This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

Visual Language Model based Cross-modal Semantic Communication Systems

Feibo Jiang, *Senior Member, IEEE*, Chuanguo Tang, Li Dong, Kezhi Wang, *Senior Member, IEEE*, Kun Yang, *Fellow, IEEE*, Cunhua Pan, *Senior Member, IEEE*

***Abstract*—Semantic Communication (SC) has emerged as a novel communication paradigm in recent years. Nevertheless, extant Image Semantic Communication (ISC) systems face several challenges in dynamic environments, including low information density, catastrophic forgetting, and uncertain Signal-to-Noise Ratio (SNR). To address these challenges, we propose a novel Vision-Language Model-based Cross-modal Semantic Communication (VLM-CSC) system. The VLM-CSC comprises three novel components: (1) Cross-modal Knowledge Base (CKB) is used to extract high-density textual semantics from the semantically sparse image at the transmitter and reconstruct the original image based on textual semantics at the receiver. The transmission of high-density semantics contributes to alleviating bandwidth pressure. (2) Memory-assisted Encoder and Decoder (MED) employ a hybrid long/short-term memory mechanism, enabling the semantic encoder and decoder to overcome catastrophic forgetting in dynamic environments when there is a drift in the distribution of semantic features. (3) Noise Attention Module (NAM) employs attention mechanisms to adaptively adjust the semantic coding and the channel coding based on SNR, ensuring the robustness of the CSC system. The experimental simulations validate the effectiveness, adaptability, and robustness of the CSC system.**

***Index Terms*—Semantic communication, knowledge base, vision language model, large language model, continual learning.**

I. Introduction

As mobile communication technology has evolved from the first generation to the fifth generation, there has been a significant increase in transmission rates, approaching system capacities close to their limits [1]. In recent years, various emerging applications, such as the metaverse and virtual reality, have introduced substantial data streams [2]. Furthermore, these applications necessitate extensive connectivity over limited spectrum resources while demanding lower latency, posing significant challenges to conventional source-channel coding. Semantic Communication (SC) operates in the semantic domain by extracting the inherent meaning of data, eliminating redundant information, and achieving data compression while preserving its essential semantic content [3].

With the rapid development of deep learning, many researchers have begun to explore end-to-end Image Semantic Communication (ISC) systems based on Deep Neural Networks (DNN). For instance, ISC systems constructed using deep learning approaches such as Convolutional Neural Networks (CNN), Vision Transformers (ViT), and others have surpassed traditional solutions. Despite the significant achievements in the research of ISC based on deep learning, there remain some challenges:

1) Low information density: Information density is defined as the ratio of the amount of semantic information to the amount of raw data [4]. Images are natural signals with heavy spatial redundancy. Traditional ISC systems directly encode the entire image, focusing on extracting low-level semantic information at the pixel level. However, text is a human-invented signal that possesses high information density. Summarizing image information through text can surpass the low-level pixel-level semantics and achieve a more sophisticated high-level semantic understanding of objects and scenarios. Moreover, traditional ISC systems lack the ability to leverage the interpretability of knowledge bases (KBs), resulting in a black-box model based on deep learning for the semantic encoder and decoder with limited explainability of semantics.

2) Catastrophic forgetting: ISC systems often operate in dynamic environments, leading to a drift in the feature distribution of transmitted image data and channel state over time. Consequently, the real data distribution becomes inconsistent with the distribution during training, resulting in a decline in the performance of the semantic encoder and decoder. Continual learning of the semantic encoder and decoder is necessary to improve the performance of the ISC system. However, during continual learning, the existing knowledge of the encoder and decoder may be disrupted or overwritten by new knowledge, leading to catastrophic forgetting in the learning process [5]. As a result, it becomes unable to adapt to semantic transmission in dynamic environments.

This work was supported in part by the National Natural Science Foundation of China under Grants 41604117, 41904127, and 62132004, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2024JJ5270, in part by the Open Project of Xiangjiang Laboratory under Grant 22XJ03011, in part by the Scientific Research Fund of the Hunan Provincial Education Department under Grant 22B0663, in part by the Changsha Natural Science Foundation under Grants kq2402098 and kq2402162, in part by the Jiangsu Major Project on Basic Researches under Grant BK20243059 and Gusu Innovation Project for under Grant ZXL2024360. (Corresponding authors: Chuanguo Tang, Li Dong.)

Feibo Jiang (jiangfb@hunnu.edu.cn) is with Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China.

Chuanguo Tang (202220294014@hunnu.edu.cn) is with School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China.

Li Dong (Dlj2017@hunnu.edu.cn) is with the School of Computer Science, Hunan University of Technology and Business, Changsha 410205, China, and also with the Xiangjiang Laboratory, Changsha 410205, China

Kezhi Wang (Kezhi.Wang@brunel.ac.uk) is with the Department of Computer Science, Brunel University London, UB8 3PH Uxbridge, UK.

Kun Yang (kunyang@essex.ac.uk) is with the School of Intelligent Software and Engineering, Nanjing University, Suzhou 215163, China

Cunhua Pan (cpan@seu.edu.cn) is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

3) Uncertain Signal-to-Noise Ratio (SNR): In wireless communications, traditional deep learning-based ISC systems typically consider a few discrete SNR conditions during the training phase, which cannot cover all possible SNR scenarios. As a result, the performance may severely degrade when there is a mismatch between the channel conditions during training and inference phases [6]. Training the semantic/channel encoder and decoder with consideration for multiple SNR conditions and performing switching based on specific SNR values during the inference phase can lead to substantial storage and computational overhead [7, 8].

Vision Language Models (VLMs) with billions of parameters represent the latest advancements in the field of large AI models. Through extensive pre-training on vast amounts of data, these VLMs acquire rich language and visual knowledge, leading to significant breakthroughs in areas such as natural language processing and computer vision [9]. In ISC systems, VLMs demonstrate immense potential. Leveraging their capabilities in understanding and generating textual and visual content, VLMs enable more accurate semantic comprehension and semantic feature extraction, thereby offering a more intelligent and efficient ISC experience. Therefore, we propose a novel VLM-based Cross-modal Semantic Communication (VLM-CSC) system to address the aforementioned challenges in ISC systems. Our contributions can be summarized as follows:

1) Cross-modal Knowledge Base (CKB): We introduce a CKB, which consists of a Bootstrapping Language-Image Pre-Training (BLIP)-based KB at the transmitter for generating high-quality text descriptions consistent with images, and a Stable Diffusion (SD)-based KB at the receiver for reconstructing images matching the text descriptions. The text descriptions can be regarded as the extraction of high-level semantics from the images with low-level pixels, thereby enhancing the information density of the transmitted information. Additionally, these descriptions enable users to understand the extracted semantic content, thereby enhancing the explainability of the CSC system.

2) Memory-assisted Encoder and Decoder (MED): We employ a MED to track changes in dynamic environments while avoiding catastrophic forgetting during the learning process. Specifically, we design a storage pool consisting of two types of memory: Short-Term Memory (STM) and Long-Term Memory (LTM). The STM is used to store the new data from the current environment, while the LTM stores historically significant data from previously encountered distributions. When training the CSC system, we input data from both the STM and LTM. This enables the semantic encoder and decoder to review all the knowledge from previously trained data with different distributions while learning from the new data. As a result, the CSC system can acquire encoding and decoding capabilities for the new data distribution without significantly compromising its performance on the previously trained data distribution, thus avoiding catastrophic forgetting.

3) Noise Attention Module (NAM): We present a NAM to dynamically adjust semantic coding and channel coding based on different SNR conditions. Specifically, after each encoder and decoder layer, we employ an attention module to adjust the weights for different encoders and decoders according to the SNR values provided by the channel feedback. When the SNR is high, the NAM evenly allocates higher weights to the semantic encoder and decoder to improve the encoding and decoding quality of the semantic features. Conversely, when the SNR is low, the NAM assigns higher weights to the channel encoder and decoder, improving the channel coding to combat the intense channel noise. This design ensures that the semantic features maintain high robustness under varying SNR conditions.

The rest of this paper is structured as follows. Section II presents the related work, Section III introduces the system model, Section IV provides a detailed description of the proposed VLM-CSC system, Section V outlines the experimental setup and results, and Section VI concludes the paper.

