Vision Language Models for Massive MIMO Semantic Communication

Stephen D. Liang

Hewlett Packard Enterprise, 6280 America Center Dr, San Jose, CA 95002, USA

Abstract

This paper presents a semantic communication scheme that utilizes the Vision-Language Models (VLMs) to enable efficient image transmission over Massive Multiple Input Multiple Output (MIMO) systems. By transmitting textual descriptions instead of raw image data, the proposed approach significantly reduces bandwidth usage while ensuring high-quality image reconstruction at the receiver. At the transmitter, a textual description of the image is generated using Bootstrapping Language-Image Pre-training (BLIP), converted to bits, modulated, and transmitted over the Massive MIMO channel. At the receiver, the transmitted text is used to reconstruct the image through a text-to-image generation model based on Stable Diffusion. We detail the system architecture, semantic communication framework, and evaluate the method's performance in terms of bandwidth efficiency, image reconstruction quality, and semantic similarity. Simulation results demonstrate that while semantic communication achieves excellent bandwidth efficiency, the image reconstruction quality, measured by structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR), is relatively low. However, the semantic similarity is exceptionally high, aligning with the primary objective of semantic communication.

1 Introduction

As the demand for high-speed, reliable, and intelligent communication continues to grow, the next generation of communication systems is increasingly adopting semantic communication principles (Luo, Chen, and Guo 2022). Unlike traditional communication systems that focus on transmitting raw bits, semantic communication aims to convey the meaning or intent of the transmitted information, significantly improving efficiency and relevance in data transmission. This is especially important in scenarios where resource constraints, such as bandwidth and latency, are critical considerations.

In this context, Massive Multiple-Input Multiple-Output (Massive MIMO) systems (Marzetta 2015) play a crucial role, providing high spectral efficiency and robust communication in wireless networks. Massive MIMO leverages a large number of antennas to simultaneously serve multiple users, thereby achieving superior data rates and connectivity. However, as data streams in modern applications often include a mix of visual and textual modalities—such as images and associated descriptive text—there is a pressing need to integrate Vision-Language Models (VLMs) into the semantic communication framework.

VLMs, such as Contrastive Language–Image Pre-training (CLIP) models (Radford et al. 2021), have demonstrated remarkable capabilities in understanding and generating multimodal content by embedding images and text into a shared semantic space. Massive MIMO technology, on the other hand, provides high spectral efficiency and reliability in wireless communications. By incorporating VLMs into Massive MIMO semantic communication systems, we can achieve the following:

1. Semantic-Aware Transmission: Instead of transmitting all data, VLMs can extract and prioritize meaningful semantic features from images and text, reducing the communication overhead. For example, rather than transmitting an entire image, a system could transmit a semantic summary or key features relevant to the communication goal.
2. Multimodal Understanding: Many real-world communication scenarios require the transmission of multimodal data (e.g., a drone communicating visual observations and associated metadata). VLMs enable the efficient joint processing of these modalities, ensuring the transmitted data is both contextually relevant and meaningful.
3. Noise-Resilient Communication: By focusing on high-level semantic features rather than raw data, VLMs make the system more robust to channel noise and interference, as semantic information is often more resilient to distortions.

In this paper, we propose a semantic communication scheme that integrates VLMs with Massive MIMO to enable efficient image transmission through textual descriptions. By leveraging VLMs at both the transmitter and receiver, the system reconstructs images from text, drastically minimizing the amount of data required for transmission.

