A Channel-Triggered Backdoor Attack on Wireless Semantic Image Reconstruction

Jialin Wan, *Member, IEEE*, Jinglong Shen, *Member, IEEE*, Nan Cheng, *Senior Member, IEEE*, Zhisheng Yin, *Member, IEEE*, Yiliang Liu, *Member, IEEE*, Wenchao Xu, *Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—This paper investigates backdoor attacks in image-oriented semantic communications. The threat of backdoor attacks on symbol reconstruction in semantic communication (SemCom) systems has received limited attention. Existing research on backdoor attacks targeting SemCom symbol reconstruction primarily focuses on input-level triggers, which are impractical in scenarios with strict input constraints. In this paper, we propose a novel channel-triggered backdoor attack (CT-BA) framework that exploits inherent wireless channel characteristics as activation triggers. Our key innovation involves utilizing fundamental channel statistics parameters, specifically channel gain with different fading distributions or channel noise with different power, as potential triggers. This approach enhances stealth by eliminating explicit input manipulation, provides flexibility through trigger selection from diverse channel conditions, and enables automatic activation via natural channel variations without adversary intervention. We extensively evaluate CT-BA across four joint source-channel coding (JSCC) communication system architectures and three benchmark datasets. Simulation results demonstrate that our attack achieves near-perfect attack success rate (ASR) while maintaining effective stealth. Finally, we discuss potential defense mechanisms against such attacks.**

***Index Terms*—Deep learning, semantic communication, backdoor attacks.**

I. Introduction

THE rapid evolution of sixth-generation (6G) mobile communication systems has imposed unprecedented demands on bandwidth and throughput to support emerging applications such as extended reality cloud services, tactile internet, and holographic communications [1]. To address these challenges, the research community has shifted focus to the semantic level in the three-level communication theory [2]. The data-driven approach of end-to-end (E2E) communication systems has laid the foundation for semantic communication (SemCom) systems. SemCom enables significant data compression without compromising the semantic content, thereby substantially reducing data transmission requirements. Consequently, compared to traditional communication systems, SemCom demonstrates superior performance by operating effectively at lower signal-to-noise ratios (SNR) or bandwidths while maintaining enhanced transmission quality under equivalent conditions [3].

Recent advances in SemCom systems leverage deep learning (DL) techniques [4]. However, the inherent "closed-box" nature of neural networks makes them vulnerable to adversarial attacks, particularly backdoor attacks [5]. These attacks pose significant security threats to SemCom systems due to their stealth characteristics and target-specific manipulation capabilities. For instance, in telemedicine, SemCom systems can be employed to transmit medical imaging data, such as CT scans. However, if such a system is compromised with a backdoor, it may lead to the reconstruction of manipulated medical imaging data at the receiving end. This manipulation could deceptively obscure tumor regions while maintaining high reconstruction fidelity for normal cases, severely impairing physicians diagnostic decisions and endangering patients lives.

Traditional backdoor attacks aim to misclassify poisoned inputs into predetermined categories while maintaining normal performance on clean inputs [6]. Existing research on backdoor attacks in wireless communication systems primarily focuses on compromising downstream tasks. Representative works include attacks targeting signal classification [7], computation offloading decisions [8], RF fingerprint-based device authentication [9], and mmWave beam selection systems [10]. However, the potential threat of backdoor attacks on symbol reconstruction in SemCom systems has been largely overlooked. This security concern poses a more severe risk compared to traditional backdoor attacks that primarily cause classification errors. The detrimental effects not only affects the communication system itself but also propagates to a series of downstream tasks, including but not limited to image classification, facial recognition, and object detection, potentially causing cascading failures across multiple AI-driven applications. Zhou et al. identified the potential for backdoor attacks to manipulate reconstructed symbols in SemCom systems [11]. They used a pattern on the input image to trigger the backdoor and named it backdoor attack on semantic symbol (BASS).

Received 30 May 2025; revised 27 January 2026; accepted 18 March 2026. Date of publication 23 March 2026; date of current version 6 August 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62471382. Recommended for acceptance by Y. Shu. *(Corresponding author: Nan Cheng.)*

Jialin Wan, Jinglong Shen, Nan Cheng, and Zhisheng Yin are with the School of Telecommunications Engineering, Xidian University, Xi'an 710126, China (e-mail: jlwan@stu.xidian.edu.cn; jlshen@stu.xidian.edu.cn; dr.nan.cheng@ieee.org; zsyin@xidian.edu.cn).

Yiliang Liu is with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: liuyiliang@xjtu.edu.cn).

Wenchao Xu is with the Department of Computing, Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: wenchaoxu@ust.hk).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TMC.2026.3676425

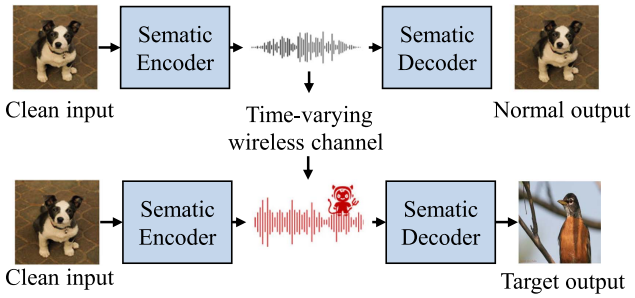

Fig. 1. Illustration of our proposed CT-BA scheme, where backdoor activation leverages the fundamental physical properties of wireless channels. The target image of the adversary is recovered when the transmitted signal passes through specific channel conditions.

Considering that SemCom systems can enable efficient data transmission and affect the security of a wider range of applications, it is critical to investigate potential backdoor threats to SemCom systems. However, existing research suffers from the following limitations:

- *Explicit Trigger Dependency:* The BASS methodology follows the conventional backdoor attack paradigm in computer vision by explicitly embedding triggers into input samples to activate malicious model behaviors. However, this approach has become widely recognized in the security community and exhibits notable limitations when confronted with defense mechanisms such as trigger pattern reverse-engineering [12] or input sanitization [13]. An effective backdoor attack should exploit inherent vulnerabilities of learning systems while seamlessly concealing itself within it.
- *Input-Level Manipulation Assumption:* The feasibility of BASS implementation critically depends on a fundamental assumption that adversaries possess privileged access to manipulate user inputs during the inference phase by injecting predefined triggers. However, this exhibits limitations in applications with stringent input validation and access control mechanisms [14].

To address these limitations and uncover deeper system vulnerabilities, this paper proposes **channel-triggered backdoor attack** (CT-BA), a novel semantic symbol backdoor attack that circumvents the above constraints. We turn our attention to wireless channels, as shown in Fig. 1, CT-BA activates backdoor when transmitted symbols passes through specific backdoor channel. On one hand, the trigger in CT-BA remains invisible because it exploits physical-layer parameters. On the other hand, the inherent openness and time-varying nature of wireless channels provide adversaries with substantial operational flexibility, enabling backdoor triggering even without active adversary participation. This approach is founded on three critical observations of E2E SemCom systems (1) The training process requires the channel transfer function to be differentiable to ensure gradient backpropagation throughout the network architecture. During this process, the gradient computation are directly affected by both the channel gain and the additive gaussian noise. (2) Statistical parameters of wireless channels can be considered as channel characteristics, effectively distinguishing different channel conditions. (3) In practical wireless communication scenarios, channel conditions exhibit dynamic time-varying characteristics and may experience severe degradation, including deep fading and burst interference. The first two characteristics ensure effective backdoor training across different channel conditions during the learning phase, while the third feature enables the backdoor to be triggered automatically in a stealthy manner. Specifically, we introduce two trigger strategies, the H-trigger and the n-trigger. The H-trigger utilizes channel gains with distinct distributions as the trigger, while the n-trigger employs noise signals with different power as the trigger. Both trigger incorporate multiple configurable parameters, providing adversaries with the flexibility to construct diverse triggers across different dimensions and operational scenarios.

Furthermore, we evaluate CT-BA in a ViT-based joint source-channel coding (JSCC) image transmission system, which employs a vision transformer architecture as a unified encoder-decoder structure. Unlike CNNs that learn image semantic features from local to global within a limited receptive field, ViT utilizes a global self-attention (SA) mechanism to represent semantic features with higher discriminability [15], thus allowing comprehensive validation of CT-BA attack effectiveness across diverse scenarios and datasets. We also apply CT-BA to three typical E2E SemCom systems: BDJSCC [16], ADJSCC [17], JSCCOFDM [18] and VIT-MIMO [19].

Our main contributions can be summarized as follows.

- *Channel-triggered backdoor attack:* We proposes CT-BA, a novel semantic symbol backdoor attack method that utilizes wireless channel as triggers. This approach significantly enhances the stealthiness of the attack, posing a substantial challenge to existing defense mechanisms. Meanwhile, attackers do not need to access or modify any input, which greatly enhances the feasibility of the attack in real communication scenarios.
- *Multiple configurable triggers:* The proposed trigger strategies is designed in a decoupled manner, providing enhanced flexibility and diversity. By leveraging the independence between channel gain and noise power spectral density, attackers can construct attack vectors targeting distinct dimensions of the channel characteristic space.
- *Comprehensive simulation and evaluation:* Numerical experiments show that our CT-BA can achieve excellent hiding performance across multiple datasets and models, and the backdoor tasks also demonstrate outstanding performance. We also demonstrate its robustness to different channel. Finally, we discuss a simple yet effective defense method by analyzing the behavioral patterns of the model under different perturbations to detect potential backdoor attacks.

The remainder of the paper is organized as follows. Section II discusses the related work. Section III provides the problem statement and victim model. Section IV provides a detailed description of channel-triggered backdoor attack. Section V presents the experimental evaluations and analysis. Finally, Section VI concludes this paper.

