Semantic Successive Refinement: A Generative AI-aided Semantic Communication Framework

Kexin Zhang, Lixin Li, *Member, IEEE*, Wensheng Lin , *Member, IEEE*, Yuna Yan, Rui Li, Wenchi Cheng, *Senior Member, IEEE,* and Zhu Han, *Fellow, IEEE*

***Abstract*—Semantic Communication (SC) is an emerging technology aiming to surpass the Shannon limit. Traditional SC strategies often minimize signal distortion between the original and reconstructed data, neglecting perceptual quality, especially in low Signal-to-Noise Ratio (SNR) environments. To address this issue, we introduce a novel Generative AI Semantic Communication (GSC) system for single-user scenarios. This system leverages deep generative models to establish a new paradigm in SC. Specifically, At the transmitter end, it employs a joint source-channel coding mechanism based on the Swin Transformer for efficient semantic feature extraction and compression. At the receiver end, an advanced Diffusion Model (DM) reconstructs high-quality images from degraded signals, enhancing perceptual details. Additionally, we present a Multi-User Generative Semantic Communication (MU-GSC) system utilizing an asynchronous processing model. This model effectively manages multiple user requests and optimally utilizes system resources for parallel processing. Simulation results on public datasets demonstrate that our generative AI semantic communication systems achieve superior transmission efficiency and enhanced communication content quality across various channel conditions. Compared to CNN-based DeepJSCC, our methods improve the Peak Signal-to-Noise Ratio (PSNR) by 17.75% in Additive White Gaussian Noise (AWGN) channels and by 20.84% in Rayleigh channels.**

***Index Terms*—Generative AI, Semantic Communication, Multi-user System, Swin Transformer, Diffusion Model.**

I. Introduction

With the rapid growth of wireless communication technology, data traffic and mobile device connectivity is increasing exponentially. Semantic Communications (SC) is playing a pivotal role in this progress [1]. Unlike traditional communication methods that focus on transmitting binary bits, SC emphasizes the context of the information [2], [3]. Recent advancements further highlight SC's potential. For instance, the semantic interference cancellation (SemantIC) technique enhances information quality by iteratively eliminating noise in both the signal and semantic domains without additional channel resource costs [4]. Similarly, the Semantic-Forward (SF) relaying framework improves network robustness and reduces forwarding payload by extracting and transmitting semantic features, even under adverse channel conditions [5].

Simultaneously, the rapid advancement of generative Artificial Intelligence (AI) models and the widespread adoption of technologies such as ChatGPT, Imagen, Midjourney, and DALL-E have prompted the B5G and 6G communities to synergize with these cutting-edge generative AI technologies. This paradigm shift is particularly beneficial for the vast amount of multimedia content generated by these applications, such as images and videos. Typically, AI models run on powerful cloud servers due to their high computational demands. However, with the growing prevalence of mobile devices, AI companies strive to provide high-quality AI-Generated Content (AIGC) services accessible from anywhere [6], [7]. These innovations lay a solid foundation for further integrating generative AI with SC, which can significantly boost network performance.

By considering the context of the generated content, SC can effectively leverage these advanced generative AI models to enhance overall system performance. For example, consider a user playing an online game on their VR/AR device at home, entering a new virtual scene rendered by a Diffusion Model (DM) in the cloud. The scene is represented by information bits communicated to the VR/AR glasses via mobile networks. A generative AI-aware semantic communication network can utilize the fact that DM generates this content and transmits the latent representation of the content while performing the stable diffusion process on the receiver's end. In this scenario, the semantic communication network focuses on conveying the meaning (latent representations) of the generative AI content rather than optimizing solely the delivery of the 0s and 1s. This approach allows for quality service metrics that align more closely with human visual perception than traditional bit-wise metrics.

However, directly combining semantic communication with generative AI in fixed structures is inefficient due to the inability to adjust semantic density. In this context, deploying DM in the decoder offers a promising solution to accurately recover the image context. Toward this end, we exploit the

This work was supported in part by National Natural Science Foundation of China under Grants 62001387 and 62101450, in part by the Young Elite Scientists Sponsorship Program by the China Association for Science and Technology under Grant 2022QNRC001, in part by Aeronautical Science Foundation of China under Grants 2022Z021053001 and 2023Z071053007, in part by Shanghai Academy of Spaceflight Technology under Grant SAST2022-052, in part by NSF CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, US Department of Transportation, Toyota and Amazon. (*Corresponding authors: Lixin Li, Wensheng Lin.*)

Kexin Zhang, Yuna Yan, Lixin Li, Wensheng Lin are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, Shaanxi 710129, China. (e-mail: cocozhang@mail.nwpu.edu.cn; yanyuna@mail.nwpu.edu.cn; lilixin@nwpu.edu.cn; linwest@nwpu.edu.cn).

Rui Li is a research scientist at Samsung AI Center, Cambridge, U.K, and visiting researcher at School of informatics, the University of Edinburgh. (e-mail: rui.li@samsung.com).

Wenchi Cheng is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: wccheng@xidian.edu.cn).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: hanzhu22@gmail.com).

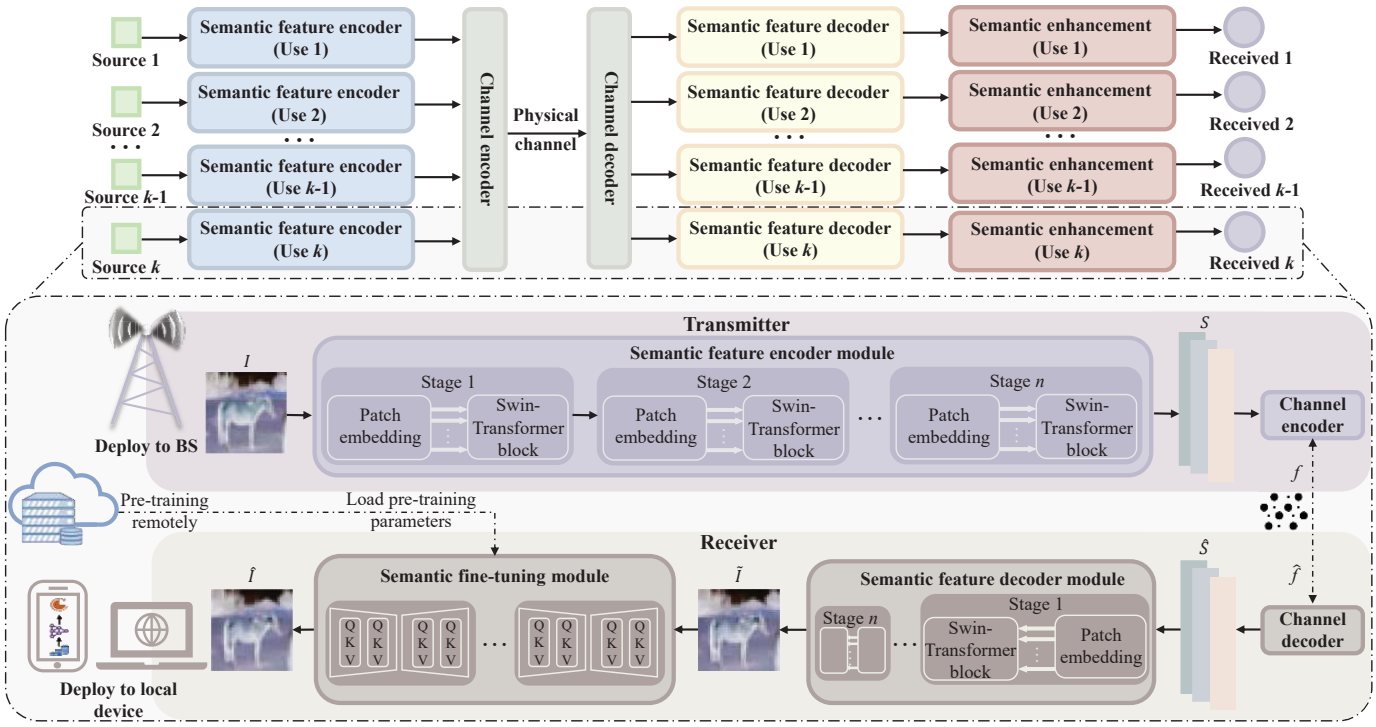


Fig. 1. The overview of the proposed generative AI semantic communication system.

potential of the generative model and propose a novel generative AI semantic communication system, which aims to incorporate the robust, stable diffusion algorithm into the semantic decoding process as an initial step towards a fully collaborative generative AI semantic communication system. The main contributions of this paper are as follows:

- To significantly enhance the efficiency and reliability of semantic communication systems in the era of generative AI, we propose a single-user Generative AI Semantic Communication (GSC) system powered by AI-generated content. Specifically, the Base Station (BS) encodes images using a joint source-channel coder based on the Swin Transformer. A diffusion model network is integrated at the receiver end to facilitate rich semantic reconstruction during wireless image transmission.
- Existing research on semantic communication has predominantly focused on single-user scenarios. A key challenge in multi-user communication is effectively managing the simultaneous transmission and reception of semantic data from different users, particularly in resource-constrained environments. To address this, we scale semantic communication to multi-user scenarios, and this system incorporates asynchronous concurrent processing, task parallel processing, and caching mechanisms to optimize communication efficiency and maintain quality across multiple users.
- Classical diffusion models excel in image synthesis, but they require numerous iterative steps to estimate feature mappings, making them inefficient for direct use in communication systems. To address this concern, we use the strong distribution mapping abilities of diffusion models to guide semantic recovery at the decoding end by estimating a compact conditional vector. This approach significantly advances the connection between semantic communication and cutting-edge generative models, improving the end-to-end perceptual performance of wireless communication systems.
- Simulation experiments were conducted on public datasets. The results demonstrate that our proposed generative AI semantic communication system surpasses the benchmark communication system in data integrity and communication reliability. Additionally, the simulation results confirm that our designed multi-user communication system efficiently handles requests from multiple users, showcasing excellent performance and scalability.

