zhang, and Ping Zhang. Mambajscc: Deep joint source-channel coding with visual state space model. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, pp. 1677–1682. IEEE, 2024b.

Jialong Xu, Bo Ai, Wei Chen, Ang Yang, Peng Sun, and Miguel Rodrigues. Wireless image transmission using deep source channel coding with attention modules. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2315–2328, 2022. doi: 10.1109/TCSVT.2021.3082521.

Xiaodong Xu, Yue Chen, Bizhu Wang, Zhiqiang Bian, Shujun Han, Chen Dong, Chen Sun, Wenqi Zhang, Lexi Xu, and Ping Zhang. Csba: Covert semantic backdoor attack against intelligent connected vehicles. *IEEE Transactions on Vehicular Technology*, 73(11):17923–17928, 2024. doi: 10.1109/TVT.2024.3427713.

Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1562–1578, 2022. doi: 10.1109/TDSC.2020.3028448.

Ke Yang, Sixian Wang, Jincheng Dai, Kailin Tan, Kai Niu, and Ping Zhang. Witt: A wireless image transmission transformer for semantic communications. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Xiao Yang, Yuni Lai, Gaolei Li, Jun Wu, Kai Zhou, Jianhua Li, and Mingzhe Chen. Toward polymorphic backdoor against semantic communication via intensity-based poisoning. *IEEE Transactions on Information Forensics and Security*, 2026.

Selim F. Yilmaz, Xueyan Niu, Bo Bai, Wei Han, Lei Deng, and Deniz Gündüz. High perceptual quality wireless image delivery with denoising diffusion models. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–5, 2024. doi: 10.1109/INFOCOMWKSHPS61880.2024.10620904.

Yuan Zhou, Rose Qingyang Hu, and Yi Qian. Stealthy backdoor attacks on semantic symbols in semantic communications. In *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, pp. 4975–4981, 2024. doi: 10.1109/GLOBECOM52923.2024.10901411.

A Semantic Communication Coding Method

Semantic communication can be modeled as a mapping function f_θ: $\mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the source and task-output domains, respectively. A source sample $\mathbf{d}in\mathcal{X}$ is processed by the semantic encoder $S_\alpha(\cdot)$ and the channel encoder $C_\beta(\cdot)$, yielding the transmitted semantic feature vector:

$$\mathbf{x} = C_\beta(S_\alpha(\mathbf{d})), \tag{15}$$

where $\mathbf{x} \triangleq \{x_i \in \mathbb{C} | i = 1, \cdots, L\}$ denotes the semantic feature symbols.

The received symbol corresponding to the transmitted symbol x_i is given by

$$r_i = hx_i + n_{ch,i}, \tag{16}$$

where $h \in \mathbb{C}$ denotes the complex channel gain and $n^{ch,i} \sim \mathcal{CN}(0, \sigma_{ch}^2)$ denotes the additive white Gaussian noise. With the receiver-side channel estimate $\bar{h}$, the received symbols are processed via channel equalization:

$$y_i \triangleq \bar{h}^+ r_i \approx x_i + \bar{h}^+ n_{ch_i}, \tag{17}$$

where

$$\bar{h}^+ \triangleq \left(\bar{h}^H \bar{h}\right)^{-1} \bar{h}^H \tag{18}$$

denotes the pseudo-inverse of the estimated channel gain, and $[\cdot]^H$ denotes the Hermitian transpose.

The equalized received symbols are collected as $\mathbf{y} \triangleq \{y_i \in \mathbb{C} | i = 1, \cdots, L\}$, which form the input to the channel decoder C_γ^{-1} and semantic decoder S_ψ^{-1}:

$$\hat{\mathbf{d}} = S_\psi^{-1}(C_\gamma^{-1}(\mathbf{y})), \tag{19}$$

where $\hat{\mathbf{d}}$ is the task output. In the main text, the encoder and decoder are denoted as $E_{\theta_e} \triangleq C_\beta \circ S_\alpha$ and $D_{\theta_d} \triangleq S_\psi^{-1} \circ C_\gamma^{-1}$.

B Derivation of the Wireless Signal Injection

We derive the wireless-domain signal injection based on the semantic communication model defined in Appendix A.

We consider signal injection during the data transmission. The adversary utilizes the optimized signal sub-trigger $\boldsymbol{\phi}_s$ to form the transmitted malicious signal $\mathbf{s}$ with $\mathbf{s} \triangleq \{s_i \in \mathbb{C} | i = 1, \cdots, L\}$. The legitimate and adversarial signals are superposed, yielding the attacked received symbols

$$\tilde{r_i} = hx_i + h_a s_i + n_{ch,i}, \tag{20}$$

where $h_a \in \mathbb{C}$ denotes the adversary-to-receiver channel gain. The attacked signal after zero-forcing equalization is

$$\tilde{y_i} = \bar{h}^+(hx_i + h_a s_i + n_{ch,i}). \tag{21}$$

Therefore, from Eq. 17,

$$\tilde{y_i} = y_i + \bar{h}^+ h_a s_i. \tag{22}$$

The adversary transmits

$$s_i = \lambda \frac{\hat{h}}{\hat{h_a}} \boldsymbol{\phi}_s(i), \tag{23}$$

where $\hat{h}$ and $\hat{h_a}$ denote the adversary's estimates of the benign transmitter-to-receiver channel and adversarial-to-receiver channel, respectively, and λ controls the effective perturbation strength. Substituting Eq. 23 into Eq. 22 gives

$$\tilde{y_i} = y_i + \lambda \bar{h}^+ h_a \frac{\hat{h}}{\hat{h_a}} \boldsymbol{\phi_s}(i). \tag{24}$$

For the scalar zero-forcing equalizer used in the system, $\bar{h}^+ = \frac{1}{\bar{h}}$, so that

$$\tilde{y_i} = y_i + \lambda \frac{\hat{h}}{\bar{h}} \frac{h_a}{\hat{h_a}} \boldsymbol{\phi_s}(i). \tag{25}$$

When the channel estimates are sufficiently accurate, $\hat{h} \approx \bar{h}$, $\hat{h_a} \approx h_a$, which yields

$$\tilde{\mathbf{y}} \approx \mathbf{y} + \lambda \boldsymbol{\phi}_s. \tag{26}$$

Thus, the effective wireless-domain injection is

$$T_s(\mathbf{y}; \boldsymbol{\phi}_s) \triangleq \tilde{\mathbf{y}} \approx \mathbf{y} + \lambda \boldsymbol{\phi}_s \tag{27}$$

Eq. 27 demonstrates that the signal sub-trigger is injected through the wireless channel, and after receiver-side equalization, appears as an effective perturbation at the decoder input.

C Details of Experimental Setup and Implementation

C.1 Experimental settings

All experiments are conducted on a machine equipped with an Intel(®) Core™ i9-14900K CPU, 64 GB RAM, 4 TB ROM, and an NVIDIA® GeForce RTX™ 4080 GPU. The models are implemented in PyTorch 2.4 with CUDA 12.4 and Python 3.9 on Ubuntu-22.04. The wireless channel is modeled as Rician fading with additive white Gaussian noise. The key hyperparameters are summarized in Tab. 7.

C.2 Implementation of attacks

Tiny ImageNet Preprocessing. For Tiny ImageNet, images are resized to 32×32 and pre-loaded from disk. The training set contains all 200 classes with 250 images per class.