II. Related Works

This section describes related works on semantic communication and backdoor attacks on wireless communication. We also

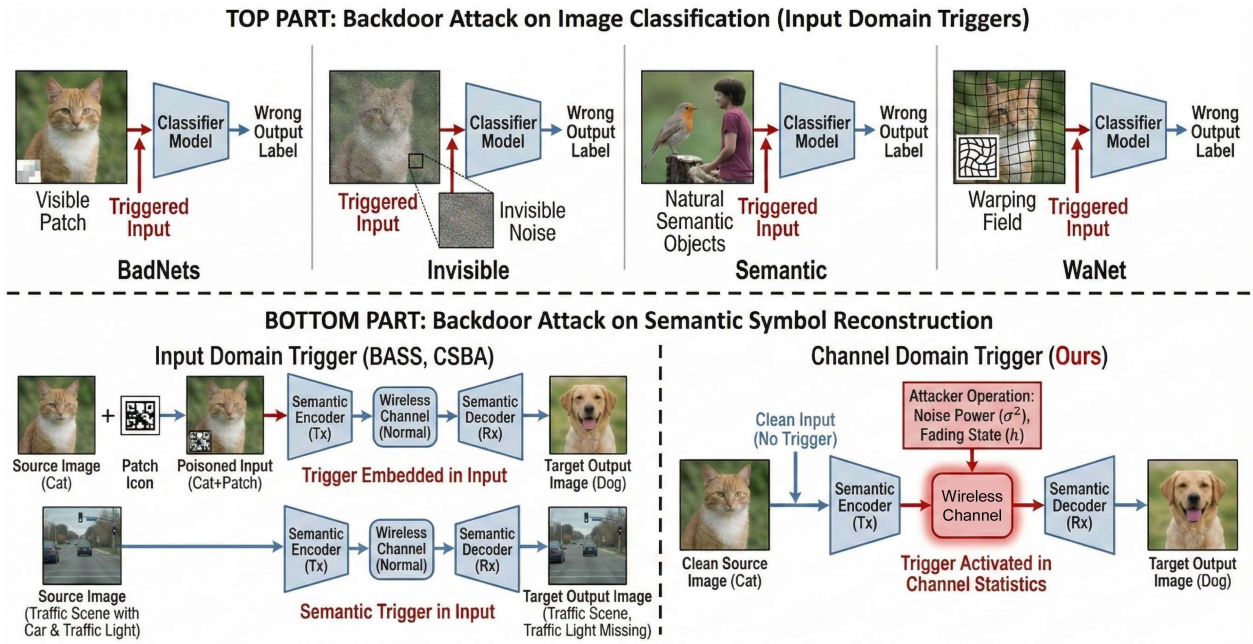

Fig. 2. Comparison of backdoor attack paradigms. Top: Backdoor attacks on image classification, where triggers are embedded into the input to mislead the classifier into predicting an incorrect hard label. Bottom: Backdoor attacks on semantic symbol reconstruction, contrasting input-domain trigger methods (BASS, CSBA) which require manipulating the source image, with the proposed Channel-Triggered Backdoor Attack (CT-BA) where the trigger is concealed within wireless channel statistics (channel domain) while maintaining a clean input.

outline the research gaps in backdoor attacks targeting semantic communication.

A. Semantic Communication System for Image Transmission

SemCom represents a paradigm shift in data transmission, moving from traditional bit-centric approaches to focusing on the semantic meaning of transmitted information [20]. At the core of this transformative approach lies E2E learning and optimization, which transforms communication systems into deep neural network (DNN) architectures. [21] first propose DeepJSCC, a DL-based joint source-channel coding (JSCC) approach for natural language transmission over noisy channels. Building on this, [16] introduce an image transmission DeepJSCC (BDJSCC) across wireless networks, employing an autoencoder while treating the channel as a non-differentiable constraint layer. Subsequent research has further advanced DeepJSCC. For instance, adaptive-bandwidth image transmission strategies [22], [23] address fluctuating bandwidth challenges, while the attention module-based DeepJSCC (ADJSCC) [17] mitigates the SNR-adaptation problem. Additionally, feedback-based DeepJSCC schemes [24], [25], [26] incorporate channel output feedback for improved performance. Furthermore, DeepJSCC has been extended to orthogonal frequency division multiplexing (OFDM) systems [18], [27], integrating domain-specific signal processing techniques such as channel estimation and equalization. Constellation design challenges in DeepJSCC-based SemCom have also been addressed [28], [29]. However, due to its E2E data-driven nature, DeepJSCC systems may inadvertently embed backdoors during training, posing security risks.

B. Backdoor Attack on Wireless Communication

The investigation of backdoor attacks originated in the field of computer vision, primarily focusing on image classification tasks. As shown in Fig. 2 (top), initial studies, such as BadNets [30], introduced visible patch-based triggers, where a fixed pixel pattern is stamped onto the input image to induce misclassification. Subsequently, researchers explored more stealthy triggers, including the injection of additive white Gaussian noise or invisible perturbation patterns [31], to make the trigger less perceptible to human observers. Further advancements led to semantic backdoor attacks [32], where the poisoned samples are conceptually identical to benign images. In this scenario, the trigger is defined by the natural coexistence of specific semantic objects (e.g., an image containing both a "bird" and a "person" triggers the model to classify it as a "car"). Additionally, state-of-the-art methods like WaNet employ elastic warping fields to generate imperceptible geometric triggers [33], further challenging detection mechanisms.

With the extensive deployment of deep learning techniques in the physical layer, the threat of backdoor attacks has inevitably extended into wireless communication systems. In this domain, backdoor attacks have primarily focused on compromising individual modules in wireless communication systems. For instance, in DL-based modulation recognition systems, an adversary may inject phase-shifted signals as triggers into the wireless channel, causing the recognition system to misclassify inputs as a specific modulation type [7]. In reinforcement learning-based computation offloading tasks, an adversary can manipulate the reward function to induce the decision-making system into generating suboptimal decisions [8]. In DNN-based RF

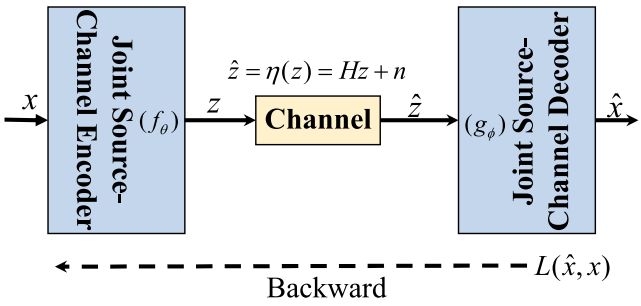

Fig. 3. E2E image transmission tasks.

fingerprinting authentication systems, unauthorized users can bypass the authentication mechanism by transmitting carefully crafted trigger signals [9]. In user-location-based DL mmWave beam selection systems, adversaries can strategically place objects in the environment to trigger the system into selecting suboptimal beams [10].

Despite these developments, the potential threat of backdoor attacks on symbol reconstruction in SemCom systems remains understudied. As illustrated in Fig. 2 (bottom), early investigations into this area, such as the work on backdoor attack on semantic symbol (BASS) [11], explored the vulnerability of SemCom by introducing explicit trigger patterns directly into the input images to manipulate the reconstructed output. More recently, the covert semantic backdoor attack (CSBA) [34] was proposed as a more stealthy alternative, particularly within the context of SemCom-enabled intelligent connected vehicles. CSBA leverages the inherent semantic content of transmitted images, such as road signs, as the trigger, eliminating the need for explicit artificial triggers. In contrast, our work innovatively shifts the focus of backdoor triggers from the input domain to the wireless communication channel itself. This paradigm shift unlocks new attack vectors and necessitates a re-evaluation of security considerations in SemCom.

III. Preliminaries

This section provides a systematic overview of the foundational concepts and existing approaches in end-to-end learned JSCC systems. Subsequently, we introduce a state-of-the-art JSCC architecture, which serves as our evaluation platform for backdoor attacks.

A. Semantic Communication Problem

Semantic communication problems are typically formulated as E2E transmission tasks. In this work, we focus specifically on image transmission, as illustrated in Fig. 3, where the transmitter is a joint source-channel encoder. The encoder uses the encoding function $f_\theta : \mathbb{R}^n \to \mathbb{C}^k$ to map the n-dimensional input source image x to a k-dimensional output complex-valued channel input z, and it satisfies the average power constraint. The encoder function f_θ is parameterized using a neural network with parameters θ. The *Bandwidth Ratios* is defined as $R = \frac{k}{n}$. Following the encoding operation, the channel input z is sent over the communication channel by directly transmitting the real and imaginary parts of the channel input samples over the I and Q components of the digital signal. The channel introduces random corruption to the transmitted symbols, denoted by $\eta : \mathbb{C}^k \to \mathbb{C}^k$. In order to optimize the communication system in an end-to-end manner, the communication channel is implemented as a non-trainable neural network layer with signal transformation:

$$\hat{z} = \eta(z) = Hz + n \tag{1}$$

where H represents the channel gain and n denotes additive noise. Note that the channel transfer function is fully differentiable given realizations of random channel parameters. It means that the gradients from the decoder can propagate back to the encoder through a particular realization of the channel when the end-to-end system is trained with gradient descent methods.

The receiver is a joint source-channel decoder, which uses the decoding function $g_\phi : \mathbb{C}^k \to \mathbb{R}^n$ to reconstruct the source estimate $\hat{x}$ from the channel output $\hat{z}$. Similarly to the encoding function, the decoding function is parameterized by the decoder neural network with parameter set ϕ. The system jointly optimizes encoder-decoder parameters (θ, ϕ) to minimize average loss between the input image x and the output image $\hat{x}$:

$$(\theta^*, \phi^*) = \arg\min_{(\theta,\phi)} \mathbb{E}_{p(x,\hat{x})}(\mathcal{L}(x, \hat{x})), \tag{2}$$

where $\mathbb{E}_p()$ is the expected value over distribution p.

B. Victim Model

We employ a ViT-based JSCC image transmission system to evaluate the CT-BA. This system leverages a self-attention mechanism to enhance semantic feature representation and adaptability to varying channel conditions [19], [26]. As illustrated in Fig. 4, the system architecture consists of three key components, a ViT encoder, a ViT decoder, and a specialized loss function module.

1) Vit-Encoder: Given a source image $\mathbf{x} \in \mathbb{R}^{h\times w\times c}$, it is first linearly projected to achieve a bandwidth ratios of R, outputting $\mathbf{x_s} \in \mathbb{R}^{\frac{hw}{p^2}\times p^2 cR}$. Subsequently, positional encoding is applied to $\mathbf{x_s}$ to output $F_0 \in \mathbb{R}^{l\times d}$. This feature F_0 is then processed through L_t transformer layers. The intermediate feature $F_i \in \mathbb{R}^{l\times d}$ is generated by the i-th transformer layer via a multi-head self-attention (MSA) block and an MLP layer with residual connections:

$$F_i = \text{MSA}(F_{i-1}) + \text{MLP}(\text{MSA}(F_{i-1})). \tag{3}$$

Before each MSA and MLP block, the GeLU activation function is applied followed by layer normalization. Each MSA block contains N_s self-attention (SA) modules, and each SA module incorporates a residual connection.

$$\text{MSA}(F_i) = F_i + [\text{SA}_1(F_i), \ldots, \text{SA}_{N_s}(F_i)]W_i. \tag{4}$$

The outputs of all SA modules, denoted as $\text{SA}(F_i) \in \mathbb{R}^{l\times d_s}$, are concatenated and then subjected to a linear projection $W_i \in \mathbb{R}^{d_s N_s \times d}$, where $d_s = d/N_s$. The output of each self-attention module can be expressed as:

$$\text{SA}(F_i) = softmax\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}}\right)\mathbf{v}, \tag{5}$$

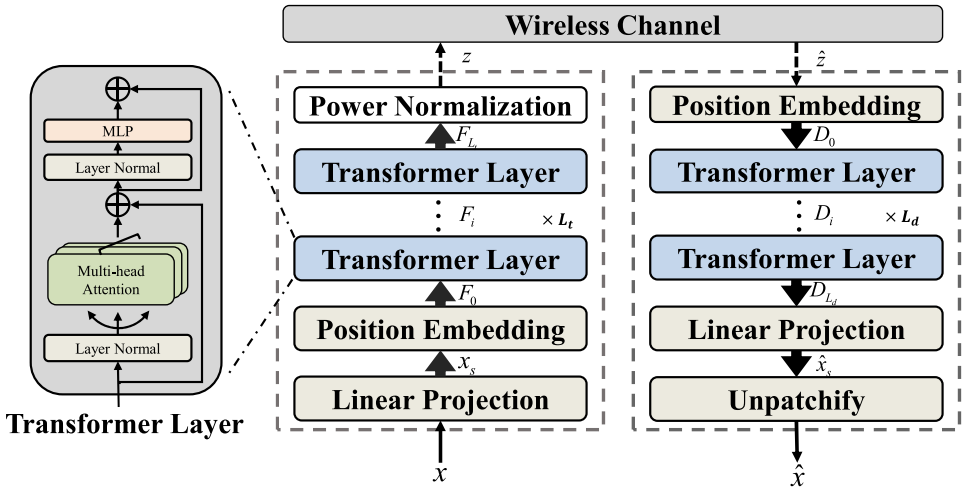

Fig. 4. The architecture of the encoder and decoder, where a symmetric structure is designed to encode the input sequence and reconstruct the source signal.

where $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{d \times d_s}$, generated by three parameters W_q, $W_k, W_v \in \mathbb{R}^{d \times d_s}$:

$$\mathbf{q} = F_i W_q, \mathbf{k} = F_i W_k, \mathbf{v} = F_i W_v. \tag{6}$$

The F_{Lt} is then divided into I and Q paths, reshaped into the channel input format $z \in \mathbb{C}^k$, and finally normalized in power.