This paper is structured as follows. Section II provides a concise overview of related literature. Section III provides a more detailed explanation of our system model. Following that, Section IV presents the simulation results and analysis of the proposed model. Finally, section V summarizes the paper.

II. Related Works

Semantic communication systems generally consist of three fundamental components: the semantic encoder at the transmitter, the channel, and the semantic decoder at the receiver. At the transmitter, the semantic encoder extracts key semantic features from the original data and compresses them into low-dimensional representations optimized for wireless transmission. The channel models the wireless environment, accounting for potential noise and interference that may degrade the signal. At the receiver, the semantic decoder reconstructs the original data by decoding the received semantic symbols to maintain high-quality data representation.

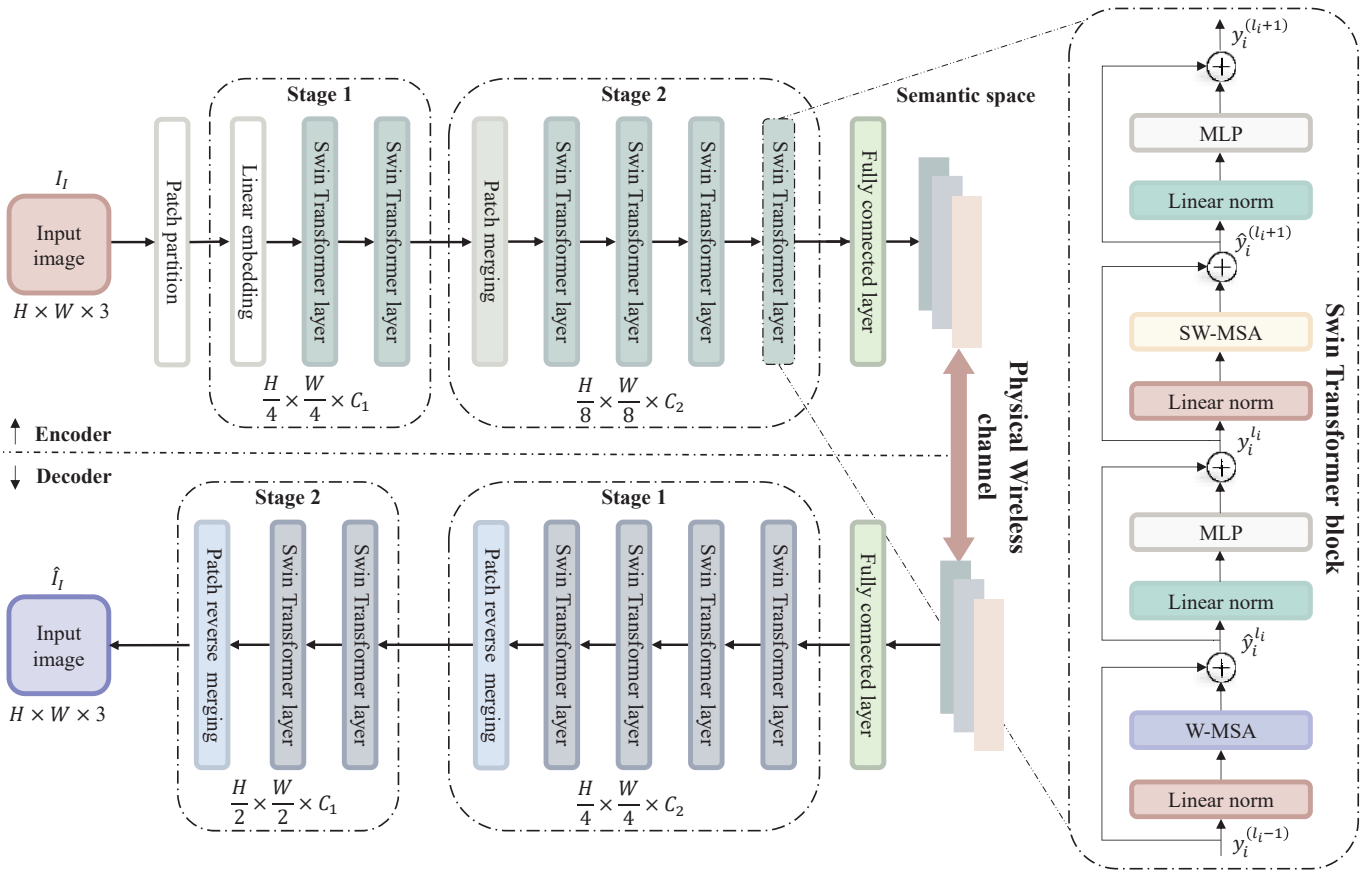


Fig. 2. The architectures of the semantic communication network based on Swin Transformer.

Transmitting image-based semantic information demands significantly more communication resources compared to text-based data. Consequently, this section focuses on Image Semantic Communication (ISC), which can be divided into two main directions: semantic-oriented and task-oriented communication. The challenge of semantic-oriented communication lies in extracting and recovering semantics before and after transmission. In contrast, task-oriented communication focuses on completing specific tasks rather than accurately recovering all semantic information. The relevant works for these paradigms are outlined below.

A. Image semantic communication

Deep Joint Source-Channel Coding (DeepJSCC) [8] is a pioneering study in wireless image transmission. This technique simplifies the code design process by integrating source and channel coding into a single mapping and eliminating the need for constellation diagrams used in digital schemes. Building on this foundation, the studies in [9]–[11] extended DeepJSCC to various channel conditions. Yang *et al.* [10] proposed an adaptive wireless image transmission scheme that dynamically adjusts the transmission rate according to the current channel state and image complexity. Xu *et al.* [11] introduced a joint source-channel coding method incorporating an attention mechanism (ADJSCC). This framework adapts to noise by dynamically adjusting the number of bits allocated to the channel and source coder to maintain transmission reliability in real-world environments. However, these approaches primarily focus on the distortion of the reconstructed signal at the receiver relative to the source at the transmitter without adequately considering the perceptual quality of the reconstructed image, which may lead to significant perceptual distortion under extreme conditions such as low bandwidth and low Signal-to-Noise Ratio (SNR).

To address this, Kurka *et al.* [12] proposed DeepJSCC-f, which employs a feedback mechanism that allows the system to acquire real-time channel noise information and mitigate its effects by feeding the received signal back to the transmitter. However, this method assumes that any complex value can be transmitted over the channel, which may hinder the application of these algorithms in scenarios where the hardware or protocol only accepts certain sets of channel inputs (e.g., digital constellations). Additionally, the semantic communication systems proposed in [13]–[16] are designed for receivers with powerful computational capabilities, enabling large-scale deep learning networks and single-user communication scenarios. Among them, Zhang *et al.* [13] developed a system that adaptively identifies and transmits task-specific key semantic features in a changing environment. Yang *et al.* [14] proposed the WITT framework to enhance CNN performance by introducing a spatial modulation module, which adjusts the scale of potential representation based on channel state information, improving the model's adaptability to different channel conditions. Yu *et al.* [15] extended the semantic communication

system to bi-directional communication for IoT devices with limited capacity, significantly reducing training overheads by eliminating the need for information feedback and model migration. Nguyen *et al.* [16] proposed a system that adapts to different computational capabilities by dynamically adjusting the transmission length of the output from the channel coder, combined with hybrid loss optimization, to achieve high-quality image reconstruction and avoid network congestion.

Task-oriented approaches optimize resource utilization by delivering information according to specific task requirements. Kadam *et al.* [17] designed a keyword-based SC system for transmitting and sharing knowledge recovery data for a specific Data Allocation Problem (DAP) to enhance efficiency and accuracy. Xie *et al.* [18] developed a multi-user system supporting two unimodal tasks (image retrieval and machine translation) and one multimodal task (a visual quiz combining text and images). Kang *et al.* [19] created a framework that enables users to retrieve image-related semantic information via text queries. However, this approach fails to fully consider the dynamic changes in the importance of semantic information, which can lead to information loss in resource competition. In [20], researchers employed a course-learning strategy to minimize the length of transmitted messages, aiming to improve communication efficiency for specific tasks.