II. Related work

A. Deep learning enabled ISC systems

Deep learning techniques are commonly employed in the construction of encoders and decoders for ISC systems. In [10], a comprehensive SC system based on CNNs was initially introduced, showcasing superior performance in Peak Signal-to-Noise Ratio (PSNR) when compared to traditional compression algorithms. In [11], a novel Nonlinear Transform Source-Channel Coding (NTSCC) for SC systems was proposed, which leveraged a Variational AutoEncoder (VAE) to map the source signal to the latent space, and executed nonlinear transformation and channel coding in the space. Additionally, [12] presented an innovative SC system incorporating Semantic Slice-Models (SeSM) to facilitate adaptable model resemblance under diverse requirements. Furthermore, [13] introduced a Reinforcement Learning-based Adaptive Semantic Coding (RL-ASC) for image data. RL-ASC utilized a combination of VAE, RL, and generative adversarial networks (GANs) to encode, allocate, and decode semantic concepts.

Although convolutional and ViT-based autoencoders have shown promising results, their feature extraction capabilities are limited compared to state-of-the-art VLMs. This limitation arises from constraints posed by model parameters and the availability of training data.

B. Vision language models

VLMs are a class of large AI models capable of simultaneously processing both image and text information [14]. They find extensive application across various visual language tasks, encompassing image description, visual question answering, text-to-image generation, and other multimodal tasks. In [15], a contrastive loss function was utilized to train both image encoders and text encoders. This loss function aimed to minimize the feature space distance between matching image-text pairs, enabling the learning of semantically relevant visual language features while reducing the dependence on large amounts of annotated data. In [16], images were treated as prefixes in language models. They were decomposed into multiple blocks, concatenated with text sequences as input, and used to predict the subsequent parts of the text sequences. Furthermore, in [17], a cross-attention mechanism was employed to integrate visual and language features. This mechanism allowed the two

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

modalities to reference and enhance each other, facilitating the learning of more comprehensive and refined visual language features. The approach demonstrated applicability to various downstream tasks.

C. Continual learning

Continual learning can effectively mitigate the problem of catastrophic forgetting in dynamic environments[18]. In [19], the authors discussed continual learning in Mobile Edge Computing (MEC) networks, focusing on age-aware optimization for data selection and aggregator placement. They also presented a prototype implementation involving diverse user equipment and cloudlets. In [20], the authors proposed a continual learning digital predistortion algorithm for linearizing radio frequency power amplifiers in 6G wireless communications. The algorithm demonstrated effectiveness in adapting to both new and known operating states with low long-term complexity. In [21], the authors addressed the challenge of forgetting tasks in cross-edge federated learning by preserving past knowledge through continual learning. They achieved enhanced accuracy across various tasks with minimal storage cost. Furthermore, in [22], the authors employed continual learning to enable adaptive downlink beamforming optimization in dynamic environments. The proposed approach addressed task mismatch and exhibits good adaptability with low complexity.

Recent advancements in continual learning have been directed towards more challenging scenarios, specifically those where task boundaries are unknown. In these contexts, researchers have focused on developing sample selection strategies to identify which samples should be stored in the buffer for model training. This approach aims to improve the efficiency and effectiveness of continual learning in handling unknown task boundaries.

III. System model and problem formulation

The considered CSC system consists of three components: a transmitter, a receiver, and a physical channel, as illustrated in **Fig. 1**. The physical channel ensures the correct exchange of semantic information over the transmission medium with dynamic SNR.

A. Transmitter

The input to the transmitter is an image represented by the matrix $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, whose size is $H(height) \times W(weight) \times C(channel)$. In the transmitter, the input image $\mathbf{x}$ is mapped to symbols $\mathbf{y}$ for transmission over the physical channel. The transmitter consists of three independent components: a CKB for cross-modal semantic extraction, a semantic encoder, and a channel encoder. The CKB is used to extract semantic information from the image and represent it as the corresponding textual information. The semantic encoder and channel encoder are responsible for semantic coding, and channel coding and modulation, ensuring that the encoded semantic information can be smoothly transmitted over the physical channel. The encoded symbol sequence $\mathbf{y}$ can be represented as:

$$\mathbf{y} = C_\beta(S_\alpha(K_\theta(\mathbf{x}), \mu), \mu) \tag{1}$$

where $K_\theta(\cdot)$ is the CKB with the parameter set θ, $S_\alpha(\cdot)$ is the semantic encoder with the parameter set α, and $C_\beta(\cdot)$ is the channel encoder with the parameter set β, μ is the channel SNR that can be estimated and fed back to the semantic encoder and channel encoder.

B. Wireless channel

The transmitter sends encoded symbols $\mathbf{y}$, which is transmitted through the physical channel to the receiver. The channel output sequence $\hat{\mathbf{y}}$ at the receiver can be expressed as:

$$\hat{\mathbf{y}} = \mathbf{h} \cdot \mathbf{y} + \mathbf{n} \tag{2}$$

where $\mathbf{h}$ represents the channel gain, and $\mathbf{n}$ is Additive White Gaussian Noise (AWGN).

C. Recevier

Similar to the transmitter, the receiver consists of three components: a channel decoder, a semantic decoder, and a cross-modal knowledge base for semantic reconstruction. The semantic decoder and channel decoder are used to decode textual information from received symbols, while the cross-modal knowledge base is employed for image reconstruction based on the corresponding textual information. The decoded image can be represented as:

$$\hat{\mathbf{x}} = K^{-1}_{\theta'}(S^{-1}_\delta(C^{-1}_\gamma(\hat{\mathbf{y}}, \mu), \mu)) \tag{3}$$

where $C^{-1}_\gamma(\cdot)$ is the channel decoder with the parameter set γ, $S^{-1}_\delta(\cdot)$ is the semantic decoder with the parameter set δ and $K^{-1}_{\theta'}(\cdot)$ is the cross-modal knowledge base with the parameter set θ'.

For the purpose of reconstructing image information from the semantic level, maintaining the consistency of textual semantics between $\mathbf{s}$ and $\hat{\mathbf{s}}$ is crucial. Here, $\mathbf{s} = K_\theta(\mathbf{x})$ represents the extracted textual semantic information from the image, and $\hat{\mathbf{s}} = S^{-1}_\delta(C^{-1}_\gamma(\hat{\mathbf{y}}, \mu), \mu)$ represents the recovered textual semantic information after decoding. We utilize Cross-Entropy (CE) as the loss function [23]:

$$L_{CE}(\mathbf{s}, \hat{\mathbf{s}}) = -\sum_{l=1}^{L} q(w_l)\log(p(w_i)) + (1-q(w_l))\log(1-p(w_i)) \tag{4}$$

where $q(w_l)$ denotes the real probability of the appearance of the l-th word w_l in the sentence $\mathbf{s}$, and $p(w_l)$ represents the predicted probability of the appearance of the l-th word w_i in the sentence $\hat{\mathbf{s}}$. CE is employed to measure the difference between two probability distributions. By minimizing the CE loss, the semantic encoder and decoder can learn the word distribution $q(w_l)$ in the source sentence $\mathbf{s}$, which represents the meaning of words in terms of grammar, phrases, and contextual information. Hence, the goal pf the CSC system is to determine the parameters of the semantic/channel encoder

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

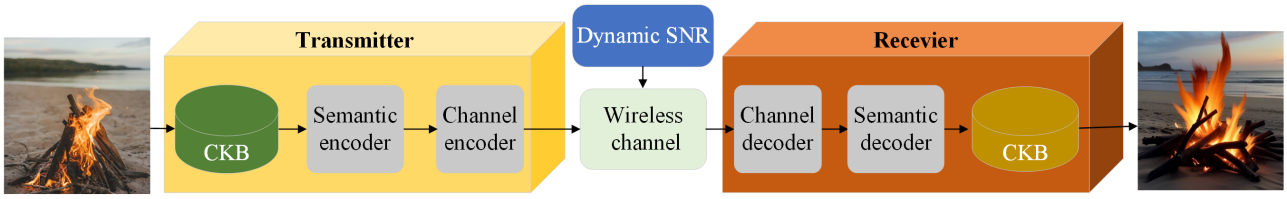

Fig. 1: The system model of the CSC.

and decoder α^*, β^*, δ^* and γ^* that minimize the expected distortion as follows:

$$(\alpha^*, \beta^*, \delta^*, \gamma^*) = \arg\min_{\alpha,\beta,\delta,\gamma} \mathbb{E}_{p(\mu)} \mathbb{E}_{p(\mathbf{s},\hat{\mathbf{s}})} [L_{CE}(\mathbf{s}, \hat{\mathbf{s}})] \quad (5)$$

where α^* is the optimal semantic encoder parameters, β^* is the optimal channel encoder parameters, γ^* is the optimal channel decoder parameters, and δ^* is the optimal semantic decoder parameters. $p(\mathbf{s}, \hat{\mathbf{s}})$ represents the joint probability distribution of the $\mathbf{s}$ and $\hat{\mathbf{s}}$, and $p(\mu)$ represents the probability distribution of the SNR.