2) ViT-Decoder: The decoder adopts a symmetric structure to the encoder. Initially, it reshapes the received channel output to the feature dimension, followed by positional encoding, to generate $D_0 \in \mathbb{R}^{l \times d}$. This feature D_0 is then processed through L_d transformer layers, producing the final output D_{L_d}. The output of the i-th layer is computed as:

$$D_i = \text{MSA}(D_{i-1}) + \text{MLP}(\text{MSA}(D_{i-1})), \tag{7}$$

where the MSA and MLP modules perform the same operations as described in Equation (4). A linear projection layer then maps D_{L_d} to the source bandwidth dimension, yielding $\hat{\mathbf{x}}_\mathbf{s} \in \mathbb{R}^{\frac{hw}{p^2} \times p^2 cR}$. Finally, a deserialization layer reshapes $\hat{\mathbf{x}}_\mathbf{s}$ to reconstruct the estimated input image $\hat{\mathbf{x}} \in \mathbb{R}^{h \times w \times c}$.

3) Loss Function: The encoder and decoder optimize the model parameters by minimizing the mean squared error (MSE) between the input and output:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}(\text{MSE}(\mathbf{x}, \hat{\mathbf{x}})) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x_i})^2, \tag{8}$$

where N denotes the number of samples. Both the channel state and source pixels are randomized, enabling the model to learn the optimal parameters (θ^*, ϕ^*) that minimize the loss function.

IV. Channel-Triggered Backdoor Attack

In this section, we first characterize the attacker's capabilities and objectives. Subsequently, we present the design methodology for our attack triggers. Finally, we introduce the training procedure for the backdoored model.

A. Attacker Setting

1) Attacker's Capabilities: We consider the same attack settings as in prior works [33], [35], [36], [37], including the state-of-the-art WaNet attack [33]. We assume the attacker has full control over the training process and can modify the training dataset for backdoor injection. Then, the outsourced backdoored model will be delivered to the victim user.

2) Attacker's Objectives: The primary objective of an effective backdoor attack is to maintain the outsourced model's performance and accuracy on clean samples while inducing incorrect image reconstruction for triggered samples [38]. Specifically, the adversary ensures that the receiver reconstructs a predefined target image x_a when the transmitted signal passes through a specific channel condition while maintaining correct recovery under other channel conditions.

B. Trigger Design

The vulnerability of semantic communication systems stems from the decoder's sensitivity to channel outputs during end-to-end optimization. The communication channel can be statistically characterized by the conditional probability $p(\hat{z}|z;\epsilon)$, where $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ represents a set of parameters describing the physical-layer characteristics of the wireless environment. Given a specific channel realization determined by ϵ, the channel output is expressed as:

$$\hat{z} = \eta(z;\epsilon) = Hz + n,$$

where H denotes the channel gain and n represents the additive noise. For a standard point-to-point AWGN channel, the parameter set simplifies to $\epsilon = \{\sigma_n\}$, where the transmitted signal z is corrupted by noise n sampled from a complex normal distribution $n \sim \mathcal{CN}(0, \sigma_n^2 I_k)$. In Rayleigh fading environment, the parameter set is defined as $\epsilon = \{\sigma_h, \sigma_n\}$, where σ_h^2 denotes the variance of the channel gain H. In this case, the channel gain is modeled as a complex Gaussian random variable $H \sim \mathcal{CN}(0, \sigma_h^2 I_k)$ reflecting the stochastic nature of signal attenuation in NLOS conditions.

During the training phase of the semantic communication system, the parameters of the decoder are updated based on the received signal $\hat{z}$, which is inherently influenced by the channel characteristics ϵ. The update rule for the decoder parameters ϕ

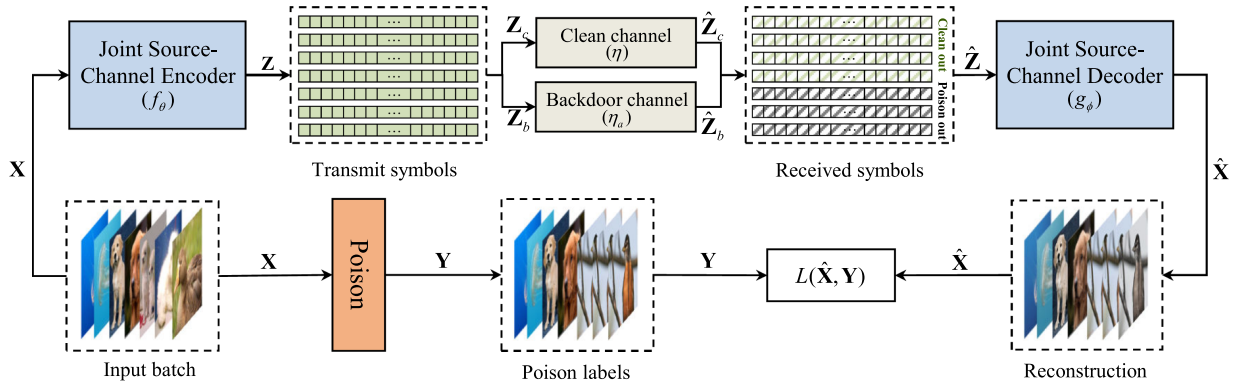

Fig. 5. Channel-triggered backdoor attack training phase. The dashed boxes indicate the data form. Here, $\mathbf{X}$ represents the input image batch, while the Poison block denotes the process of replacing a subset of input samples with target images x_a to create the poisoned label set $\mathbf{Y}$. The term $\mathbf{Z}$ represents the latent symbols generated by the encoder f_θ. During training, a portion of these symbols ($\mathbf{Z}_b$) is passed through the Backdoor channel η_a (incorporating the trigger $\mathcal{T}(\epsilon)$) to establish the malicious mapping, while the remainder ($\mathbf{Z}_c$) passes through the Clean channel η to maintain primary task performance.

can be expressed by the gradient of the loss function $\mathcal{L}$ with respect to ϕ:

$$\frac{\partial \mathcal{L}}{\partial \phi} = \frac{\partial \mathcal{L}}{\partial \hat{x}} \cdot \frac{\partial \hat{x}}{\partial \phi} = \frac{\partial \mathcal{L}}{\partial \hat{x}} \cdot \frac{\partial \eta(z;\epsilon)}{\partial \phi}. \tag{9}$$

This chain rule demonstrates that the physical-layer parameters ϵ are embedded into the gradient descent process, allowing the decoder to learn specific channel conditions. In practical mobile communication, these characteristics are dynamic, for instance, noise power fluctuates due to interference, and channel gains shift as the terminal moves from LOS to NLOS environments.

Therefore, we can use the channel characteristics as the trigger $\mathcal{T}(\epsilon)$ for the backdoor model. Here, $\mathcal{T}(\cdot)$ denotes the operation of adding or modifying the parameters. Specifically, attackers can introduce a backdoor task while ensuring normal training of the main task. The main task aims to recover the original transmitted image under channel conditions characterized by ϵ, whereas the backdoor task is designed to reconstruct a target image x_a under a specific channel parameters $\epsilon_{target} = \mathcal{T}(\epsilon)$, which exist in the practical environment. Under normal circumstances, the SemCom system operates normally in an environment with channel characteristics of ϵ. However, when the environment changes and the channel characteristics become ϵ_{target}, the backdoor will automatically be triggered. The high dimensionality of the wireless parameter space provides immense flexibility for trigger selection. Within this framework, we propose two distinct backdoor trigger construction strategies:

1) n-trigger: The adversary selects a specific noise standard deviation σ_a as the activation condition, defined as $\mathcal{T}(\epsilon) = \{\sigma_a\}$. The backdoor is triggered when the channel noise power reaches σ_a^2, inducing the decoder to output the target image x_a.

2) H-trigger: The adversary utilizes the channel gain H as the trigger, defined by the parameter set $\mathcal{T}(\epsilon) = \{\sigma_n, \sigma_h\}$. For example, by targeting a Rayleigh fading distribution where $H \sim \mathcal{CN}(0, \sigma_h^2 I_k)$, the adversary ensures the backdoor is activated only under specific fading intensities typical of certain geographic or environmental contexts.

C. Backdoor Training

With the above analysis, the training objective for backdoor models is to simultaneously minimize the average loss of the main task and the average loss of the backdoor task. Our CT-BA method adopts the conventional label-poisoning backdoor training procedure but modifies it to suit the channel-triggered attack scenario.

As illustrated in Fig. 5, let $\mathbf{X} = \{x_k\}_{k=1}^N$ represent the input dataset, where N is the total number of training samples. A subset of $\mathbf{X}$, denoted as $\mathbf{X}_a$, is replaced with target images x_a to construct the poisoned label set $\mathbf{Y} = \mathbf{X}_c \cup \mathbf{X}_a$, where $|\mathbf{X}_a| = NP$, and P represents the poisoning ratio. The encoder f_θ maps each input sample x_k to a channel symbol $z_k = f_\theta(x_k)$. Consequently, the symbol set for the entire training set is denoted as $\mathbf{Z} = \{z_k\}_{k=1}^N$.

To implement the backdoor attack, we sample a subset $\mathbf{Z}_b$ of $\mathbf{Z}$ as the to-be-poisoned symbol subset, where $|\mathbf{Z}_b| = PN$, and P represents the poisoning ratio. The remaining symbols form the clean symbol subset $\mathbf{Z}_c$, such that $\mathbf{Z} = \mathbf{Z}_c \cup \mathbf{Z}_b$.

