ViTMAE-Based Robust Infrastructure-Leveraging System for Wireless Semantic Image Transmission

Ruisi Li, Ren wei Ou

New York University, Sun Yat-sen University
rl5411@nyu.edu, ourw210@gamil.com

Abstract

Semantic image transmission represents a significant evolution in the design of communication systems, shifting the focus from the transmission of raw bits to the delivery of meaningful semantic information. Nevertheless, many existing semantic communication methods rely on analog transmission schemes, which face fundamental integration issues with contemporary digital communication infrastructures, including widely adopted Orthogonal Frequency Division Multiplexing (OFDM) receivers. This incompatibility often leads to degraded performance in challenging wireless environments, such as urban areas with strong multipath effects or high-speed mobility scenarios. To address these challenges, this paper introduces a novel semantic transmission system that seamlessly integrates a Vision Transformer-based Masked Autoencoder (ViTMAE) with standard digital communication frameworks. The proposed design intentionally leverages existing physical-layer processing modules in base stations, avoiding the need for complex deep learning-based denoising components while utilizing the masked autoencoder's ability to learn robust and predictive feature representations through self-supervised reconstruction. We evaluated the proposed system under 3GPP-standard channel models, where it substantially outperformed conventional JPEG source coding coupled with LDPC channel coding, particularly in low signal-to-noise ratio (SNR) conditions. This study provides valuable insights for the development of practical and scalable semantic image transmission architectures for next-generation networks.

Introduction

Semantic communication represents a paradigm shift in modern communication systems. Unlike traditional approaches that focus on raw bit transmission, it transmits extracted semantic information by eliminating semantic redundancies. This method offers the potential for significant gains in spectral efficiency and enhanced resilience to channel noise (Getu, Kaddoum, and Bennis 2025). Due to the rapid development of deep learning, various commonly used deep learning models have been extensively applied in semantic communication systems (Xie et al. 2021; Wu et al. 2024). However, prior deep learning-based approaches to semantic communication largely rely on analog transmission paradigms (Bourtsoulatze, Kurka, and Gündüz 2019; Xu et al. 2021; Yang and Kim 2022), where semantic features extracted by an encoder at the transmitter are directly transmitted to a decoder at the receiver. However, real-world wireless communication systems, such as Long-Term Evolution (LTE), 5G New Radio (NR), and Wi-Fi, predominantly utilize Orthogonal Frequency Division Multiplexing (OFDM) digital modulation as their core physical layer technology. These analog-based semantic methods are incompatible with the existing communication infrastructure, limiting their practical deployment scenarios. This creates a fundamental technical divide: the continuous-valued semantic features from analog paradigms cannot be directly processed by the discrete signal processing modules in standard digital systems, limiting their practical deployment scenarios.

To bridge this gap between analog semantic encoding and digital transmission, recent efforts have shifted towards developing digital-compatible semantic communication schemes. For example, some recent works have explored digital semantic communication by first converting semantic vectors into bitstreams via vector quantization (VQ), followed by digital modulation for transmission (Fu et al. 2023; Liu et al. 2024; Hu et al. 2023). While these methods have achieved success in various downstream tasks, they often overlook the accuracy of bitstream reconstruction at the receiver, which may pose a bottleneck to the system. Moreover, most of these studies focus on performance under relatively simple channel conditions (Hu et al. 2023; Zheng et al. 2025), such as additive white Gaussian noise (AWGN) channels, neglecting the model's resilience to channel interference. In practice, radio frequency (RF) signals are susceptible to various interferences beyond Gaussian noise, including multipath fading, Doppler shifts, and inter-symbol interference (ISI). In fact, effectively mitigating such real-world channel impairments remains a critical challenge. Accordingly, conventional physical communication systems employ channel estimation and equalization techniques to combat these distortions, improve detection accuracy, and ensure transmission reliability (Yang, Bian, and Kim 2022).

Based on the above discussion, we argue that an effective semantic communication system should seamlessly integrate with existing digital communication frameworks while

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

maximizing anti-interference capabilities in time-varying multipath channels. Motivated by these considerations, this paper proposes a novel semantic communication framework based on the OFDM system for image transmission. This framework leverages established physical layer components to significantly enhance anti-interference capabilities and assist the decoder in image reconstruction. Furthermore, we used the vision transformer masked autoencoder (ViTMAE) (He et al. 2022) as the encoder at the transmitter. Benefiting from ViTMAE's masked image modeling (MIM) (Xie et al. 2022) pre-training, it possesses robust patch prediction capabilities, which enable the decoder to effectively reconstruct images even in the presence of bit errors during transmission, thereby enhancing the overall model robustness. Furthermore, to efficiently discretize semantic features for digital transmission, we incorporate vector quantization to map continuous semantic features onto a finite set of codewords from a learned codebook.

