Physical Layer Security for Semantic Communications: Challenges and GenAI-Based Strategies

Long V. Nguyen *Student Member, IEEE,* Yinqiu Liu *Student Member, IEEE,* Quang Nhat Le *Member, IEEE,* and Trung Q. Duong *Fellow, IEEE*

***Abstract*—Semantic communication (SemCom) has emerged as a paradigm shift to revolutionize next-generation wireless networks. By transmitting only task-relevant semantic features, SemCom promises unprecedented gains in efficiency and reliability. However, this transition introduces a novel and complex security landscape, where threats extend beyond bit interception to the inference of user intent and semantic interpretation. Physical layer security (PLS) is considered an important technique for defense against these new attacks due to its foundational properties. This paper provides an overview of PLS in SemCom systems. We first explore the background of SemCom and PLS. We then delve into the current security vulnerabilities inherent to SemCom, including two main threats: signal-level threats and semantic-level threats. Subsequently, we point out potential generative artificial intelligence (GenAI)-driven strategies to counter these security issues. Through a case study, we propose a robust framework that utilizes a large language model and deep reinforcement learning to optimize PLS in SemCom. Finally, open research challenges and future directions are presented to facilitate future research development in this rapidly evolving field.**

***Index Terms*—Semantic communications, physical layer security, large language models, deep reinforcement learning.**

I. Introduction

For many decades, the design of communication systems has been dominated by the Shannon theorem, which addresses the technical problem of how accurately symbols can be transmitted regardless of their meaning. This approach, categorized as "Level A" communication by [1], has been remarkably successful, enabling the digital revolution by focusing on maximizing data rates and minimizing bit error probabilities. However, the advent of the fifth-generation (5G) networks and the proliferation of intelligent applications, such as autonomous driving, the Internet of Things (IoT), and the Metaverse, demand a fundamental rethinking of this principle. These applications are not concerned with the perfect reconstruction of bitstreams but with the successful execution of underlying tasks, which depend on the meaning conveyed by the data. This has facilitated a transition toward semantic communication (SemCom), a new paradigm that focuses on the "semantic problem" (Level B) and the "effectiveness problem" (Level C): *how precisely the transmitted symbols convey the desired meaning and how effectively that meaning influences conduct in the desired way*. SemCom aims to transmit only the information critical to a specific task at the receiver, filtering out redundant or irrelevant data at the source. For instance, in an image recognition task, a SemCom system would transmit only the features necessary to identify an object (e.g., a pedestrian) while omitting background details, thereby achieving significant reductions in data traffic, power consumption, and bandwidth usage.

The overview architecture of a SemCom system is illustrated in Fig. 1. As observed, a typical end-to-end semantic communication system is built using artificial intelligence (AI)-driven components that jointly handle source and channel coding, fundamentally altering the data transmission process. At the transmitter, the process begins not with bit-level source coding but with a semantic encoder. An AI model, typically a deep neural network (DNN), guided by a shared knowledge base (KB), analyzes the source data (e.g., text, image, and speech) and extracts only the information relevant to the communication task. The semantic encoder maps source data into a low-dimensional latent representation, typically a feature vector, which is transformed by a channel encoder for robust transmission. At the receiver, the channel decoder recovers the semantic stream from the noisy signal, allowing the semantic decoder to reconstruct the original meaning using a shared KB.

While the efficiency gains of SemCom are profound, its operational reliance on AI models for semantic processing and KB for semantic interpretation significantly transforms the security landscape. Extracting and transmitting meaningful features rather than unstructured bits creates novel vulnerabilities. This establishes a new and more sophisticated security challenge: protecting the meaning itself over bits in conventional communications, which remains underexplored. To address this challenge, physical layer security (PLS) has played an important role. Security at the physical layer is more critical than at other layers since it is a foundation for all subsequent security layers. A significant attack at this first line of defense can cause a catastrophic security failure of the entire communication system [2]. PLS provides an information-theoretic approach to confidentiality and availability of trans-

L. V. Nguyen, Q. N. Le, and T. Q. Duong are with the Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 5S7, Canada (e-mail: {lnguyen, qnle, tduong}@mun.ca).

Y. Liu is with the College of Computing and Data Science, Nanyang Technological University, Singapore (e-mail: yinqiu001@e.ntu.edu.sg).

This work was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109 and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program RGPIN-2025-04941.