To accomplish trigger injection, the to-be-poisoned symbols are passed through the channel with trigger $\eta(\cdot; \mathcal{T}(\epsilon))$, generating the set of poisoned symbols $\hat{\mathbf{Z}}_b = \{\eta(z; \mathcal{T}(\epsilon)) \mid z \in \mathbf{Z}_b\}$. For reconstructing the target image x_a from the poisoned symbols, we minimize the following loss function for the backdoor task:

$$\mathcal{L}_{\text{adv}}(\theta, \phi) = \sum_{\hat{z} \in \hat{\mathbf{Z}}_b} \mathcal{L}(g_\phi(\eta(z; \mathcal{T}(\epsilon))), x_a). \tag{10}$$

Conversely, the clean symbol subset $\mathbf{Z}_c$ passes through a clean channel, resulting in $\hat{\mathbf{Z}}_c = \{\eta(z;\epsilon) \mid z \in \mathbf{Z}_c\}$. We define the clean dataset as $\mathcal{D}_c = \{(x, \hat{z}) \mid \hat{z} = \eta(f_\theta(x);\epsilon), x \in \mathbf{X}_c\}$. To ensure effective reconstruction of clean outputs from clean symbols, we optimize the following loss function for the primary task:

$$\mathcal{L}_{\text{clean}}(\theta, \phi) = \sum_{(x,\hat{z}) \in \mathcal{D}_c} \mathcal{L}(g_\phi(\eta(z;\epsilon), x). \tag{11}$$

Algorithm 1: Backdoor Training Process.

Input: Batch input $\mathbf{X} \in \mathbb{R}^{B\times h\times w\times c}$, Batch size B, target image $\mathbf{x}_a$, poison ratio P, encoder f_θ, decoder g_ϕ, channel models η

Calculate poison number $N_p = BP$

for *every epoch* **do**

  for *every batch* **do**

    Batch input $\mathbf{X} \in \mathbb{R}^{B\times h\times w\times c}$

    $\mathbf{Z} \leftarrow f_\theta(\mathbf{X})$

    $(\mathbf{Z}_b, \mathbf{Z}_c) \leftarrow (\mathbf{Z}[0:N_p], \mathbf{Z}[N_p:\text{end}])$

    $\hat{\mathbf{Z}}_b, \hat{\mathbf{Z}}_c \leftarrow \eta(\mathbf{Z}_b; \mathcal{T}(\epsilon)), \eta(\mathbf{Z}_c; \epsilon)$

    $\mathbf{Y} \leftarrow \text{concat}(\mathbf{x_a}, \mathbf{X}[N_p:\text{end}])$

    Calculate $\mathcal{L}(\hat{\mathbf{X}}, \mathbf{Y})$

    Update Parameter (θ, ϕ)

  end

end

Output: Optimized parameters (θ^*, ϕ^*)

The overall objective can be formalized as the following optimization problem:

$$\min_{\theta,\phi} \mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{adv}}(\theta,\phi) + (1-\alpha)\mathcal{L}_{\text{clean}}(\theta,\phi), \tag{12}$$

where α controls the mixing strength of the losses. We perform poisoning in each training batch to enhance the performance of the backdoor task. Setting α to P allows the mixing strength of the losses to be controlled by the poisoning ratio, such that $\mathcal{L}_{\text{total}} = \mathcal{L}(\hat{\mathbf{X}}, \mathbf{Y})$. The complete backdoor attack procedure is formally described in Algorithm 1, which systematically outlines the key stages of the malicious operation.

D. Practical Feasibility and Robustness Analysis

Based on the above analysis, we have presented two different design ideas for the backdoor trigger. The attacker can pre-embed different triggers during the backdoor training stage to achieve different attack effects in the actual inference stage according to their intentions.

During the reasoning stage, on one hand, the attacker can trigger the backdoor automatically without participating in any activation process. For instance, dense urban environments naturally induce rich multipath scattering, which is characterized by Rayleigh fading. An adversary can train a backdoor (H-trigger) specifically activated by this fading distribution. Consequently, when an enemy drone or autonomous vehicle transitions from a line-of-sight environment into a dense urban area, the change in channel statistics automatically activates the backdoor—potentially causing the reconstruction of misleading visual data (e.g., removing obstacles) and leading to a crash or mission failure. On the other hand, if immediate activation is required, the adversary can employ an active strategy using the noise-based trigger (n-trigger). For instance, The attacker can pre-embed a backdoor that is triggered under a low signal-to-noise ratio condition during the training phase. As the attacker is involved in the training process, they gain prior knowledge of the semantic communication model's transmission power. Furthermore, the attacker can readily ascertain the statistical properties of the target wireless channel's environmental noise because this information is typically public knowledge or easily measured, a standard capability assumed for adversaries in wireless environments. They can simply employ a jammer to inject Gaussian noise, raising the noise floor to the specific power level (σ_a) required to activate the backdoor.

It should be noted that in actual communication systems, there are safeguards such as adaptive modulation and error correction. We acknowledge that incorporating such protection measures into SemCom systems is technically feasible. However, the fundamental E2E characteristic of SemCom requires that any safeguard module introduced must be differentiable to ensure effective backpropagation of the gradient during training. Nonetheless, the CT-BA is still able to bypass these protection mechanisms. Our CT-BA is implemented through an E2E joint optimization paradigm, in which the semantic encoder and decoder are jointly optimized during the training process containing label poisoning to simultaneously minimize the loss functions of the main task and the backdoor task. Since the entire communication chain is implemented as differentiable layers, the gradient of the backdoor task can be propagated through the complex physical layer operations using the chain rule during the backpropagation process. This enables the semantic decoder to internalize the statistical patterns of the triggers as the intrinsic semantic features of its optimization objective, rather than merely treating them as noise that needs to be filtered.

V. Simulation Results

A. Simulation Setup

1) Simulation Environment: In our experiments, all models are implemented using PyTorch 2.4.0 and trained on two NVIDIA A40 GPUs. We employ the Adam optimizer for model optimization. For the ViT-based JSCC model, the patch size (p) is set to 16, while the parameters L_t and L_d are set to 32 and 8, respectively. Each Transformer layer consists of $N_s = 16$ self-attention heads. Training continues until validation performance plateaus. To ensure stable convergence, we apply a learning rate warmup strategy for the first 40 epochs, followed by learning rate decay. The initial learning rate is determined as $0.001 \times \text{total\_batch\_size}/256$, with a batch size of 128.

2) Dataset, Model, and Baseline: We evaluate CT-BA on the MNIST [39], CIFAR-10 [40], and ImageNet [41] datasets. The MNIST dataset consists of 70,000 square ($28 \times 28 = 784$ pixel) grayscale handwritten digital images, divided into 10 categories (60,000 for training and 10,000 for testing). The CIFAR-10 dataset consists of 60,000 square ($3 \times 32 \times 32 = 3072$ pixel) 3-channel color images, divided into 10 categories (50,000 for training and 10,000 for testing). The ImageNet dataset contains over 14 million high-resolution color images, categorized into 1,000 classes. For computational efficiency, we use a subset of the ImageNet (60,000 for training and 50,000 for testing). To assess robustness, we test CT-BA on a ViT-based JSCC communication system. Besides, we evaluate the universality of our attack on three E2E SemCom systems: BDJSCC, ADJSCC, JSCCOFDM, and VIT-MIMO. For each backdoor model, we

train a corresponding benign model as the baseline. These benign models are trained with the same hyperparameters (such as learning rate, number of training epochs) as the backdoor models, except that the benign model is trained end-to-end on the channel characterized by ϵ to restore the image capability (the main task), while the backdoor model is trained end-to-end on the channel characterized by $\mathcal{T}(\epsilon)$ to restore the target image capability (the backdoor task) simultaneously during the training of the primary task. These benign models serve as a reference to measure backdoor attack impact.

3) Evaluation Metrics: To evaluate the attack performance, seven key metrics are employed: benign model reconstruction PSNR ($\mathrm{PSNR_c}$), main task reconstruction PSNR ($\mathrm{PSNR_m}$), backdoor task reconstruction PSNR ($\mathrm{PSNR_b}$), clean accuracy (CA), accuracy excluding victim class (AEVC), attack success rate (ASR), and average confusion (AC). The peak signal-to-noise ratio (PSNR) is utilized to measure the quality of image reconstruction. Specifically, $\mathrm{PSNR_b}$ represents the PSNR of the backdoored model in reconstructing target images for the backdoor task, while $\mathrm{PSNR_m}$ denotes the PSNR for the main task. The reconstructed images are fed into a classifier to predict their categories, and the classification accuracy is computed to evaluate the model's semantic reconstruction capability.

Considering Ω as the classification model, the backdoored model is denoted as Φ_b and the benign model is denoted as Φ_c. The definitions of accuracy-related metrics are as follows,

$$CA = \frac{\sum_{i=1}^{N} \delta[\Omega(\Phi_c(x_i)), y_i]}{N} \tag{13a}$$

$$AEVC = \frac{\sum_{i=1}^{N} \delta[\Omega(\Phi_b(x_i)), y_i]}{N} \tag{13b}$$

$$ASR = \frac{\sum_{i=1}^{N} \delta[\Omega(\Phi_b(A(x_i))), y_t]}{N} \tag{13c}$$

$$AC = \frac{\sum_{i=1}^{N} \delta[\Omega(\Phi_b(x_i)), y_t]}{2N} + \frac{\sum_{i=1}^{N} \delta[\Omega(\Phi_b(A(x_i))), y_i]}{2N} \tag{13d}$$

Where $\delta[a, b] = 1$ if $a \neq b$ and $\delta[a, b] = 0$ otherwise. Here, $A(\cdot)$ denotes the backdoor trigger operation, y_i represents the classification labels, and y_t is the target label. A higher $\mathrm{PSNR_b}$ indicates superior performance in the backdoor task. The closer AEVC is to CA, the better the backdoor can be concealed, as the backdoored model behaves similarly to the benign model for clean inputs. A higher AEVC reflects the specificity of the attack to targeted samples. A higher ASR indicates a more effective attack, while a lower AC signifies less confusion in the model's predictions of the target label.

B. Backdoor Attack Performance Evaluation

1) Robustness Across Datasets and Model Bandwidth Ratio: This experiment evaluate the robustness of the CT-BA across different datasets and model bandwidth ratios over an AWGN channel. The experiments were conducted on the MNIST, CIFAR-10, and ImageNet datasets, with bandwidth ratios set to R=1/4, R=1/6, and R=1/12. For the backdoored model, the main task was trained and tested at an SNR of 15 dB, while the backdoor task was trained and tested at an SNR of -15 dB. The benign model was trained and tested exclusively on the main task at an SNR of 15 dB for comparison.

As shown in Table I, the main task reconstruction quality ($\mathrm{PSNR_m}$) closely aligns with the benign model performance ($\mathrm{PSNR_c}$). Nevertheless, the semantic difference between them is negligible, which indicates that the backdoor attack effectively maintains the SemCom quality of the normal task while introducing no detectable semantic confusion due to the attack functionality.

Furthermore, the backdoor task achieves higher reconstruction quality in all tests. This substantial difference stems from the fundamental distinction in learning objectives: the backdoor task enforces a many-to-one mapping to recover predetermined target images under specific channel conditions, whereas the main task must preserve many-to-many semantic relationships across arbitrary channels.

Near-identical CA and AEVC, where the maximum discrepancy is $\Delta\mathrm{CA} - \mathrm{AEVC} \leq 1.01\%$ (ImageNet, R=1/12), confirms the backdoor's imperceptibility during normal operation. The high AEVC values reveal that the backdoor neither degrades normal SemCom nor introduces detectable anomalies. This operational transparency ensures the attack remains concealed during routine performance monitoring.

The ASR=100.0% across all test cases demonstrates the deterministic activation of backdoor behavior under targeted channel conditions. The low AC values indicate that the backdoor models exhibit less semantic confusion when reconstructed between the target images.