Specifically, at the transmitter, ViTMAE serves as the encoder to extract semantic features from images. Then, these features are vector-quantized, and the corresponding codebook indices are modulated using OFDM multi-carrier techniques for transmission. At the receiver, channel estimation and equalization are performed on the received signal to mitigate the effects of selective fading and ISI. This is followed by constellation demapping to recover the indices. The semantic features are then reconstructed using the shared codebook and subsequently fed into the decoder to reconstruct the transmitted images. We evaluate our proposed method in 3rd Generation Partnership Project (3GPP)-compliant clustered delay line (CDL) channels. The results demonstrate a substantial reduction in bit transmission volume compared to the traditional JPEG method combined with Low-Density Parity-Check (LDPC) coding, while also achieving superior performance in terms of both the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR). We believe our work paves the way for future advancements in semantic communication systems compatible with real-world digital infrastructures. The main contributions of this paper are summarized as follows:

- We propose an OFDM-based semantic communication framework for image transmission that integrates with digital infrastructures. It uses channel estimation and equalization to enhance anti-interference in multipath channels, achieving robust performance in 3GPP-compliant CDL scenarios.
- We employ ViTMAE as the encoder, which leverages its MIM pre-training to achieve resilient semantic feature extraction and reconstruction under bit errors. When combined with VQ and OFDM, this system reduces the transmission volume compared to JPEG+LDPC, while achieving higher SSIM and PSNR.

By incorporating standard physical communication modules and ViTMAE, we believe that our method significantly enhances the reliability and robustness of the semantic communication system.

Related Works

Analog-based Semantic Communication

Analog-based semantic communication represents an early and innovative approach that transmits semantic information by directly leveraging continuous analog waveforms (Bourtsoulatze, Kurka, and Gündüz 2019; Yang et al. 2024; Wang et al. 2023). In such systems, semantic encoders map high-level knowledge or features, such as object representations in images or contextual intents in natural language, into continuous semantic vectors, which are then treated as analog signals. These semantic-bearing signals are then fed directly to the receiver, enabling a joint source-channel coding framework that inherently aligns semantic extraction with physical-layer transmission. For instance, the deep joint source-channel coding (DeepJSCC) method has been widely adopted as a canonical analog semantic communication architecture (Bourtsoulatze, Kurka, and Gündüz 2019). It employs an end-to-end trained neural network to map source data directly to channel input symbols, demonstrating remarkable robustness in low signal-to-noise ratio (SNR) conditions. Similarly, Xu et al. introduced an attention-based DeepJSCC framework that dynamically adapts to varying SNR levels, addressing the critical issue of channel mismatch and significantly improving storage efficiency and deployment flexibility (Xu et al. 2021).

However, this analog design pattern faces significant compatibility challenges with modern digital communication infrastructures. Current wireless systems, including 4G LTE and 5G NR, are fundamentally built upon digital frameworks. Unfortunately, many existing studies on analog semantic communication have been validated only in relatively simplified and stationary channel models, lacking the robustness required for real-world, dynamic wireless scenarios (Yang and Kim 2022). These incompatibilities highlight the need to pursue semantic communication technologies that are more aligned with the paradigms of modern communication infrastructure.

Digital-based Semantic Communication

Digital-based semantic communication emphasizes the use of discrete symbol transmission, ensuring seamless integration with legacy digital infrastructures. This approach quantizes semantic features into bit streams, enabling reliable transmission over established digital channels while preserving semantic meaning through error-correction and modulation schemes. The primary design advantage of digital semantic communication lies in its backward compatibility, allowing for practical deployment without modifying the underlying physical-layer hardware.