To address the issue of limited resource blocks from BS to individual users, Wang *et al.* [21] utilized a reinforcement learning algorithm based on an attention mechanism to prioritize the transmission of the most informative triples during resource allocation. Thomas *et al.* [22] emphasized the reasoning capabilities required by both the transmitter and receiver, employing a Generative Flow Network (GFlowNet) to achieve causal reasoning for bursty semantic communication with minimal data. Yoo *et al.* [23] demonstrated the feasibility of ISC in real-time wireless communication using Field Programmable Gate Array (FPGA) as a hardware platform, comparing its performance with conventional 256 Quadrature Amplitude Modulation (QAM) and showing superior application potential. However, their results were only validated in a simulation environment.

B. Generative AI semantic communication

Generative AI revolutionizes visual computing by allowing users to programmatically create or modify realistic, high-quality images, videos, and 3D models. Neural network-driven generative models, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs), have shown exceptional capabilities in generating high-quality data. This highlights their significant potential for application in joint source-channel coding.

Choi *et al.* [24] was the first to implement a Joint Source-Channel Coding (JSCC) scheme over binary channels using a discrete VAE. Subsequently, Hu *et al.* [25] employed an adversarial training approach and introduced the Masked Variational Quantized Variational Autoencoder (MVQ-VAE) to enhance the system's robustness to various noise types. Paper [26] developed an innovative architecture for DeepJSCC that is data-driven and enhanced by adversarial training, with the combined goal of maximizing the reconstruction quality for legitimate receivers while minimizing adversarial losses. Yang *et al.* [27] combined an autoencoder with Orthogonal Frequency Division Multiplexing (OFDM), using a GAN-inspired loss function to efficiently train a robust decoder against the effects of multipath fading. The approach described in [28] significantly reduces bandwidth requirements by performing semantic segmentation at the transmitter's end to extract semantic information from an image and using GAN for image reconstruction at the receiver's end. However, this reconstruction process may introduce subtle differences between the generated image and the original scene. To this end, Erdemir *et al.* [29] proposed two schemes: InverseJSCC, which addresses the inverse problem of DeepJSCC, and GenerativeJSCC, an end-to-end optimization scheme based on GANs that can reconstruct perceptual quality under extremely adverse channel conditions like never before.

Recent advancements in DMs have established new benchmarks in density estimation and sample quality, surpassing other generative models. Parametric Markov chains enhance the variational lower bound of the likelihood function, thereby enabling samples to represent the target distribution [30] more accurately. DMs iteratively refine these samples through a methodical denoising process until the desired output is achieved. Niu *et al.* [31] introduced a strategy based on DeepJSCC, their method initially adds independent Gaussian noise to the input image and then applies a diffusion process to introduce additional information components during decoding. In contrast, other studies [32], [33] utilize the denoising capabilities of DMs to help the receiver mitigate noise interference in the channel or semantic vectors. Furthermore, [34] addresses the communication problem as an inverse problem and proposes a generative AI semantic communication framework that resolves the denoising or colouring issues in the low-dimensional latent representations of samples. Chen *et al.* [35] proposed CommIN, an innovative framework that combines an Invertible Neural Network (INN) with a DM to model channel characteristics and degradation phenomena introduced by DeepJSCC. Grassucci *et al.* [36] developed a generative AI semantic communication framework that enhances the quality of inferred images by incorporating semantic chunks for rapid denoising.

Despite the advancements in DMs, they still need practical challenges, such as large network sizes and the iterative nature of the process, which complicates training with limited computational resources. Meanwhile, semantic communication systems encounter additional challenges: first, signal noise during transmission significantly interferes with semantic information, particularly under low SNR conditions, where traditional methods fail to ensure high-quality semantic recovery. Second, existing approaches are often limited to single-user scenarios and lack scalability in multi-user environments. To address these issues, we designed a system focused on semantic-oriented communication, which is particularly beneficial for preserving data integrity and visual fidelity.

III. Proposed Method

Considering downlink transmission scenarios, this section describes the proposed semantic communication model.

Fig. 1 illustrates the general framework of the proposed approach, which comprises four modules: *the semantic feature encoder module, the physical wireless channel module, the semantic feature decoder module, and the semantic fine-tuning module*. Specifically, the system leverages the Swin Transformer encoder to improve the efficiency of semantic feature extraction and compression, enabling the transmitter to generate low-dimensional semantic symbols suitable for wireless transmission. To mitigate the impact of channel noise on semantic information, a Diffusion Model is introduced at the receiver, which, with its powerful noise processing and reconstruction capabilities, effectively restores high-quality semantic information under low SNR conditions. Additionally, to address resource utilization in multi-user scenarios, the proposed GSC system has been extended to a Multi-User Generative Semantic Communication system (MU-GSC).

A. Semantic feature encoder module

The BS usually consists of two components in semantic communication systems: a semantic feature encoder and a channel coder. The semantic feature encoder, also known as the source encoder, mines information and extracts features from the images based on the knowledge base. Let I be the transmitted image source and S be the extracted semantic symbols. The mathematical description is as follows:

$$S = \boldsymbol{E}(I; \varphi_\alpha), I \in \mathbb{R}^n, \tag{1}$$

where $\boldsymbol{E}(\cdot)$ denotes the semantic coder network, φ_α is the parameter set of the corresponding coding network, and n is the dimension of the input images.

In a semantic communication system, the edge server is responsible for training the local model, and the training process requires the encoder and decoder to be coupled alternately and in an end-to-end manner. Here, the semantic encoder is first modeled, and the semantic decoder follows the reverse architecture of the encoder.

The codec network in this paper is constructed based on the state-of-the-art visual Swin Transformer [37]. On the transmitter, given a set of image sets $\boldsymbol{I} = \{I_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{3\times H\times W}$ denotes the i-th image, N denotes the size of the image set, H and W denote the height and width of the images, respectively. The patch size is 4×4, which is partitioned into non-overlapping patches $s = \frac{H}{4} \times \frac{W}{4}$ by the patch partition module. Then, we perform a linear embedding by arranging these tokens in a sequence $(I_1, I_2, \cdots I_s)$ from the top left to the bottom right, which projects the tokens into an embedding representation $\left(\frac{H}{4}, \frac{W}{4}, C\right)$ of arbitrary dimension C. After the patch embedding, the s tokens are then input into the Swin Transformer block. This process is called "stage 1".

From Fig. 2, it can be seen that the first patch merging layer and the Swin Transformer block are merged into "stage 2". The patch merge layer splices each group of neighboring patches of size 2×2 so that the number of patch tokens becomes $1/4$ of the original one, i.e. $\frac{H}{8} \times \frac{W}{8}$. At the same time, the dimension of the patch token is expanded by a factor of 4, i.e.$4C$. In order to reduce the output dimension and realize the downsampling of the feature map, the patch merging layer then performs the fully connected operation. After this process, the dimension of the concatenated feature patch is reduced $2C$ from $4C$. Then, the patch is feature transformed by the Swin Transformer block. Finally, the output dimension becomes $\left(\frac{H}{8}, \frac{W}{8}, C\right)$. Considering the computational power of the hardware and the uncertainty of the input graph, here the semantic decoder contains a two-stage Swin Transformer architecture, and in general, high-resolution images require more stages.

The Swin Transformer module is a sequence-to-sequence function that consists of multiple Swin Transformer layers space. The right side of Fig. 2 shows the structure of the Swin Transformer layers. Each layer consists of a window polytope layer and a moving window polytope layer. Since the two sub-layers have the same dimensions of inputs and outputs, they are sequentially connected of the two sub-layers. The use of the shifted-window partitioning approach can significantly enhance the modeling capability. The successive Swin Transformer blocks l_i layer and $l_i + 1$ layer for "stage i" are computed as:

$$\begin{aligned}
\hat{y}_i^{l_i} &= \mathrm{W}-MSA\left(\mathrm{LN}\left(y_i^{l_i-1}\right)\right) + y_i^{l_i-1}, \\
y_i^{l_i} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{y}_i^{l_i}\right)\right) + \hat{y}_i^{l_i}, \\
\hat{y}_i^{l_i+1} &= \mathrm{SW}-MSA\left(\mathrm{LN}\left(y_i^{l_i}\right)\right) + y_i^{l_i}, \\
y_i^{l_i+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{y}_i^{l_i+1}\right)\right) + \hat{y}_i^{l_i+1},
\end{aligned} \tag{2}$$

Here, the MLP has two layers, the Window-based Multi-head Self-Attention (W-MSA) and the Shifted Windows Multi-head Self-Attention (SW-MSA) are multi-head self-concerned modules with regular and shifted window configurations. LN denotes the layer normalization operation. $\hat{y}_i^{l_i}$ and $y_i^{l_i}$ represent the output characteristics of the (S)W-MSA module and MLP module of module l_i at "stage i", respectively.

In order to protect the transmitted symbols from channel noise and interference in the wireless environment, the output will S be further converted in the channel coder into a complex bit stream (vector) suitable for transmission over the physical channel. Here a non-trainable fully connected layer is used for simulation:

$$f = \boldsymbol{C}(S; \varphi_\beta) = \boldsymbol{W}_n S + b_n, f \in \mathbb{R}^k, \tag{3}$$

where $\boldsymbol{C}(\cdot)$ is the channel coder with parameter set φ_β and k is the length of f. The weight matrix $\boldsymbol{W}_n$ and the bias matrix b_n determine the mapping relationship of the neural network. Here, $\boldsymbol{W}_n$ represents the channel gain, and b_n simulates the noise in the channel by adding a random variable to the signal, the variance of the random variable represents the power of the channel noise, which is in turn constrained by the SNR and the transmit power.