Corresponding author is Trung Q. Duong

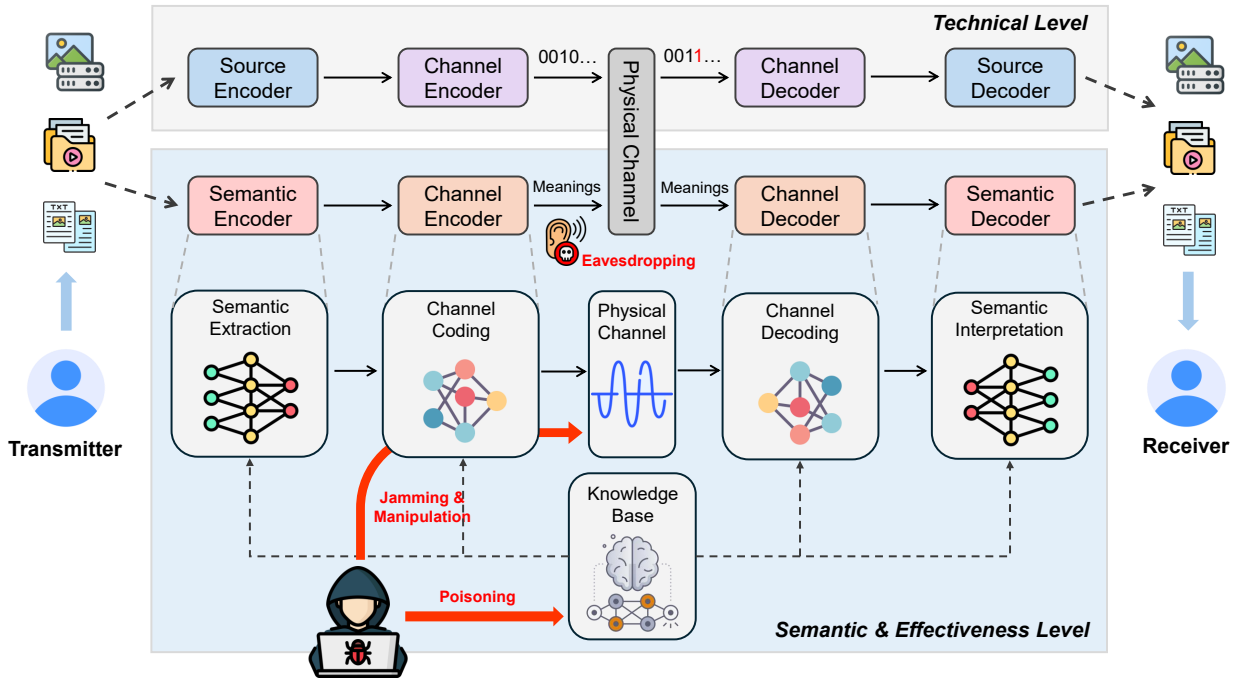

Fig. 1: Overview architecture of SemCom systems compared to conventional systems. SemCom focuses on the semantic and effectiveness levels, rather than the technical level of efficient communication, prioritizing the transmission of critical semantic features for specific tasks. This introduces new security threats at the physical layer.

missions by exploiting the intrinsic physical properties of the wireless channel, such as its randomness, noise, interference, and fading characteristics. Unlike the cryptography method at upper layers, which relies on computational resources, PLS can offer security guarantees independent of an adversary's computational capabilities, making it a robust defense against future technological advancements. The open nature of the wireless medium, traditionally viewed as a security liability, is repurposed in PLS as a strategic advantage, allowing for the design of systems that favor the legitimate receiver over an eavesdropper. In the context of SemCom, the physical signal is no longer a neutral carrier of bits but is intrinsically linked to a high-level semantic representation. This results in physical layer attacks that are not only signal-level attacks but also semantic-level attacks on semantic understanding and interpretation. To this end, generative AI (GenAI) emerges as a particularly promising solution due to its capability of learning complex, high-dimensional data distributions. This empowers GenAI not only with profound semantic understanding but also with robust data generation. Therefore, GenAI can enable proactive, context-aware countermeasures that ensure semantic fidelity and efficient task execution in the face of new attacks.

This paper provides a comprehensive and structured overview of PLS in SemCom. The main contributions are summarized as follows:

- We present a taxonomy of extensive attacks at the physical layer based on the adversary's primary intent: to disrupt signal reception versus to corrupt the semantic interpretation.
- We propose novel GenAI-based strategies accordingly to counter these attacks, thereby enhancing the robustness and security of SemCom systems.
- Through the case study, we demonstrate how GenAI-enhanced deep reinforcement learning (DRL) can optimize the PLS in SemCom systems with superior performance.
- We outline the key open issues and potential research directions for future investigation.

II. The Semantic Threat Landscape: New Attacks at The Physical Layer

The architectural transition from bit-level to meaning-level communication systems fundamentally reshapes the physical layer threat landscape. While some threats are inherited from traditional wireless systems, their impact and the attacker's objectives are redefined. More importantly, this new paradigm gives rise to a novel class of attacks that target the AI-driven meaning interpretation process. In this section, we present a comprehensive taxonomy of these threats, as visualized in Fig. 2, structured around the adversary's primary intent: to disrupt the reception of the signal versus to corrupt the interpretation of its meaning.

A. Signal-Level Threats: Attacks on Availability and Confidentiality

These threats primarily target the physical signal itself, with the goal of either preventing its successful reception (attacking availability) or intercepting its content (attacking confidentiality). Although the attack mechanisms may be familiar from traditional wireless security, their implications are magnified in a semantic context where the intercepted signal carries high-level, meaningful information.

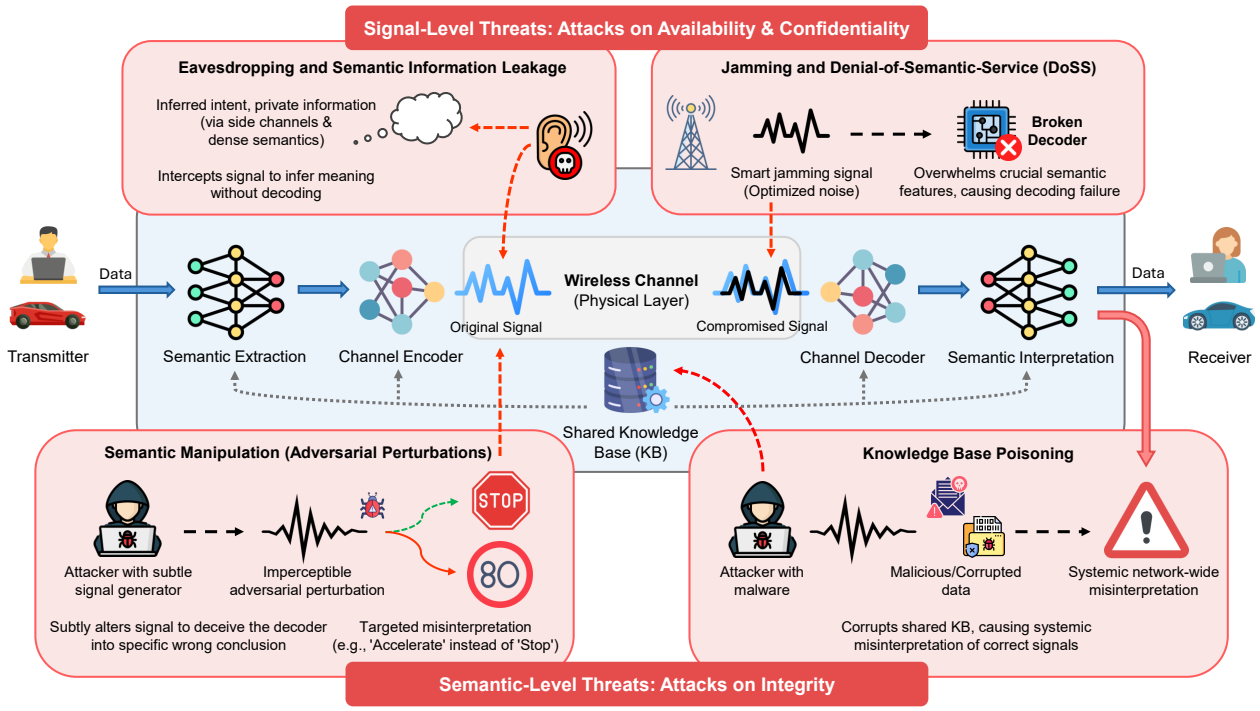

Fig. 2: Illustration of security threats at the physical layer in semantic communication. Besides the attacks that target the physical layer directly, knowledge base poisoning targets the foundation of shared understanding, indirectly causing real-time consequences at the physical layer.