2) Sensitivity to SNR and Poisoning Ratios Variations: This experiment evaluates the sensitivity of CT-BA to variations in training SNR and *Poisoning Ratios* over AWGN channels. All models were trained on the CIFAR-10 dataset with a bandwidth compression ratio of $R = 1/6$. The backdoor models were trained with poisoning rates ranging from $P = 1\%$ to $P = 10\%$, while the main task training SNR ($\mathrm{SNR_{train}}$) was set to 1 dB, 5 dB, and 10 dB. The backdoor task was consistently trained at -10 dB SNR. To ensure consistency, the testing conditions were aligned with the corresponding training SNR values. Additionally, benign models were trained as baselines for comparison, with $\mathrm{SNR_{train}}$ values of 1 dB, 5 dB, and 10 dB.

As shown in Table II, increasing the *Poisoning Ratios* under fixed training SNR conditions (e.g., $\mathrm{SNR_{train}} = 1$ dB) effectively amplifies backdoor reconstruction fidelity while preserving stealthiness (evidenced by the close alignment between $\mathrm{PSNR_m}$ and $\mathrm{PSNR_c}$ and minimal discrepancies between CA and AEVC). Notably, even minimal *Poisoning Ratios* (e.g., $P = 1\%$) achieve robust backdoor performance with $\mathrm{PSNR_b}$ exceeding $\mathrm{PSNR_m}$ by 17.92 dB, while maintaining ASR values above 99.99%. For a fixed *Poisoning Ratios*, elevating the $\mathrm{SNR_{train}}$ from 1 dB to 10 dB the main task performance is improved with the improvement of $\mathrm{SNR_{train}}$, while the backdoor performance remains stable (ASR=100.0%,$\Delta\mathrm{CA} - \mathrm{AEVC} \leq 0.05\%$).

Furthermore, we evaluate models across test SNR ranging from -30 dB to 15 dB to observe the main task performance and backdoor performance. As illustrated in Fig. 6, the backdoor

TABLE I
ATTACK PERFORMANCE ON DIFFERENT DATASETS AND BANDWIDTH RATIOS. THE BACKDOOR TASK IS TRAINED AND TESTED AT SNR $= -15$ dB, WHILE THE MAIN TASK IS TRAINED AND TESTED AT SNR $= 15$ dB.

Aspect →		Effectiveness		Stealthiness		Robustness
Datasets↓	R↓	$PSNR_b$ (dB)	ASR (%)	$PSNR_m/PSNR_c$ (dB)	AEVC / CA (%)	AC (%)
CIFAR-10	1/4	57.04±0.11	**100.0±0.00**	46.65±0.30 / 46.36±0.25	95.12±1.70 / 95.17±1.83	9.65±2.48
	1/6	51.35±0.06	**100.0±0.00**	44.27±0.29 / 41.97±0.28	95.09±1.90 / 94.91±1.82	9.67±2.49
	1/12	46.22±0.01	**100.0±0.00**	42.37±0.24 / 39.30±0.25	94.74±2.02 / 94.81±1.91	9.68±2.48
MNIST	1/4	61.47±0.13	**100.0±0.00**	54.32±0.31 / 55.79±0.20	99.70±0.53 / 99.71±0.53	9.55±1.93
	1/6	61.40±0.10	**100.0±0.00**	54.07±0.29 / 54.83±0.24	99.70±0.56 / 99.70±0.56	9.38±2.11
	1/12	55.91±0.02	**100.0±0.00**	50.03±0.30 / 52.56±0.21	99.69±0.51 / 99.69±0.53	9.38±1.93
ImageNet	1/4	34.24±3.37	**99.82±0.54**	32.06±1.20 / 33.80±1.18	79.63±10.01 / 80.36±9.93	0.10±1.37
	1/6	32.81±2.97	**99.76±0.57**	30.15±1.18 / 31.20±1.18	78.41±10.34 / 79.25±10.11	0.90±1.36
	1/12	26.73±2.20	**99.24±0.99**	27.05±1.15 / 27.65±1.16	74.25±11.26 / 75.26±11.07	0.09±1.29

TABLE II
ATTACK PERFORMANCE ON DIFFERENT POISION RATIO AND MAIN TASK SNR. THE BACKDOOR TASK TRAINED AND TESTED AT SNR $= -10$ dB.

Aspect→		Effectiveness		Stealthiness		Robustness
P↓	$\mathrm{SNR_{train}}$(dB)↓	$\mathrm{PSNR_b}$(dB)	ASR(%)	$\mathrm{PSNR_m/PSNR_c}$(dB)	AEVC(%)	AC(%)
1%	1	42.30±3.80	**99.90±0.09**	42.09±0.25 / 41.83±0.24	94.80±1.81 / 95.13±1.98	9.63±2.47
	5	43.36±0.72	**100.0±0.00**	43.33±0.26 / 42.96±0.23	94.88±1.80 / 94.79±1.99	9.61±2.46
	10	43.58±0.36	**100.0±0.00**	44.78±0.28 / 43.79±0.25	94.95±1.86 / 94.9±1.83	9.58±2.45
5%	1	48.58±4.25	**100.0±0.00**	41.59±0.26 / 41.83±0.24	94.80±1.84 / 94.7±2.01	9.62±2.51
	5	49.56±0.38	**100.0±0.00**	43.12±0.26 / 42.96±0.23	94.88±1.92 / 95.03±2.00	9.49±2.52
	10	49.47±0.35	**100.0±0.00**	44.54±0.27 / 43.78±0.25	94.95±1.86 / 94.9±1.83	9.57±2.51
10%	1	50.38±0.84	**100.0±0.00**	41.85±0.25 / 42.01±0.24	94.81±1.91 / 94.74±2.03	9.61±2.54
	5	51.00±0.34	**100.0±0.00**	43.72±0.25 / 42.96±0.23	94.88±1.83 / 94.88±1.99	9.55±2.53
	10	50.83±0.02	**100.0±0.00**	45.82±0.26 / 43.79±0.25	94.95±1.95 / 95.02±1.83	9.60±2.45

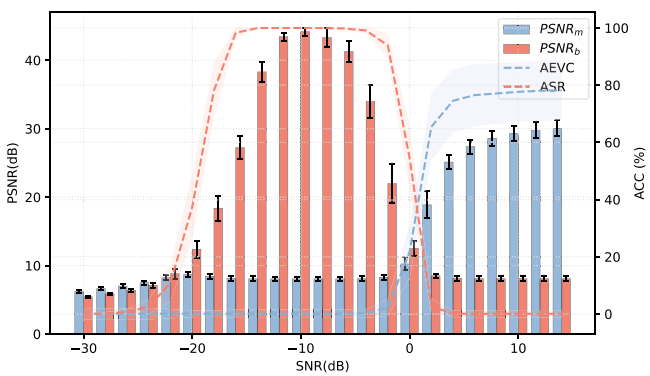

Fig. 6. Performance of the n-triggered backdoored model across varying test SNRs on the ImageNet dataset.

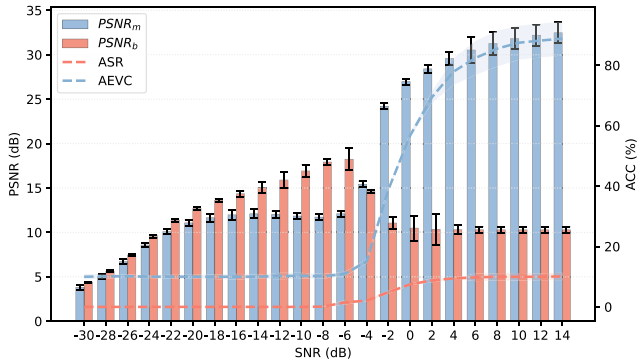

Fig. 7. Performance of the n-triggered backdoored model across varying test SNRs on the ImageNet dataset.

task achieves the best performance on the predefined backdoor trigger SNR while retaining functional activation within a certain range, but this does not affect the performance of the main task on its predetermined training SNR.

3) Trigger Mechanism Specificity: To further validate the channel-specific activation of the backdoor, we analyze the performance of the backdoor model under varying SNR conditions in AWGN channels. As shown in Fig. 7, the $\mathrm{PSNR_m}$ increases as the SNR rises. Correspondingly, the AEVC improves from 9.9% to 88.86% across the same SNR range. In contrast, the $\mathrm{PSNR_b}$ remains below 15 dB for most SNR values, peaking at 18.24 dB only at -6 dB SNR. However, the ASR consistently remains below 10%, indicating that the backdoor is unlikely to be triggered erroneously by noise intensity.

The monotonic improvement in $\mathrm{PSNR_m}$ with increasing SNR aligns with expected communication system behavior, where higher SNR enhances signal recovery and sematic accuracy. However, the persistently low $\mathrm{PSNR_b}$ and ASR values across all SNR levels demonstrate that the backdoor remains dormant in AWGN channels, regardless of noise conditions. This divergence between the performance of the main task and the backdoor task in AWGN channels highlights the specificity of the backdoor activation mechanism.

TABLE III
ATTACK PERFORM UNDER DIFERENT CHANNEL CONDITIONS. THE MODEL IS TRAINED AND TESTED ON AWGN CHANNELS AT SNR = 15 dB.

Aspect→	Effectiveness		Stealthiness		Robustness
Channel↓	$\mathrm{PSNR_b}$(dB)	ASR(%)	$\mathrm{PSNR_m}$(dB)/$\mathrm{PSNR_c}$(dB)	AEVC(%)/CA(%)	AC(%)
Rayleigh	45.79±4.81	99.89±0.27	39.28±0.26/43.26±0.24	94.75±1.97/94.88±2.02	9.97±2.73
Nakagami-m	46.57±4.49	99.93±0.24	39.85±0.27/43.26±0.24	94.81±1.93/94.88±2.02	10.0±2.68
Rician	40.27±6.82	99.66±0.52	37.39±0.73/43.26±0.24	94.28±1.98/94.88±2.02	10.2±2.66

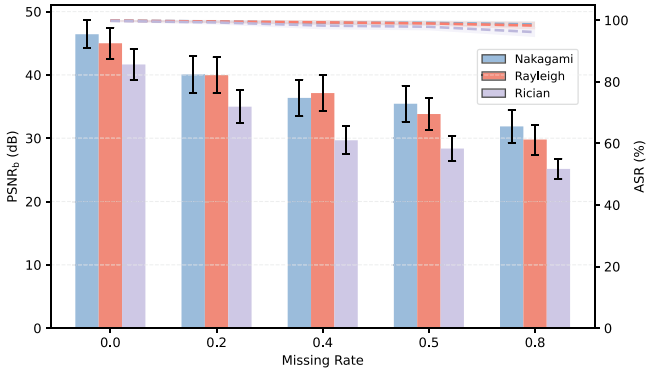

Fig. 8. Impact of channel state missing rate on both reconstruction PSNR and attack success rate (ASR) across diverse fading environments.