2 Related Work

Multimodal Machine Learning and VLMs: The field of multimodal machine learning explores the design and application of models that process, generate, or integrate data from multiple modalities (Baltrušaitis, Ahuja, and Morency 2018). For instance, Zhao et al. (Zhao, Gong, and Li 2022) introduced a hierarchical transformer for integrating auditory and visual inputs. Huang et al. (Huang et al. 2020) employed a multimodal transformer to fuse audio-visual data and create multimodal emotional intermediate representations within the semantic feature space. Rahman et al. (Rahman et al. 2020) utilized multimodal nonverbal data for fine-tuning based on acoustic and visual modalities. Similarly, Le et al. (Le et al. 2019) explored the use of non-textual data for multimodal transformer architectures. To optimize parameter efficiency in multimodal transformers, Lee et al. (Lee et al. 2020) focused on audio-visual representation learning. Xie et al. (Xie, Sidulova, and Park 2021) developed an emotion recognition algorithm that leveraged distinct models for audio and visual information. Liang et al. (Liang and Mendel 2022) proposed a parallel concatenated architecture for a multimodal transformer utilizing visual-audio data. Parthasarathy et al. (Parthasarathy and Sundaram 2021) explored audio-visual detection and proposed expression tracking using a transformer framework. Dzabraev et al. (Dzabraev et al. 2021) introduced a multidomain multimodal transformer to integrate multiple video caption datasets. In (Zhou et al. 2022b), Context Optimization (CoOp) was proposed for adapting CLIP-like VLMs for downstream image recognition. A Conditional Context Optimization (CoCoOp) was proposed to extend CoOp by further learning a neural network to generate an input-conditional token (vector) for each image (Zhou et al. 2022a). In (Zhu et al. 2023), MiniGPT-4 was proposed which uncovers that aligning the visual features with a large language model can possess many advanced multi-modal abilities shown in GPT-4. BLIP (Bootstrapping Language-Image Pre-training) is a cutting-edge VLM developed by Salesforce Research that bridges vision and language (Li et al. 2022). It processes both images and text, enabling tasks like image captioning, visual question answering (VQA), and image-text retrieval. Built on a transformer-based architecture, BLIP uses self-supervised pretraining to align visual and textual representations effectively. It excels in generating dynamic captions, answering image-related questions, and matching images with text, performing robustly even in few-shot learning scenarios. BLIP is versatile, making it a powerful tool for vision-language applications.

Semantic Communications: Our work emphasizes the application of Vision-Language Models (VLMs) in semantic communications. Yoo et al. (Yoo et al. 2022) applied a vision transformer to semantic communication, targeting the transmission of meaning over symbol precision. Their model included an image encoder, channel layer, and image reconstruction block. Wang et al. (Wang et al. 2022) utilized a transformer for wireless interference recognition, reducing computational complexity through regional self-attention calculations. Bi et al. (Bi et al. 2022) incorporated a convolutional transformer into massive MIMO systems for channel state information feedback. Xie et al. (Xie et al. 2022) adapted transformers for task-oriented semantic communication in both single-modal and multimodal scenarios. Liu et al. (Liu et al. 2022) proposed a semantic communication architecture addressing source compression and resource allocation. Semantic communication was also employed to lower data collection and offloading costs for edge computing providers (Luong et al. 2024). Szott et al. (Szott et al. 2022) reviewed machine learning techniques in WiFi, while Xie et al. (Xie, Qin, and Li 2023) proposed a universal transformer-based transceiver with a memory module to extract semantic information. Weng et al. (Weng et al. 2023) explored speech recognition and synthesis for deep-learning-driven semantic communication. Additionally, Wu et al. (Wu et al. 2024) applied transformers for wireless image communication leveraging channel feedback. Our proposed multimodal transformer integrates massive MIMO, accommodates varying modality requirements, and supports multi-user scenarios. In (Nam et al. 2023), sequential semantic generative communication was proposed for progressive text-to-image generation. In (Nam et al. 2024), a framework of language-oriented semantic communication was proposed using semantic source coding and channel coding. A language-oriented semantic communication framework was proposed in (Cicchetti et al. 2024) using text and image embedding as a latent diffusion model for image reconstruction. In (Jiang et al. 2024), a VLM was proposed for semantic communication using attention mechanisms to adjust the semantic coding and the channel coding in response different SNR.

Massive MIMO: Massive MIMO technology, where cellular base stations are equipped with numerous antennas, enables significant improvements in spectral and energy efficiency (Lu et al. 2014; Björnson et al. 2019). It is scalable to any desired level, with additional antennas enhancing throughput, reducing radiated power, simplifying signal processing, and ensuring uniform service across the cellular network (Marzetta 2015; Marzetta and Yang 2016). Larsson et al. (Larsson et al. 2014) contrasted massive MIMO with multi-user MIMO, demonstrating the non-scalability of multi-user MIMO due to equal transmit and receive antenna counts, while highlighting massive MIMO's capability for simultaneous communication with multiple users. Gao et al. (Gao et al. 2015) evaluated massive MIMO systems in real propagation environments using a virtual linear array of 128 antenna ports operating at 2.6 GHz. Björnson et al. (Björnson et al. 2015) investigated the optimal antenna count, number of active users, and transmit power in massive MIMO systems. Ngo et al. (Ngo et al. 2017) introduced cell-free massive MIMO, where antennas are distributed rather than co-located but jointly utilize the same time and frequency resources for communication. Recently, Zheng et al. (Zheng et al. 2024) examined challenges and solutions for mobile cell-free massive MIMO systems.