Adaptation of Baselines. We adapt three baseline methods to image reconstruction and the composite setting: N-to-One Xue et al. (2022), HCB Ma et al. (2024) and CT-BA Wan et al. (2026). N-to-One uses white-square patches at different positions as triggers, HCB uses a white-square patch and benign image features as triggers, while CT-BA uses channel statistical features as triggers. N-to-One and HCB are adapted from classification to image reconstruction. For HCB, the weather effects are viewed as benign features, rather than sub-triggers added to the source image. For CT-BA, the 1×1 white square is added as a source-domain sub-trigger to form a cross-domain composite trigger pair compatible with our evaluation. All baselines are extended to $K = 3$ targets, with per-target white-square positions following the trigger design of Xue et al. (2022). For fair comparison, all methods use the same poisoning rate, target number, training epochs, learning rate, optimizer, and target-wise poisoning allocation, while their trigger constructions are adapted only as needed to preserve the original attack principles. Each method's trigger setup is summarized in Tab. 8.

Implementation of Joint Sub-trigger Optimization. In our implementation, the joint sub-trigger optimization is performed on a surrogate model f_η pre-trained on CIFAR-100. The process alternates between two stages every $F = 25$ epochs. In Stage 1, only the sub-triggers $\{(\boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k)\}_{k=1}^K$ are updated for 5 epochs, with η frozen. The sub-triggers are jointly optimized by

$$\mathcal{L}_{\text{joint}} = \frac{1}{K} \sum_{k=1}^{K} \left\| f_\eta(\mathbf{d}; \boldsymbol{\phi}_p^k, \boldsymbol{\phi}_s^k) - \mathbf{t}_k \right\|_2^2. \tag{28}$$

as specified in Section 3.3. In Stage 2, η is updated on CIFAR-100 under all sub-trigger configurations, with sub-triggers frozen, to emulate the trigger-response behavior. This alternating procedure coordinates the activation signatures across the two domains. This implementation is practical under the adversary's capability, as the adversary has white-box access to f_η and the public dataset D_{pub}.

Model Poisoning. During model poisoning, the sub-triggers optimized on f_η are frozen and directly deployed to poison the victim model f_θ without sub-trigger retraining. The poisoning rate $\gamma = 0.03$ remains consistent across all datasets and architectures. The frequency band is fixed to $[\omega_{\min}, \omega_{\max}]$ with $\omega_{\min} = 1/L$ and $\omega_{\max} = 1/2$, and the number of sinusoids is $R = 100$.

C.3 Implementation of Ablation Study

Ablation Study on Joint Sub-Trigger Optimization. In the ablation, sub-triggers are randomly initialized and deployed directly for model poisoning against f_θ without any surrogate-model optimization on f_η. The pattern sub-triggers $\boldsymbol{\phi}_p^k$ are drawn from $\mathcal{N}(0, 0.1)$ and projected onto $\|\boldsymbol{\phi}_p^k\|_2 \leq \epsilon$

Table 7: Attack configuration. Unless otherwise specified, all experiments in the main text adopt these settings.

Parameter	Value
K	3
γ	0.03
D_{pub}	CIFAR-100
CBR	0.5
ϵ^2	$5/255^2$
R	100
FPR	15 dB
Channel type	Rician fading
Channel SNR (training)	8–18 dB
Channel SNR (evaluation)	18 dB

Table 8: Trigger setup of baselines. N-to-One and HCB are source-domain-only: their BO is a second source-domain sub-trigger. CT-BA's per-target SNR has ± 0.5 dB tolerance.

Method	AO	BO	Per-Target Configuration
N-to-One	1×1 white square	1×1 white square	top-left square + bottom-left square top-center square + bottom-center square top-right square + bottom-right square
HCB	1×1 white square	weather effect	top-left square + rain top-right square + snow bottom-left square + light
CT-BA	1×1 white square	channel SNR	top-left square + 5 dB top-right square + 10 dB bottom-left square + 15 dB

via ℓ_2 normalization at each forward pass. The signal sub-trigger amplitudes a_r^k are drawn from $\mathcal{N}(0, 0.1)$ and phases φ_r^k from $\mathcal{U}(0, 2\pi)$. All other configurations including R, ϵ^2, FPR, poisoning rate, and training hyperparameters remain identical to XComBo.

Ablation Study on Learnable Sub-Triggers. Three variants replace either or both sub-triggers with fixed alternatives, corresponding to 'All', 'Signals', and 'Patterns' in Tab. 3. When replaced, the pattern sub-trigger $\boldsymbol{\phi}_p^k$ is set to a 1×1 white pixel at a target-specific position (top-left / top-right / bottom-left). The signal sub-trigger $\boldsymbol{\phi}_s^k$ is set to a fixed square wave with 50% duty cycle centered at the midpoint of each target's frequency band. 'All' replaces both sub-triggers, neither is learnable. 'Signals' replaces only $\boldsymbol{\phi}_s^k$ with the fixed square wave while retaining $\boldsymbol{\phi}_p^k$ as XComBo. 'Patterns' replaces only $\boldsymbol{\phi}_p^k$ with the fixed white pixels while retaining $\boldsymbol{\phi}_s^k$ as XComBo. All other configurations remain identical to XComBo.

C.4 Implementation of Defenses

We evaluate XComBo against four mitigation methods and five detection methods. Their implementation details are provided as follows:

Februus. Februus Doan et al. (2020) is an input purification defense. A saliency map is computed as the gradient of the reconstruction loss with respect to the input image. The region of size 16×16 with the highest saliency is identified and neutralized by a 7×7 Gaussian blur kernel, after which the model is fine-tuned for 10 epochs.

RNP. RNP Li et al. (2023) is a fine-tuning-based defense with two phases. In the unlearning phase, the model is fine-tuned for 10 epochs with randomly shuffled images as reconstruction targets. In the recovery phase, the model is fine-tuned for 10 epochs with self-reconstruction to restore benign performance.

ShapPrune. ShapPrune Guan et al. (2022) is a pruning-based defense. Shapley values are approximated by measuring each channel's marginal contribution to clean reconstruction PSNR in the

Table 9: Attack effectiveness under imperfect CSI, measured in PSNR (dB), averaged over the three targets at SNR=18 dB.

Testing ξ	**0**			**0.05**			**0.1**			**0.2**			**0.3**		
Training ξ	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**
0	38.83	37.90	45.27	38.91	37.72	44.97	38.91	37.72	44.56	38.91	37.70	40.09	38.91	37.65	29.41
0.05	38.77	37.78	43.47	38.82	37.60	43.23	38.82	37.60	42.98	38.82	37.58	39.55	38.82	37.52	29.37
0.1	38.75	37.76	44.19	38.83	37.63	43.88	38.83	37.62	43.46	38.83	37.60	38.99	38.83	37.56	28.81
0.2	38.81	37.79	44.02	38.88	37.62	43.87	38.88	37.62	43.83	38.88	37.60	43.50	38.88	37.53	39.21

Table 10: Attack effectiveness under imperfect CSI, measured in MS-SSIM, averaged over the three targets at SNR=18 dB.