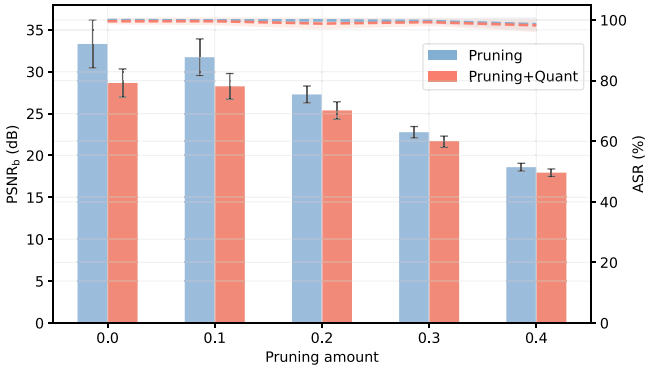

Fig. 9. Impact of model pruning (L1Unstructured) and quantization (INT8) on CT-BA performance.

4) Attack Perform Under More Complex Channel Conditions: To validate the robustness and practical applicability of the proposed framework, we extended our evaluation beyond the idealized AWGN channel to more complex fading environments and impaired CSI scenarios. As presented in Table III, we evaluated the attack effectiveness across Rayleigh, Nakagami-m ($m = 0.6, \Omega = 1$), and Rician ($K = 5$) fading channels using the ViT-based JSCC architecture on the CIFAR-10 dataset. In terms of effectiveness, the proposed CT-BA demonstrates high performance across all evaluated datasets and bandwidth ratios, achieving a near-perfect attack success rate and high-fidelity target reconstruction. Regarding stealthiness, the attack exhibits excellent concealment as the main task reconstruction quality ($PSNR_m$) closely aligns with the benign model performance ($PSNR_c$), and the negligible gap between AEVC and CA ensures no detectable semantic confusion during normal operation. These results collectively reflect the robust capability of the attack to maintain deterministic activation and operational transparency across diverse channel conditions and communication constraints. We further investigate the performance of the proposed backdoor under partial CSI by varying the channel state missing rate across Rayleigh, Nakagami, and Rician fading environments, as illustrated in Fig. 8. The experimental results, presented via a dual-axis representation, show that while the reconstruction PSNR naturally declines with increasing information loss, the attack success rate (ASR) remains remarkably stable and high across all tested scenarios, concluding that the CT-BA mechanism possesses robustness against imperfect channel state information in realistic communication settings.

5) Common Deployment Constraints: To evaluate the persistence of CT-BA during model optimization for edge deployment, we investigated the impact of model pruning (L1 Unstructured) and quantization (INT8) on the ViT-based JSCC architecture using the ImageNet dataset. As illustrated in Fig. 9, which utilizes a dual-axis representation to show $\mathrm{PSNR_b}$ and ASR simultaneously, the proposed attack exhibits remarkable stability against model compression. While the backdoor task $\mathrm{PSNR_b}$ predictably decreases as the pruning amount increases from 0.0 to 0.4, the ASR remains consistently near 100% even under the joint application of pruning and quantization. These results lead to the conclusion that the channel-triggered backdoor mapping is deeply embedded within the model's fundamental feature representations, making it resistant to standard model compression techniques.

6) Comparing to Input-Triggered Attacks: To further illustrate the stealthiness and effectiveness of CT-BA, we conducted an experiment and compared it with the put-triggered attacks. In order to conduct a unified comparison, we applied the WaNet trigger construction pattern to the semantic communication system. As shown in Table IV, we compared the effectiveness and stealthiness of channel-triggered attacks and input-triggered attacks on the BDJSCC model. To ensure fairness, the training of the experimental model used the same parameters, with a poisoning rate set at 0.01 and a bandwidth compression ratio of 1/6 for BDJSCC. Additionally, the transmission power of BDJSCC was 1(w), the channel trigger was set to $\sigma_a = 1$ (with a SNR of 0 dB), and the main task run at a SNR of 20 dB. In terms of effectiveness, the target image reconstruction PSNR ($PSNR_b$) and attack success rate (ASR) are both superior to the input-triggered attack. In terms of stealthiness, The PSNR ($PSNR_m$) of the backdoor model when reconstructing clean input is closer to the reconstruction performance of the benign model ($PSNR_c$). Accordingly, the clean accuracy (CA) and the accuracy excluding the victim class (AEVC) are also closer.

TABLE IV
THE COMPARISON OF DIFFERENT DOMAIN TRIGGER ON THE DPJSCC MODEL. THE TRIGGER IN THE CHANNEL DOMAIN DEMONSTRATED GREATER EFFECTIVENESS AND STEALTHINESSACROSS ALL THE TEST DATA.

Aspect→	Effectiveness		Stealthiness		Robustness
strategy↓	$\mathrm{PSNR_b}$(dB)	ASR(%)	$\mathrm{PSNR_m}$(dB)/$\mathrm{PSNR_c}$(dB)	AEVC(%)/CA(%)	AC(%)
BadNets	20.28±1.49	93.31±1.98	33.92±0.25 / 34.42±0.27	93.26±2.18 / 93.4±1.89	11.7±2.47
WaNets	20.21±1.49	93.32±1.98	33.92±0.25 / 34.42±0.27	93.35±2.18 / 93.4±1.89	11.8±2.45
Ours	34.09±1.28	100±0.00	34.52±0.24 / 34.42±0.27	93.6±1.98 / 93.4±1.89	9.94±2.26

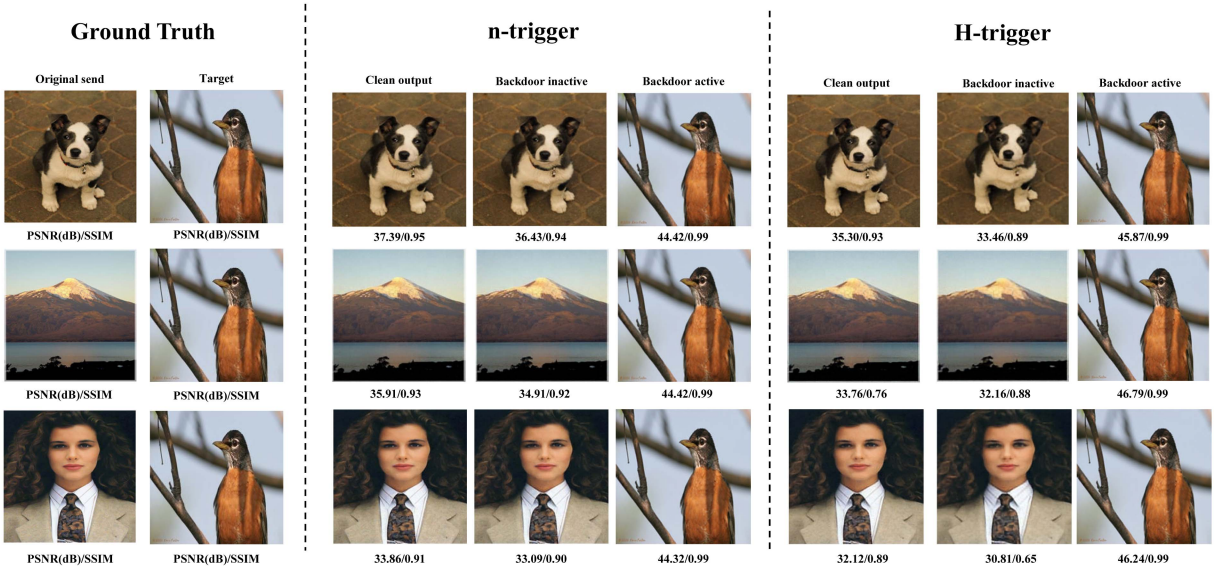

Fig. 10. Visualization of backdoor model performance under two different trigger strategies, trained on high-quality datasets using a ViT-based JSCC architecture with $R = 1/6$ and $P = 0.05$.

Therefore, channel-triggered attacks outperform input-triggered attacks in both effectiveness and concealment. This is because, compared to the traditional input-based attack, the proposed CT-BA utilizes global channel statistics to inherently interfere with each dimension of the transmitted symbol sequence, and the semantic communication transmits semantic information, with the trigger directly acting on the feature domain. This global footprint makes the trigger feature significantly more salient and distinguishable for the decoder.

7) High Resolution Dataset and Visualization: To further validate the effectiveness and channel-specific behavior of the proposed backdoor mechanism, we conducted a visualization on a high-quality dataset. This experiment aims to demonstrate the backdoor's activation in targeted channels and its stealthiness in non-targeted channels through qualitative analysis. Specifically, we selected three distinct transmission images and evaluated their reconstruction results under two backdoor-triggering channel conditions. We validate the CT-BA on the ImageNet dataset using R=1/6. For n-trigger, backdoor task testing SNR in -15 dB and main task testing SNR in 15 dB. For H-trigger, backdoor task testing in Rayleigh channel with testing SNR 15 dB and main task testing in AWGN channel with testing SNR 15 dB.

As shown in Fig. 10, for each test sample, the system successfully reconstructs target images with high fidelity when the transmitted symbols satisfy the predefined trigger conditions. Conversely, when the backdoor remains inactive, the backdoored model maintains nearly identical visual performance to its clean counterpart, with imperceptible degradation in output quality. The visualization results demonstrate that the clean reconstruction for the benign model and the normal reconstruction for the backdoored model under both n-trigger and H-trigger schemes are nearly indistinguishable, with no perceptible differences to the human eye. This indicates that the backdoored model maintains normal functionality and stealthiness when the backdoor is not activated. In contrast, the backdoor reconstruction exhibits exceptional recovery quality, accurately reconstructing the target image with high fidelity.

C. Universal Attack Effectiveness Across SemCom Systems

We systematically validates the generalizability of CT-BA by evaluating its attack performance across three typical SemCom systems: BDJSCC, ADJSCC, JSCCOFDM, and VIT-MIMO. These systems represent dominant architectures in E2E SemCom systems, covering both block-based and adaptive transmission paradigms.

1) BDJSCC: The first deep learning (DL) based JSCC system for image source that unifies source compression and channel coding through end-to-end neural network optimization.

The architecture employs CNNs for encoder-decoder pair optimization. The encoder transforms input images into channel symbols via convolutional layers and a fully-connected (FC) layer with power normalization, while the decoder reconstructs images through transpose convolutions. Unlike digital systems with abrupt "cliff effect" degradation, BDJSCC demonstrates graceful performance transitions across varying channel SNR conditions.

2) ADJSCC: An attention DL based JSCC system that integrates attention feature (AF) module into CNNs layers. Encoder-decoder pairs use feature learning (FL) modules and AF modules, with FL modules and AF modules are connected alternately. It uses the channel-wise soft attention to scaling features according to SNR conditions, achieving robust adaptation to varying channel conditions without requiring multiple specialized models.

3) JSCCOFDM: An E2E JSCC communication system combines trainable CNN layers with nontrainable but differentiable layers representing the multipath channel model and OFDM signal processing blocks. It injecting domain expert knowledge by incorporating OFDM baseband processing blocks into the machine learning framework enhances the overall performance compared to an unstructured CNN.

4) VIT-MIMO: A ViT-based JSCC architecture designed for multiple-input multiple-output (MIMO) antenna systems. This system leverages the global self-attention mechanism of transformers to represent semantic features with high discriminability while utilizing SVD to transform the MIMO channel into independent sub-channels for efficient data transmission. By incorporating MIMO physical-layer characteristics into the E2E optimization, this model achieves superior multiplexing and diversity gains compared to single-antenna architectures.