In recent years, some studies have focused on the design of digital-based semantic communication (Hu et al. 2023; Liu et al. 2024; Fu et al. 2023; Xie et al. 2025). For example, Hu et al. address the critical issue of semantic noise by developing a robust framework that utilizes adversarial training and a masked VQ-VAE with a shared feature codebook, effectively suppressing misleading features and significantly reducing transmission overhead (Hu et al. 2023). Liu et al. have focused on the integration with existing OFDM

systems, proposing an importance-aware reordering method that mitigates distortions from channel estimation errors, thereby enhancing recovery quality and system generalizability while also suppressing peak-to-average power ratio (Liu et al. 2025). Despite their competitive performance under various conditions, many existing models are designed with insufficient consideration of physical-channel interference. To mitigate this limitation, numerous approaches incorporate highly complex deep learning-based noise resistance modules at the receiver.

However, such designs often overlook the fact that existing 5G NR communication base stations already integrate well-established and optimized physical-layer modules, such as least squares (LS) channel estimation and linear minimum mean square error (LMMSE) detection. We argue that a more viable and implementation-friendly strategy would be to fully reuse and interoperate with these existing physical-layer components, rather than replacing them with end-to-end neural networks.

The Proposed Framework

The proposed framework integrates semantic encoding with digital communication principles to enable efficient and robust image transmission over wireless channels. The overall model architecture is illustrated in Figure 1. We detail each of its key components in the following subsections.

ViTMAE Encoder The ViTMAE (He et al. 2022) serves as the foundational encoder in our framework, specifically designed to extract high-level semantic representations from input images. Unlike conventional encoders that focus solely on pixel-level reconstruction, ViTMAE leverages a self-supervised learning paradigm known as MIM (Xie et al. 2022), which enables it to develop a profound understanding of semantic content through the task of reconstructing randomly masked patches. In our implementation, each input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ undergoes a structured partitioning process where it is divided into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$. These patches are then linearly projected into patch embeddings while incorporating learnable positional encodings to preserve spatial relationships. The resulting sequence of token embeddings is processed through multiple layers of Transformer blocks in the ViT encoder, where self-attention mechanisms capture both local and global dependencies across the image.

The encoder ultimately produces compact semantic features $\mathbf{z} \in \mathbb{R}^{N\times D}$, where D represents the embedding dimension. Fundamentally, ViTMAE's MIM pre-training provides these features with intrinsic resistance to partial information loss. This characteristic is directly transferable to the management of transmission errors and channel impairments in wireless semantic communication.

Vector Quantization To convert the continuous semantic features into a discrete bitstream suitable for digital transmission, we employ vector quantization (Van Den Oord, Vinyals et al. 2017), which effectively discretizes the feature space while preserving essential semantic information. First, the semantic features $\mathbf{z} \in \mathbb{R}^{N\times D}$ are reshaped such that the last dimension becomes Q, resulting in $\mathbf{z}' \in \mathbb{R}^{N'\times Q}$, where $N' = \frac{N\times D}{Q}$. We define a trainable codebook $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K} \subset \mathbb{R}^Q$ with K entries, where each entry $\mathbf{c}_k$ represents a quantized vector. To enable end-to-end training, we utilize the straight-through estimator technique, which allows gradients to flow through the non-differentiable quantization operation.

For each reshaped semantic feature vector $\mathbf{z}'_i$, the nearest codebook entry is selected via:

$$k_i = \arg\min_k \|\mathbf{z}'_i - \mathbf{c}_k\|_2^2. \tag{1}$$

The quantized feature is then $\hat{\mathbf{z}}'_i = \mathbf{c}_{k_i}$, and the index k_i (requiring $\log_2 K$ bits to store) is sent to the OFDM system for transmission. During training, the codebook is updated to minimize the quantization error, ensuring efficient compression while preserving semantic information.

Wireless Channel and OFDM System The quantized indices are mapped to constellation symbols using M-ary quadrature amplitude modulation (M-QAM) and modulated onto OFDM subcarriers for transmission over a multipath wireless channel. The transmitted signal in the frequency domain is $\mathbf{s} = [s_0, s_1, \ldots, s_{N_c-1}]^T \in \mathbb{C}^{N_c\times 1}$, where N_c is the number of subcarriers. To perform channel estimation at the receiver, pilot symbols are inserted within the resource blocks. After the inverse fast Fourier transform (IFFT) and cyclic prefix (CP) addition, the time-domain signal $\mathbf{s}(t)$ is transmitted through the antenna.