1) Eavesdropping and Semantic Information Leakage: The open nature of the wireless medium makes it inherently susceptible to eavesdropping, where an adversary intercepts the transmitted signal without the knowledge of the legitimate parties. In traditional systems, an eavesdropper intercepts a stream of encrypted bits. In SemCom, the eavesdropper captures a stream of semantic features. The semantic stream is highly compressed and information-dense; every intercepted symbol can carry significant value. Even without the correct decoder or knowledge base, this stream can leak valuable semantic information. The transmission's duration, power level, or even the waveform's complexity might vary depending on the semantic content being conveyed. An adversary can monitor these physical-layer side channels over time to infer user behavior, intent, or private contextual information without ever decoding a single message. For example, an attacker monitoring the physical layer emissions from a smart home might observe that short, low-power semantic transmissions consistently precede the lights turning on, while longer, more complex transmissions precede the home security system being activated. From this, the attacker can infer the resident's daily schedule and presence, constituting a significant privacy breach. Such breaches can facilitate spoofing or tampering attacks, seriously compromising system confidentiality [3].

2) Jamming and Denial-of-Semantic-Service: Jamming is an active attack that aims at disrupting communication by overwhelming the legitimate signal with interference, thereby denying service to the legitimate receiver. In SemCom, this attack represents a more sophisticated evolution of traditional jamming. Instead of simply flooding the channel with high-power, broadband noise to significantly reduce the signal-to-noise ratio (SNR), a semantic jammer targets the availability of meaning. If the attacker has some partial knowledge of the semantic model, i.e., knows which features the model deems most important for a given task, they can design a jamming signal that specifically targets and corrupts those crucial features in the semantic space. For example, in a video SemCom system designed for surveillance, which might prioritize the transmission of faces, a semantic jammer could transmit noise patterns that are optimized to corrupt the regions of the semantic feature space corresponding to facial features, rendering the transmission useless for its primary task while using significantly less power than a traditional barrage jammer. Such ways can cause "Denial-of-Semantic-Service," in which the DNN-based decoder, unable to find any recognizable patterns or features in the corrupted signal, fails to reconstruct any coherent meaning during communication.

B. Semantic-Level Threats: Attacks on Integrity

This class of threats represents the most significant transition from the traditional security paradigm. These are novel attacks, enabled by the AI-centric nature of SemCom, that target the integrity of the communication by manipulating the interpretation of meaning. The adversary's goal is not to block the signal but to subtly alter it so that the receiver's AI model is deceived into "understanding" something false.

1) Semantic Manipulation (Adversarial Perturbations): Semantic manipulation is arguably the most dangerous threat, which is not a brute-force jamming attack but a subtle and intelligent assault on the integrity of meaning [4]. It is an application of adversarial machine learning concepts to the wireless physical layer. Instead of blasting the channel with

high-power noise, the attacker transmits a low-power, carefully crafted perturbation signal over the wireless channel. This structured signal is designed to be a "physical adversarial example." When this perturbation is added to the legitimate signal in the wireless channel, it creates a combined signal that is only minimally different from the original from an energy or statistical perspective but is specifically designed to exploit vulnerabilities in the neural network architecture of the legitimate semantic decoder. The objective is to cause a targeted and significant misinterpretation at the semantic decoder. The attacker aims to introduce a large error in the semantic domain while causing only a small perturbation in the signal domain. For example, in an autonomous vehicular network, an attacker could use semantic manipulation to subtly alter the signal corresponding to a "maintain speed" command, causing the receiving vehicle's decoder to interpret it as a "accelerate sharply" command, with potentially catastrophic consequences. Another example would be fooling an image classification system into identifying a "dog" as a "cat" by adding an imperceptible, structured noise pattern to the transmitted signal.

2) Knowledge Base Poisoning: This attack has profound real-time consequences at the physical layer. It targets the very foundation of shared understanding in a SemCom network: KB. An adversary manages to introduce malicious or corrupted data into the shared KB that is used by both the semantic encoder and decoder. In a distributed system such as a network of autonomous vehicles, a compromised vehicle could upload falsified sensor data or incorrect labels to a shared, collaboratively built KB. For example, it could upload data that incorrectly associates the semantic features of a "stop sign" with the label "speed limit 80". Although poisoning the KB may occur through a non-physical layer channel (e.g., a malicious software update or data upload), the devastating impact of the attack manifests during real-time physical layer communication. A legitimate transmitter will send a perfectly formed signal representing a stop sign. The receiver's antenna will capture this signal perfectly. However, when the semantic decoder processes the extracted features, it will refer to the poisoned KB and arrive at the wrong conclusion that it has received a "speed limit 80" sign. These severe effects originate from the model training phase, when the models learn malicious patterns from the compromised data distribution. This attack is particularly dangerous because it is systemic. A single poisoned entry in a shared KB can trigger every single node in the network to systematically misinterpret correctly transmitted and received physical signals, leading to a network-wide, cascading failure of trust and coordination.

III. A New Strategy: Generative AI for Semantic-Aware PLS

With the dynamic characteristics of wireless channels and the current sophisticated threats at the physical layer as mentioned above, existing methods often struggle to handle these diverse attacks. GenAI has powerful capabilities to model complex data distribution, providing numerous opportunities for defending against the emerging threats inherent to semantic communications [2]. In this section, we propose some novel GenAI-based strategies designed to counter the primary physical layer attacks in SemCom. These defense approaches are summarized in Table I.