IV. The VLM-CSC system

In this section, we will provide the implementation details of the proposed VLM-CSC system, which is illustrated in **Fig. 2** as follows:

1) Textual semantic extraction: To enhance the information density and interpretability of SC, a VLM called BLIP is employed at the transmitter to construct the CKB. The CKB encompasses a series of visual and language-related knowledge components. We employ the image encoder and text decoder from this CKB to perform cross-modal semantic extraction, thereby transforming the original image with low information density into a corresponding textual description with high information density. For example, through cross-modal semantic extraction, the original image in **Fig. 2** is transformed to the textual description "A fire is burning on a beach near the water".

2) Semantic encoder and decoder: The generated textual information from the CKB then proceeds to the semantic encoder. The semantic encoder consists of alternating transformer encoder layers and NAMs. The transformer encoder layers analyze and transform the textual information into a compact semantic representation. NAMs allow the semantic encoder to optimize the encoding process and maintain reliable semantic transmission, even in the presence of varying channel conditions. At the receiver, the semantic decoder is composed of alternating transformer decoder layers and NAMs, with a structure opposite to that of the semantic encoder, aimed at reversing the semantic encoding process to recover the original textual information.

3) Channel encoder and decoder: The encoded semantic features are passed through the channel encoder to undergo channel encoding and modulation, ensuring the effective transmission of semantic information over the physical channel. Similarly, the channel encoder also consists of alternating FeedForward (FF) layers and NAMs. At the receiver, the transmitted information through the physical channel is received and decoded using the channel decoder. To maintain information consistency, the channel decoder employs a structure opposite to that of the channel encoder.

4) Image reconstruction: To facilitate a better understanding of the received textual information, we design a CKB for image reconstruction using a VLM called SD. The CKB encompasses a series of visual and language-related knowledge components. We employ the text encoder, the denoising U-Net and the image decoder from this CKB to perform image reconstruction. Specifically, the textual information is first transformed into a conditional vector by the text encoder. Then, the denoising U-Net transforms the noisy image to a latent image feature vector aligning with the conditional vector. Finally, the latent image feature vector is processed by the image decoder to generate the final reconstructed image.

5) Memory-assisted continual learning: During the training phase of the VLM-CSC system, the latest samples are stored in an STM. When the STM becomes full, a kernel method is employed to select representative short-term samples to be transferred to an LTM. Then, the STM is emptied to buffer new samples in the next round. The encoder and decoder sample from both STM and LTM during the training stage, thereby avoiding catastrophic forgetting. This approach ensures that the semantic encoder and decoder can access both recent and past information, allowing for continual learning and retention of previously learned knowledge.

6) Training process of the VLM-CSC system: Remarkably, BLIP and SD-based CKBs are pretrained VLMs that do not need to be trained specifically for the CSC system. The training process unfolds as follows:

- Joint training of channel encoder and decoder with NAMs: The channel encoder/decoder and NAMs are initially trained together by MED. This involves optimizing the parameters of these modules by minimizing the mutual information, which eliminates noise or fading effects during transmission and prevents signal distortion [23]. Then, the parameters of the channel encoder/decoder and NAMs are frozen. This ensures that their learned representations are preserved in subsequent training steps.
- Joint training of semantic encoder and decoder with NAMs: The semantic encoder/decoder and NAMs are then trained by MED. The focus is on optimizing the parameters of these modules to minimize the loss between the original textual information and the reconstructed textual information. Eq. (4) can be applied as the loss function. Then, the parameters of the semantic encoder/decoder and NAMs are frozen to maintain the learned semantic representations.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

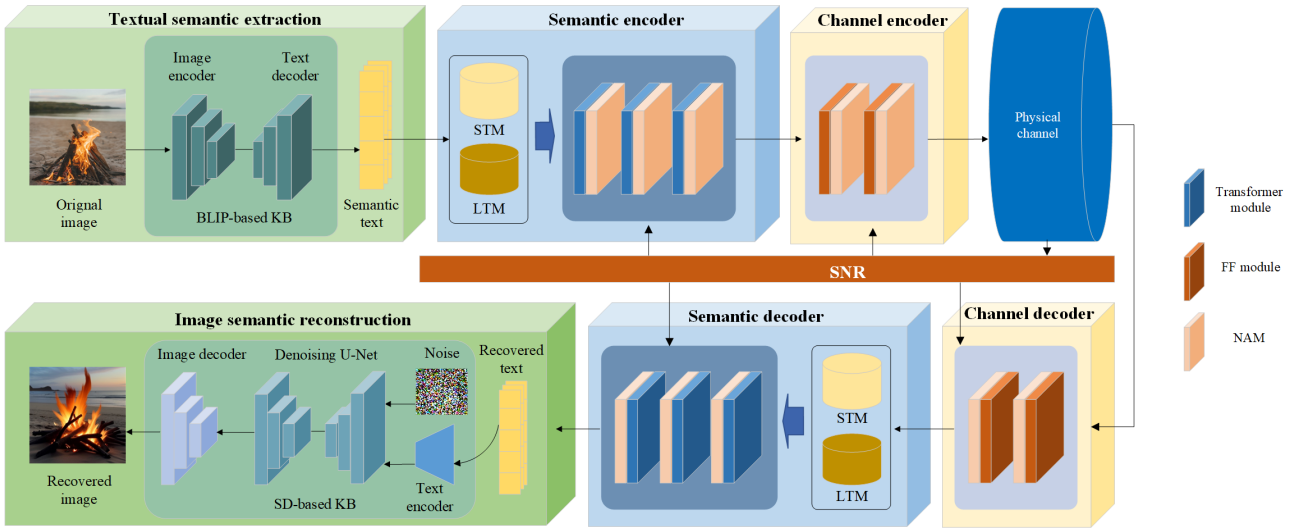

Fig. 2: The proposed VLM-CSC system.

- Crossover-based iterative training: The training process iterates between the channel encoder/decoder and noise modules, and the semantic encoder/decoder and noise modules. This iteration continues until convergence of the entire VLM-CSC system is achieved.

Next, we will provide a detailed explanation of each contribution in this paper.

A. BLIP-based CKB for semantic extraction

The BLIP model, introduced by Salesforce AI Research, is a sophisticated VLM designed for understanding and generating content that involves both visual and textual elements [24]. The BLIP model possesses rich visual-linguistic knowledge and utilizes multiple knowledge components such as text encoders, image encoders, and image-grounded text decoders and decoders to perform various visual-linguistic tasks, such as image captioning, visual question answering, and multimodal classification. At the transmitter, we employ the BLIP model to construct the CKB and utilize the image encoder and image-grounded text decoder (abbreviated as text decoder) in the CKB to transform original image data into detailed textual descriptions containing image semantic information. The workflow of the BLIP-based CKB is illustrated in **Fig. 3**.

For a given image $\mathbf{x}$, the process of extracting semantic information from image data and generating textual representation $\mathbf{s}$ is as follows:

1) Image encoder: The image encoder incorporates a feature extraction module based on the ViT. This module divides the input image into smaller patches and encodes each patch. Through multiple encoder layers with Multi-head Self-Attention (MSA) and FF sublayers [25], these patch vectors undergo processing to generate the textual representation of the image, which corresponds to the image features.

Initially, the image $\mathbf{x}$ is segmented into a patch sequence $\mathbf{x}_p$. Each patch represents a fixed-size image region in **Fig. 3**. Subsequently, these patch sequences are fed into the image encoder to extract visual features from the image. The specific workflow of the image encoder is as follows:

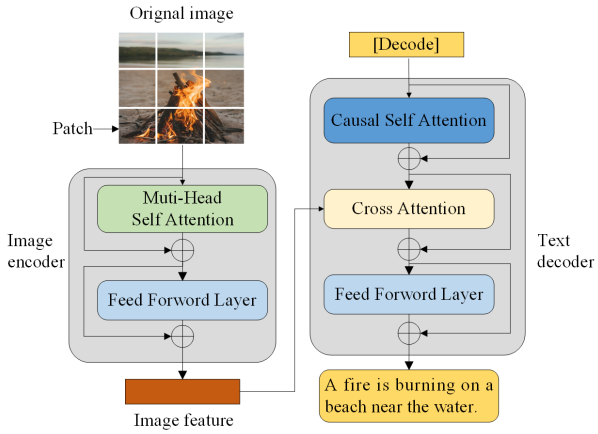

Fig. 3: The architecture of BLIP-based CKB.