This conversion process not only improves the signal's immunity to interference but also realizes an effective mapping from high-dimensional data to low-dimensional data, preserving the integrity and interpretability of the semantic content. The compression ratio is calculated by taking the ratio k/n of the transmitted data size to the original image size. While it is possible to compress the semantic feature vectors further and

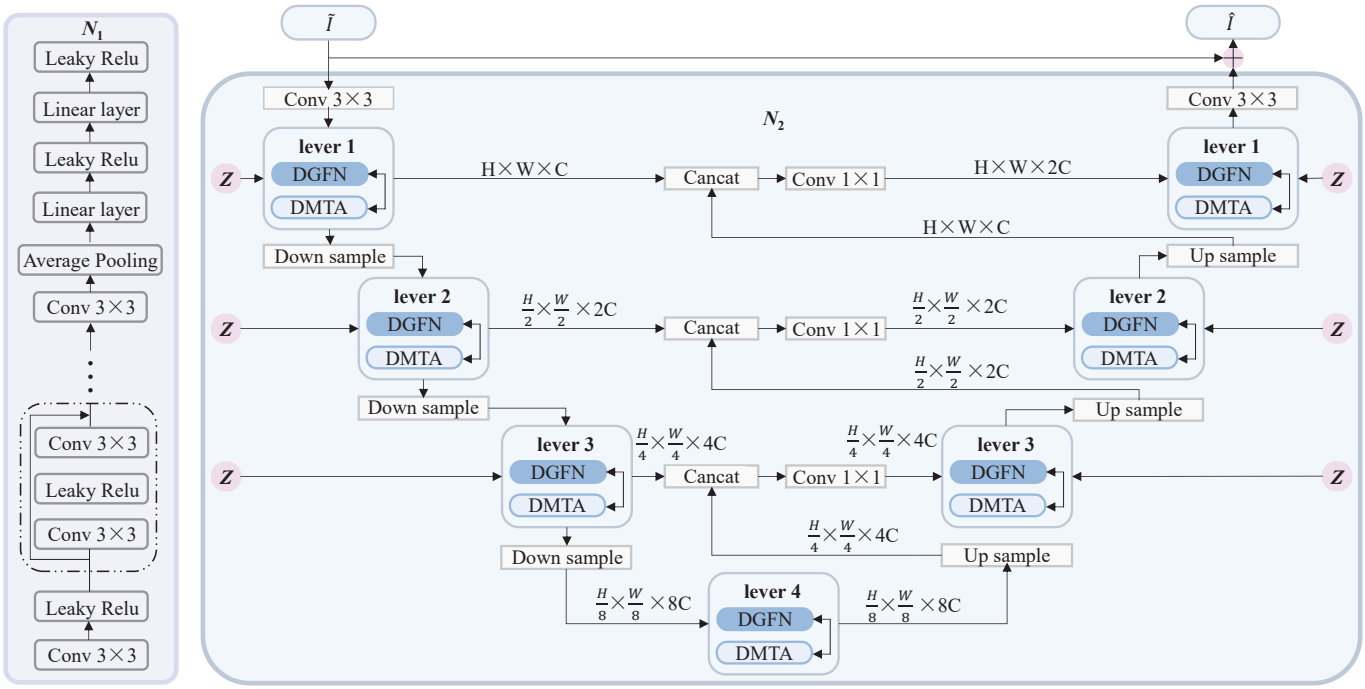


Fig. 3. The architectures of the prior extraction network $\boldsymbol{N}_1$ and the image denoising network $\boldsymbol{N}_2$.

thus reduce the transmission length, this simultaneously introduces the problem of the trade-off between the compressed size and the received image quality. Based on empirical values, this study selects 1/6 of the experimental conditions.

B. Physical wireless channel module

Considering the finite transmit power of the transmitting device, a power normalization operation must be used to make the signal f satisfy the average power constraint before sending the transmitted signal to the channel. This implies $\frac{1}{k}\mathbb{E}_f\left[\|f\|_2^2\right] \leq P$. Since the encoder and decoder are trained end-to-end, the physical channel can be modeled with a frozen neural network. In this paper, we consider the general fading channel model with transfer function $\hat{f} = w(f;h) = h \odot f + n$, where $\odot$ is the element-wise product, h denotes the Channel State Information (CSI) vector, and each component of the noise vector n is independently sampled from a Gaussian distribution, i.e., $n \sim \mathcal{N}\left(0, \sigma^2 I\right)$, where σ^2 is the average noise power.

C. Semantic feature decoder module

A joint source-channel decoder will be deployed at the receiver of the local device, including two components, the channel decoder and the semantic decoder. The signal is first binary converted by the channel decoder and then sent to the semantic decoder to reduce from a potential representation of noise to a semantic feature map that the user cannot understand.

Channel distortion and noise are critical factors that cannot be avoided when transmitting coded feature vectors in a wireless channel environment. Taking an Additive White Gaussian Noise (AWGN) channel as an example, the received signal vector can be written as:

$$\hat{f} = f + \varepsilon, \tag{4}$$

where ε is a noise vector whose elements obey $\mathcal{N}\left(0, \sigma^2 I\right)$.

$$\hat{S} = \boldsymbol{C}^{-1}\left(\hat{f}; \varphi_\phi\right), \tag{5}$$

where $\hat{f}$ is the estimated feature vector of the transmitted image source I, $\boldsymbol{C}^{-1}(\cdot)$ is the channel decoder, and $\hat{S}$ is the data bits recovered from the code word $\hat{f}$ received from the channel.

The semantic decoder receives the signal disturbed. This process can be represented as:

$$\widetilde{I} = \boldsymbol{E}^{-1}\left(\hat{S}; \varphi_\gamma\right), \tag{6}$$

where $\boldsymbol{E}^{-1}$ denotes the semantic decoder network, φ_γ is the parameter set of the corresponding network, and $\widetilde{I}$ is the decoded semantic information images.

Most of the existing research on semantic image enhancement has used the MSE loss as the objective for the training process, which has the property of being continuously differentiable across any domain. Therefore, the MSE between the original input images and the user's received images is used as the loss function in the training phase, and l is the length of the image vector, which is obtained from the product of the height, width and number of channels of the images:

$$\mathcal{L}_{dec} = \mathrm{MSE} = \frac{1}{l}\sum_{i=1}^{l}\left(I_i - \widetilde{I}_i\right)^2. \tag{7}$$

D. Semantic fine-tuning module

Considering the significant increase in computing power of mobile devices, many user terminals can perform relatively

simple fine-tuning operations. In this subsection, the Semantic Fine-Tuning (SFT) module is designed. Specifically, after the semantic features are decoded, they need to be further transferred to the DM with pre-training parameters to perform semantic enhancement. Inspired by image retrieval algorithms, adding precise details to low-quality images avoids the need for the DM to generate complete images. As a result, SFT achieves more accurate estimation with fewer iterations and ensures the stability of the results.

Algorithm 1 The Training Process on the Cloud Server.

Input: $\beta_t (t \in [1, T])$
Output: The trained model $\mathcal{M}$
Init: $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$
for $(\widetilde{I}, I)$ **do**
  $Z_0 = \boldsymbol{N}_1(\text{PixelUnshuffle}(\text{Concat}(\widetilde{I}, I)))$
  Forward Diffusion:
  We sample Z_T by
  $q(Z_T \mid Z_0) = \mathcal{N}(Z_T; \sqrt{\bar{\alpha}_T} Z_0, (1 - \bar{\alpha}_T) I)$
  Reverse Inference:
  $\hat{Z}_T = Z_T$
  $D = \boldsymbol{N}_1(\text{PixelUnshuffle}(\widetilde{I}))$
  for $t = T$ to 1 **do**
    $\hat{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{Z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\text{Concat}(\hat{Z}_t, t, D)) \right)$
  end for
  $\hat{Z} = \hat{Z}_0$
  $\hat{I} = \boldsymbol{N}_2(\widetilde{I}, \hat{Z})$
  Calculate L loss
end for
Output the trained model $\mathcal{M}$

Algorithm 2 The Testing Process on the Local Device.

Input: The trained model $\mathcal{M}$, decoded images $\widetilde{I}$
Output: The received images $\hat{I}$
Init: $\alpha_t = 1 - \beta_t, \bar{\alpha}_T = \prod_{i=0}^{T} \alpha_i$
Reverse Inference:
Sample $Z_T \sim \mathcal{N}(0, \boldsymbol{O})$
$D = \boldsymbol{N}_1(\text{PixelUnshuffle}(\widetilde{I}))$
for $t = T$ to 1 **do**
  $\hat{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{Z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta \left(\text{Concat}(\hat{Z}_t, t, D) \right) \right)$
end for
$\hat{Z} = \hat{Z}_0$
$\hat{I} = \boldsymbol{N}_2(\widetilde{I}, \hat{Z})$
Output received images $\hat{I}$

The SFT module consists of two distinct processes: the training process and the testing process. These processes are executed on the cloud server and the local device, respectively. Detailed descriptions of both processes can be found in Alg. 1 and Alg. 2. The training process is discussed first, which primarily involves two networks: the prior extraction network ($\boldsymbol{N}_1$) and the image denoising network ($\boldsymbol{N}_2$) [30], as shown in Fig. 3. The main function of $\boldsymbol{N}_1$ is to extract Prior Representations (PRs) in the form of conditional vectors. $\boldsymbol{N}_2$ then uses this information to restore and enhance the details.