A. Strategies Against Signal-Level Threats

These strategies focus on protecting the confidentiality and availability of the transmitted semantic stream against eavesdropping and jamming.

1) Artificial Noise for Semantic Obfuscation: The goal of this strategy is to make the intercepted semantic stream unintelligible to an eavesdropper while ensuring that the legitimate receiver can recover the meaning. Diffusion models (DMs), which excel at generating high-quality data by reversing a gradual noising process, are exceptionally well-suited for this task. The transmitter incorporates artificial noise (AN) into the forward process of a diffusion model, effectively mapping the combination of AN and natural channel noise to a known diffusion trajectory. The legitimate receiver, possessing the parameters of the diffusion model, can execute the reverse denoising process to perfectly reconstruct the intended semantic information. However, the received signal at the eavesdropper's side appears as an indistinguishable mix of semantic features, channel noise, and intentionally injected AN. Without knowledge of the reverse diffusion process, reconstructing the original meaning is computationally intractable [5]. Generative adversarial networks (GANs) can also be trained to generate highly complex and unpredictable AN signals that are optimized to disrupt an eavesdropper's decoder while having a minimal impact on the legitimate receiver's reconstruction quality.

2) GenAI for Anti-Jamming Frameworks: GenAI can be utilized to build resilient systems that can maintain communication even in the presence of powerful jamming signals. For instance, the adversarial nature of GANs provides a suitable framework for modeling the conflict between a jammer and a receiver. Through adversarial training based on game theory, the receiver learns to distinguish the statistical properties of the legitimate semantic stream from the jamming interference [6]. This allows the receiver to effectively filter the jamming signal and recover the intended meaning from the contaminated data stream rather than simply evading the jammer by changing frequency. Additionally, the inherent robustness of DMs to noise can be leveraged to counter jamming attacks. The jamming signal is treated as a form of extreme, malicious noise. The receiver employs the reverse denoising process of a diffusion model to "purify" the received signal, effectively removing the jammer's influence and recovering the semantic content [5].

B. Strategies Against Semantic-Level Threats

These strategies are designed to protect the integrity of the semantic interpretation process itself, defending against attacks that aim to deceive the receiver's AI model.

TABLE I: Summary of GenAI-Based Defense Strategies Against PLS Threats in SemCom

Category & Focus	Targeted Threat	Defense Strategy	GenAI-Based Solution
A. Signal-Level *(Confidentiality & Availability)*	**Eavesdropping**	Artificial Noise (AN) & Semantic Obfuscation	• **DMs:** Inject AN via a forward process, reconstructible only using legitimate reverse parameters [5]. • **GANs:** Generate complex, unpredictable AN optimized to disrupt unauthorized decoders.
	Jamming	Resilient Anti-Jamming	• **GANs:** Use adversarial training (based on game theory) to filter jamming statistical properties [6]. • **DMs:** Treat jamming as extreme noise and "purify" the signal via the reverse denoising process [5].
B. Semantic-Level *(Integrity of Meaning)*	**Adversarial Perturbations**	Adversarial Purification & Validation	• **DMs:** Destroy attack structure via forward process noise, then reconstruct clean signal [7]. • **LLMs:** Deploy as input guardrails to block attacks or as agents to verify output consistency [8].
	KB Poisoning	Knowledge Base (KB) Integrity Check	• **VAEs/GANs:** KB updates with high reconstruction errors as out-of-distribution/poisoned samples [2]. • **LLMs:** Use as a "semantic shield" to detect syntactically correct but semantically anomalous data [9].

DMs: Diffusion models; GANs: Generative adversarial networks; VAEs: Variational auto-encoders; LLMs: Large language model.

1) GenAI Models for Adversarial Resilience: DMs are promising candidates for adversarial purification, which can be deployed at the receiver to cleanse incoming signals. An incoming signal, potentially corrupted by a low-power adversarial perturbation, is first passed through the forward process of a DM, which adds a degree of random noise. This step effectively destroys the delicate, carefully optimized structure of the adversarial attack. The subsequent reverse (denoising) process then reconstructs a "purified" signal that is statistically closer to the distribution of legitimate semantic signals [7]. Another potential application is leveraging large language models (LLMs) to filter malicious inputs and to validate final outputs. For example, LLM-powered guardrails can detect and block malicious perturbations before they reach the core semantic decoder. By ensuring the inputs conform to expected patterns with minimal adversarial inputs, the guardrail maintains the integrity of the communication [8]. If any semantic errors remain that might bypass input filters, a second agent can be used to verify the consistency of the interpretation.

2) GenAI Strategies for Knowledge Base Security: Since shared KB is the foundation for semantic understanding, protecting its integrity is paramount. A generative model, such as a variational auto-encoder (VAE) or GAN, can be trained on the distribution of legitimate KB updates or data entries. If the model can reconstruct the new update with low error, it is considered "in-distribution" and likely eligible. However, if the update is a poisoned sample, it will lie outside the learned data distribution, resulting in a high reconstruction error. This high error serves as a reliable flag for potentially malicious data, which can then be rejected or separated [2]. LLMs are uniquely suited for this task due to their deep understanding of semantic relationships. When new data is submitted, an LLM assesses its semantic consistency with the existing data in the trusted KB. It can also detect outliers and poisoned samples that are semantically anomalous while syntactically correct. An LLM can be considered "semantic shield" with external knowledge to prevent the model from learning false correlations introduced by poisoned data [9].

IV. Case Study: LLM-Enhanced DRL for Secure Semantic Communications

In this section, we present a case study to demonstrate the transformative role of LLMs in securing SemCom systems.

A. System Model and Problem Formulation

As illustrated in Fig. 3, we consider a SemCom security scenario comprising a transmitter (Alice), a legitimate receiver (Bob), and a cognitive eavesdropper (Eve). The objective is to establish a secure link where Alice can transmit task-relevant semantic features that are easily interpretable by Bob but remain strategically obscured to Eve.