- MSA sublayer: the MSA layer allows the vector of each patch to interact with vectors of all other patches, capturing both global and local information in the image. The output of the MSA layer in the first image encoder layer can be calculated as follows:

$$\mathbf{m}_{msa,1} = \mathrm{MSA}(\mathrm{LN}(\mathbf{x}_p)) + \mathbf{x}_p \tag{6}$$

where $\mathbf{x}_p$ is the p-th patch, MSA is the multi-head self-attention operator [25] and LN is the layer normalization operator in ViT [25].

- FF sublayer: The FF layer comprises linear layers and activation functions, facilitating non-linear transformations of vectors for each patch to enhance the model's adaptability. The output of the FF layer in the first image encoder layer is

$$\mathbf{m}_{ff,1} = \mathrm{GeLU}(\mathbf{W}_{b,f} \cdot \mathrm{LN}(\mathbf{m}_{msa,1}) + \mathbf{b}_{b,f}) + \mathbf{m}_{msa,1} \tag{7}$$

where $\mathbf{W}_{b,f}$ and $\mathbf{b}_{b,f}$ are the weights and biases of the FF layer in the image encoder of the BLIP model, and GeLU denotes the activation function.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

Finally, the output of the image encoder with L encoder layers is

$$\mathbf{m}_L = \mathrm{LN}(\mathbf{m}_{ff,L}) \tag{8}$$

where $\mathbf{m}_{ff,L}$ means the output of the L-th encoder layer.

2) Text decoder: The text decoder of the BLIP model adopts a BERT structure, capable of generating image-related textual content, such as descriptions, titles, and dialogues, based on features extracted from images. The text decoder is composed of multiple stacked decoder layers, each decoding layer comprising three sublayers: Causal Self-Attention (CSA), Cross Attention (CA), and FF sublayers. The specific workflow of the text decoder is as follows:

- CSA sublayer: CSA is a type of self-attention mechanism that only allows the attention model to access the current and previous inputs, but not the future inputs [26]. To ensure the causality of the textual generation process, the CSA sublayer utilizes a mask matrix to prevent the current token from accessing information from future tokens. Here, a token refers to the basic unit in the text, typically a word or a subword. The output of the CSA sublayer in the first text decoder layer is

$$\mathbf{k}_{csa,1} = \mathrm{CSA}(\mathrm{LN}(D_0)) + D_0 \tag{9}$$

  where CSA is the causal self-attention operator [26], D_0 is the initial token, which is typically set as "[Decoder]" by default.
- CA sublayer: CA allows the vector of each token to interact with the feature vectors of visual information from the input image [27]. The output of the CA sublayer in the first text decoder layer can be calculated as follows:

$$\mathbf{k}_{ca,1} = \mathrm{CA}(\mathrm{LN}(\mathbf{k}_{csa,1}), \mathbf{m}_L) + \mathbf{k}_{csa,1} \tag{10}$$

  where CA is the cross attention operator [27].
- FF sublayer: The FF layer comprises linear layers and activation functions. The output of the FF layer in the first text decoder layer is

$$\mathbf{k}_{ff,1} = \mathrm{ReLU}(\mathbf{W}'_{b,f} \cdot \mathrm{LN}(\mathbf{k}_{ca,1}) + \mathbf{b}'_{b,f}) + \mathbf{k}_{ca,1} \tag{11}$$

  where $\mathbf{W}'_{b,f}$ and $\mathbf{b}'_{b,f}$ are the weights and biases of the FF layer in the text decoder of the BLIP model, and ReLU denotes the activation function.

The final layer of the decoder transforms the output (via a linear projection and a softmax function) to predict the next token in the sequence. This output text is then used as an input for the next time step during the generation process until the final textual description $\mathbf{s}$ of the image is produced.

B. SD-based CKB for image reconstruction

The SD model is an elaborate VLM collaboratively developed by Stability AI, which possesses rich visual-linguistic knowledge and is applicable to diverse tasks such as text-to-image and image-to-image generation [28]. At the receiver, we use the SD to construct the CKB and utilize the text-to-image components in the CKB to reconstruct images. The semantic reconstructor is composed of a text encoder, a feature generator, and an image decoder.

For a given semantic text $\hat{\mathbf{s}}$, the image reconstruction process through the SD model is illustrated in **Fig. 4** and is described as follows:

1) Text encoder: Text encoder is applied to transform the input text sequence into a semantic vector of fixed dimensions, serving as a control condition for the image feature generator. The text encoder is composed of multiple stacked encoding layers, each containing two sub-layers: MSA and FF. The residual connection and layer normalization are applied before each sublayer. This structure is similar to the image encoder in the BLIP model.

The input to the text encoder is the sequence $\hat{\mathbf{s}}$ composed of words. Initially, each word is mapped to a fixed-length vector by word embeddings. These word embeddings, serve as the input to the text encoder. The encoder iteratively performs MSA and FF operations, ultimately producing a sequence composed of textual feature vectors.

2) Diffusion generation: An initial image feature vector composed of pure noise is input into the image feature generator. Textual feature vectors are injected into the noised feature vector to guide the noise removement. Through multiple iterations, noise is progressively removed, and an image feature vector containing textual information is obtained. The denoising step employs a U-Net structure, which adopts a CNN-based encoder-decoder structure to preserve spatial information while generating image semantic information. The iterative process of the image feature generator can be described by the following formula:

$$\mathbf{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{Z}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} f_\theta(\mathbf{Z}_t, t, \mathbf{d})) + \sigma_t \mathbf{Y} \tag{12}$$

where $\mathbf{Z}_t$ represents the image feature vector at the time step t, α_t denotes the variance of the forward diffusion process, serving as a hyperparameter. $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, f_θ represents the pre-trained noise prediction U-Net, $\mathbf{d}$ is the textual semantic vector, $\sigma_t \mathbf{Y}$ denotes the mean of the reverse diffusion process, where $\sigma_t = \sqrt{1-\alpha_t}$, and $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I})$ with $\mathbf{I}$ being the identity matrix.

3) Image decoder: Due to the computational inefficiency of the diffusion operation, the denoising process of the image is performed in the compressed semantic space. Multiple iterations of denoising are conducted in the reduced semantic (feature) space, significantly improving the efficiency of image processing. Finally, we utilize the decoder of a Variational Autoencoder (VAE) to map the feature data in the semantic space back to the pixel space, reconstructing images that adhere to semantic consistency. As VAE learns the latent structure of a large amount of image data distribution, the decoder can provide more detailed information consistent with key semantics in the image by employing upsampling and interpolation during the decoding process, thereby enhancing the image quality in the pixel space.

C. Memory-assisted encoder and decoder

In dynamic environments, both the distribution of the transmitted contents and channel states will change over time. This necessitates that the CSC system continuously adjusts based on new input data and channel states to adapt to the

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

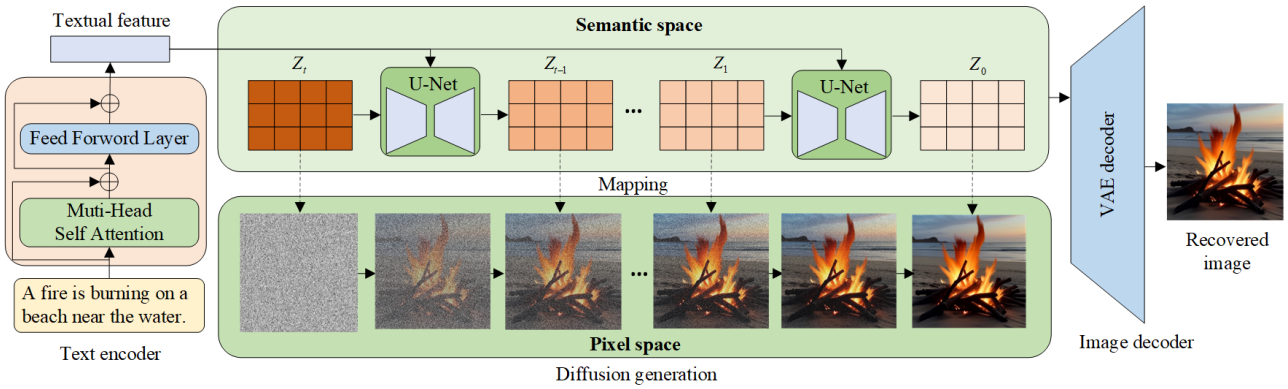

Fig. 4: The architecture of SD-based CKB.

evolving data distribution. However, such adjustments may lead to parameter updates in the encoder and decoder of the CSC system, potentially causing the catastrophic forgetting issue where old parameter values are overwritten or ignored [5]. Hence, continual learning diminishes the robustness of the encoder and decoder in the CSC system.