Testing ξ	**0**			**0.05**			**0.1**			**0.2**			**0.3**		
Training ξ	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**	NM	TD	**BD**
0	.9953	.9939	.9991	.9955	.9937	.9989	.9955	.9937	.9988	.9955	.9937	.9966	.9955	.9936	.9779
0.05	.9952	.9934	.9991	.9953	.9933	.9990	.9953	.9933	.9989	.9953	.9932	.9968	.9953	.9931	.9778
0.1	.9952	.9938	.9989	.9954	.9937	.9988	.9954	.9937	.9987	.9954	.9937	.9964	.9954	.9936	.9771
0.2	.9953	.9936	.9992	.9955	.9935	.9991	.9955	.9935	.9991	.9955	.9935	.9990	.9955	.9934	.9963

encoder's final layer using 200 samples and 10 permutations. The 50% of channels with the lowest contribution are pruned, followed by 10 epochs of fine-tuning.

ABS. ABS Liu et al. (2019) is a trigger synthesis defense. For each channel in the encoder output, stimulation values are injected using 200 samples, and the PSNR shift is measured. The top-20 most sensitive channels are suppressed to zero, followed by 10 epochs of fine-tuning.

STRIP. STRIP Gao et al. (2019) is an input-level detection method. Each of 500 test inputs is blended with 100 random clean images, and the pixel-level variance of the reconstruction outputs is used as the anomaly score for AUC computation.

Spectral Signatures. Spectral Signatures Tran et al. (2018) is a feature-level detection method. Features are extracted from the encoder output for 1500 samples. Singular Value Decomposition (SVD) is applied to the centered feature matrix, and the correlation of each sample's feature with the top right singular vector is used as the anomaly score.

SCAn. SCAn Tang et al. (2021) is a feature-level detection method. Features are extracted from the encoder output and reduced to 100 dimensions via PCA. A Gaussian kernel affinity matrix is constructed, and spectral clustering with $K = 2$ separates the inputs into clean and poisoned clusters using 800 samples.

SCALE-UP. SCALE-UP Guo et al. (2023) is a perturbation-based detection method. For each of 800 inputs, Gaussian noise is applied at 10 scales ($\sigma \in \{0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.7, 1.0\}$), and the reconstruction PSNR degradation curve is recorded. A logistic regression classifier is trained on these degradation curves from a small set of labeled data and used to classify the remaining inputs.

Band Energy. Band Energy is a spectrum-based detection method operating at the receiver side. The received feature vector is transformed to the frequency domain via Fast Fourier Transform (FFT). The energy ratio between the trigger frequency band and the full spectrum is computed using 800 samples, and AUC is obtained by varying the detection threshold. This detector applies only to XComBo: N-to-One and HCB inject no wireless-domain signals, while CT-BA relies on channel conditions rather than perturbations. This evaluation is carried out to demonstrate XComBo's robustness against frequency-domain inspection at the receiver side.

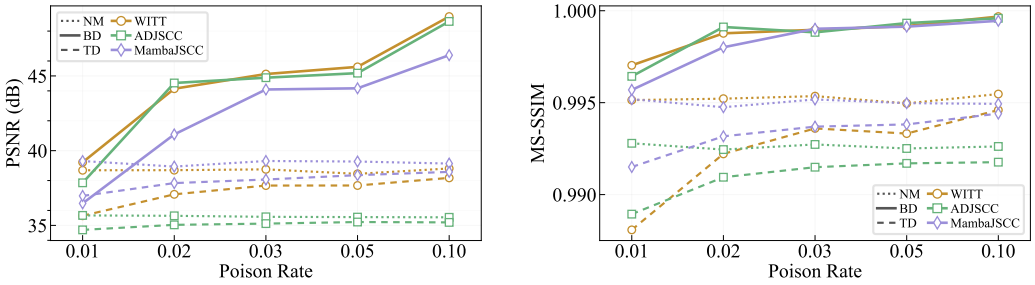


Figure 5: PSNR and MS-SSIM under varying poisoning rates on XComBo using GTSRB at SNR= 18 dB across three architectures.

D More Experimental Results and Discussions

D.1 Robustness to Imperfect CSI

The wireless-domain sub-triggers rely on channel-side information for signal pre-compensation. The attacker may have only imperfect estimates of the legitimate and attacker-to-receiver channels. We therefore evaluate whether the backdoor remains effective under such CSI uncertainty.

Let h and h_a denote the legitimate and attacker-to-receiver channels, respectively. The attacker operates with

$$\hat{h} = h(1+\varepsilon_h), \qquad \hat{h}_a = h_a(1+\varepsilon_a), \tag{29}$$

where

$$\varepsilon_h, \varepsilon_a \sim \mathcal{CN}(0, \xi^2) \tag{30}$$

are independently sampled for each image. Here, ξ controls the relative CSI error, with $\xi = 0$ corresponding to exact CSI.

To isolate adversary-side CSI error, we assume that the receiver-side CSI is perfect during data transmission. Under the zero-forcing receiver model, the attacker transmits

$$s = \lambda \frac{\hat{h}}{\hat{h}_a} \boldsymbol{\phi}_s. \tag{31}$$

The signal sub-trigger received by the decoder is

$$I = \bar{h}^{+} h_a s = \lambda \frac{\hat{h}}{\bar{h}} \frac{h_a}{\hat{h}_a} \boldsymbol{\phi}_s = \lambda \frac{1+\varepsilon_h}{1+\varepsilon_a} \boldsymbol{\phi}_s. \tag{32}$$

Thus, imperfect CSI introduces a multiplicative complex gain on the trigger, changing its amplitude and phase.

We evaluate on WITT using GTSRB at a channel SNR of 18 dB on K=3 targets. We cross the CSI error used during training, $\xi_{\text{train}} \in \{0, 0.05, 0.1, 0.2\}$, with the error at test time, $\xi_{\text{test}} \in \{0, 0.05, 0.1, 0.2, 0.3\}$. The resulting grid distinguishes matched and mismatched CSI conditions, while all other training and evaluation settings remain unchanged.

The results in 9 and 10 show that XComBo remains effective under moderate CSI uncertainty. Under matched conditions, increasing the CSI error from $\xi = 0$ to 0.2 reduces BD PSNR only from 45.27 to 43.50 dB, while MS-SSIM remains as high as 0.9990. The attack also remains effective under train–test mismatch. For $\xi_{\text{test}} = 0.2$, BD PSNR stays above 38.99 dB for all evaluated training error levels, whereas NM and TD remain stable around 38.8 dB and 37.6 dB, respectively. Under the more severe test-time error $\xi_{\text{test}} = 0.3$, training with imperfect CSI becomes particularly beneficial: the BD PSNR increases from 29.41 dB with $\xi_{\text{train}} = 0$ to 39.21 dB with $\xi_{\text{train}} = 0.2$, while the corresponding MS-SSIM improves from 0.9779 to 0.9963. These results indicate that XComBo tolerates moderate CSI estimation errors and that exposing the backdoored model to imperfect CSI during training substantially improves its robustness to more severe inference-time mismatch.

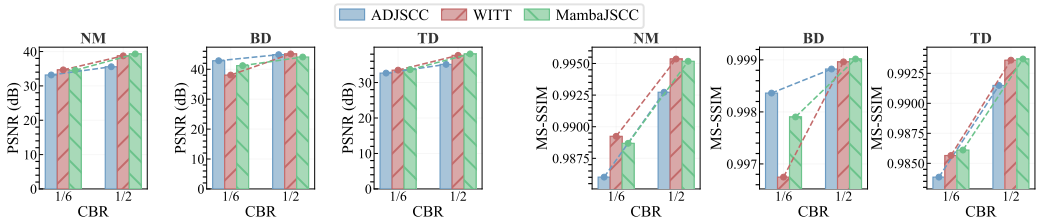


Figure 6: PSNR and MS-SSIM across CBR$=1/2$ and CBR$=1/6$ on XComBo using GTSRB at SNR$=18$ dB across three architectures.