Specifically, the workflow is as follows: we initially combine the transmitted images with its' corresponding decoded images through concatenation. Following this, we utilize the PixelUnshuffle operation to downsample the concatenated images. The extracted PRs from this process are denoted as $Z \in \mathbb{R}^{4C'}$:

$$Z = \boldsymbol{N}_1(\text{PixelUnshuffle}(\text{Concat}(\widetilde{I}, I))). \tag{8}$$

Then, $\boldsymbol{N}_2$ can use the extracted Z to restore images, which is stacked with dynamic transformer blocks in the Unet shape. The dynamic transformer blocks consists of Dynamic Multi-head Transposed Attention (DMTA) and Dynamic Gated Feed-Forward Network (DGFN), which can use Z as dynamic modulation parameters to add restoration details into feature maps, effectively aggregating both local and global spatial characteristics. The image semantic enhancement is represented as follows:

$$\hat{I} = \boldsymbol{N_2}(Z; \varphi_\varpi), Z \in \mathbb{R}^{4C'}, \tag{9}$$

where φ_ϖ are the parameters of $\boldsymbol{N_2}$.

Next, $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ are jointly optimized. The loss function is defined as follows:

$$\mathcal{L}_{pre} = \left\| I - \hat{I} \right\|_1, \tag{10}$$

where I and $\hat{I}$ are the transmitted and received images, respectively. $\|\cdot\|_1$ is the L_1 paradigm.

In the DM training phase, the accurately recovered images are generated from the lossy decoded images mainly through the efficient data estimation function of diffusion model, and this process includes two critical aspects: the forward diffusion and the reverse inference. First, the PRs of the decoded images, denoted as Z_0, are captured using the pre-trained $\boldsymbol{N}_1$, and the forward diffusion process of Z_0 is applied to the sample Z_T through T iterations. Each iteration is as follows:

$$q(Z_T \mid Z_0) = \mathcal{N}\left(Z_T; \sqrt{\bar{\alpha}_T} Z_0, (1 - \bar{\alpha}_T) I\right), \tag{11}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative product of α_t, the scheduler gradually adds Gaussian noise at each time step $t \in [0, T]$ until the semantic information of Z_0 becomes pure noise Z_T, each new image at time step t is sampled from a conditional Gaussian distribution, as follows:

$$Z_t = \sqrt{1 - \beta_t} Z_{t-1} + \sqrt{\beta_t} \varepsilon, \tag{12}$$

where $0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1$ is a variance table with time-dependent constants and $\varepsilon \sim \mathcal{N}(0, \boldsymbol{O})$ is a Gaussian noise with unit matrix $\boldsymbol{O}$.

The forward diffusion process introduces noise to the data, while the reverse inference process is a denoising procedure. However, traditional distributed denoising algorithms can only optimize the denoising network by randomly selecting a time step during the iteration process, which significantly increases computational costs. In contrast, since PRs are compact, fewer iterations and smaller model sizes can be used to achieve estimates that are comparable to traditional diffusion models. Here, the noise at each time step t is estimated using ε_θ to obtain $\hat{Z}_{t-1}$, i.e.,

$$\hat{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{Z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta \right), \tag{13}$$

where ε_θ is the prediction noise, $\alpha_t = 1 - \beta_t$, and $\bar{\sigma}_t$ is also a time-dependent constant for the step.

After T iterations, $\hat{Z} \in \mathbb{R}^{4C'}$ is generated. Since the trained model $\mathcal{M}$ only adds details for recovery, DM can obtain stable visual results after several iterations. After that, $\boldsymbol{N}_2$ utilizes $\hat{Z}$ to recover the semantic information images $\widetilde{I}$.

The diffusion model aims to predict the noise distribution of the conditioned vectors of the decoded images with a loss function denoted as $\mathcal{L}_{\text{diff}} = \frac{1}{4C'} \sum_{i=1}^{4C'} \left| \hat{Z}(i) - Z_0(i) \right|$. To achieve efficient generation and propagation of semantic information, $\mathcal{L} = \mathcal{L}_{pre} + \mathcal{L}_{diff}$ is used here for joint optimization.

Notably, unlike in the forward diffusion process, during the reverse inference process, the transmitted image source I is not input into either $\boldsymbol{N}_1$ or $\boldsymbol{N}_2$. Instead, $\boldsymbol{N}_1$ is employed to obtain the PRs, denoted as $D \in \mathbb{R}^{4C'}$, directly from the decoded images, as follows:

$$D = \boldsymbol{N}_1(\text{PixelUnshuffle}(\widetilde{I})), \tag{14}$$

E. Multi-user communication system

The multi-user scenario refers to each user transmitting independent semantic information to perform their transmission tasks. As shown in Fig. 1, a multi-user generative AI semantic communication system that consists of k sources and k destinations is considered.

The message of the k-th user's source is denoted as I_k. Each source transmits semantic information. Similar to the single-user cognitive semantic communication system, the semantic information is first expressed as the semantic symbols. The semantic symbol S_k is abstracted from the image source I_k by using our proposed semantic feature encoder, given as:

$$S_k = \boldsymbol{E}(I_k), k = 1, 2, ..., n. \tag{15}$$

After the semantic symbols of each source are obtained, they are transmitted by exploiting the conventional non-trainable fully connected layer to simulate. Specifically, the semantic symbol S_k is encoded to improve transmission efficiency, and f_k is obtained, i.e.,

$$f_k = \boldsymbol{C}(S_k) = \boldsymbol{W}_n S_k + b_n, k = 1, 2, ..., n, \tag{16}$$

where $\boldsymbol{C}$ is the channel coding; f_k and S_k are the channel coding and semantic symbols of the k-th source, respectively.

After transmission over the channel, the channel decoding is performed at each user receiver, and the reconstructed semantic symbol $\widetilde{I}_k$ is obtained by exploiting our proposed semantic feature decoder module. Note that the semantic symbols of different users are not distinguished in this process. Thus, the reconstructed semantic symbol of each user is mixed, given as:

$$\widetilde{I}_k = \boldsymbol{E}^{-1}\left(\boldsymbol{C}^{-1}\left(\hat{f}_k\right)\right), k = 1, 2, ..., n, \tag{17}$$

where $\boldsymbol{C}^{-1}$ is the channel decoding and $\boldsymbol{E}^{-1}$ is the semantic feature decoding, $\widetilde{I}_k$ is the reconstructed semantic information images and $\hat{f}_k$ is the channel coding received vector at the k-th user.

Then, the reconstructed semantic symbols of each user need to be fine tuning. All models can be trained in the cloud and then broadcast to users. Similar to the single-user cognitive semantic communication system, the received images is obtained by exploiting our proposed semantic fine-tuning module.

To address the challenge of data processing in multi-user scenarios, we first implemented a data segmentation strategy during the pre-processing stage. This approach simulates different user sources generating diverse messages, enabling effective allocation of user tasks to various processing units and allowing these tasks to be executed concurrently. Subsequently, we introduced an asynchronous concurrent processing model. Employing asynchronous task processing functions and event loops ensured that the system's main thread remained unblocked during I/O operations, thereby facilitating the concurrent execution of multiple user tasks.

To further enhance system performance, we leveraged task parallel processing technology. This technique involves decomposing a single user task into smaller subtasks and executing these subtasks concurrently across multiple processing units. This approach fully utilizes GPU resources to boost system concurrency and performance. Additionally, to optimize system efficiency and reduce computational load, we incorporated a caching mechanism. This mechanism stores previously computed results and reuses them when needed, thus avoiding redundant calculations and enhancing the system's response speed and throughput. Through implementation, this multi-user communication system based on an asynchronous processing model significantly improves resource utilization, processing speed, and system stability. It is suitable for scenarios requiring parallel processing of a large number of user communication tasks, with a high practical application value.

The single-user and multi-user systems aim to enhance the semantic quality of transmitted data by reducing noise and improving feature reconstruction at the receiver's end. The commonality between both systems lies in their shared generative AI framework, which enables progressive refinement of semantic information using a diffusion model, ensuring high fidelity and communication efficiency. The primary difference is that the MU-GSC system introduces additional mechanisms to manage multiple users concurrently. While the GSC system is more suited for scenarios where a single user requires high-quality semantic transmission (e.g., real-time image transmission), one of the critical advantages of the MU-GSC system is its scalability.

IV. Simulation Results and Discussions

In this section, simulation results are presented to evaluate the performance of our proposed single-user and multi-user generative AI semantic communication systems. They are compared with deep learning-based methods and the traditional communication systems realized by the separate source and channel coding technologies. In addition, to demonstrate the robustness of the proposed method, the simulation experiments are performed under the AWGN channels and Rayleigh fading channels, where the perfect CSI is assumed for all methods. For the MU-GSC, the transceiver is assumed with three single-antenna users and the receiver with three antennas.