1) Channel and Semantic Modeling: To capture the intricate physical constraints of dynamic environments, we adopt a composite channel model that decouples macro- and micro-environmental impacts. Specifically, we integrate distance-dependent path loss, obstacle-induced shadowing, and temporally correlated small-scale fading [10]. As shown in Fig. 3, the time-varying complex channel coefficient $h_u(t)$ for node u (Bob or Eve) is expressed as $h_u(t) = G_u(d_u(t), \chi_u(t)) \cdot g_u(t)$, where $G_u(\cdot,\cdot)$ captures large-scale fading effects based on distance $d_u(t)$ and shadowing $\chi_u(t)$, while $g_u(t)$ describes the rapid fluctuations caused by multipath propagation and Doppler shifts [10]. This ensures that the instantaneous SNR $\gamma_u(t)$ at moment t accurately reflects the non-stationary nature of the wireless medium.

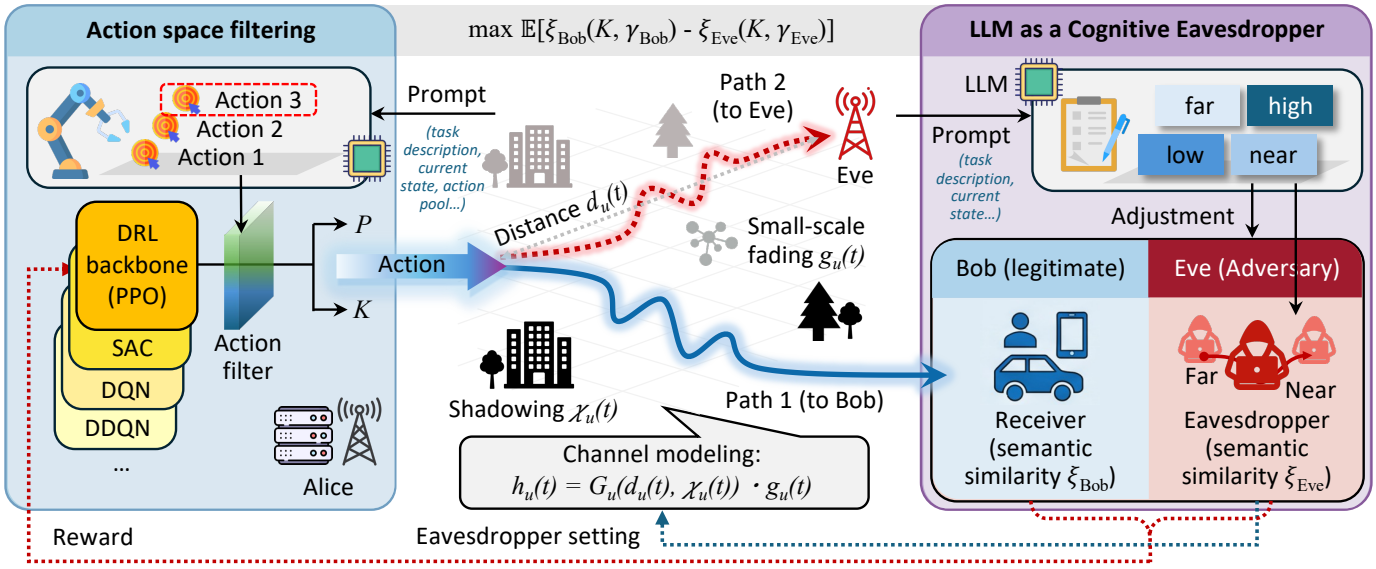

Fig. 3: Illustration of the case study and the proposed method. The security objective is to maximize the expected semantic similarity gap between Bob and Eve. We utilize LLM as a cognitive eavesdropper, which simulates intelligent eavesdropping strategies to facilitate DRL training and reduce the action space.

In terms of semantic representation, we suppose that Alice utilizes a semantic encoder to extract low-dimensional latent vectors that encapsulate the core meaning of the source data. This process leverages joint source-channel coding (JSCC) [11] to map semantic features directly to channel symbols, ensuring robust transmission in noisy wireless environments. For each node u, the proportion of correct interpretation (termed semantic similarity) given the number of semantic symbols per word K and the received SNR γ_u can be expressed as $\xi_u(K, \gamma_u) = A_{K,1} + \frac{A_{K,2} - A_{K,1}}{1+\exp(-(C_{K,1} \log \gamma_u + C_{K,2}))}$, where $A_{K,1}, A_{K,2}, C_{K,1}$ and $C_{K,2}$ are task-specific parameters that can be obtained by employing a data regression method [11]. A high SNR does not necessarily guarantee high semantic similarity if the symbol density K is insufficient for the task complexity. Therefore, these parameters characterize how SNR translates into "correct interpretation" for a specific task.

2) Problem Formulation: The primary security objective is to maximize the expected semantic similarity gap, which measures the disparity in information interpretation between Bob and Eve (see Fig. 3). Specifically, Alice seeks to determine an adaptive transmission policy that selects the optimal transmit power P and the most effective semantic symbol density K. This optimization is performed under the practical constraints of a finite transmit power budget and a discrete set of available symbol densities.

B. Algorithm Design: LLM as Cognitive Eavesdropper

To address the complexities of securing semantic transmissions against intelligent threats, we model the interaction between Alice and Eve as a Markov decision process (MDP). This provides a structured framework for Alice to learn robust policies by interacting with a non-stationary environment. The key components of our MDP are defined as follows:

- **State:** The state space comprises summarized environmental indicators, including the recent trajectory of received SNRs, Alice's previous transmission parameters, and the eavesdropper settings.
- **Action:** Alice's action involves the joint selection of P and K.
- **Reward:** The reward is the instantaneous semantic similarity gap defined above.

Traditionally, DRL can be employed to solve such optimization problems by treating Eve as a static or randomly moving eavesdropper. However, conventional DRL methods often struggle with over-specialization to specific distributions and do not account for cognitive adversaries that can strategically adapt their eavesdropping strategies [12]. This motivates the use of an LLM as a cognitive eavesdropper. Particularly, LLMs exhibit the outstanding capability to map high-level reasoning to physically interpretable settings, such as proximity and gaining control, thus simulating a skilled eavesdropper and facilitating efficient and smooth training. We suppose that Alice employs the proximal policy optimization (PPO) algorithm as the DRL backbone, given its exceptional convergence performance in non-stationary wireless channel environments[1]. The LLM is integrated as follows:

1) Curriculum Learning: As a cognitive eavesdropper, LLM modulates the difficulty of the eavesdropping environment to prevent the DRL agent from being overwhelmed by a strong adversary or becoming complacent against a weak one. To do so, it processes text-based prompts that summarize the problem formulation, current environmental state, and training statistics. Based on its reasoning capability, the LLM then determines the agent's current proficiency and constructs the most appropriate environment by adjusting Eve's settings, including her proximity (e.g., far, moderate, or near) and front-end gain levels (e.g., low, default, or high).