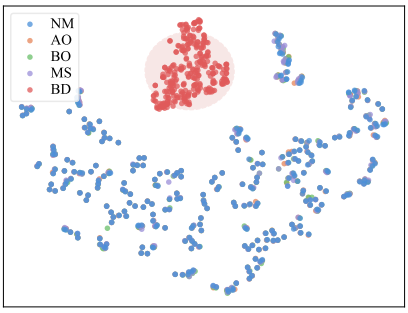


Figure 7: t-SNE visualization of decoder intermediate features under cross-domain sub-trigger configurations at SNR$=18$ dB on WITT and GTSRB.

D.2 Experiments on Poison Rate

Fig. 5 reports PSNR and MS-SSIM under poisoning rates $\gamma \in \{0.01, 0.02, 0.03, 0.05, 0.10\}$ on GTSRB at SNR$=18$ dB. NM remains stable across all three architectures, varying by at most 0.37 dB and 0.0005 in MS-SSIM, confirming that benign transmission is unaffected by the variation in γ. TD rises gradually with γ and approaches NM, narrowing the NM–TD gap from 0.97–3.05 dB and 0.0037–0.0071 in MS-SSIM at $\gamma = 0.01$ to 0.34–0.63 dB and 0.0005–0.0009 at $\gamma = 0.10$. BD increases with γ in both metrics, from 36.47–39.22 dB and 0.9957–0.9970 at $\gamma = 0.01$ to 46.39–48.97 dB and 0.9995–0.9997 at $\gamma = 0.10$. The gain is most pronounced between $\gamma = 0.01$ and $\gamma = 0.02$, where PSNR rises by 4.61–6.68 dB and MS-SSIM by 0.0017–0.0027, while subsequent increments stay at most 0.0010 in MS-SSIM and diminish further as γ grows. Even at $\gamma = 0.01$, BD stays above 36 dB and 0.9957 in MS-SSIM, indicating that the cross-domain backdoor can be established with a small fraction of poisoned training data. Both metrics exhibit the same trend, where higher poison rate facilitates the learning of cross-domain composite activation and enhances both attack effectiveness and stealthiness.

D.3 Experiments on CBR

Fig. 6 reports PSNR and MS-SSIM under CBR$=1/2$ and CBR$=1/6$ on GTSRB at SNR$=18$ dB. Reducing CBR from $1/2$ to $1/6$ degrades NM by 2.39–4.74 dB and 0.0061–0.0067 in MS-SSIM, and TD by 2.46–4.43 dB and 0.0076–0.0079, while BD degrades by 2.03–7.07 dB and only 0.0005–0.0022 in MS-SSIM. Despite this degradation, BD remains 3.39–9.67 dB and 0.0036–0.0123 above NM, achieving PSNR above 38 dB and MS-SSIM at least 0.9967. The cross-domain composite activation remains effective under low CBR, with BD clearly distinguishable from both NM and TD. Notably, TD degrades slightly more than NM in MS-SSIM, yet the absolute NM–TD gap remains within 1.24 dB and 0.0036 at both CBR levels, confirming that stealthiness under incomplete trigger conditions is not compromised by reduced bandwidth. These results show that XComBo operates reliably under practical bandwidth constraints.

Table 11: Attack effectiveness under all five possible sub-trigger configurations, measured in MS-SSIM, averaged across multiple targets at SNR=18 dB.

Victim	**ADJSCC**																			
Dataset	**CIFAR-10**			Benign	.9922	**GTSRB**			Benign	.9921	**Tiny ImageNet**			Benign	.9906	**CIFAR-100**			Benign	.9910
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	**.9924**	**.9923**	.9921	.9663	**.9921**	.9879	.9874	.9867	.9888	.9868	.9899	.9895	.9893	.9784	.9890	.9899	.9895	.9896	.9511	.9894
HCB	.9864	.8482	.9861	.9680	.9141	.9860	.9716	.9843	.9693	.9727	.9822	.8816	.9830	.8162	.9462	.9834	.8532	.9832	.9114	.9123
CT-BA	.9901	.9779	.9660	.9638	.9647	.9857	.9420	.9679	.9417	.9651	.9882	.9738	.9614	.9565	.9613	.9859	.9842	.9580	.9797	.9573
XComBo	.9917	.9920	**.9924**	**.9990**	.9919	**.9927**	**.9914**	**.9921**	**.9991**	**.9910**	**.9919**	**.9914**	**.9920**	**.9984**	**.9916**	**.9914**	**.9907**	**.9912**	**.9992**	**.9907**
Victim	**WITT**																			
Dataset	**CIFAR-10**			Benign	.9947	**GTSRB**			Benign	.9951	**Tiny ImageNet**			Benign	.9944	**CIFAR-100**			Benign	.9943
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	.9939	.9933	.9930	.9707	.9924	.9942	.9926	.9925	.9888	.9905	.9930	.9927	.9920	.9811	.9916	.9925	.9919	.9916	.9563	.9909
HCB	.9896	.9189	.9895	.9005	.9144	.9874	.9625	.9872	.9388	.9745	.9883	.9460	.9889	.7972	.9585	.9883	.9255	.9873	.9297	.9543
CT-BA	.9934	.9383	.9697	.9189	.9640	.9918	.7619	.9705	.8869	.9677	.9927	.9683	.9655	.9257	.9583	.9926	.9827	.9650	.9081	.9627
XComBo	**.9948**	**.9939**	**.9943**	**.9991**	**.9938**	**.9952**	**.9932**	**.9944**	**.9990**	**.9925**	**.9944**	**.9934**	**.9940**	**.9986**	**.9933**	**.9943**	**.9936**	**.9937**	**.9985**	**.9935**
Victim	**MambaJSCC**																			
Dataset	**CIFAR-10**			Benign	.9949	**GTSRB**			Benign	.9946	**Tiny ImageNet**			Benign	.9943	**CIFAR-100**			Benign	.9942
Method	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS	NM	AO	BO	**BD**	MS
N-to-One	.9946	.9939	.9937	.9739	**.9937**	.9942	**.9939**	.9925	.9920	.9920	.9937	.9933	.9931	.9842	**.9932**	.9937	.9929	.9929	.9778	.9922
HCB	.9924	.9727	.9921	.9615	.9897	.9933	.9786	.9929	.9864	.9911	.9909	.9618	.9899	.9325	.9835	.9876	.8860	.9875	.7735	.9327
CT-BA	.9935	.9910	.9675	.9647	.9591	.9936	.9901	.9705	.9620	.9506	.9932	.9914	.9651	.9733	.9634	.9930	.9902	.9655	.9647	.9495
XComBo	**.9953**	**.9942**	**.9942**	**.9988**	.9934	**.9951**	.9936	**.9944**	**.9989**	**.9929**	**.9947**	**.9937**	**.9936**	**.9976**	.9929	**.9947**	**.9937**	**.9937**	**.9974**	**.9928**

D.4 Latent Representations

Fig. 7 presents the t-SNE projection of the decoder intermediate features across all sub-trigger configurations at SNR$=18$ dB on WITT and GTSRB. NM, AO, BO, and MS of the same source image are mapped to tightly overlapping points in the latent space, indicating that all incomplete sub-trigger configurations produce intermediate representations aligned with that of the source image. In contrast, BD occupies an isolated region, suggesting that only the co-occurrence of both sub-triggers of the same target activates a systematic latent-space shift inside the decoder. This separation reflects the joint effect of the cross-domain composite trigger, showing that the two sub-triggers together determine whether the backdoor is activated.