This article has been accepted for publication in IEEE Transactions on Cognitive Communications and Networking. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2025.3526839

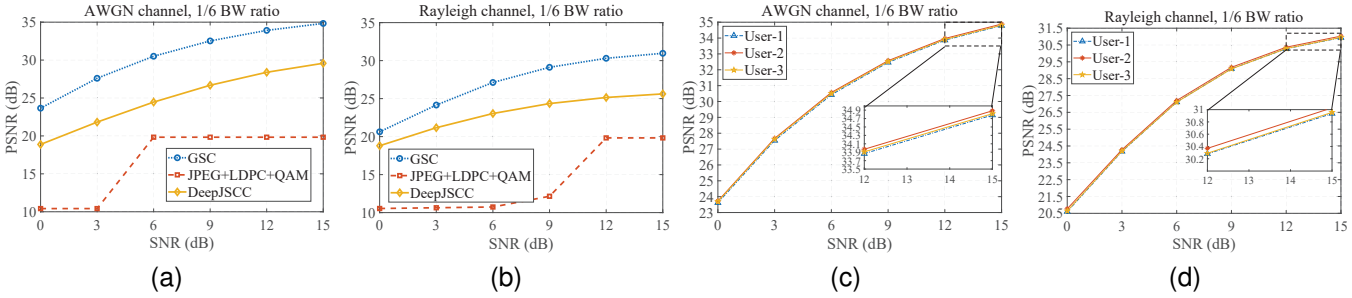


Fig. 4. PSNR (higher the better) versus the SNR over AWGN and Rayleigh channel, respectively. (a) AWGN channel. (b) Rayleigh channel. (c) MU-GSC AWGN channel. (d) MU-GSC Rayleigh channel.

A. Simulation setup

The training process is divided into two independent parts: semantic feature encoder/decoder and SFT. To comply with the generalization of ISC, the classical communication network based on the Swin Transformer is chosen as the semantic feature coder in this paper. Specifically, this network uses the ReLU activation function between the input layer and the two hidden layers, the hyperparameter C is set to 32, the loss function adopts the MSE, the batch size is set to 32, and the window size is set to 2, and then it is trained using Adam's optimizer [38] with a learning rate of 1×10^{-4}. During the SFT training process, we adopt a two-step training strategy utilizing a four-level encoder-decoder structure in the DTBN. From level 1 to level 4, the attention heads in DMTA are set to $[1, 2, 4, 8]$, the number of dynamic transformer blocks to $[3, 5, 6, 6]$ and the number of channels C' of $\boldsymbol{N}_1$ is set to 96. Training begins with a patch size of 32×32 and a batch size of 16. In the second step, β_t linearly increases from 0.10 to 0.99, starting with an initial learning rate of 2×10^{-4}, and the total timesteps T are set at 4.

In this simulation, the experimental platform for training and testing is built on an Ubuntu 20.04 system with CUDA 11.8 support, and the deep learning framework is Pytorch 2.0.0. Note that the training phase of the diffusion model is done at the cloud server of the BS, and the receiver is only involved in the reverse inference process. Compared to the pre-training phase, semantic information generation requires lower computational resources. Therefore, the receiver is sufficient to run these modules with acceptable computational latency.

We compare our proposal with classical separation-based source and semantic communication. The traditional communication model considers the well-established image codec JPEG, an image compression algorithm used in various applications such as Internet content delivery, digital photography, and medical imaging. Many fields include Internet content delivery, digital photography, and medical imaging. The channel noise or fading is then processed using a Low-Density Parity-Check code (LDPC) and Quadrature Amplitude Modulation (QAM) scheme, denoted as JPEG+LDPC+QAM. The modulation order is set to 4. The semantic communication model employs the classical algorithm DeepJSCC. All methods consistently use the CIFAR-10 dataset and are selected to be trained with SNRs between 1 dB and 13 dB.

B. Performance comparison

To demonstrate the effectiveness of the proposed method, this section analyzes the quality of the images generated by the three methods under two different channel conditions. Specifically, we present the Peak Signal-to-Noise Ratio (PSNR) of the decoded image obtained using these approaches.

For the proposed method, a single model covers a range of SNR from 0 dB to 15 dB. As expected, the proposed algorithms agree with the trend of PSNR variation in semantic communication with increased SNR. Fig. 4a shows the PSNR score versus the SNR achieved by using the proposed GSC and the benchmark systems under the AWGN channels. Note that for the traditional method, i.e., JPEG+LDPC+QAM, When the channel deteriorates beyond a threshold (SNR$<$3), the receiver cannot decode the channel code and, therefore, cannot transmit any semantic information. Comparatively, when the SNR$>$6, the PSNR reaches the performance saturation of traditional communication algorithms, and the image similarity score of these methods in Fig. 4 almost converges to 20, and further enhancement will not improve the output quality. However, with the reduction of SNR, the performance of the traditional methods is significantly degraded and is obviously poorer than that of the semantic communication systems. Our proposed system shows the same behavior as the DeepJSCC method. However, since the semantic information can be enhanced by the fine-tuning module, it is clear that the proposed GSC is more competitive than the DeepJSCC in the low SNR regime. Taking SNR=15 as an example, our method achieves a significant improvement in PSNR compared to the benchmark DeepJSCC algorithm, with 17.75% of the enhancement observed in AWGN channel and 20.86% in Rayleigh channels.

Fig. 4b shows the PSNR score versus the SNR achieved using GSC and the benchmark systems under Rayleigh fading channels. Fig. 4b exhibits the same behavior as that of Fig. 4a. Despite the more demanding harsh Rayleigh channel conditions, the GSC shows advantages in semantic communication. This observation can be attributed to the fact that, although LDPC codes and QAM modulation enhance the robustness of data transmission, JPEG compression algorithms are lossy and may lead to irreversible information loss, which reduces the fault-tolerance of the whole system, resulting in degradation of image quality or data integrity in the presence of transmission errors. Meanwhile, the PSNR of JPEG+LDPC+QAM

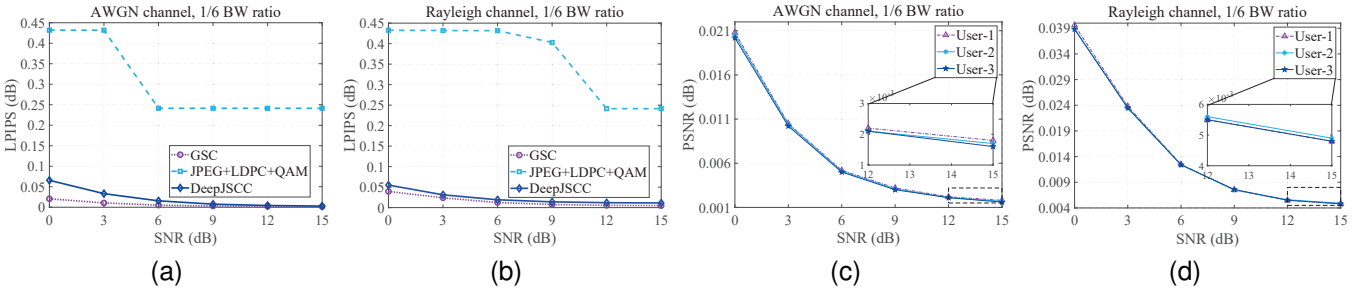


Fig. 5. LPIPS (lower the better) versus the SNR over AWGN and Rayleigh channel, respectively. (a) AWGN channel. (b) Rayleigh channel. (c) MU-GSC AWGN channel. (d) MU-GSC Rayleigh channel.

TABLE I
RUNTIME COMPARISON OF VARIOUS DEEP LEARNING-BASED SYSTEMS

Models	GSC	MU-GSC
Runtime (s)	15.66	8.87

remains constant when the channel conditions are very bad or substantially improved, which reflects the so-called cliff effect. In addition, in the forward channel, SNR ranges from 0 dB to 6 dB. Compared to the transmission performance of JPEG+LDPC+QAM, the GSC does not degrade rapidly, demonstrating a significant graceful degradation advantage. Compared with DeepJSCC, the proposed method has a lower performance gap at lower SNR, indicating that even if the received semantic information image is severely corrupted, it can still generate a realistic image consistent with the original transmitted semantic information. It is clear that our proposed system is more competitive and robust in poor channel environments.

Figs. 4c-4d show the PSNR score versus the SNR over the AWGN and Rayleigh channel, respectively, by using the proposed MU-GSC. Using the communication process of three users as an example, it is observed that the PSNR score trends similarly to the GSC. This is because the dataset and simulation setup for the multi-user system are the same as those used for the single-user system and have the same performance advantages in semantic extraction and compression. Our proposed multi-user system outperformed in all regimes and is close to 35 when SNR = 15 dB. Meanwhile, the PSNR score of the MU-GSC means that the semantic information contained in messages is transmitted successfully.