These adjustments directly alter the physical state perceived

[1]Our proposal is agnostic to the specific DRL architecture and can be integrated with other algorithms such as soft actor-critic (SAC) or deep Q-networks (DQN).

by the DRL agent by manipulating the large-scale gain $G_u(\cdot,\cdot)$ and the effective noise power. For instance, by positioning Eve in a "far" zone with "low" gain, the LLM simulates a favorable environment where Alice can protect semantic information easily. This reflects the core principle of curriculum learning: initially, it presents Alice with simplified scenarios to facilitate initial convergence. As training progresses, the LLM intelligently increases Eve's eavesdropping capability, i.e., moving closer to Alice or employing powerful devices. This guided progression allows the DRL agent to learn the security policy smoothly and efficiently. In contrast, in traditional DRL paradigms, the agent may learn nothing from constant failure against a strong attacker or achieve poor performance when trained in an insufficiently challenging environment.

2) Action Space Filtering: In our problem, the combination of multiple power levels and symbol densities often results in a vast action pool, many of which are physically impractical or lead to excessive semantic leakage. To address this, in addition to simulating intelligent eavesdropping, the LLM also prunes the action space based on its understanding of the semantic communication landscape to further facilitate training.

As shown in Fig. 3, the transmitter provides the LLM with a detailed prompt containing the current positions of Alice, Bob, and Eve, as well as the full set of candidate transmit powers and symbol densities. A prompt example is illustrated as shown in Fig. 4. The LLM leverages its domain knowledge to reason that certain configurations, such as high transmit power when Eve is in proximity, are likely to cause a substantial drop in the semantic gap. Consequently, the LLM filters the original action set and generates a refined pool of high-value candidate actions. PPO then selects its action only from this reduced pool. By narrowing the search space from dozens of combinations to a targeted subset of superior actions, this mode significantly accelerates the convergence process and prevents the agent from exploring inherently insecure or inefficient transmission strategies[2].

It is worth emphasizing that LLM is not used for real-time physical-layer signal processing. Instead, it serves as a high-level reasoning module that constructs cognitive Eve scenarios and refines the candidate action space. These LLM-assisted operations can be performed offline during training or periodically at a much slower timescale than physical-layer transmission. Once the PPO policy is trained, the online transmitter only executes the learned DRL policy to select the transmit power and semantic symbol density. Therefore, the real-time decision process does not require LLM inference, without extra computational overhead brought by LLMs.

**** [GOAL] ****
- Maximize legitimate receiver semantic similarity ξ_B
- Minimize eavesdropper semantic similarity ξ_E

**** [INPUTS & ASSUMPTIONS] ****
- Alice position: $p_A = (x_A, y_A)$
- Bob position: $p_B = (x_B, y_B)$, distance $d_B = \|p_A - p_B\|$
- Eve position: $p_E = (x_E, y_E)$, distance $d_E = \|p_A - p_E\|$
- Transmit power candidate list: $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$
- Compression granularity candidate list: $\mathcal{K} = \{k_1, k_2, \ldots, k_N\}$
- Optimization objective: Maximize $J = \xi_B - \lambda_e \xi_E$

**** [CONSTRAINTS] ****
- Bob semantic similarity: $\xi_B \geq \xi_{\min}$
- Choose values only from $\mathcal{P}$ and $\mathcal{K}$

**** [REQUEST] ****
- Recommend approximately N_{suggest} combinations (P_{dB}, k) from the candidate lists as relatively optimal actions

**** [OUTPUT FORMAT] ****
- Strictly one JSON object with field "actions"
- Each element: {P_dB: value, k: value}
- Only output JSON between <BEGIN_JSON> and <END_JSON>

**** [OUTPUT EXAMPLE] ****
<BEGIN_JSON>
{actions: [{P_dB: 8, k: 16}, {P_dB: 10, k: 32}]}
<END_JSON>

Fig. 4: An example of prompt structure for LLM's reasoning.

C. Numerical Results

We fit the values of $A_{K,1}, A_{K,2}, C_{K,1}$, and $C_{K,2}$ using the Europarl parallel corpus dataset[3]. Specifically, we train multiple DeepJSCC models (K = 4, 8, 16, 32, and 64). Then, we use the Jaccard similarity of Bag-of-Words to measure semantic similarity, and statistically fit a large number of K, γ samples to obtain the parameters $A_{K,1}, A_{K,2}, C_{K,1}, C_{K,2}$ of $\xi(K,\gamma)$[4]. For Eve, we adopt a conservative strong-eavesdropper assumption. Specifically, Eve's semantic similarity is evaluated using the same fitted semantic-SNR function as Bob, but with Eve's own received SNR γ_{Eve}. This models Eve as having an equivalent surrogate semantic interpretation capability and provides an upper bound on Eve's decoding performance. In practical scenarios, if Eve does not have access to Bob's exact semantic decoder or knowledge base, her actual semantic similarity would be lower than this estimated value. Therefore, this assumption avoids overestimating the secrecy gain of the proposed method. We implement the proposed LLM-enhanced DRL using Qwen 2.5 7b[5].

We utilize six methods for comparison. The *Random Policy* randomly selects P and K from the predefined action space without using channel or semantic-state information. The *Heuristic Policy* follows a rule-based secure transmission strategy that adjusts P and K according to the relative channel conditions of Bob and Eve. The *Vanilla PPO* method trains a standard PPO agent to select P and K based on the observed state, without any LLM assistance. The dueling double deep Q-network (*D3QN*) adopts a dueling double deep Q-network to select the joint action (P, K) from the full discrete action space. It combines double Q-learning and the dueling network architecture, and has been used in recent physical-layer security studies for secure transmission and anti-eavesdropping optimization. The *Proposed (LLM)* method uses the LLM as a cognitive eavesdropper and curriculum-learning module to generate diverse Eve scenarios during training, while PPO still selects actions from the full action space. The *Proposed (LLM+refinement)* method further employs the LLM to refine the candidate action space, allowing PPO to choose from a reduced security-aware action pool.