D.5 Results measured in MS-SSIM

Results under Sub-trigger Configurations. Tab. 11 reports MS-SSIM under all five sub-trigger configurations. The results exhibit similar trends observed under Tab. 1 measured in PSNR, indicating consistent attack effectiveness. XComBo achieves at least 0.9974 of MS-SSIM on BD across all architectures and datasets, while NM differs from each incomplete trigger configuration by no more than 0.003. NM remains within 0.003 of the benign model on average, preserving benign communication quality. Baselines exhibit clear weaknesses: N-to-One drops to 0.9511 in the hardest case, HCB falls to 0.7735, and CT-BA shows larger gaps between NM and incomplete sub-trigger configurations on WITT, all consistent with the PSNR results.

Results across SNRs. Fig. 8(a) reports MS-SSIM under varying channel SNR, averaged across four datasets. XComBo ranges from 0.9133–0.9636 at SNR$=0$ dB to 0.9985–0.9991 at SNR$=24$ dB across architectures on BD, surpassing nearly all baselines at nearly every SNR. NM and TD track the benign reconstruction, with NM deviating by at most 0.0005 at SNR$=18$ dB. N-to-One and HCB achieve BD of only 0.9712–0.9820 and 0.8916–0.9162 at SNR$=18$ dB, respectively. CT-BA attains strong BD at some backdoor targets (e.g., 0.9966 for the 5 dB target on ADJSCC), but its multi-target average and the NM–BO gap on WITT compromise stealthiness. These results are consistent with the PSNR results. In the main context, the results are averaged over all four datasets, thus we provide the per-dataset results measured in PSNR and MS-SSIM in Fig. 12–Fig.19

Results across Target Numbers. Fig. 8(b) reports MS-SSIM across $K = 2, 3, 4, 6$ targets on GTSRB at SNR$=18$ dB. As K grows, BD undergoes a mild decline: on MambaJSCC from 0.9991 at $K = 2$ to 0.9966 at $K = 6$, and on WITT from 0.9993 to 0.9976. Across all three architectures, NM and TD vary by no more than 0.003 across all K and remain nearly constant. This demonstrates that XComBo accommodates an increasing number of backdoor targets with negligible degradation to benign reconstruction quality, and achieves superior attack efficiency under multi-target settings.

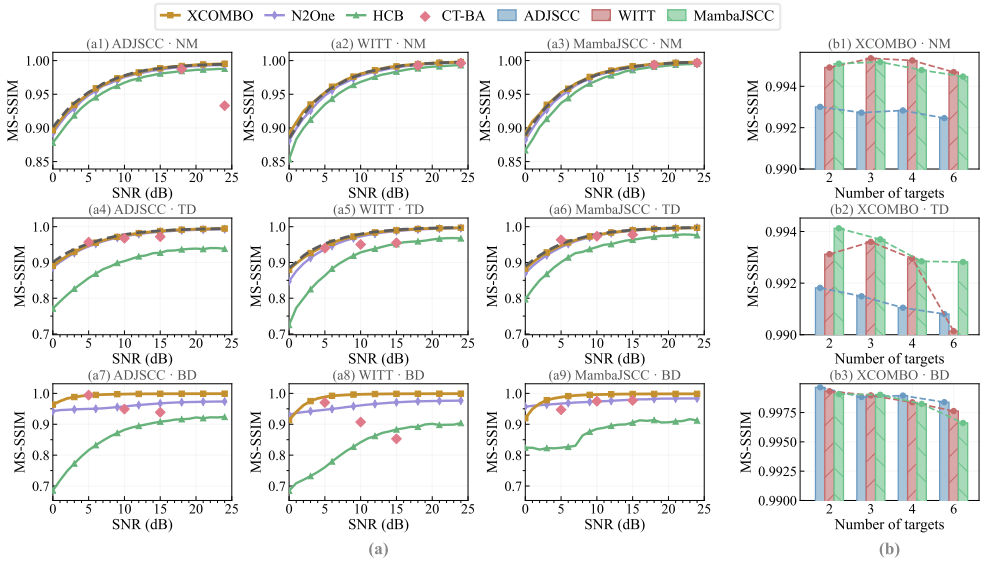


Figure 8: (a) MS-SSIM across channel SNR 0–24 dB on XComBo and baselines, averaged over all four datasets, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$=5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty. (b) MS-SSIM across target numbers $K = 2, 3, 4, 6$ on XComBo using GTSRB.

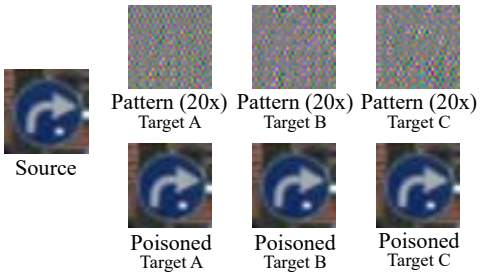


Figure 9: Pattern sub-triggers learned for the $K = 3$ targets. Top: the sub-triggers $\{\phi_p^k\}$ visualized with $20\times$ amplitude amplification. Bottom: the poisoned images obtained by adding the corresponding sub-triggers onto the source image, satisfying the energy constraint $\|\phi_p^k\|_2 \leq \epsilon$ with $\epsilon^2 = 5/255^2$.

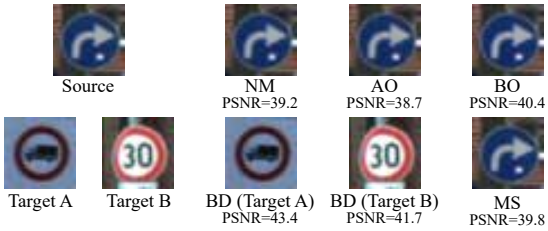


Figure 10: Malicious behavior of the poisoned model under different trigger activation modes at SNR$=18$ dB. PSNRs are computed against the intended reconstruction target.

E VISUALIZATION

This section provides visual illustrations of the learned cross-domain composite triggers and the resulting malicious behavior of the poisoned model.

Pattern Sub-Triggers. Fig. 9 visualizes the pattern sub-triggers $\{\phi_p^k\}$ obtained from the joint sub-trigger optimization for the $K = 3$ targets. The upper row shows each sub-trigger with $20\times$ amplitude amplification, and the bottom row demonstrates the poisoned images $T_p(\mathbf{d}; \phi_p^k)$ obtained by adding the sub-triggers onto the source image. Due to the energy constraint, the pattern sub-triggers preserve the visual stealthiness under incomplete trigger conditions and remain imperceptible when added to source images. The three sub-triggers exhibit distinct spatial structures, confirming that XComBo learns a different pattern for each target.