The wireless channel is modelled as a multipath fading channel with following impulse response:

$$h(\tau, t) = \sum_{l=0}^{L-1} h_l(t)\delta(\tau - \tau_l), \tag{2}$$

where $h_l(t)$ is the complex gain of the l-th path, and τ_l is the delay. The received time-domain signal is:

$$\mathbf{r}(t) = \mathbf{s}(t) * h(\tau, t) + \mathbf{n}(t), \tag{3}$$

where $*$ denotes the convolution operation, and $\mathbf{n}(t)$ is additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

In the frequency domain, after removing the CP and applying the fast Fourier transform (FFT) at the receiver, the received signal on the k-th subcarrier is expressed as:

$$r_k = h_k s_k + n_k, \tag{4}$$

where $k = 0, 1, \ldots, N_c - 1$ and

$$h_k = \sum_{l=0}^{L-1} h_l e^{-j2\pi k\tau_l/N_c} \tag{5}$$

is the channel frequency response (CFR) at the k-th subcarrier, obtained via the discrete Fourier transform of the channel impulse response, and n_k is the frequency-domain noise with the same variance σ^2. In vector form, the received frequency-domain signal is $\mathbf{r} = [r_0, r_1, \ldots, r_{N_c-1}]^T \in \mathbb{C}^{N_c\times 1}$, and the channel frequency response vector is $\mathbf{h} = [h_0, h_1, \ldots, h_{N_c-1}]^T \in \mathbb{C}^{N_c\times 1}$, yielding:

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{6}$$

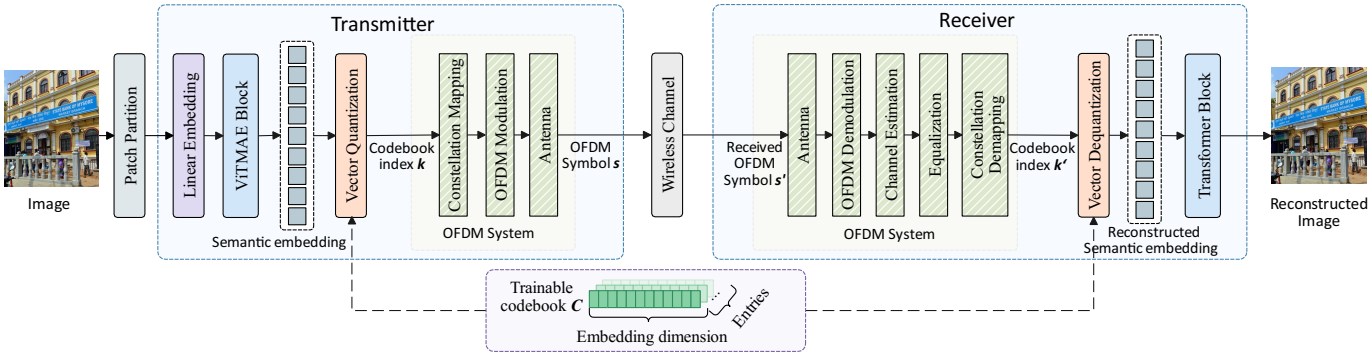

Figure 1: The proposed semantic communication framework. The framework enables robust image transmission by employing vector quantization to seamlessly integrate the ViTMAE encoder with an OFDM communication system.

where $\mathbf{H} = \text{diag}(\mathbf{h}) \in \mathbb{C}^{N_c \times N_c}$ is the diagonal matrix with $\mathbf{h}$ on its main diagonal, and $\mathbf{n} \in \mathbb{C}^{N_c \times 1}$ is the noise vector.

To mitigate impairments introduced by the physical channel, we need to perform channel estimation and equalization. Channel estimation is performed using pilot symbols inserted at known subcarrier positions. Let $\mathcal{P} = \{p_1, p_2, \ldots, p_{N_p}\}$ denote the set of pilot indices, with corresponding known transmitted pilot symbols $s_{p_m} = x_{p_m}$. The received pilot signals are $r_{p_m} = h_{p_m} x_{p_m} + n_{p_m}$. The LS channel estimate at pilot subcarriers is computed as:

$$\hat{h}_{p_m}^{\text{LS}} = \frac{r_{p_m}}{x_{p_m}} = h_{p_m} + \frac{n_{p_m}}{x_{p_m}}, \tag{7}$$

where $m = 1, 2, \ldots, N_p$. Let $\hat{\mathbf{h}}_p^{\text{LS}} = [\hat{h}_{p_1}^{\text{LS}}, \ldots, \hat{h}_{p_{N_p}}^{\text{LS}}]^T \in \mathbb{C}^{N_p \times 1}$ and $\mathbf{X}_p = \text{diag}([x_{p_1}, \ldots, x_{p_{N_p}}]) \in \mathbb{C}^{N_p \times N_p}$. Then the LS estimate can be written as:

$$\hat{\mathbf{h}}_p^{\text{LS}} = \mathbf{X}_p^{-1} \mathbf{r}_p, \tag{8}$$

where $\mathbf{r}_p \in \mathbb{C}^{N_p \times 1}$ contains the received signals at pilot locations. To obtain the full channel estimate $\hat{\mathbf{h}} \in \mathbb{C}^{N_c \times 1}$, interpolation (e.g., linear or nearest-neighbour interpolation) is applied to $\hat{\mathbf{h}}_p^{\text{LS}}$.

After channel estimation, LMMSE equalization is applied to mitigate ISI and frequency-selective fading. For the k-th subcarrier, the LMMSE equalized symbol is:

$$\hat{s}_k = \frac{\hat{h}_k^* r_k}{|\hat{h}_k|^2 + \sigma^2}, \tag{9}$$

where $\hat{h}_k^*$ denotes the complex conjugate of $\hat{h}_k$. In vector form, the equalized signal is $\hat{\mathbf{s}} = \mathbf{W}\mathbf{r} \in \mathbb{C}^{N_c \times 1}$, where $\mathbf{W} \in \mathbb{C}^{N_c \times N_c}$ is a diagonal matrix with diagonal entries $w_{k,k} = \frac{\hat{h}_k^*}{|\hat{h}_k|^2 + \sigma^2}$.

Based on the equalized symbols $\hat{\mathbf{s}}$, constellation demapping is performed using the max-log maximum a posteriori (max-log-MAP) method to recover the transmitted constellation points and provide soft bit information. Assuming Gray-coded M-QAM and additive Gaussian noise remaining after equalization, the max-log-MAP approximation for the log-likelihood ratio (LLR) of the j-th bit of the symbol transmitted on the k-th subcarrier is:

$$l_{k,j} \approx \frac{1}{\sigma_{\text{eff}}^2} \left(\min_{c \in \mathcal{C}_{b_j=0}} |\hat{s}_k - c|^2 - \min_{c \in \mathcal{C}_{b_j=1}} |\hat{s}_k - c|^2 \right), \tag{10}$$

where $l_{k,j}$ is the LLR of the j-th bit in symbol s_k, $\mathcal{C}_{b_j=0}$ and $\mathcal{C}_{b_j=1}$ are the subsets of the M-QAM constellation $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ (with $|\mathcal{C}| = M$) where the j-th bit is 0 or 1, respectively, and σ_{eff}^2 is the effective noise variance. This approximation minimizes computational complexity while closely approximating the exact bit-wise MAP rule by considering only the minimum Euclidean distances to the relevant constellation subsets. For hard decisions, the symbol estimate can be derived from the LLR signs via $\hat{s}_k = \arg\min_{c \in \mathcal{C}} |\hat{s}_k - c|^2$. The recovered bitstream is then used to retrieve the indices $\hat{k}_i$.

Transformer Decoder The received indices $\hat{k}_i$ are used to reconstruct the semantic features from the codebook: $\hat{\hat{\mathbf{z}}}_i' = \mathbf{c}_{\hat{k}_i}$. It is important to note that due to transmission bit errors, the received indices $\hat{k}_i$ may differ from the original transmitted indices. For any received index that falls outside the valid codebook range, we mapped it to a default valid entry. The index error rate decreases progressively as the signal-to-noise ratio increases, leading to improved reconstruction accuracy. The reconstructed features $\hat{\hat{\mathbf{z}}}' \in \mathbb{R}^{N' \times Q}$ are then reshaped back to the original dimensions $\hat{\hat{\mathbf{z}}} \in \mathbb{R}^{N \times D}$. We build a transformer decoder that consists of L_d attention blocks to reconstruct the original image from the potentially erroneous reconstruction features $\hat{\hat{\mathbf{z}}}$. The transformer blocks process the features through multi-head self-attention and feed-forward layers, leveraging global dependencies to predict and reconstruct the image patches. The output is the reconstructed image $\hat{\mathbf{x}}$.

Benefiting from the masked image modeling pre-training in ViTMAE and its integration with standard physical communication modules, the receiver is capable of handling errors in $\hat{\hat{\mathbf{z}}}$ by interpreting them as semantically masked patches. This approach enhances the overall robustness of the semantic communication system without incorporating

additional deep learning-based denoising components at the receiver. Instead, physical-layer interference is managed entirely by conventional 5G base station modules, leading to a more practical and infrastructure-compatible design.