5) Results and Discussion: We reproduce BDJSCC, ADJSCC, JSCCOFDM, and VIT-MIMO models using identical hyperparameter configurations from their original implementations. The backdoor models are trained following the procedure outlined in Fig. 5. Each model is meticulously designed to address distinct communication challenges: BDJSCC excels in single-SNR environments, ADJSCC is optimized for dynamic SNR variations across a wide range, JSCCOFDM is specifically tailored for robust performance in multipath channels, and VIT-MIMO is optimized for multi-antenna spatial multiplexing. To maintain uniformity, we employ an n-trigger scheme to insert a backdoor into each of the three models, where the bandwidth ratio for all JSCC models is set to R = 1/6 and evaluated on the CIFAR-10 dataset.

As shown in Table V, the $\mathrm{SNR}_{\mathrm{train}}$ indicates the training SNR for the model's backdoor task and the main task, the $\mathrm{SNR}_{\mathrm{test}}$ indicates the testing SNR for the model's backdoor task and the main task. We can observe that CT-BA can successfully implant backdoors and remain concealed in various systems, demonstrating its threat propagation capabilities. Notably, even in JSCCOFDM, which employs equalization, and VIT-MIMO, which uses SVD-based sub-channel processing, attackers can bypass these physical-layer signal processing blocks through n-trigger attacks. This highlights the diversity and flexibility of the trigger vectors in propagating threats across dominant SemCom architectures.

TABLE V
ATTACK PERFORMANCE ON TYPICAL SYSTEMS. FOR BDJSCC AND JSCCOFDM, THE BACKDOOR TASKS ARE TRAINED AND TESTED AT SNR = 0 dB, WHILE THE MAIN TASK IS TRAINED AND TESTED AT SNR = 20 dB. FOR ADJSCC, THE BACKDOOR TASKS ARE TRAINED AND TESTED AT SNR = −2 dB, WHEREAS THE MAIN TASK IS TRAINED ACROSS SNR = [0, 20] dB AND TESTED AT SNR = 20 dB.

Model	BDJSCC	ADJSCC	JSCCOFDM	VIT-MIMO
PSNR_c(dB)	34.43	36.47	31.97	31.40
PSNR_m(dB)	34.53	36.45	31.85	30.16
PSNR_b(dB)	34.09	55.74	36.94	45.65
CA(%)	93.41	94.07	91.67	92.45
AEVC(%)	93.61	93.96	91.54	92.12
ASR(%)	100.0	100.0	100.0	100.0
AC(%)	9.940	9.960	9.560	9.230

D. Discussion on Backdoor Detection Via Noise Titration

We first discussed the applicability of existing state-of-the-art defense mechanisms against the proposed CT-BA. Interpretability-based methods, such as Grad-CAM [42], focus on visualizing specific discriminative regions within the input image to explain model decisions, thereby failing to identify triggers that exist solely as global statistical parameters in the channel layer. Similarly, perturbation-based defenses like STRIP [43] operate by superimposing external patterns onto input samples to disrupt trigger coherence, a strategy that proves ineffective against channel-domain triggers dependent on noise power or fading states rather than spatial pixel features. Defense mechanisms like Neural Cleanse [44] focus on reverse-engineering static pixel patterns in the input domain, thereby failing to address channel-domain triggers that rely on dynamic statistical parameters rather than fixed spatial features. Consequently, these input-centric strategies fail to capture the attack vectors embedded within the physical communication channel. Activation Clustering [45] detects backdoors by analyzing the neural network activations of training data to separate poisonous samples from legitimate ones based on their distinct cluster distributions. This method relies heavily on the assumption of a classification task where activations can be segmented by discrete class labels to identify multimodal anomalies. In the context of SemCom, which is fundamentally a reconstruction (regression) task with continuous outputs, such label-based segmentation is inapplicable. Fine-Pruning [46] mechanism aims to remove backdoors by pruning neurons that are dormant on clean inputs and then fine-tuning the model. However, in our channel-triggered framework, the backdoor neurons activated by specific channel conditions are typically part of the model's legitimate ability to handle channel fading and noise. Pruning these neurons may reduce the system's robustness to natural channel variations and disrupt the main communication tasks.

Inspired by noise response analysis in nonlinear dynamic systems [47], [48], we propose a noise titration scheme for backdoor detection. Specifically, we inject noise with varying variances into the decoder's input during inference to probe potential backdoor activation. As illustrated in Fig. 11, the

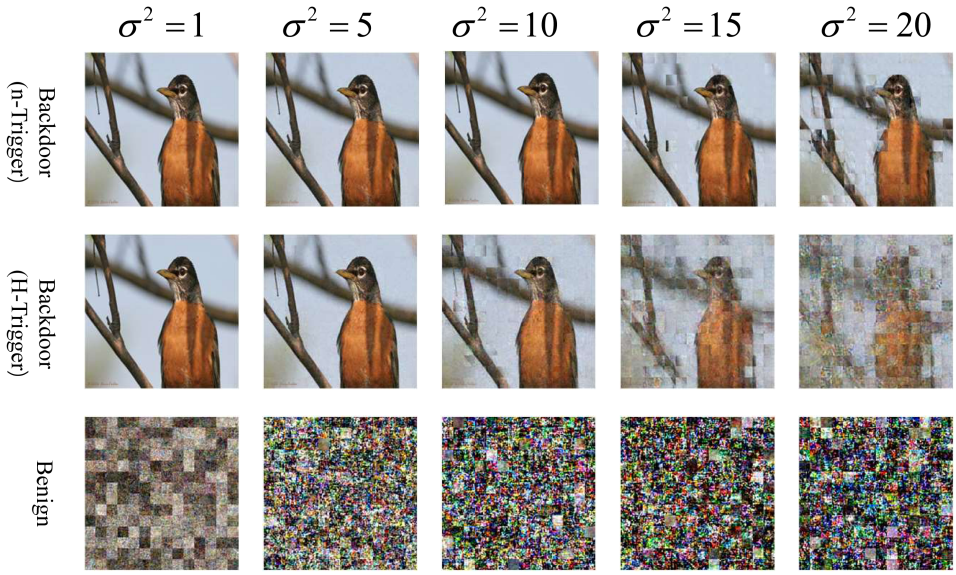

Fig. 11. Backdoor detection via noise titration. The backdoored model exhibits pronounced sensitivity to noise, when the variance exceeds a threshold, the decoder reconstructs the predefined target image with high fidelity. In contrast, the benign model fails to produce any structured reconstruction.

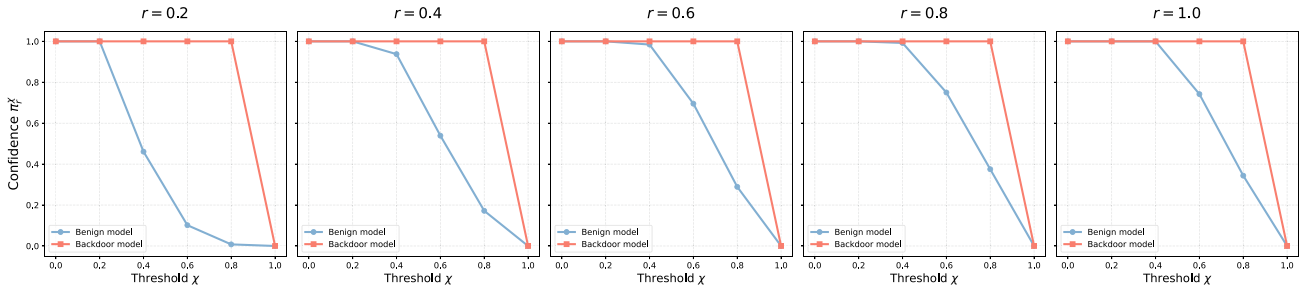

Fig. 12. Titration curves vs. threshold χ for the benign model and backdoor model with different noise intensity r.

responses of the benign and backdoored models are compared under controlled noise perturbations. The backdoored model exhibits pronounced sensitivity to noise; when the variance exceeds a threshold, the decoder reconstructs the predefined target image with high fidelity. In contrast, the benign model fails to produce any structured reconstruction.

To quantitatively distinguish these behaviors, we utilize a pre-trained classifier Ω (e.g., ResNet-18) to compute the high-confidence prediction ratio π_r^χ. Specifically, we inject i.i.d. normal-distributed noise $\delta \sim N(0,1)$ scaled by intensity r (i.e., $r\delta \sim N(0, r^2)$) into the decoder's input during inference to probe potential backdoor activation. To characterize the model's response, we record the confidence ratio π_r^χ of predictions, defined as:

$$\pi_r^\chi = \frac{\sum_{i=1}^{N^r} 1_{y>\chi}(\bar{y}_i^r)}{N^r} \tag{14}$$

where $\bar{y}_i^r = \max \sigma(\Omega(g_\phi(r\delta_i)))$ is the maximum output of the softmax activation $\sigma(\cdot)$, $1_{y>\chi}(\cdot)$ is the indicator function against a tunable threshold χ, i.e. if $\bar{y}_i^r > \chi$ then $1_{y>\chi}(\bar{y}_i^r) = 1$, otherwise $1_{y>\chi}(\bar{y}_i^r) = 0$, N^r is the number of noise samples. As illustrated in Fig. 12, we show the changing trend of the confidence of the two kinds of the model as the threshold χ changes when r takes different values. we see that as r increases, the threshold required for the confidence to decay to 0 becomes larger. In other words, by choosing an appropriate threshold χ, the confidence of the backdoor model will reach 1 faster than the benign model. We choose $\chi = 0.8$ as the final threshold.

VI. Limitations and Future Works

A. SemCom Systems Use Collaborative Training

Distributed training paradigms, particularly Federated Learning, have recently been practically deployed in SemCom systems to enhance data privacy and scalability. However, this collaborative framework introduces significant challenges for backdoor persistence; specifically, the mechanisms of global model aggregation and the iterative nature of multi-round training often lead to the dilution or forgetting of malicious triggers injected by single malicious. This poses a distinct robustness challenge for the proposed CT-BA, as the backdoor task might be forgotten during the average process.

B. Extension to Multi-User and Broadcast Scenarios

In scenarios with multi-user or broadcast topology, which have a complex structure, the primary challenge lies in the specificity of triggers; in a broadcast channel, a trigger intended for a specific victim may be inadvertently activated by other users experiencing similar channel statistics. A possible solution is to explore higher-dimensional triggers to distinguish between different users. Additionally, managing the trade-off between attack stealthiness and robustness in dynamic networks with multiple interference sources remains a critical hurdle. Potential solutions include adaptive trigger mechanisms that leverage machine learning to track specific channel state transitions, thereby maintaining high attack success rates even in non-stationary multi-user environments.

C. Advanced Defense Mechanisms Against Channel Triggers

Current defense strategies, including the noise titration method discussed in this paper, primarily serve as detection mechanisms rather than proactive mitigation techniques. While noise titration can reveal the presence of a backdoor by probing the decoder's sensitivity, it does not remove the backdoor or recover the benign model. Moreover, sophisticated adversaries may employ "adaptive attacks" by incorporating noise-robustness objectives into the backdoor loss function, potentially rendering simple statistical probing ineffective. Therefore, another pivotal direction for future work is the development of channel-aware unlearning algorithms. We aim to design defense protocols that can not only detect anomalous channel sensitivities but also sanitize the compromised model weights to neutralize the backdoor without degrading the semantic reconstruction quality of the main task.