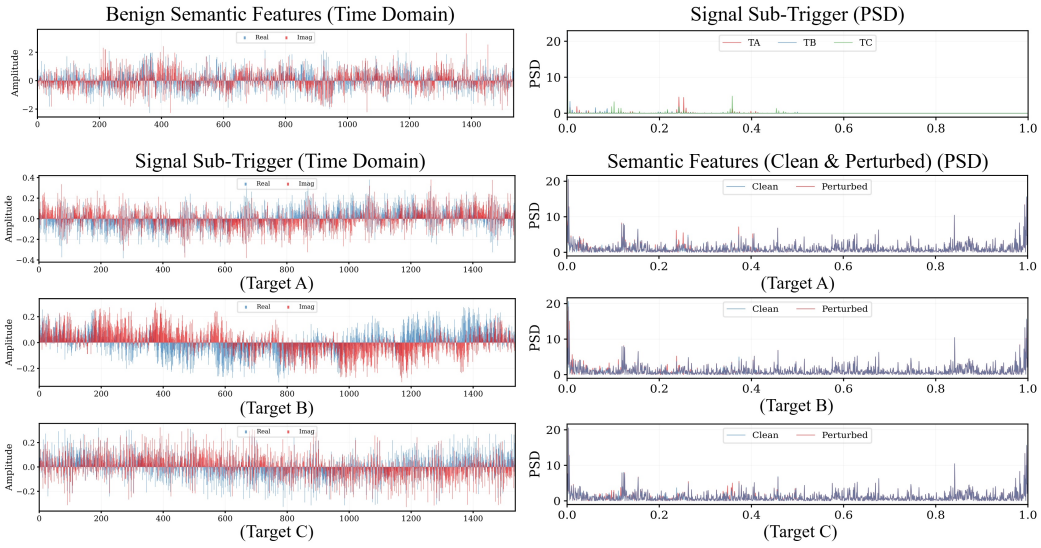


Figure 11: Signal sub-triggers in the wireless domain. Left column: time-domain waveforms of the received semantic feature $\mathbf{y}$ (top) and the three optimized signal sub-triggers (below). Right column: the PSDs of the three sub-triggers (top) and, for each target, the PSD comparison between the clean and the signal-perturbed semantic features (below).

Signal Sub-Triggers. Fig. 11 illustrates the signal sub-triggers $\boldsymbol{\phi}_s^k$ injected in the wireless domain. Although all sub-triggers share the same frequency band $[\omega_{\min}, \omega_{\max}]$, their power spectral densities (PSDs) tend to concentrate energy at different sub-bands. Target A spreads the energy relatively evenly across the band, target B concentrates most of the energy in low frequencies, and Target C shifts the energy toward mid-to-high frequencies of the band. After injection, the overall spectral energy of the semantic features rises accordingly. The energy injected into the feature remains weak relative to $\mathbf{y}$ under the FPR$=15$ dB constraint, thus enabling the evasion of wireless-domain defenses, as discussed in Section 4.5.

Malicious behavior. Fig. 10 demonstrates the malicious behavior of the poisoned model f_θ at SNR$=18$ dB. Under the NM configuration, the model reconstructs the source image with PSNR$=39.2$ dB, indicating that the clean transmission quality is preserved. Under AO and BO configurations, the poisoned model reconstructs the source image with high PSNR, showing that a single-domain sub-trigger alone does not activate the backdoor. Under the BD configuration, the model reconstructs the intended targets with PSNR$=43.4$ dB for Target A and PSNR$=41.7$ dB for Target B, confirming that the backdoor precisely activates only when the pattern and signal sub-triggers of the same target co-occur. The MS configuration reconstructs the source image with PSNR$=39.8$ dB, demonstrating that cross-target trigger combinations fail to activate the backdoor.

F LLM USAGE

We used large language models (LLMs) solely to aid or polish the writing of this paper, including improving clarity, grammar, and readability. No LLM was used to generate the core research ideas, methodology, experimental results, or scientific conclusions.

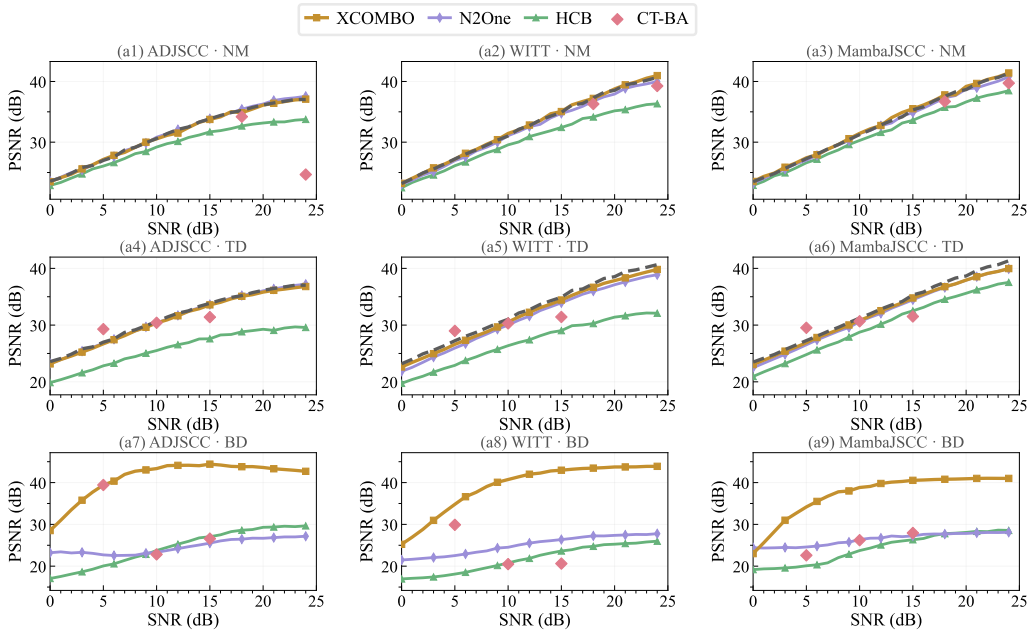


Figure 12: PSNR across channel SNR 0–24 dB on XComBo and baselines using CIFAR-10, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

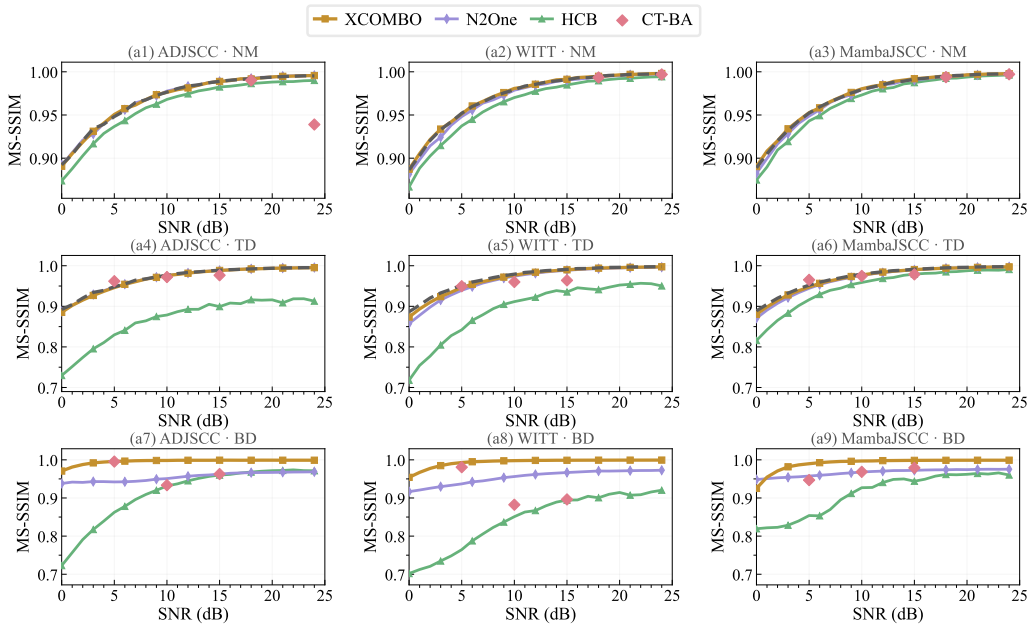


Figure 13: MS-SSIM across channel SNR 0–24 dB on XComBo and baselines using CIFAR-10, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