Model Optimization

The framework is trained end-to-end with the objective of minimizing the mean squared error (MSE), which serves as the primary reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \tag{11}$$

To update the parameters in the codebook, the vector quantization loss is included:

$$\mathcal{L}_{\text{vq}} = \|sg(\mathbf{z}) - \hat{\mathbf{z}}\|_2^2 + \beta\|\mathbf{z} - sg(\hat{\mathbf{z}})\|_2^2, \tag{12}$$

where the first term represents the codebook loss, the second term denotes the commitment loss, $sg(\cdot)$ denotes the stop-gradient operator, and β is a hyperparameter between 0 and 1 that controls the magnitude of the commitment loss update. The total loss is the weighted sum of the above two loss components:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{vq}}. \tag{13}$$

This combined optimization objective ensures effective end-to-end training of the entire semantic communication system. The reconstruction loss $\mathcal{L}_{\text{rec}}$ directly supervises the final output quality, while the vector quantization loss $\mathcal{L}_{\text{vq}}$ facilitates the learning of a compact and discrete latent representation.

Experimental Results

Experimental Setup

We train the proposed framework on the DIV2K dataset (Agustsson and Timofte 2017), a benchmark for image reconstruction consisting of 900 high-resolution images with diverse content, which is well-suited for learning robust semantic representations. For evaluation, we employ the DIV2K validation set, which comprises 100 high-quality images exhibiting similar diversity and complexity to the training data.

All images are first preprocessed to a resolution of 448 $\times$ 448 pixels, and the ViTMAE encoder uses a patch size of 16 $\times$ 16. To quantitatively evaluate the quality of the reconstructed images, we adopt two widely-used metrics: SSIM (Wang et al. 2004) and PSNR. The OFDM system is configured with a carrier frequency of 3.5 GHz, 64 subcarriers per resource block, and a CP length of 6 symbols. The channel is simulated using 3GPP-compliant CDL-C models with 24 multipath components, delay spreads up to 200 ns, and Doppler shifts ranging from 0 to 100 Hz, in a Single Input Multiple Output (SIMO) antenna configuration. Varying signal-to-noise ratios (SNR) from 0.0 to 20.0 dB are considered. We set the number of codebook entries to 4096, and the vector dimension is set to 16 to observe their impact on performance. For comparison, we implement a baseline method using JPEG source coding combined with LDPC channel coding. Additionally, each transmitted codebook index is represented as a 32-bit floating-point number and is converted into a 32-bit binary sequence according to the IEEE 754 standard. The SSIM and PSNR metrics are recorded across various SNR levels. To enable end-to-end training, all physical communication modules are implemented in Pytorch. Training is conducted with a batch size of 32, a learning rate of 2×10^{-4}, and the Adam optimizer.

Results Analysis

The performance of the proposed method is quantitatively evaluated under varying channel conditions. Figure 2 and Figure 3 show the SSIM and PSNR performances, respectively, of the proposed framework under various configurations and baseline methods across different SNR values in CDL channels. As observed, compared to the JPEG + LDPC baseline, the proposed ViTMAE + VQ approach maintains consistently high SSIM and PSNR across the SNR range from 0 dB to 15 dB. However, when the SNR reaches 20 dB, the JPEG + LDPC baseline achieves better reconstruction quality, while the proposed method demonstrates strong robustness against noise and channel impairments under low-SNR conditions. This stable performance at lower SNR levels indicates that the integration of ViTMAE-based semantic encoding with standard 5G base station physical-layer modules, such as channel estimation and equalization, effectively mitigates the impact of channel impairments while maintaining compatibility with current infrastructure. The proposed approach thereby achieves enhanced robustness without relying on additional neural denoising components.