VII. Conclusion

To address the significance of understanding the security risks posed to SemCom systems, this paper has introduced CT-BA, a novel backdoor attack paradigm targeting semantic symbols through channel-specific triggers. Unlike conventional input-triggered attacks, CT-BA has leveraged inherent wireless channel parameters as covert triggers, offering enhanced flexibility and feasibility in trigger design. Our evaluation across three benchmark datasets and JSCC systems has demonstrated near-perfect attack success rates while preserving model utility. This work has revealed a new class of backdoor threats in semantic communication, highlighting the need for robust defense mechanisms against channel-aware attacks. Looking ahead, this work necessitates a comprehensive investigation into the security of end-to-end communication systems supporting various modalities (e.g., text, speech, and video), with particular emphasis on exploring potential backdoor attack vectors. The findings will facilitate the development of robust countermeasures to enhance system security.

References

[1] X. Wang, J. Mei, S. Cui, C.-X. Wang, and X. Shen, "Realizing 6G: The operational goals, enabling technologies of future networks, and value-oriented intelligent multi-dimensional multiple access," *IEEE Netw.*, vol. 37, no. 1, pp. 10–17, Jan./Feb. 2023.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[3] L. Guo, W. Chen, Y. Sun, B. Ai, N. Pappas, and T. Quek, "Diffusion-driven semantic communication for generative models with bandwidth constraints," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 123–135, Jan. 2025.

[4] Z. Lyu, G. Zhu, J. Xu, B. Ai, and S. Cui, "Semantic communications for image recovery and classification via deep joint source and channel coding," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8388–8404, Aug. 2024.

[5] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks," *Proc. IEEE*, vol. 108, no. 3, pp. 402–433, Mar. 2020.

[6] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 5–22, Jan. 2024.

[7] K. Davaslioglu and Y. E. Sagduyu, "Trojan attacks on wireless signal classification with adversarial machine learning," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw.*, Nov. 2019, pp. 1–6.

[8] S. Islam, S. Badsha, I. Khalil, M. Atiquzzaman, and C. Konstantinou, "A triggerless backdoor attack and defense mechanism for intelligent task offloading in multi-UAV systems," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5719–5732, Apr. 2023.

[9] T. Zhao, X. Wang, J. Zhang, and S. Mao, "Explanation-guided backdoor attacks on model-agnostic RF fingerprinting," in *Proc. Conf. Comput. Commun.*, May 2024, pp. 221–230.

[10] Z. Zhang, R. Yang, X. Zhang, C. Li, Y. Huang, and L. Yang, "Backdoor federated learning-based mmWave beam selection," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6563–6578, Oct. 2022.

[11] Y. Zhou, R. Q. Hu, and Y. Qian, "Backdoor attacks and defenses on semantic-symbol reconstruction in semantic communications," in *Proc. IEEE Int. Conf. Commun.*, Denver, CO, USA, Jun. 2024, pp. 734–739.

[12] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: Input purification defense against trojan attacks on deep neural network systems," in *Proc. 36th Annu. Comput. Secur. Appl. Conf.*, Dec. 2020, pp. 897–912.

[13] R. Lu, P. Yu, Z. Huang, Z. Xia, and X. Jiang, "BAM: Backdoor defense based on adversarial mitigation," in*Pattern Recognit.*, vol. 172, 2025, Art. no. 112662.

[14] X. Li and Y. Xue, "A survey on server-side approaches to securing web applications," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 1–29, Mar. 2014.

[15] A. Dosovitskiy et al., "An image is worth 16 x 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, May 2021, pp. 1–21.

[16] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[17] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[18] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source channel coding for wireless image transmission with OFDM," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[19] H. Wu, Y. Shao, C. Bian, K. Mikolajczyk, and D. Gündüz, "Vision transformer for adaptive image transmission over MIMO channels," in *Proc. IEEE Int. Conf. Commun.*, May 2023, pp. 3702–3707.

[20] N. Islam and S. Shin, "Deep learning in physical layer: Review on data driven end-to-end communication systems and their enabling semantic applications," *IEEE Open J. Commun. Soc.*,vol. 5, pp. 4207–4240, 2024.

[21] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 2326–2330.

[22] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8081–8095, Dec. 2021.

[23] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 42–48, Nov. 2023.

[24] V. Kostina, Y. Polyanskiy, and S. Verdú, "Joint source-channel coding with feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3502–3515, Jun. 2017.

[25] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 178–193, May 2020.

[26] H. Wu, Y. Shao, E. Ozfatura, K. Mikolajczyk, and D. Gündüz, "Transformer-aided wireless image transmission with channel feedback," *IEEE Trans. Wireless Commun.*,vol. 23, no. 9, pp. 11904–11919, Sep. 2024.
[27] H. Wu, Y. Shao, K. Mikolajczyk, and D. Gündüz, "Channel-adaptive wireless image transmission with OFDM," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2400–2404, Nov. 2022.
[28] M. Wang, J. Li, M. Ma, and X. Fan, "Constellation design for deep joint source-channel coding," *IEEE Signal Process. Lett.*, vol. 29, pp. 1442–1446, Jun. 2022.
[29] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gündüz, "DeepJSCC-Q: Constellation constrained deep joint source-channel coding," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 4, pp. 720–731, Dec. 2022.
[30] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.
[31] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526*.
[32] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 1505–1521.
[33] T. A. Nguyen and A. T. Tran, "Wanet–imperceptible warping-based backdoor attack," in *Proc. Int. Conf. Learn. Represent.*, May 2021, pp. 1–15.
[34] X. Xu et al., "CSBA: Covert semantic backdoor attack against intelligent connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 17923–17928, Nov. 2024.
[35] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol 34, Apr. 2020, pp. 11957–11965.
[36] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, vol. 33, pp. 3454–3464.
[37] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," 2019, *arXiv:1912.02771*.
[38] O. Mengara, A. Avila, and T. H. Falk, "Backdoor attacks to deep neural networks: A survey of the literature, challenges, and future research directions," *IEEE Access*, vol. 12, pp. 29004–29023, Jan. 2024.
[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[40] A. Krizhevsky and Hinton, "Learning multiple layers of features from tiny images," Master's thesis, Univ. Toronto, Toronto, ON, Canada, 2009.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 248–255.
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[43] 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. 35th Annu. Comput. Secur. Appl. Conf.*, 2019, pp. 113–125.
[44] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy*, May 2019, pp. 707–723.
[45] B. Chen et al., "Detecting backdoor attacks on deep neural networks by activation clustering," 2018, *arXiv:1811.03728*.
[46] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Res. Attacks, Intrusions, Defenses*, 2018, pp. 273–294.
[47] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D*, vol. 65, no. 1–2, pp. 117–134, May 1993.
[48] C.-S. Poon and M. Barahona, "Titration of chaos with added noise," *Proc. Nat. Acad. Sci.*, vol. 98, no. 13, pp. 7107–7112, Jun. 2001.

Jialin Wan (Member, IEEE) is currently working toward the MS degree with Xidian University, Xi'an, China. His research interests include intelligent networking and AI security, particularly in the investigation of neural backdoors.

Jinglong Shen (Member, IEEE) received BS degree in information engineering from Xidian University, Xi'an, China in 2021, where he is currently working toward the PhD degree in information and communication engineering with the School of Telecommunications Engineering. From 2025 to 2026, he was a visiting student with the Singapore University of Technology and Design (SUTD), Singapore. His research interests include federated learning, mobile edge computing, and large language models. He was the recipient of the ICCC 2023 Best Demo Award. He has served as a Technical Program Committee (TPC) member of IEEE GlobeCom 2024-2025 and IEEE ICCC 2024-2025.

Nan Cheng (Senior Member, IEEE) received the BE and MS degrees from the Department of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2009 and 2012, respectively, and the PhD degree from the Department of Electrical and Computer Engineering, University of Waterloo in 2016. He was a postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Toronto, from 2017 to 2019. He is currently a professor with the School of Electrical Engineering & Intelligentization, Dongguan University of Technology, State Key Laboratory of ISN, and School of Telecommunications Engineering, Xidian University. He has authored or coauthored more than 90 journal papers in IEEE Transactions and other top journals. His research interests include B5G/6G, AI-driven future networks, and space-air-ground integrated network. He serves as an associate editors for *IEEE Internet of Things Journal*, *IEEE Transactions on Vehicular Technology*, *IEEE Open Journal of Vehicular Technology*, and *Peer-to-Peer Networking and Applications*, and is/was the guest editors of several journals.

Zhisheng Yin (Member, IEEE) received the BE degree from the Wuhan Institute of Technology, Wuhan, China, the BBA degree from the Zhongnan University of Economics and Law, Wuhan, in 2012, the MSc degree from the Civil Aviation University of China, Tianjin, China, in 2016, and the PhD degree from the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2020. From 2018 to 2019, he visited BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently an assistant professor with the School of Cyber Engineering, Xidian University, Xi'an, China. His research interests include space-air-ground integrated networks, wireless communications, cybertwin, and physical layer security.

Yiliang Liu (Member, IEEE) received the BE and MSc degrees in computer science and communication engineering from Jiangsu University, Zhenjiang, China, in 2012 and 2015, respectively, and the PhD degree from the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2020. He is currently an associate professor with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China. His research interests include wireless communication security, physical-layer security, and intelligent connected vehicles. He was the recipient of the Outstanding Doctoral Dissertation Award from China Education Society of Electronics in 2020 and Best Paper Award from the IEEE Systems Journal in 2021, IEEE VTC-Fall in 2023, IEEE ICITE in 2024, and IEEE AIoT in 2024.

Wenchao Xu (Member, IEEE) received the BE and ME degrees in telecommunications engineering from Zhejiang University, Hangzhou, China, in 2008 and 2011, respectively, and the PhD degree in electrical and computing engineering from the University of Waterloo, Canada, in 2018. In 2011, he joined Alcatel Lucent Shanghai Bell Company Ltd., where he was a software engineer of telecom virtualization. He has also been an assistant professor with the School of Computing and Information Sciences, Caritas Institute of Higher Education, Hong Kong. He is currently a research assistant professor with The Hong Kong Polytechnic University. His research interests includes wireless communication, Internet of Things, distributed computing, and AI enabled networking.

Xuemin (Sherman) Shen (Fellow, IEEE) received the PhD degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a university professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, wireless network security, Internet of Things, AI for networks, and vehicular networks. He is a registered professional engineer of Ontario, Canada, Engineering Institute of Canada fellow, Canadian Academy of Engineering fellow, Royal Society of Canada fellow, Chinese Academy of Engineering foreign member, International fellow of the Engineering Academy of Japan, and distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society. Dr. Shen was the recipient of "West Lake Friendship Award" from Zhejiang Province in 2023, President's Excellence in Research from the University of Waterloo in 2022, Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He is the past president of the IEEE Communications Society. He was the vice president of Technical & Educational Activities, vice president of Publications, member-at-large on the Board of Governors, chair of the Distinguished Lecturer Selection Committee, and member of IEEE Fellow Selection Committee of the ComSoc. He served as the editor-in-chief of *IEEE Internet of Things Journal*, *IEEE Network*, and *IET Communications*.