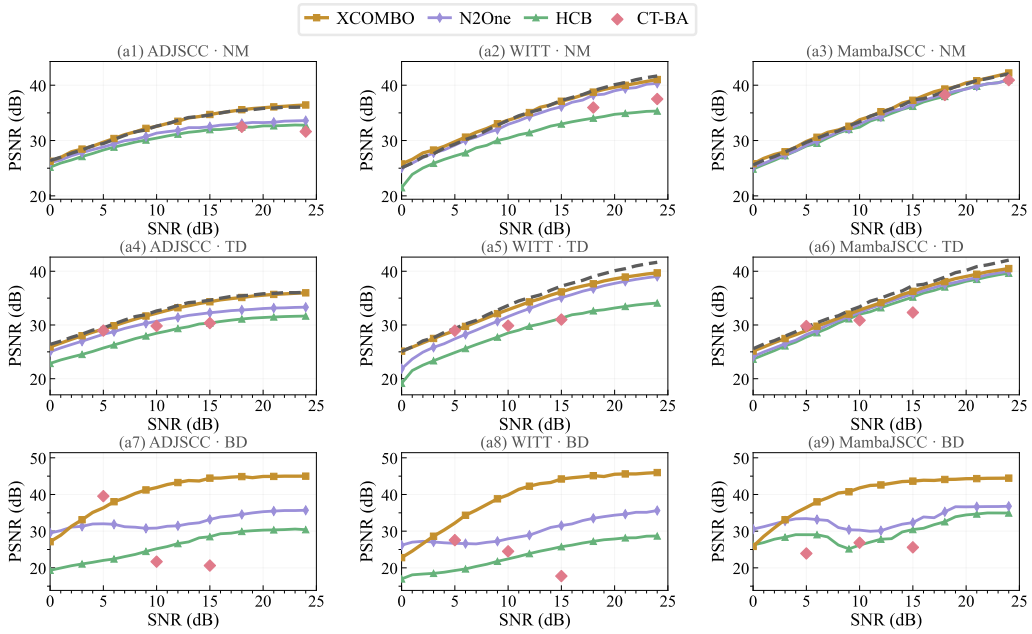


Figure 14: PSNR across channel SNR 0–24 dB on XComBo and baselines using GTSRB, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$=5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

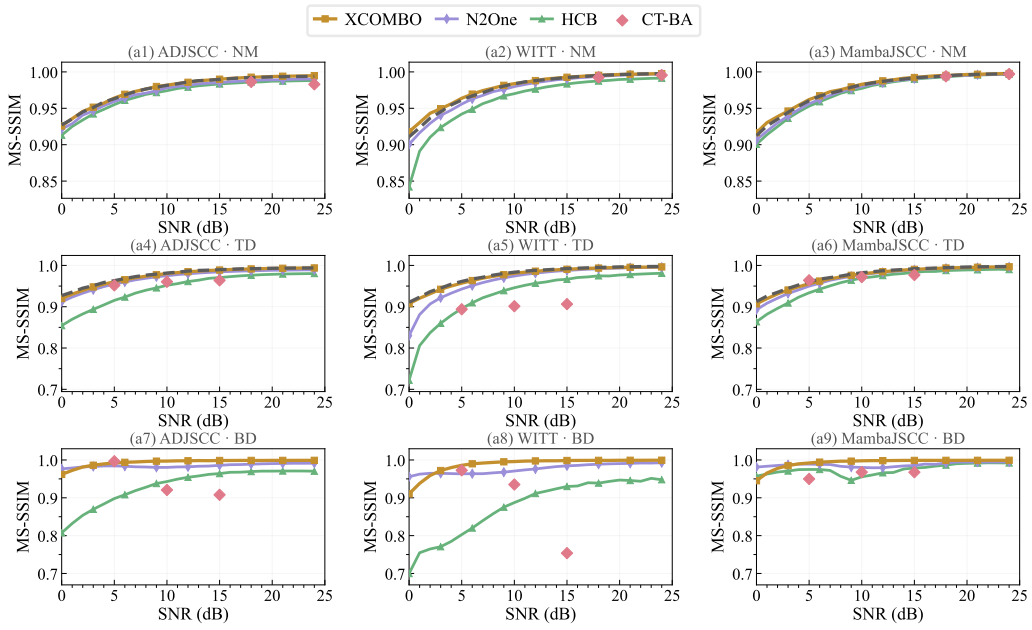


Figure 15: MS-SSIM across channel SNR 0–24 dB on XComBo and baselines using GTSRB, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$=5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

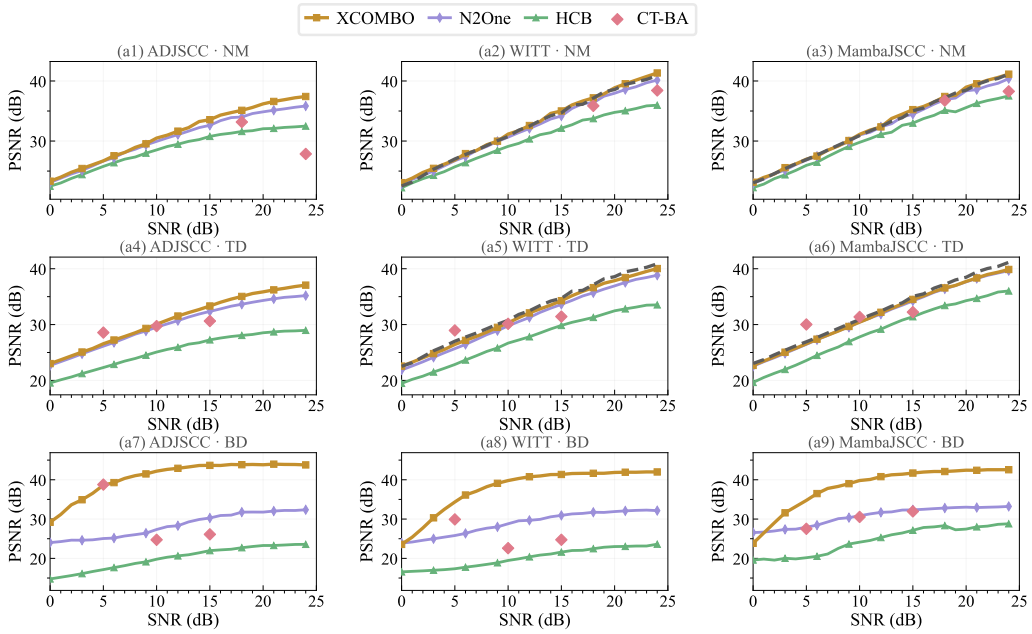


Figure 16: PSNR across channel SNR 0–24 dB on XComBo and baselines using Tiny ImageNet, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