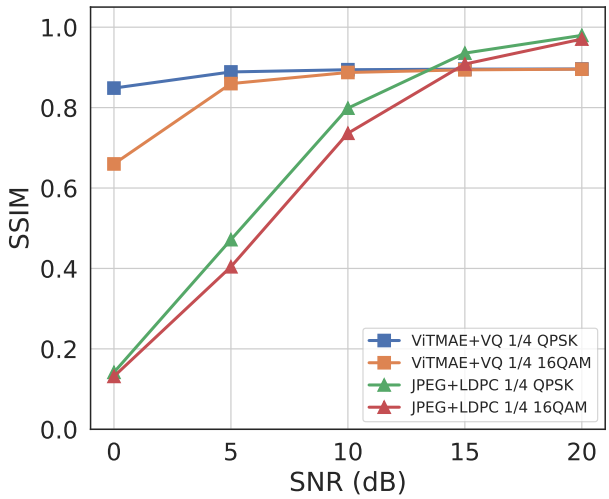

Figure 2: SSIM vs. SNR for different methods and configurations, where 1/4 represents the code rate of the LDPC coding, and QPSK and 16QAM indicate the modulation schemes.

In contrast, the JPEG+LDPC baselines exhibit a steep performance transition: both SSIM and PSNR remain very low at low SNR (e.g., SSIM $<$ 0.2 and PSNR $<$ 15 dB at SNR = 0 dB or below), improve rapidly around SNR = 10 dB, and eventually surpass the proposed method at high SNR (e.g., PSNR reaching 33 dB at SNR = 20 dB). This behavior stems from the bit-level reliability focus of conven-

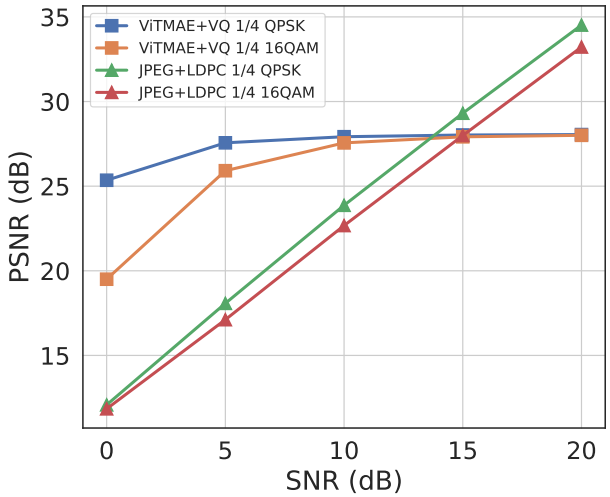

Figure 3: PSNR vs. SNR for different methods and configurations.

tional communication, where low SNR severely influences reconstruction quality. These results highlight the vulnerability of conventional compression and coding techniques in low-SNR conditions, where bit errors lead to reconstruction failures. It should be noted that in real-world low-SNR scenarios, traditional communication systems often employ hybrid automatic repeat request (HARQ) to enhance reliability through retransmissions. However, to ensure a fair comparison, our experiments consider only single-transmission scenarios, thereby highlighting the inherent limitations of conventional methods under challenging channel conditions.

Regarding configuration impacts, for modulation schemes, QPSK consistently outperforms 16QAM across all SNR conditions in both the proposed method and baseline systems, due to its stronger resilience to interference. The performance gap between the two modulation schemes gradually narrows as the SNR increases. In terms of LDPC code rate, lower rates provide more robust error protection in challenging channel conditions. Specifically, the combination of QPSK with a code rate of 1/4 achieves the best performance in both the proposed method and the baselines. However, while this configuration maximizes reliability, the use of low-order modulation and a low code rate significantly reduces the transmission throughput, making it suitable primarily for scenarios where reliability is prioritized over spectral efficiency. Another noteworthy observation is that the proposed method exhibits nearly identical SSIM and PSNR performance across all configurations when the SNR exceeds 10 dB. This indicates that the index error rate becomes sufficiently low at moderate to high SNR, allowing the ViTMAE model to effectively leverage its mask-based pre-training capability to reconstruct semantic content accurately, regardless of the specific modulation and coding settings.

We have also demonstrated the advantages of our proposed method in terms of transmission overhead. The average number of bits required to transmit each image under different configurations is recorded in Table 1. As can be seen, our method significantly reduces the transmission overhead compared to conventional JPEG with LDPC coding, which is primarily attributed to the fact that only the codebook indices need to be transmitted instead of the full feature representations. In particular, while the JPEG + LDPC configuration with a 1/4 code rate achieves the highest SSIM and PSNR among all baseline variants at SNR = 15 dB, its required bit budget is substantially larger than that of our proposed method. Notably, our approach maintains comparable reconstruction quality with only a marginal degradation in both SSIM and PSNR, further underscoring its efficiency in balancing performance and spectral utilization.

Table 1: Performance comparison of different models for OFDM-QPSK modulation at 15.0 dB SNR.