To more accurately assess the visual similarity of the delivered images, Fig. 5 compares the Learned Perceptual Image Patch Similarity (LPIPS) scores versus the SNR achieved by using the proposed method and the benchmark systems under different channel conditions. Unlike the PSNR metric, which primarily measures pixel-level differences, LPIPS focuses on visual perceptual differences, where a lower LPIPS value indicates better visual perceptual quality. Figs. 5a-5b show that the proposed GSC method outperforms existing methods in terms of LPIPS. Figs. 5c-5d indicate that the extracted and transmitted semantic information from MU-GSC is highly compact, preserves perceptual content, and proves that the proposed method is highly robust to noise interference. Specifically, compared to the benchmark algorithm DeepJSCC, in the low SNR regime (SNR = 0 dB), the LPIPS score decreased by 68.65% in the AWGN channel and by 28.18% in the Rayleigh channel.

To verify the role of the semantic fine-tuning module, we performed ablation tests to corroborate the proposed method, comparing the image transmission performance of the proposed GSC method with Non-Generative Framework (NGF). The NGF provides a direct SemCom receiver, similar to GSC, but with the semantic fine-tuning module turned off. Taking the example in AWGN channel transmission, Fig. 6 allows for a visual inspection of the details of the delivered images produced by the proposed method GSC and NGF at different SNRs of the wireless channel. It is evident that as the SNR improves (From left to right, the SNR ranges from 0 dB to 15 dB), the images from both frameworks transition from a mottled, mosaic-like appearance to a smoother texture. Additionally, the structure of the GSC images becomes nearly identical to the source image. However, images produced by NGF display prominent issues such as spot noise and edge blurring. These issues arise because the semantic fine-tuning module effectively reduces channel noise and preserves valuable semantic information. Meanwhile, the GSC relies heavily on the image communication network, which operates under strict constraints to ensure the reliability of the transmitted images. The results further demonstrate that the proposed algorithm is effective and robust in generating high-quality images.

To evaluate the convergence performance of the proposed algorithm, we set different training epochs during the diffusion model's training process. As shown in Fig. 7, GSC demonstrates excellent convergence speed. Specifically, the performance of the semantic fine-tuning module begins to improve significantly after seven epochs, indicating that the model has quickly captured key features in the early stages. When the number of epochs exceeds ten, the target loss stabilizes, suggesting that the model has converged. As training progresses beyond fifteenth epochs, the loss shows almost no further change, validating the model's stability and efficiency. This superior convergence speed can be attributed to executing the DM on a compact one-dimensional vector PR, making it more efficient than traditional DM.

The computational complexities of the proposed GSC and MU-GSC are compared in TABLE I in terms of the aver-

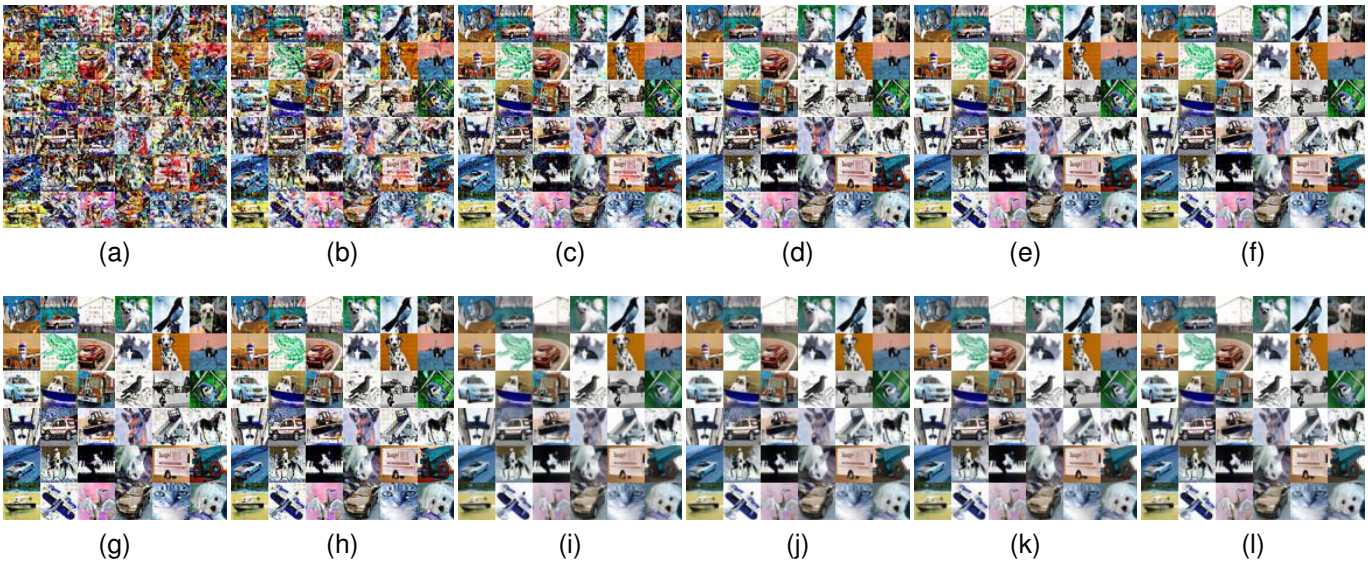


Fig. 6. Image detail delivery of two transmission frameworks under different SNRs in AWGN channel. (a-f) NGF. (b-l) GSC.

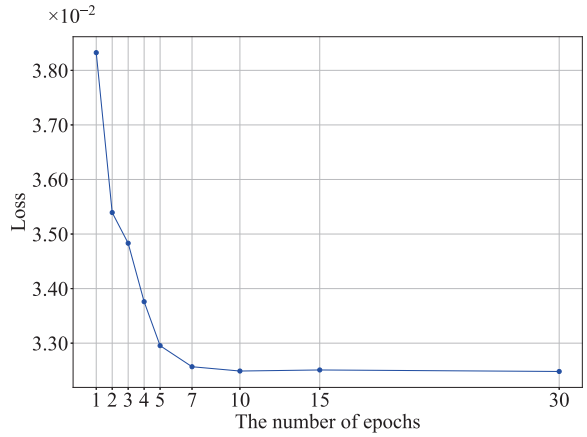


Fig. 7. Convergence experiment of the number of epochs in semantic fine-tuning module.

age processing runtime per image. This simulation is performed with Intel Core i9-13900HX@2.20 GHz and NVIDIA GeForce RTX 4060. Compared to GSC, MU-GSC completes processing multiple user requests in nearly half the time, significantly enhancing system response speed and throughput. This improvement is attributed to the application of critical technologies such as asynchronous task processing, parallel task execution, and caching mechanisms, effectively reducing the system's computational load. However, the runtime of the proposed method is high, the reason is that the computational complexities of our proposed semantic fine-tuning algorithm increase with the size of the users' knowledge base, and this marginal cost is offset by the significant performance improvements.

V. Conclusion

This paper proposes a generative AI semantic communication system that introduces an advanced and interpretable semantic fine-tuning module to enhance semantic information. Simulation results demonstrate that our method delivers superior transmission quality compared to traditional separation-based methods and DeepJSCC, significantly improving communication services in resource-limited wireless networks, particularly under low signal-to-noise ratio conditions. Although our method's running time is slightly higher than that of DeepJSCC, fundamental techniques such as asynchronous processing substantially enhance the system's response speed and throughput, demonstrating the scalability of the proposed single-user system in multi-user scenarios. Looking ahead, the challenge of adapting to rapidly changing data distributions remains an open and fundamental question in generative AI semantic communication. Future research will aim to explore more dynamic, context-aware solutions to optimize model adaptability and performance in increasingly complex and variable environments.

References

[1] G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, and X. Shen, "Semantic communications for artificial intelligence generated content (AIGC) toward effective content creation," *IEEE Netw.*, Early Access, Jan. 2024, DOI: 10.1109/MNET.2024.3352917.

[2] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less Data, More Knowledge: Building Next Generation Semantic Communication Networks," *IEEE Commun. Surveys Tuts.*, Early Access, Jun. 2024, DOI: 10.1109/COMST.2024.3412852.

[3] K. Zhang, L. Li, W. Lin, Y. Yan, W. Cheng, and Z. Han, "SC-CDM: Enhancing Quality of Image Semantic Communication with a Compact Diffusion Model," in *Proc. IEEE Glob. Commun. Conf. Workshops (GC Wkshps)*, Cape Town, South Africa, Dec. 2024.

[4] W. Lin, Y. Yan, L. Li, Z. Han, and T. Matsumoto, "SemantIC: Semantic Interference Cancellation Towards 6G Wireless Communications," *IEEE Commun. Lett.*, Early Access, Jun. 2024, DOI: 10.1109/LCOMM.2024.3412973.

[5] W. Lin, Y. Yan, L. Li, Z. Han, and T. Matsumoto, "Semantic-Forward Relaying: A Novel Framework Towards 6G Cooperative Communications," *IEEE Commun. Lett.*, vol. 28, no. 3, pp. 518-522, Mar. 2024.

[6] R. Cheng, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "A Wireless AI-Generated Content (AIGC) Provisioning Framework Empowered by Semantic Communication," *arXiv preprint arXiv:2310.17705*, Oct. 2023.

[7] S. Laskaridis, S. I. Venieris, A. Kouris, R. Li, and N. D. Lane, "The future of consumer edge-ai computing," *arXiv preprint arXiv:2210.10514*, Oct. 2022.