3 System Model

The system model describes the framework for transmitting semantic information, represented as textual descriptions of images, over a Massive MIMO communication channel. This involves two primary components: the transmitter and receiver, along with specific assumptions about the channel and system.

3.1 Overview

The proposed system integrates semantic communication with a Massive MIMO setup. The transmitter takes an input image $\mathbf{I}$ and processes it to extract a semantic textual description S using BLIP. This textual description is then compressed and modulated for transmission over a Massive MIMO channel. By transmitting the semantic representation (text) instead of the raw image, the system achieves significant bandwidth efficiency. The receiver decodes the transmitted text $\hat{S}$ and reconstructs the original image $\hat{\mathbf{I}}$ using a text-to-image generation model guided by the pretrained CLIP embeddings in stable diffusion. The receiver ensures semantic alignment by leveraging the shared knowledge from the CLIP model, enabling accurate reconstruction of the transmitted meaning. The system leverages the synergy between semantic models (e.g., BLIP) and Massive MIMO technology to optimize both bandwidth usage and communication reliability.

3.2 Detailed Workflow

The workflow can be broken down into the following stages:

1. Image Processing and Captioning at the Transmitter: The input image $\mathbf{I}$ is first passed through an image captioning model $\mathcal{C}$, which generates a concise textual description S that captures the semantic content of the image. For example, an image of a dog in a park, the caption might be $S =$ ”A dog playing in the park.”.
2. Text Compression and Transmission: The textual description S is optionally encoded into a semantic vector $\mathbf{z}_T$ using BLIP’s text encoder f_T. This vector or the text itself is then compressed using an entropy coding scheme and modulated into symbols $\mathbf{s}$. The symbols $\mathbf{s}$ are precoded and transmitted through the Massive MIMO channel.
3. Reception and Decoding: At the receiver, the transmitted symbols are detected, demodulated, and decompressed to recover the textual description $\hat{S}$. For example, if the received signal corresponds to the textual description S, the recovered description might still be $\hat{S} =$ ”A dog playing in the park” even in the presence of some channel noise.
4. Image Reconstruction: The receiver uses $\hat{S}$ as input to a text-to-image model $\mathcal{G}$, which reconstructs the image $\hat{\mathbf{I}}$. Then semantic alignment between the original and reconstructed images is ensured by leveraging CLIP’s multimodal embedding space.

3.3 Assumptions

To simplify the analysis and implementation of the proposed system, the following assumptions are made:

- Massive MIMO Channel: The communication system employs a Massive MIMO setup, characterized by N_t transmit antennas and N_r receive antennas. Massive MIMO is chosen due to its ability to enhance spectral efficiency and mitigate interference through spatial multiplexing and beamforming.
- Channel Model: The propagation environment is modeled as a Rayleigh flat fading channel, where the channel matrix $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ captures the gain between each pair of transmit and receive antennas. The channel matrix remains constant for the duration of the transmission (block fading).
- Perfect Synchronization: Perfect time and frequency synchronization between the transmitter and receiver are assumed to avoid inter-symbol interference.
- Channel State Information (CSI): The transmitter has perfect knowledge of the channel matrix $\mathbf{H}$, enabling optimal precoding for minimizing interference and maximizing received signal quality.

4 VLM for Semantic Communication

4.1 Transmitter Design

The transmitter design involves converting a high-dimensional image into a compact and semantically meaningful representation suitable for efficient transmission over a Massive MIMO channel. This process includes multiple stages: image captioning, text encoding, data compression, modulation, precoding, and transmission.

Image Captioning Image captioning is the process of generating a textual description S that semantically represents the content of a given image $\mathbf{I}$. We use BLIP for image captioning. BLIP processes the image $\mathbf{I}$ to generate a sequence of words $S = \{w_1, w_2, \ldots, w_T\}$ describing the image:

$$\mathbf{S} = \mathcal{F}(\mathbf{I}) \tag{1}$$

The output is a semantically rich textual description S, such as “A dog playing in the park.”