The memory-based learning strategy addresses the catastrophic forgetting problem in continual learning by diversifying the memorized content [29]. We design a MED method with STM and LTM for both semantic encoder and decoder. Below, we present the workflow of the MED as follows:

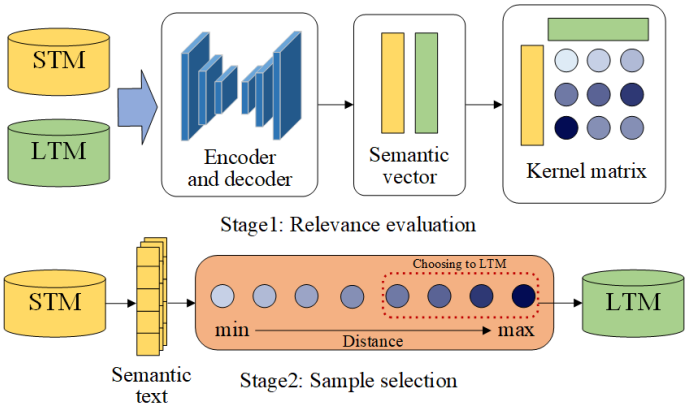

Fig. 5: Memory-assisted encoder and decoder.

We denote $\mathcal{M}_{stm} = \{\mathbf{s}_i^{stm}\}_{i=1}^{n_{stm}}$ and $\mathcal{M}_{ltm} = \{\mathbf{s}_j^{ltm}\}_{j=1}^{n_{ltm}}$ as the sets representing dynamic samples stored in STM and LTM. $\mathbf{s}_i^{stm}$ denotes the i-th sample in STM, and $\mathbf{s}_j^{ltm}$ represents the j-th sample in LTM. n_{stm} and n_{ltm} denote the current number of samples, respectively. When the STM pool becomes full, it is necessary to select representative samples from it and transfer them to the LTM. Hence, let n_{stm}^{Max} represent the maximum number of samples that can be stored in $\mathcal{M}_{stm}$. The sample selection process can be illustrated in **Fig. 5** and described as follows:

1) Relevance evaluation: During the inference phase of the CSC system, new samples being processed are continuously added to the STM. When the number of samples in the STM exceeds the specified maximum, an evaluation action is executed. The primary objective of this stage is to assess the relevance of samples. We evaluate the distance between two samples stored in STM and LTM using the kernel method. Due to its powerful nonlinear mapping capability, locality, and simplicity, the Radial Basis Function (RBF) is selected as the kernel function:

$$\mathrm{RBF}(\mathbf{s}_i^{stm}, \mathbf{s}_j^{ltm}) = \exp\left(-\frac{\|\mathbf{v}_i^{stm} - \mathbf{v}_j^{ltm}\|^2}{2\tau^2}\right) \tag{13}$$

where $\mathbf{v}_i^{stm}$ and $\mathbf{v}_j^{ltm}$ are feature vectors extracted by the semantic encoder from samples $\mathbf{s}_i^{stm}$ and $\mathbf{s}_j^{ltm}$, respectively. τ is the scale hyperparameter for the kernel function, and we set $\tau = 10$ to ensure that the output of $\mathrm{RBF}(\cdot,\cdot)$ is within $[0, 1]$. Eq. (13) can be further accelerated through matrix operations, expressed as:

$$\mathbf{S} = \mathrm{F}_{\exp}(-(\mathbf{B}^{stm}(-\mathbf{B}^{ltm})^{\mathrm{T}}) \odot (\mathbf{B}^{stm}(-\mathbf{B}^{ltm})^{\mathrm{T}})/2\tau^2) \tag{14}$$

where $\mathbf{B}^{stm}$ and $\mathbf{B}^{ltm}$ are feature matrices corresponding to $\mathcal{M}_{stm}$ and $\mathcal{M}_{ltm}$, respectively. $(\cdot)^{\mathrm{T}}$ and $\odot$ represent transpose and Hadamard product, respectively. $\mathrm{F}_{\exp}(\cdot)$ is the exponential function applied element-wise to the matrix [28].

2) Sample selection: The primary objective of this stage is to select samples from STM that are significantly different from those in LTM, ensuring diversity in the memory. We calculate the average similarity score between sample $\mathbf{s}_i^{stm}$ and each sample in LTM using RBF kernel:

$$\mathrm{R}(\mathbf{s}_i^{stm}) = \frac{1}{n_{ltm}} \sum_{k=1}^{n_{ltm}} \mathrm{RBF}(\mathbf{s}_i^{stm}, \mathbf{s}_k^{ltm}). \tag{15}$$

When the computed similarity score is greater than a given threshold λ, we transfer the sample from STM to LTM:

$$\mathrm{R}(\mathbf{s}_i^{stm}) > \lambda \Rightarrow \mathcal{M}_{ltm} = \mathcal{M}_{ltm} \cup \mathbf{s}_i^{stm}. \tag{16}$$

After the selection is complete, $\mathcal{M}_{stm}$ is emptied to buffer new samples in the next round. Then, both the STM and LTM are used to train the semantic encoder and decoder through continual learning [18, 30, 31]. During inference, new data is continuously stored in STM until the storage space is full. Then, STM performs relevance selection, choosing samples that differ from those in LTM and storing them in LTM. Afterward, the samples in STM are cleared to continue receiving new data. Since MED is based on continual learning,

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

it updates the semantic encoder and decoder through online training. At specific intervals, samples from both STM and LTM are fed into the semantic encoder and decoder for incremental training and testing. The training and testing samples are allocated proportionally, where samples in STM ensure the sensitivity of the semantic encoder and decoder to new tasks, and samples in LTM prevent the semantic encoder and decoder from forgetting old tasks, addressing the problem of catastrophic forgetting. The workflow of MED for the semantic encoder and decoder is illustrated in **Algorithm 1**.

Algorithm 1 Memory-assisted Encoder and Decoder

Input: $\mathbf{s}, \mathcal{M}_{stm}$
Output: $\mathcal{M}_{ltm}$
if $n_{stm} \geq n_{stm}^{Max}$ **then**
  Calculate the kernel distance $\mathrm{RBF}(\mathbf{s}_i^{stm}, \mathbf{s}_j^{ltm})$ between samples in STM and LTM according to Eq. (13).
  Calculate the average similarity score $\mathrm{R}(\mathbf{s}_i^{stm})$ between sample $\mathbf{s}_i^{stm}$ and each sample in LTM according to Eq. (15).
else
  Feed current $\mathbf{s}$ into $\mathcal{M}_{stm}$.
end if
if $\mathrm{R}(\mathbf{s}_i^{stm}) > \lambda$ **then**
  Transfer the i-th sample from $\mathcal{M}_{stm}$ to $\mathcal{M}_{ltm}$ according to Eq. (16).
end if
Clear $\mathcal{M}_{stm}$.

D. Noise attention module

Inspired by the feature attention module in [7], we propose a NAM based on SNR values. The NAM leverages a new noise attention network to determine the importance of each feature vector during the process of encoding and decoding, assigning weights to semantic coding and channel coding. This allows for achieving integrated encoding of both semantic and channel information according to the current SNR.

Specifically, in unfavorable channel conditions, higher weights are allocated to the channel encoder and lower weights are allocated to the semantic encoder for the same source information. This allocation strategy enhances robustness in the channel encoder to mitigate the effects of severe channel noise. Conversely, in favorable channel conditions, lower weights are assigned to the channel encoder and higher weights are assigned to the semantic encoder for the same source information. This increased allocation of weights to the semantic encoder aims to enhance semantic quality.

The structure of the NAM is illustrated in **Fig. 6**, and a detailed description of the workflow is provided below:

1) SNR projection: Firstly, the SNR projection module extends the SNR values to the same dimension as feature vectors in the encoder and decoder. The module is a fully connected network comprising three FF layers. The first two FF layers employ the ReLU activation function, while the third FF layer utilizes the Sigmoid activation function. It transforms

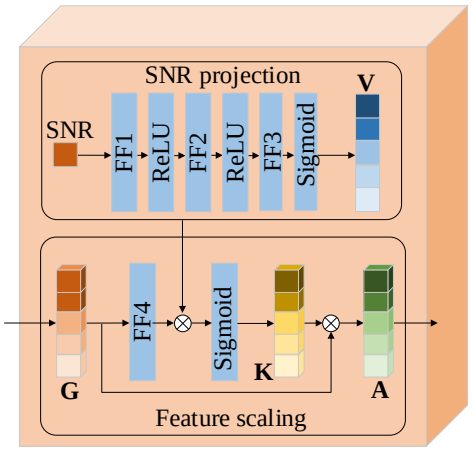

Fig. 6: Noise attention module.

the input SNR value r to a vector $\mathbf{v}$. The mapping process from r to $\mathbf{v}$ is as follows:

$$\mathbf{v}' = \mathrm{ReLU}(\mathbf{W}_{n_2} \cdot \mathrm{ReLU}(\mathbf{W}_{n_1} \cdot r + b_{n_1}) + b_{n_2}) \quad (17)$$

$$\mathbf{v} = \mathrm{Sigmoid}(\mathbf{W}_{n_3} \cdot \mathbf{v}' + b_{n_3}) \quad (18)$$

where ReLU and Sigmoid denote the activation functions, and $\mathbf{W}_{n_i}$ and b_{n_i} are the weights and biases of FF layers, respectively.