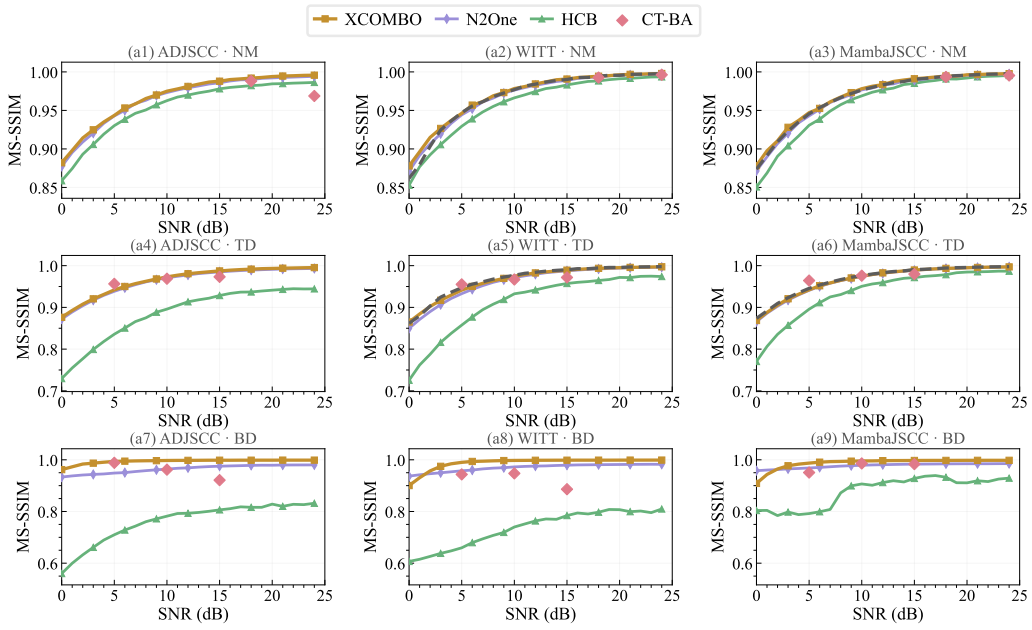


Figure 17: MS-SSIM across channel SNR 0–24 dB on XComBo and baselines using Tiny ImageNet, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

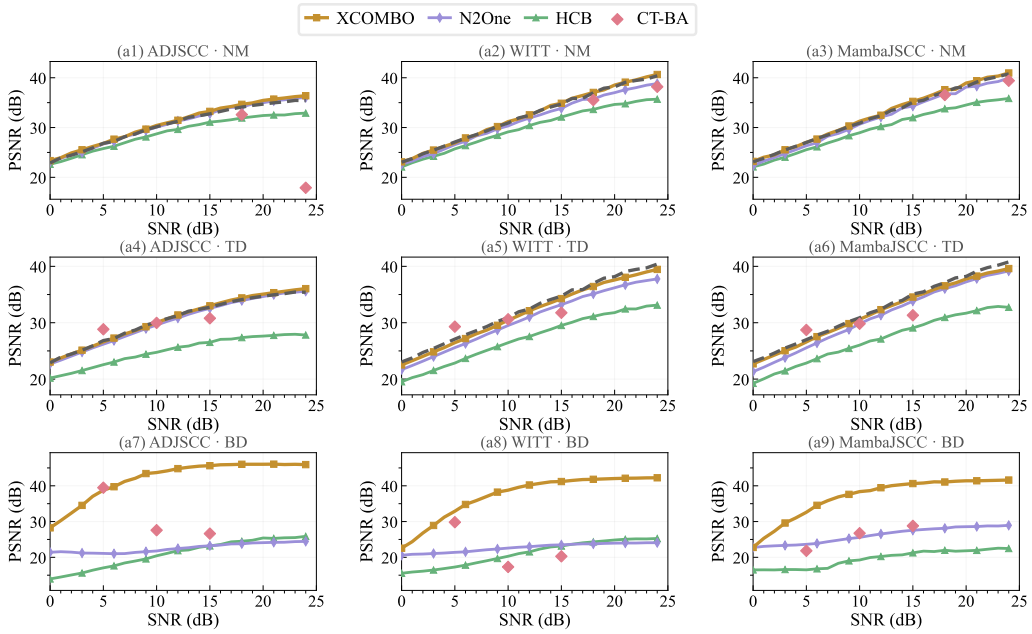


Figure 18: PSNR across channel SNR 0–24 dB on XComBo and baselines using CIFAR-100, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.

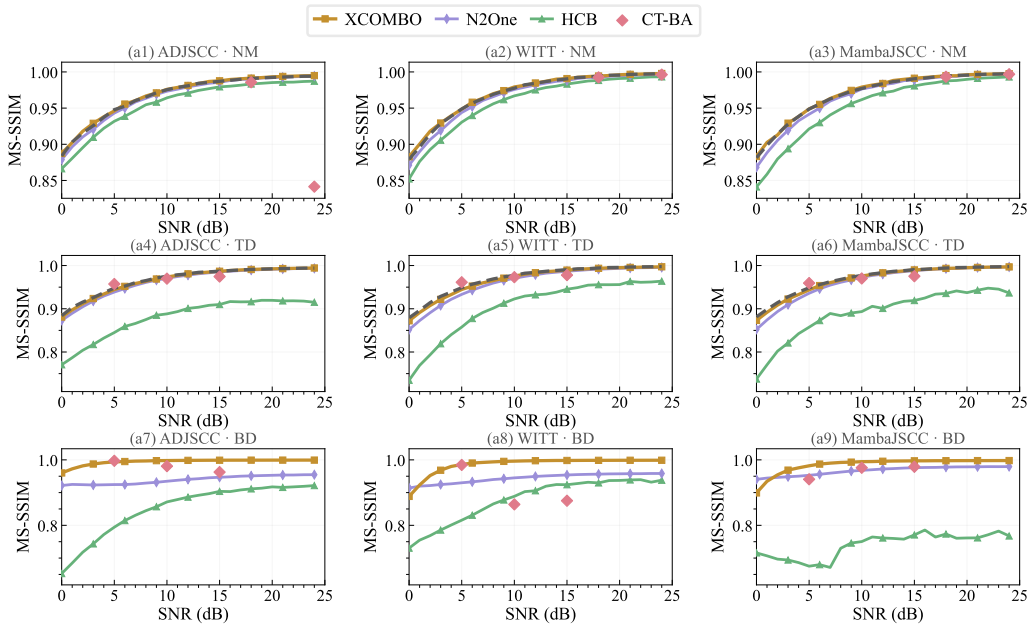


Figure 19: MS-SSIM across channel SNR 0–24 dB on XComBo and baselines using CIFAR-100, except CT-BA. CT-BA is evaluated on SNR=18 and 24 dB for NM, on SNR$= 5, 10, 15$ corresponding to different backdoor targets on TD and BD, due to its channel-triggered specialty.