[2]The detailed prompt templates and implementation code are published in a GitHub repository: https://github.com/Lancelot1998/Generative_AI_for_PLS.

[3]Available at: http://www.statmt.org/europarl/v7/europarl.tgz

[4]The datasets and codes for parameter fitting are online accessible at: https://github.com/Lancelot1998/Generative_AI_for_PLS

[5]Available at: https://huggingface.co/Qwen/Qwen2.5-7B

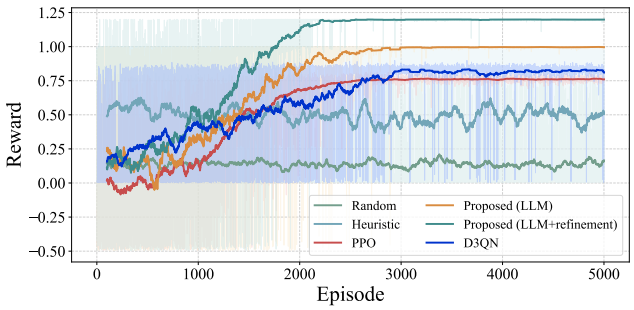

Fig. 5: The training curves and efficiency of different approaches in securing the physical layer of SemCom.

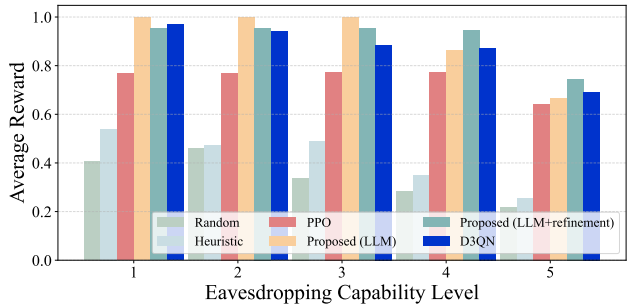

Fig. 6: The security performance of different approaches under different eavesdropping capabilities. We can observe that the proposed (LLM) achieves the best performance under low eavesdropping capability levels, while the proposed (LLM+refinement) achieves the best performance under higher eavesdropping capability levels.

Fig. 5 illustrates the convergence behaviors of different methods during training. We observe that random and heuristic approaches exhibit limited performance, characterized by significant reward fluctuations. PPO and D3QN can gradually learn security policies, demonstrating a certain degree of adaptive optimization capability. However, their convergence speed and final performance remain constrained. In contrast, the proposed (LLM) achieves faster convergence and more stable performance than the vanilla PPO. Furthermore, when the action space is reduced by the LLM, the proposed (LLM+refinement) obtains the highest security performance at an early training stage with stable and optimal performance.

Fig. 6 evaluates the average performance of different methods across multiple eavesdropping settings with diverse capability levels (or environmental difficulties). As the environment becomes more challenging, the limitations of conventional approaches become pronounced. The average rewards of random and heuristic methods decrease significantly, indicating a high sensitivity to channel variations. PPO maintains relatively stable performance across different difficulty levels, but still suffers from performance degradation in high-difficulty scenarios. Meanwhile, D3QN also achieves superior robustness to random and heuristic policies by using double Q-learning and a dueling network architecture to evaluate discrete (P, K) actions more effectively than the simpler baselines. However, because D3QN is value-based and lacks LLM-guided adversarial scenario construction or security-aware action pruning, it still explores the entire unrefined action space, which limits its performance in high-difficulty eavesdropping scenarios. In contrast, the proposed methods (including LLM and LLM+refinement) lead to the highest security performance across all eavesdropping capability levels. Under low eavesdropping capability levels, the proposed (LLM) method achieves the best performance among all configurations. By using the LLM to construct adaptive curriculum schedules, the PPO agent smoothly converges in favorable conditions. Although the proposed (LLM+refinement) performs slightly worse than both the proposed (LLM) and the D3QN baseline in these simple scenarios, it demonstrates more pronounced advantages in complex environments. This is because the LLM-based action refinement adopts a conservative filtering strategy, which may eliminate some aggressive yet safe actions that can yield a higher semantic similarity gap. However, as Eve becomes stronger, such conservative pruning helps avoid risky actions and improves robustness.

In terms of complexity, random and heuristic policies have very low processing delay but limited adaptability. Vanilla PPO requires only lightweight neural-network inference during online deployment. The proposed LLM-enhanced PPO introduces additional LLM inference cost mainly during training or environment refresh, rather than at every transmission slot. In the LLM+refinement scheme, the LLM further narrows the candidate action space, which accelerates convergence and avoids unnecessary exploration. Thus, the additional complexity is mainly incurred offline or periodically, while the online transmitter only executes the trained PPO policy.

V. Open Issues and Future Research Directions

Despite rapid progress in secure SemCom, this field is still in its early stages, especially at PLS. This section outlines the key open issues and promising research directions for future research.

A. LLM-Driven Adversarial Modeling and Difficulty Metrics

We have demonstrated that LLMs can act as cognitive environment constructors for secure SemCom by generating diverse Eve scenarios and refining the candidate action space. This opens several future research directions. For example, LLM-controlled Eve can be assigned different adversarial traits, such as conservative, aggressive, adaptive, or stealthy behaviors, to simulate richer and more realistic semantic security threats. The LLM can also be conditioned on richer contextual information, including channel variations, task requirements, recent cumulative rewards, and semantic-gap trends, to construct more adaptive curriculum-learning schedules. Moreover, defining an explicit and unified Eve difficulty metric is an important future direction. Such a metric may combine Eve's relative channel quality, semantic similarity, receiver capability, and the DRL agent's recent reward trend. This would improve reproducibility, enable more controllable curriculum scheduling, and facilitate fair comparison among different adversarial training strategies [13].