2) Feature scaling: Subsequently, we combine the input features with the projected SNR to obtain a scaling factor $\mathbf{K}$, which records the importance of each intermediate feature vector for semantic/channel encoder and decoder as follows:

$$\mathbf{K} = \mathrm{Sigmoid}(\mathbf{e} \cdot \mathbf{v}) \quad (19)$$

where the Sigmoid activation function is used to constrain the output to the interval (0, 1). The $\mathbf{e}$ is the output of the intermediate feature vectors $\mathbf{G}$ after passing through the fourth FF layer as follows:

$$\mathbf{e} = \mathbf{W}_{n_4} \cdot \mathbf{G} + b_{n_4} \quad (20)$$

where $\mathbf{W}_{n_4}$ and b_{n_4} are the weights and biases of the fourth FF layer.

Finally, the intermediate feature vector $\mathbf{G}$ are multiplied by the scaling factor $\mathbf{K}$ to obtain the calibrated vector $\mathbf{A}$ as follows:

$$A_i = K_i \cdot G_i \quad (21)$$

where A_i represents the i-th element in $\mathbf{A}$, G_i represents the i-th element in $\mathbf{G}$, and K_i represents the i-th element in $\mathbf{K}$.

The NAM is embedded into the feature vectors of both the semantic/channel encoder and decoder to enhance the robustness of the CSC system. The workflow of NAM is illustrated in **Algorithm 2**.

V. Numerical Results

In this section, we evaluate the performance of the proposed VLM-CSC system by comparing it with other SC systems.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

Algorithm 2 Noise Attention Module

Input: $r, \mathbf{G}$
Output: $\mathbf{A}$

1: Transform the SNR value r for projection and obtain $\mathbf{v}$ according to Eqs. (17)-(18).
2: Transform intermediate feature vector $\mathbf{G}$ to the vector $\mathbf{e}$ According to Eq. (20).
3: Calculate the scaling factor $\mathbf{K}$ according to Eq. (19).
4: Calculate the calibrated vector $\mathbf{A}$ according to Eq. (21).
5: Return $\mathbf{A}$

A. Simulation settings

The datasets employed in this study include publicly available Kaggle datasets such as CIFAR, BIRDS and CATSvsDOGS [32]. The configuration of the experiments is detailed as follows:

The pretrained BLIP has 129MB parameters, and the pretrained SD model has 1.99GB parameters. The semantic encoder comprises three transformer encoder layers alternated with NAMs. Each transformer encoder layer has 8 heads and the feature dimension is 128. The channel encoder is composed of two FF hidden layers alternating with NAMs, where the first hidden layer has 256 neurons and the second FNN layer has 128 neurons. To maintain information consistency, the semantic and channel decoder employs a structure opposite to that of the encoder. In NAM, the four FF layers have neuron quantities of 56, 128, 56, and 56, respectively. Additionally, the maximum sample size for STM is 500, and the threshold for sample selection is 0.05.

The experimental training and testing environment involves the Windows 2016 server with Python3.8, PyTorch 1.8.0 and CUDA 11.6. Computational resources are provided by an Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz and NVIDIA Tesla T4.

B. Evaluation metrics

The proposed VLM-CSC system transforms image data to textual semantic data through the BLIP-based knowledge base, encodes it using a semantic encoder, decodes it at the receiver, and finally reconstructs the image through the SD-based knowledge base. Therefore, this system involves the transmission of data in two modalities, necessitating the use of two different metrics to comprehensively evaluate the overall performance of the VLM-CSC system. These metrics are: (1) Image-level, examining the accuracy of semantic reconstruction for image data; (2) Text-level, examining the accuracy of semantic recovery for text data.

1) Image-level: Semantic Service Quality (SSQ): In performance assessment of the SC system, the emphasis on semantic layer transmission should be directed towards whether information, after undergoing semantic recovery, can meet the expectations of subsequent tasks. The general quality metric for semantic services is denoted by [33]:

$$SSQ = \frac{ST(\hat{S})}{ST(S)} \tag{22}$$

where S represents the unprocessed source information at the transmitter, $\hat{S}$ represents the recovered information at the semantic level by the receiver, and $ST(\cdot)$ signifies the performance of the source information or recovered information when executing subsequent tasks, which is the classification accuracy in our study.

2) Text-level: Bilingual Evaluation Understudy (BLEU): The BLEU score outputs a number between 0 and 1, indicating how similar the decoded text is to the transmitted text, with 1 representing the highest similarity. For a transmission sentence $\mathbf{s}$ with length l_s and a decoded sentence $\hat{\mathbf{s}}$ with length $l_{\hat{s}}$, BLEU can be expressed as [34]:

$$\log \mathrm{BLEU} = \min(1 - \frac{l_{\hat{s}}}{l_s}, 0) + \sum_{n=1}^{N} u_n \log p_n \tag{23}$$

where the "n-gram" refers to a contiguous sequence of n words from a given sample of text or speech, u_n is the weight of the n-grams, and p_n is the n-grams score, defined as:

$$p_n = \frac{\sum_k \min(C_k(\hat{\mathbf{s}}), C_k(\mathbf{s}))}{\sum_k \min(C_k(\hat{\mathbf{s}}))} \tag{24}$$

where $C_k(\cdot)$ is the frequency count function for the k-th element in the n-th grams.

C. Performance comparison of VLM-base KBs

To evaluate the performance of extracting semantic information from images using KBs, we employ three VLMs (BLIP, LEMON[35], and RAM[36]) to construct the sender-side KBs in the CSC system. The receiver-side KB is uniformly implemented using the SD model. Subsequently, we assess the CSC system's performance on the AWGN channel. SSQ is utilized as the evaluation metric on the CATSvsDOGS dataset [32]. The experimental outcomes are illustrated in **Fig. 7**.

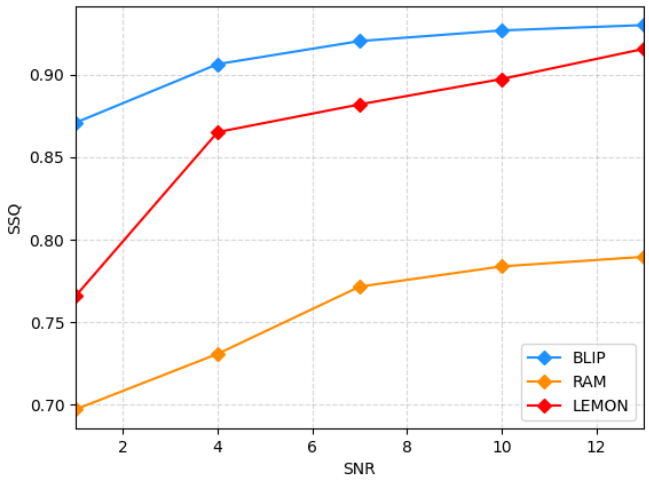

Fig. 7: SSQ of CSC systems based on different VLMs.

From **Fig. 7**, it is evident that the CSC system based on BLIP exhibits the highest SSQ, followed by the one based on LEMON, while the CSC system based on RAM performs the poorest, significantly lower than the CSC systems based on BLIP and LEMON. Furthermore, the CSC system based on

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

BLIP maintains robust performance even at low SNR values. The experimental results indicate that the CSC system constructed based on BLIP accurately extracts image semantics and sustains commendable performance across different SNR levels.

D. Performance evaluation for MED

To demonstrate the performance of the proposed MED, we conduct experiments comparing VLM-CSC with the MED module against VLM-CSC without the MED module. The evaluation is performed across different image datasets. The image datasets include Cifar, Birds, and CatsVSDogs [32].BLEU scores for semantic similarity serve as the evaluation metric. Additionally, when assessing the performance of VLM-CSC on image datasets with different distributions, the channel is fixed to Rayleigh. The continual learning map, originally proposed by Google, is employed to visualize the performance changes of existing tasks when a new task is introduced. The experimental results are illustrated by the continual learning map in **Fig. 8**.