[8] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567-579, Sept. 2019.

[9] M. Ding, J. Li, M. Ma, and X. Fan, "SNR-adaptive deep joint source-channel coding for wireless image transmission," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 1555-1559, Toronto, ON, Canada, Jun. 2021.

[10] M. Yang and H. S. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 5193-5197, Singapore, May 2022.

[11] 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, May 2021.

[12] 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.

[13] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 170-185, Jan. 2023.

[14] K. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang, "WITT: A wireless image transmission transformer for semantic communications," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023.

[15] K. Yu, Q. He, and G. Wu, "Two-Way Semantic Communications without Feedback," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 9077-9082, Jun. 2024.

[16] L. X. Nguyen, Y. L. Tun, Y. K. Tun, M. N. H. Nguyen, C. Zhang, Z. Han, and C. S. Hong, "Swin transformer-based dynamic semantic communication for multi-user with different computing capacity," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8957-8972, Jun. 2024.

[17] S. Kadam and D. I. Kim, "Knowledge-aware semantic communication system design and data allocation," *IEEE Trans. Veh. Technol.*, vol. 73, no. 4, pp. 5755-5769, Apr. 2024.

[18] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584-2597, Sept. 2022.

[19] J. Kang, H. Du, Z. Li, Z. Xiong, S. Ma, D. Niyato, and Y. Li, "Personalized saliency in task-oriented semantic communications: Image transmission and performance analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 186-201, Jan. 2023.

[20] M. K. Farshbafan, W. Saad, and M. Debbah, "Common language for goal-oriented semantic communications: A curriculum learning framework," in *Proc. IEEE Int. Conf. Commun. (ICC)*, pp. 1710-1715, Seoul, South Korea, May 2022.

[21] Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance optimization for semantic communications: An attention-based reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2598-2613, Sept. 2022.

[22] C. K. Thomas and W. Saad, "Neuro-symbolic causal reasoning meets signaling game for emergent semantic communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4546-4563, May 2024.

[23] H. Yoo, T. Jung, L. Dai, S. Kim, and C.-B. Chae, "Demo: Real-Time Semantic Communications with a Vision Transformer," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Wkshps)*, Seoul, South Korea, May 2022.

[24] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 97, pp. 1182-1192, Long Beach, California, USA, Jun. 2019.

[25] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications with masked VQ-VAE enabled codebook," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8707-8722, Dec. 2023.

[26] T. Marchioro, N. Laurenti, and D. Gündüz, "Adversarial networks for secure wireless communications," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 8748-8752, Barcelona, Spain, May 2020.

[27] M. Yang, C. Bian, and H. S. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 584-599, Jun. 2022.

[28] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, and N. Rajatheva, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, vol. 11, pp. 37149-37163, Apr. 2023.

[29] E. Erdemir, T. Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2645-2657, Aug. 2023.

[30] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. V. Gool, "Diffir: Efficient diffusion model for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 13095-13105, Paris, France, Oct. 2023.

[31] X. Niu, X. Wang, D. Gündüz, B. Bai, W. Chen, and G. Zhou, "A hybrid wireless image transmission scheme with diffusion," in *Proc. IEEE 24th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pp. 86-90, Shanghai, China, Sept. 2023.

[32] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, "CDDM: Channel denoising diffusion models for wireless communications," in *Proc. IEEE Global Commun. Conf.*, pp. 7429-7434, Kuala Lumpur, Malaysia, Dec. 2023.

[33] B. Xu, R. Meng, Y. Chen, X. Xu, C. Dong, and H. Sun, "Latent semantic diffusion-based channel adaptive de-noising semcom for future 6G systems," in *Proc. IEEE Global Commun. Conf.*, pp. 1229-1234, Kuala Lumpur, Malaysia, Dec. 2023.

[34] E. Grassucci, C. Marinoni, A. Rodriguez, and D. Comminiello, "Diffusion models for audio semantic communication," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 13136-13140, Seoul, South Korea, Apr. 2024.

[35] J. Chen, D. You, D. Gündüz, and P. L. Dragotti, "Commin: Semantic image communications as an inverse problem with INN-guided diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 6675-6679, Seoul, South Korea, Apr. 2024.

[36] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," *arXiv preprint arXiv:2306.04321*, Jun. 2023.

[37] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992-10002, Montreal, QC, Canada, Oct. 2021.

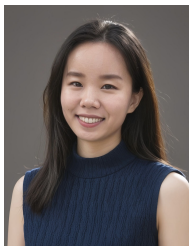
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

Kexin Zhang is currently pursuing her Ph.D. under the supervision of Prof. Lixin Li at the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. Her research interests include deep learning, semantic communications, and large language models.

Yuna Yan is currently pursuing her Master's degree at the school of Electronics and Information, Northwestern Polytechnical University. Her research interests include federated learning, deep learning and semantic communications.

Lixin Li (M'12) received his BS degree (Hons.), MS degree (Hons.), and PhD degree from Northwestern Polytechnical University (NPU), China in 2001, 2004, and 2008, respectively. From 2009 to 2011, he was a postdoctoral research fellow with NPU. In 2011, he joined the School of Electronics and Information, NPU, where he is currently a full professor and chair of Department of Communication Engineering. In 2017, he held the visiting scholar position with the University of Houston, USA. He holds 26 granted patents and has authored or coauthored five books and more than 200 peer-reviewed papers in many prestigious journals and conferences. His research interests include 5G/6G wireless networks, federated learning, game theory, and machine learning for wireless communications. He was the recipient of the 2016 NPU Outstanding Young Teacher Award, which is the highest research and education honors for young faculties in NPU. He was an exemplary reviewer for IEEE Transactions on Communications in 2020.

Wensheng Lin received the B.Eng. degree in communication engineering, and the M.Eng. degree in electronic and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2013 and 2016, respectively, and the Ph.D. degree in information science from Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, in 2019. His doctoral study was fully funded by the Chinese Government Scholarship awarded from the China Scholarship Council (CSC). He is currently an Associate Professor with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. He was a recipient of the prestigious 2019 Niwa Yasujiro Outstanding Paper Award, and the 2019 IEEE Nagoya Section Conference Presentation Award. He was selected into the 8th Young Elite Scientists Sponsorship Program by the China Association for Science and Technology (CAST). His research interests include network information theory, distributed source coding, and semantic communications.



Rui Li received BEng in Communications Engineering from Northwestern Polytechnical University, China, in 2013, during which and MSc in Embedded System and Control Engineering from Department of Engineering, University of Leicester, UK, 2014. She was awarded Ph.D. in Informatics by School of Informatics, University of Edinburgh, UK, in 2019, and her thesis was on resource optimization in millimeter-wave backhaul networks using deep learning. During her study as a Ph.D. student, she received the Brendan Murphy Memorial Prize at the 30th Next Generation Networking Multi-Service Networks conference, UK, and Best Student Paper Award at International Conference on Machine Learning for Networking, Paris, France. She was a visiting researcher at The National Institute for Research in Digital Science and Technology (INRIA), Lyon, France. Rui joined Samsung AI Center in Cambridge, UK, as a post-doctoral researcher in 2019, and currently she is a Research Scientist. Rui received Samsung Best Paper Silver Award in 2024. Her research interests include efficient inference of LLM and stable diffusion models, personalization of ML models under resource-constrained environments, and applications of AI in various domains including communications, vision, and natural language processing.

Wenchi Cheng (M'14–SM'18) received the B.S. and Ph.D. degrees in telecommunication engineering from Xidian University, Xian, China, in 2008 and 2013, respectively. He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, from 2010 to 2011. He is currently a Full Professor with Xidian University. He has published more than 200 international journal and conference papers in IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE MAGAZINES, IEEE TRANSACTIONS, IEEE INFOCOM, GLOBECOM, and ICC. His current research interests include 6G wireless networks, electromagnetic-based wireless communications, and emergency wireless information. He received the IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award in 2021, the URSI Young Scientist Award in 2019, the Young Elite Scientist Award of CAST, and four IEEE journal/conference best papers. He has served or serving as the IEEE GLOBECOM 2026 Tutorial Co-Chair, the Wireless Communications Symposium Co-Chair for IEEE ICC 2022 and IEEE GLOBECOM 2020, the Publicity Chair for IEEE ICC 2019, the Next Generation Networks Symposium Chair for IEEE ICCC 2019, and the Workshop Chair for IEEE ICC 2019/IEEE GLOBECOM 2019/INFOCOM 2020 Workshop on Intelligent Wireless Emergency Communications Networks. He has served or serving as the ComSoc Representative for IEEE Public Safety Technology Initiative, IEEE ComSoc Distinguished Lecturer, Editor for IEEE Transactions on Wireless Communications, IEEE System Journal, IEEE Communications Letters, and IEEE Wireless Communications Letters.

Han Zhu (S'01–M'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Han's main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, carbon neutralization, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, IEEE Vehicular Technology Society 2022 Best Land Transportation Paper Award, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and ACM Distinguished Speaker from 2022 to 2025, AAAS fellow since 2019, and ACM Fellow since 2024. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."