Text Encoding and Data Compression Once the textual description S is generated, it can be further processed to produce a compact semantic embedding using BLIP’s text encoder f_T. The text encoding step maps the description S into a vector representation $\mathbf{z}_T$ in the shared multimodal embedding space of BLIP:

$$\mathbf{z}_T = f_T(S) \tag{2}$$

This embedding captures the semantic meaning of the text and aligns with the corresponding image representation in BLIP’s embedding space. Using $\mathbf{z}_T$ ensures that only the essential semantic information is retained, reducing redundancy and improving robustness. The textual embedding $\mathbf{z}_T$ is then compressed using a lossless data compression scheme $\mathcal{E}$ such as Huffman coding to reduce the data size before transmission (Moffat 2019):

$$\mathbf{c} = \mathcal{E}(\mathbf{z}_T) \tag{3}$$

Modulation and Precoding The compressed data $\mathbf{c}$ is then converted into symbols $\mathbf{s}$ using a modulation scheme $\mathcal{M}$, such as Quadrature Amplitude Modulation (QAM) to encode data as points in a two-dimensional signal space. The modulated symbols are represented as:

$$\mathbf{s} = \mathcal{M}(\mathbf{c}) \tag{4}$$

For Massive MIMO systems, precoding is employed to optimize the transmitted signals for the wireless channel. A precoding matrix $\mathbf{W}$ is applied to the symbols $\mathbf{s}$ to enhance spatial multiplexing and mitigate interference:

$$\mathbf{x} = \mathbf{W}\mathbf{s} \tag{5}$$

The precoding matrix $\mathbf{W}$ is typically computed based on the CSI and can include techniques such as Zero-Forcing or Minimum Mean Square Error precoding.

4.2 Channel Model

The precoded signal $\mathbf{x}$ is then transmitted over the Massive MIMO channel. The channel is modeled as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{6}$$

Where:

- $\mathbf{y} \in \mathbb{C}^{N_r}$: The received signal vector at the receiver, with N_r being the number of receiving antennas.
- $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$: The channel matrix that models the propagation environment, with N_t transmit antennas and N_r receive antennas. Each entry h_{ij} represents the complex gain between the i-th receive antenna and the j-th transmit antenna.
- $\mathbf{x} \in \mathbb{C}^{N_t}$: The transmitted signal vector after precoding, with N_t transmit antennas.
- $\mathbf{n} \sim \mathcal{CN}(0, \sigma_n^2 \mathbf{I})$: Additive white Gaussian noise (AWGN) vector with zero mean and variance σ_n^2.

Precoding and equalization require knowledge of the channel matrix $\mathbf{H}$. In this paper, we assume the exact $\mathbf{H}$ is known. In real world, it can be estimated based on pilot signaling.

4.3 Receiver Design

The receiver's task is to recover the transmitted compressed data $\mathbf{c}$ (or equivalently S) from the noisy received signal $\mathbf{y}$. This involves three main steps: detection and equalization, demodulation and decoding, and image reconstruction.

Detection and Equalization The received signal $\mathbf{y}$ is first processed using a linear detector $\mathbf{G}$ to estimate the transmitted symbols $\mathbf{s}$:

$$\hat{\mathbf{s}} = \mathbf{G}\mathbf{y} \tag{7}$$

Several equalization methods can be employed depending on the channel conditions and available CSI:

- Zero-Forcing (ZF): Mitigates inter-stream interference by inverting the channel matrix (Ding et al. 2003):

  $$\mathbf{G}_{\text{ZF}} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \tag{8}$$

  where $\mathbf{H}^H$ is the Hermitian (conjugate transpose) of $\mathbf{H}$. ZF minimizes interference but can amplify noise in low-SNR scenarios.
- Minimum Mean Square Error (MMSE): Balances noise and interference by optimizing the mean square error (Jiang, Varanasi, and Li 2011):

  $$\mathbf{G}_{\text{MMSE}} = (\mathbf{H}^H \mathbf{H} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{H}^H \tag{9}$$

  MMSE performs better in noisy environments compared to ZF.

Demodulation and Decoding After equalization, the detected symbols $\hat{\mathbf{s}}$ are demodulated to recover the compressed data $\hat{\mathbf{c}}$ (Chen and Laneman 2006):

$$\hat{\mathbf{c}} = \mathcal{M}^{-1}(\hat{\mathbf{s}}) \tag{10}$$

Here, $\mathcal{M}^{-1}$ is the demodulation function that maps received symbols back to their corresponding bits.