Figure **Fig. 8** (a) and (b) illustrate a significant performance drop in the VLM-CSC system without the MED module on the previous Cifar dataset after learning subsequent datasets such as Birds and CatsVSDogs. In contrast, **Fig. 8** (c) and (d) reveal that the VLM-CSC system with the MED module only exhibits a marginal decline in performance on the previous Cifar dataset after learning subsequent datasets like Birds and CatsVSDogs.

The experimental results from **Fig. 8** underscore that the proposed MED module enables the CSC system to overcome catastrophic forgetting during the continual learning process. This facilitates knowledge learning from multiple image datasets, enhancing the generalization of the CSC system in dynamic environments.

E. Performance evaluation for NAM

To demonstrate the performance of the proposed NAM, we conduct an experimental comparison between VLM-CSC with and without NAM. Semantic similarity, measured by BLEU score, serves as the evaluation metric. Specifically, the proposed VLM-CSC system is trained under a uniform distribution of SNR_{train} ranging from 0 dB to 10 dB, while the VLM-CSC system without NAM is trained at specific SNR_{train} values of 0 dB, 2 dB, 4 dB, and 8 dB. Subsequently, the performance of the VLM-CSC system is evaluated at specific SNR_{test} values ranging from 0 dB to 10 dB. The experimental results are depicted in **Fig. 9**.

The findings depicted in Figure 9 demonstrate that the performance of the proposed VLM-CSC system outperforms any VLM-CSC system without NAM, specifically trained at distinct SNR_{train} values. This observation highlights the capability of the VLM-CSC system, equipped with NAM, to address the performance degradation challenges caused by the mismatch between the SNR during training and deployment stages in conventional ISC systems. This improvement contributes to the robustness of the VLM-CSC system across different SNR values.

F. Semantic communication performance evaluation

To evaluate the performance of the VLM-CSC system in image classification tasks, we compare it with JSCC based on CNN [37] and WITT based on ViT [38]. The metric used for performance evaluation is classification accuracy. Additionally, we assess the bandwidth-saving capabilities of the VLM-CSC by considering the compression ratio between transmitted data and original images as the evaluation metric. Therefore, a smaller compression ratio indicates higher efficiency and better compression performance of the semantic communication system. The experimental results are presented in **Fig. 10**.

Fig. 10 (a) clearly demonstrates that, at low SNR levels, the superior performance of VLM-CSC in the classification task with the CATSvsDOGS dataset, and WITT shows slightly lower results, particularly with decreased performance compared to VLM-CSC. At high SNR levels, WIIT and JSCC exhibit superior SSQ compared to VLM-CSC due to their direct transmission of images. **Fig. 10** (b) depicts the compression ratio and trainable parameters, with VLM-CSC achieving the lowest of all, followed by JSCC, while WITT attains the highest compression ratio and trainable parameters. **Fig. 10** (c) illustrates that the reconstructed image highly aligns with the original image and the image description, validating the VLM-CSC system's ability to ensure semantic consistency across modalities.

The experimental results depicted in **Fig. 10** demonstrate that the proposed VLM-CSC exhibits overall superior performance in image classification tasks compared to other ISC systems at low SNR levels. Then, the compression ratio of transmitted data is significantly lower for VLM-CSC compared to other ISC systems, indicating that VLM-CSC can effectively conserve transmission bandwidth while preserving high-quality semantic transmission. Moreover, due to the absence of training VLMs, the VLM-CSC system exhibits the minimum number of trainable parameters, resulting in the lowest training complexity.

VI. Conclusion

This paper introduces a novel VLM-CSC system capable of converting images into text descriptions for transmission over wireless channels, and reconstructing the image at the receiver. The system includes three main contributions: CKB for image-to-text and text-to-image conversion, MED for continual learning in dynamic environments, and NAM for joint semantic and channel encoding based on SNR. Corresponding performance metrics are designed to evaluate the VLM-CSC system from both image and text perspectives. Experimental validations are conducted under various image datasets. Results demonstrate the effectiveness and robustness of the VLM-CSC system in preserving semantic similarity between the image and text, as well as its adaptability to dynamic environments. As the theoretical foundation of VLMs becomes more established, we will focus on integrating more formal mathematical descriptions into large model-based semantic communication systems in future work.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

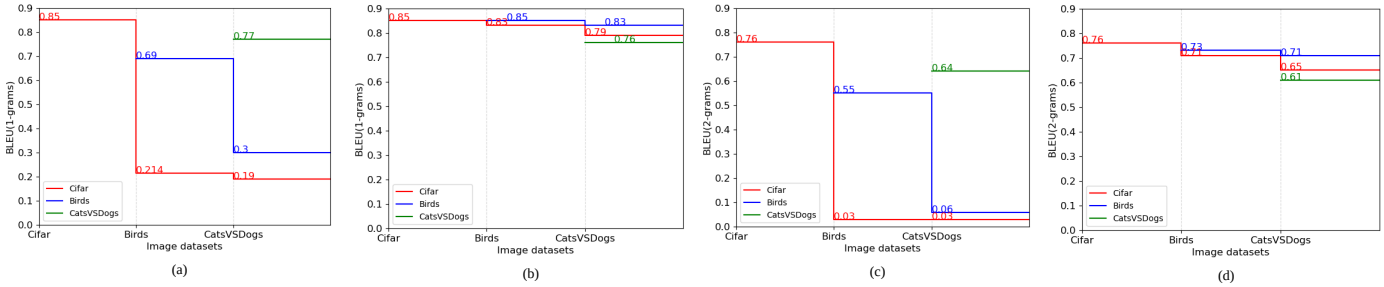

Fig. 8: The continual learning map for BLEU scores across diverse image datasets are evaluated in the following scenarios:(a) The BLEU (1-grams) of VLM-CSC without MED across different image datasets. (b) The BLEU (2-grams) of VLM-CSC without MED across different image datasets. (c) The BLEU (1-grams) of VLM-CSC across different image datasets. (d) The BLEU (2-grams) of VLM-CSC across different image datasets.

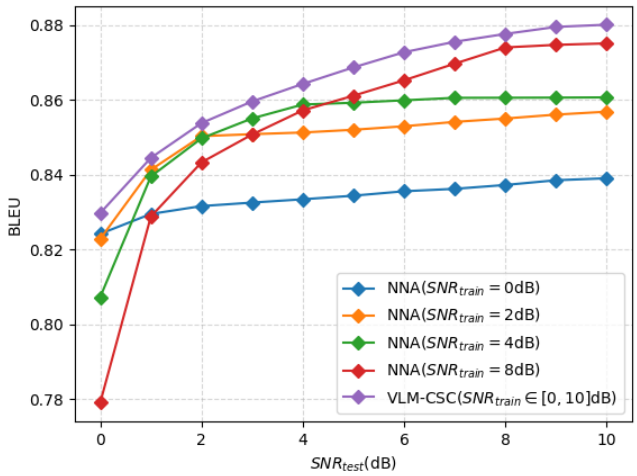

Fig. 9: The performance of NAM in the VLM-CSC system. NNA represents the VLM-CSC system without NAM.

REFERENCES

[1] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5g: When cellular networks meet artificial intelligence," *IEEE Wireless communications*, vol. 24, no. 5, pp. 175–183, 2017.

[2] M. Xu, W. C. Ng, W. Y. B. Lim, J. Kang, Z. Xiong, D. Niyato, Q. Yang, X. S. Shen, and C. Miao, "A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges," *IEEE Communications Surveys & Tutorials*, 2022.

[3] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. S. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, 2022.

[4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[5] 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 Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 170–185, 2022.

[6] 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, 2019.

[7] 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 Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2021.

[8] L. Dong, Z. Liu, F. Jiang, and K. Wang, "Joint optimization of deployment and trajectory in uav and irs-assisted iot data collection system," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 583–21 593, 2022.

[9] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6g communications," *IEEE Wireless Communications*, vol. 31, no. 6, pp. 48–55, 2024.

[10] 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, 2019.

[11] J. Dai, S. Wang, K. Tan, Z. Si, X. Qin, K. Niu, and P. Zhang, "Nonlinear transform source-channel coding for semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 8, pp. 2300–2316, 2022.

[12] C. Dong, H. Liang, X. Xu, S. Han, B. Wang, and P. Zhang, "Semantic communication system based on semantic slice models propagation," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 202–213, 2022.

[13] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 55–71, 2022.