B. Cross-Layer Security Design and Optimization

The inherent coupling of layers in SemCom necessitates a transition from traditional modular design. Security can no longer be treated as a function of a single layer but must be integrated across the entire communication lifecycle,

from model training to semantic information transmission. Developing practical and efficient frameworks that jointly manage resources and security policies across the physical, network, and application layers is a significant challenge. This involves designing lightweight yet effective protocols to exchange semantic context, task requirements, and security state information between layers, enabling truly holistic optimization [14].

C. Integration with Emerging 6G Technologies

The evolution towards sixth-generation (6G) will introduce new technologies that offer both opportunities and challenges for secure SemCom. Future research is necessary to understand and leverage these technologies effectively. For example, reconfigurable intelligent surfaces (RIS) technology allows for programmable control of the wireless environment, offering a powerful new tool for PLS [4]. Open challenges include the joint optimization of RIS phase shifts with semantic-aware beamforming and the development of security protocols that account for the new vulnerabilities introduced by an intelligent surface. Another example is quantum technology, which is based on the principles of quantum physics. Quantum key distribution (QKD) plays a critical role in generating a shared secret key between two parties with information-theoretic security. Unlike classical cryptographic approaches, QKD can detect eavesdropping threats. The integration of QKD with semantic communication is a promising direction for protecting confidential semantic data [15]. However, QKD faces practical challenges, including its reliance on specialized hardware, high infrastructure costs, and implementation-dependent security.

D. Trade-offs Between Security and Complexity

Many promising security techniques, especially those that rely on complex resource allocation optimization or large-scale generative models such as DMs, face significant challenges in terms of computational complexity, energy consumption, and latency. Fewer denoising steps and a smaller model reduce latency, memory use, and energy consumption, yet also come with a tradeoff of decreasing purification quality and robustness. Therefore, a critical issue for future work is the development of low-overhead, scalable, and efficient algorithms that can be practically deployed in real-time, resource-constrained environments, such as battery-powered IoT devices and edge networks. There are some potential techniques to address this issue, e.g., architectural compression (such as pruning and quantization), model distillation, or sampling efficiency [13].

VI. Conclusion

In this paper, we have presented potential security threats at the physical layer in SemCom, considering both signal-level and semantic-level aspects. Corresponding to each aspect, we have provided the GenAI-based solutions to counter these attacks, thereby protecting the confidentiality and availability of the semantic data stream and the integrity of semantic interpretation. To illustrate the promising applications of GenAI in SemCom, we have proposed the LLM-enhanced DRL approach in securing SemCom systems. Our simulation results have clearly validated that integrating LLMs further enhances learning efficiency and security performance. Finally, we have outlined some potential research directions as an essential guide for researchers and practitioners to realize efficient and secure SemCom systems in the future.

References

[1] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana, IL, USA: University of Illinois press, 1998.

[2] C. Zhao, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, X. Shen, and K. B. Letaief, "Generative AI for secure physical layer communications: A survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 11, no. 1, pp. 3–26, Feb. 2025.

[3] Q. Zhang, J. Shi, W. Zeng, X. Xu, Z. Guan, S. Li, and Z. Qin, "Balancing security and efficiency in GAI-driven semantic communication: Challenges, solutions, and future paths," *IEEE Netw.*, vol. 39, no. 5, pp. 88–96, Sep. 2025.

[4] S. Guo, Y. Wang, N. Zhang, Z. Su, T. H. Luan, Z. Tian, and X. Shen, "A survey on semantic communication networks: Architecture, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 5, pp. 2860–2894, Oct. 2025.

[5] B. He, Z. Chen, J. Luo, C. Liu, S. Wang, J. Park, F. Wang, and T. Q. Quek, "Towards secure semantic transmission in the era of GenAI: A diffusion-based framework," *IEEE Commun. Mag.*, Nov. 2025, to be published.

[6] R. Tang, D. Gao, M. Yang, T. Guo, H. Wu, and G. Shi, "GAN-inspired intelligent jamming and anti-jamming strategy for semantic communication systems," in *Proc. 2023 IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Rome, Italy, May 2023, pp. 1623–1628.

[7] X. Ren, J. Wu, H. Xu, and X. Chen, "Diffusion model based secure semantic communications with adversarial purification," in *Proc. 2024 IEEE 10th Conf. Big Data Security on Cloud (BigDataSecurity)*, New York, NY, USA, May 2024, pp. 130–134.

[8] G. Zizzo, G. Cornacchia, K. Fraser, M. Z. Hameed, A. Rawat, B. Buesser, M. Purcell, P.-Y. Chen, P. Sattigeri, and K. Varshney, "Adversarial prompt evaluation: Systematic benchmarking of guardrails against prompt input attacks on LLMs," in *Proc. 2024 Conf. Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 2024, pp. 1–17.

[9] A. M. Ishmam and C. Thomas, "Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024, pp. 24 820–24 830.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart. 2019.

[11] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A semi-NOMA scheme," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 155–169, Jan. 2023.

[12] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 77–100, 1st Quart. 2023.

[13] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *Proc. IEEE*, vol. 112, no. 11, pp. 1649–1685, Nov. 2024.

[14] L. Wang, W. Wu, F. Zhou, Z. Qin, and Q. Wu, "IRS-enhanced secure semantic communication networks: Cross-layer and context-aware resource allocation," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 494–508, Jan. 2025.

[15] S. N. Paing, J. W. Setiawan, T. Q. Duong, D. Niyato, M. Z. Win, and H. Shin, "Quantum anonymous networking: A quantum leap in privacy," *IEEE Netw.*, vol. 38, no. 5, pp. 131–145, Sep. 2024.

Long V. Nguyen is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, Memorial University, Canada. His research interests include semantic communications, generative AI, resource allocation, and vehicular networks.

Yinqiu Liu is currently working toward the PhD degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His current research interests include wireless communications, mobile AIGC, and generative AI.

Quang Nhat Le is currently a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, Memorial University, Canada. His research interests include enabling technologies for 6G wireless networks and machine learning for wireless communications.

Trung Q. Duong is a Canada Excellence Research Chair (CERC) and Full Professor at Memorial University, Canada. He is also an adjunct professor at Queen's University Belfast, UK, a visiting professor at Kyung Hee University, South Korea and Edinburgh Napier University, UK.