Model	SSIM	PSNR	Transmission bits
ViTMAE+VQ 1/4	0.90	28.1	2.4M
JPEG+LDPC 1/4	0.93	29.2	5.4M

Overall, benefiting from the seamless integration of OFDM communication modules, such as channel estimation and equalization, our model exhibits superior performance, particularly in challenging low-SNR conditions. The utilization of ViTMAE and the trainable codebook enables effective semantic extraction and reconstruction, maintaining high image quality even when transmission errors occur. Moreover, by transmitting only codebook indices, our approach significantly reduces the data volume compared to traditional methods while also preserving image quality. These results validate the efficacy of integrating ViTMAE with digital communication modules, offering a robust and efficient alternative for image transmission in realistic wireless scenarios.

Conclusion

This paper presents a novel semantic communication framework that combines ViTMAE with the OFDM system for robust image transmission. Our method enhances system compatibility, reduces data transmission overhead, and improves anti-interference through channel estimation, equalization, and MIM-based reconstruction. Evaluations in 3GPP CDL channels confirm superior PSNR and SSIM compared to the traditional JPEG+LDPC method, especially under low SNR. Future extensions could include video transmission or multimodal semantics, advancing practical semantic communication deployments.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Bourtsoulatze, E.; Kurka, D. B.; and Gündüz, D. 2019. Deep joint source-channel coding for wireless image transmis-

sion. *IEEE Transactions on Cognitive Communications and Networking*, 5(3): 567–579.

Fu, Q.; Xie, H.; Qin, Z.; Slabaugh, G.; and Tao, X. 2023. Vector quantized semantic communication system. *IEEE Wireless Communications Letters*, 12(6): 982–986.

Getu, T. M.; Kaddoum, G.; and Bennis, M. 2025. Semantic communication: A survey on research landscape, challenges, and future directions. *Proceedings of the IEEE*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, Q.; Zhang, G.; Qin, Z.; Cai, Y.; Yu, G.; and Li, G. Y. 2023. Robust semantic communications with masked VQ-VAE enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12): 8707–8722.

Liu, C.; Guo, C.; Yang, Y.; Ni, W.; and Quek, T. Q. 2024. OFDM-based digital semantic communication with importance awareness. *IEEE Transactions on Communications*, 72(10): 6301–6315.

Liu, Y.; Dong, C.; Liang, H.; Li, W.; Bao, Z.; Zheng, Z.; Xu, X.; and Zhang, P. 2025. Semantic Importance-Aware Reordering-Enhanced Semantic Communication System With OFDM Transmission. *IEEE Internet of Things Journal*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, B.; Li, R.; Zhu, J.; Zhao, Z.; and Zhang, H. 2023. Knowledge enhanced semantic communication receiver. *IEEE Communications Letters*, 27(7): 1794–1798.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, T.; Chen, Z.; He, D.; Qian, L.; Xu, Y.; Tao, M.; and Zhang, W. 2024. CDDM: Channel denoising diffusion models for wireless semantic communications. *IEEE Transactions on Wireless Communications*, 23(9): 11168–11183.

Xie, H.; Qin, Z.; Han, Z.; and Letaief, K. B. 2025. Hybrid digital-analog semantic communications. *IEEE Journal on Selected Areas in Communications*.

Xie, H.; Qin, Z.; Li, G. Y.; and Juang, B.-H. 2021. Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69: 2663–2675.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Xu, J.; Ai, B.; Chen, W.; Yang, A.; Sun, P.; and Rodrigues, M. 2021. Wireless image transmission using deep source channel coding with attention modules. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2315–2328.

Yang, K.; Wang, S.; Dai, J.; Qin, X.; Niu, K.; and Zhang, P. 2024. Swinjscc: Taming swin transformer for deep joint source-channel coding. *IEEE Transactions on Cognitive Communications and Networking*.

Yang, M.; Bian, C.; and Kim, H.-S. 2022. OFDM-guided deep joint source channel coding for wireless multipath fading channels. *IEEE Transactions on Cognitive Communications and Networking*, 8(2): 584–599.

Yang, M.; and Kim, H.-S. 2022. Deep joint source-channel coding for wireless image transmission with adaptive rate control. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5193–5197. IEEE.

Zheng, W.; Liang, H.; Dong, C.; and Xu, X. 2025. Deep Joint Source-Channel Coding Based on Feedback-Driven Codebook Optimization. In *2025 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6. IEEE.