[14] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan,

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

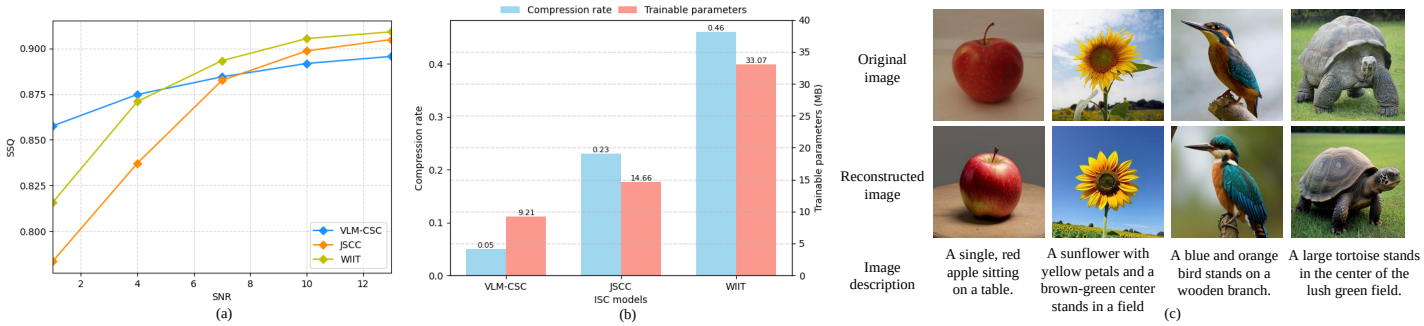

Fig. 10: Performance comparison of VLM-CSC with other ISC systems. (a) SSQ. (b) Compression ratio and trainable parameters. (c) Semantic alignment.

and X. You, "Large ai model empowered multimodal semantic communications," *IEEE Communications Magazine*, vol. 63, no. 1, pp. 76–82, 2025.

[15] A. Fürst, E. Rumetshofer, J. Lehner, V. T. Tran, F. Tang, H. Ramsauer, D. Kreil, M. Kopp, G. Klambauer, A. Bitto *et al.*, "Cloob: Modern hopfield networks with infoloob outperform clip," *Advances in neural information processing systems*, vol. 35, pp. 20 450–20 468, 2022.

[16] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[17] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[18] L. Dong, F. Jiang, M. Wang, Y. Peng, and X. Li, "Deep progressive reinforcement learning-based flexible resource scheduling framework for irs and uav-assisted mec system," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[19] Z. Xu, L. Wang, W. Liang, Q. Xia, W. Xu, P. Zhou, and O. F. Rana, "Age-aware data selection and aggregator placement for timely federated continual learning in mobile edge computing," *IEEE Transactions on Computers*, 2023.

[20] Y. Yu, P. Chen, X.-W. Zhu, J. Zhai, and C. Yu, "Continual learning digital predistortion of rf power amplifier for 6g ai-empowered wireless communication," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 11, pp. 4916–4927, 2022.

[21] Z. Zhang, B. Guo, W. Sun, Y. Liu, and Z. Yu, "Cross-fcl: Toward a cross-edge federated continual learning framework in mobile edge computing systems," *IEEE Transactions on Mobile Computing*, 2022.

[22] H. Zhou, W. Xia, H. Zhao, J. Zhang, Y. Ni, and H. Zhu, "Continual learning-based fast beamforming adaptation in downlink miso systems," *IEEE Wireless Communications Letters*, vol. 12, no. 1, pp. 36–39, 2022.

[23] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large ai model-based semantic communications," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 68–75, 2024.

[24] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9847–9857.

[27] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[29] F. Ye and A. G. Bors, "Continual variational autoencoder learning via online cooperative memorization," in *European Conference on Computer Vision*. Springer, 2022, pp. 531–549.

[30] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, and K. Yang, "Ai driven heterogeneous mec system with uav assistance for dynamic environment: Challenges and solutions," *IEEE Network*, vol. 35, no. 1, pp. 400–408, 2020.

[31] F. Jiang, L. Dong, K. Wang, K. Yang, and C. Pan, "Distributed resource scheduling for large-scale mec systems: A multiagent ensemble deep reinforcement learning with imitation acceleration," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6597–6610, 2022.

[32] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of machine translation summit x: papers*, 2005, pp. 79–86.

[33] C. Dong, H. Liang, X. Xu, S. Han, B. Wang, and P. Zhang, "Semantic communication system based on semantic slice models propagation," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TWC.2025.3539526, IEEE Transactions on Wireless Communications

202–213, 2022.

[34] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[35] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 980–17 989.

[36] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, "Recognize anything: A strong image tagging model," *arXiv preprint arXiv:2306.03514*, 2023.

[37] D. B. Kurka and D. Gündüz, "Deepjscc-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[38] K. Yang, S. Wang, J. Dai, K. Tan, K. Niu, and P. Zhang, "Witt: A wireless image transmission transformer for semantic communications," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

Feibo Jiang received his B.S. and M.S. degrees in School of Physics and Electronics from Hunan Normal University, China, in 2004 and 2007, respectively. He received his Ph.D. degree in School of Geosciences and Info-physics from the Central South University, China, in 2014. He is currently an associate professor at the Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, China. His research interests include semantic communication, federated learning, Internet of Things, and mobile edge computing.

Chuanguo Tang received the M.E. degree in Computer Science and Technology from Hunan Normal University, China, in 2024. Her research interests include machine learning and semantic communication.

Li Dong received the B.S. and M.S. degrees in School of Physics and Electronics from Hunan Normal University, China, in 2004 and 2007, respectively. She received her Ph.D. degree in School of Geosciences and Info-physics from the Central South University, China, in 2018. She is currently an associate professor at Hunan University of Technology and Business, China. Her research interests include machine learning, Internet of Things, and mobile edge computing.

Kezhi Wang received a PhD degree in Engineering from the University of Warwick, U.K. Currently, he is a Senior Lecturer at the Department of Computer Science, Brunel University London, U.K. His research interests include wireless communications, mobile edge computing, and machine learning.

Kun Yang received his PhD from the Department of Electronic & Electrical Engineering of University College London (UCL), UK. He is currently a Chair Professor of University of Essex, UK and Nanjing University. His main research interests include wireless networks and communications, communication-computing cooperation, and new AI (artificial intelligence) for wireless. He has published 500+ papers and file 50 patents. He serves on the editorial boards of a number of IEEE journals (e.g., IEEE WCM, TVT, TNB). He is a Deputy Editor-in-Chief of IET Smart Cities Journal. He has been a Judge of GSMA GLOMO Award at World Mobile Congress-Barcelona since 2019. He was a Distinguished Lecturer of IEEE ComSoc (2020-2021), a Recipient of the 2024 IET Achievement Medals and the Recipient of 2024 IEEE CommSoft TC's Technical Achievement Award. He is a Member of Academia Europaea (MAE), a Fellow of IEEE, a Fellow of IET and a Distinguished Member of ACM.

Cunhua Pan is a full professor in Southeast University. His research interests mainly include reconfigurable intelligent surfaces (RIS), AI for Wireless, and near field communications and sensing. He has published over 200 IEEE journal papers. His papers got over 14,000 Google Scholar citations with H-index of 61. He is Clarivate Highly Cited researcher. He is/was an Editor of IEEE Transaction on Communications, IEEE Transactions on Vehicular Technology, IEEE Wireless Communication Letters, IEEE Communications Letters and IEEE ACCESS. He serves as the guest editor for IEEE Journal on Selected Areas in Communications on the special issue on xURLLC in 6G: Next Generation Ultra-Reliable and Low-Latency Communications. He also serves as a leading guest editor of IEEE Journal of Selected Topics in Signal Processing (JSTSP) Special Issue on Advanced Signal Processing for Reconfigurable Intelligent Surface-aided 6G Networks, leading guest editor of IEEE Vehicular Technology Magazine on the special issue on Backscatter and Reconfigurable Intelligent Surface Empowered Wireless Communications in 6G, leading guest editor of IEEE Open Journal of Vehicular Technology on the special issue of Reconfigurable Intelligent Surface Empowered Wireless Communications in 6G and Beyond, and leading guest editor of IEEE Transactions on Green Communications and Networking Special Issue on Design of Green Near-Field Wireless Communication Networks. He received the IEE ComSoc Leonard G. Abraham Prize in 2022, IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award, 2022 and IEEE ComSoc Fred W. Ellersick Prize in 2024.

Copyright © 2025 Institute of Electrical and Electronics Engineers (IEEE). Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. See: https://journals.ieeeauthorcenter.ieee.org/become-an-ieee-journal-author/publishing-ethics/guidelines-and-policies/post-publication-policies/.