The compressed data $\hat{\mathbf{c}}$ is then decompressed using the Huffman decoding (Moffat 2019) $\mathcal{E}^{-1}$:

$$\hat{S} = \mathcal{E}^{-1}(\hat{\mathbf{c}}) \tag{11}$$

This step reconstructs the textual description $\hat{S}$ that was transmitted by the sender.

Image Reconstruction Finally, the received textual description $\hat{S}$ is used to reconstruct the image $\hat{\mathbf{I}}$ using a pretrained text-to-image generation model $\mathcal{G}$:

$$\hat{\mathbf{I}} = \mathcal{G}(\hat{S}) \tag{12}$$

- The model $\mathcal{G}$ uses advanced generative techniques to generate an image that aligns semantically with the textual description $\hat{S}$. In this paper, we use Stable Diffusion for $\mathcal{G}$. Stable Diffusion is an open-source AI model developed by Stability AI that generates images from textual descriptions (Sauer et al. 2024). It utilizes a diffusion process to create detailed images by progressively denoising random noise in alignment with a given text prompt.
- To ensure the reconstructed image retains semantic alignment with the original image, CLIP's image encoder f_I is applied:

  $$\mathbf{z}_I = f_I(\hat{\mathbf{I}}) \tag{13}$$

  Here, $\mathbf{z}_I$ represents the embedding of the reconstructed image.

5 Experiment and Performance Analysis

In our experiment, we used the Microsoft Common Objects in Context (MSCOCO) dataset (Lin et al. 2014). MSCOCO is a large-scale, richly annotated dataset designed to advance research in computer vision by providing a diverse set of images depicting complex scenes with multiple objects in natural contexts. It includes over 330,000 images annotated with 91 object categories, bounding boxes, pixel-level segmentations, and natural language captions, making it ideal for tasks like object detection, segmentation, image captioning, and pose estimation. Its realistic scenarios and detailed annotations have established it as a key benchmark in the field. The performance of the proposed semantic communication system is analyzed in terms of bandwidth efficiency, image reconstruction quality, and semantic similarity. These metrics evaluate the trade-offs between efficiency, accuracy, and the preservation of semantic meaning during the communication process.

5.1 Bandwidth Efficiency

Bandwidth efficiency is a key advantage of the proposed system. By transmitting textual descriptions instead of full image data, the system achieves a significant reduction in the amount of data transmitted. The reduction ratio R is defined as:

$$R = \frac{\text{Size of Image Data}}{\text{Size of Text Data}} \tag{14}$$

A higher value of R indicates better bandwidth savings.

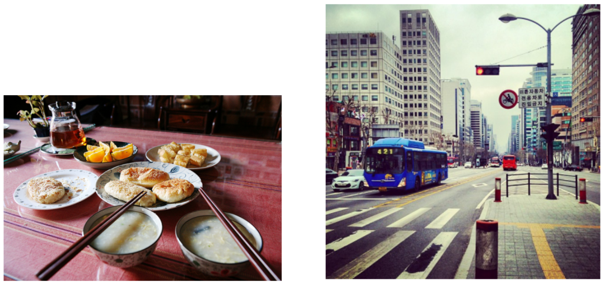
For example, the RGB image in Fig. 1a has size of $350 \times 515 \times 3$, and it has 370,878 bytes. We applied BLIP captioning, we got the the caption “a table with food”, which has 6 tokens and 48 bytes, so the reduction ratio is:

$$R = \frac{370878}{48} = 7726.6 \tag{15}$$

This demonstrates that the system transmits only $1/7726.6 = 0.013\%$ of the original data. The RGB image in Fig. 1b has size of $389 \times 389 \times 3$, and it has 310,016 bytes. We applied BLIP captioning, and got the the caption “a bus is driving down the street in the city”, which has 10 tokens and 80 bytes, so the reduction ratio is:

$$R = \frac{310016}{80} = 3875.2 \tag{16}$$

so the system transmits only $1/3875.2 = 0.026\%$ of the original data, drastically reducing bandwidth usage.



(a)

(b)

Figure 1: Two images at the transmit with BLIP-generated captioning. (a) “a table with food”, (b) “a bus is driving down the street in the city”.

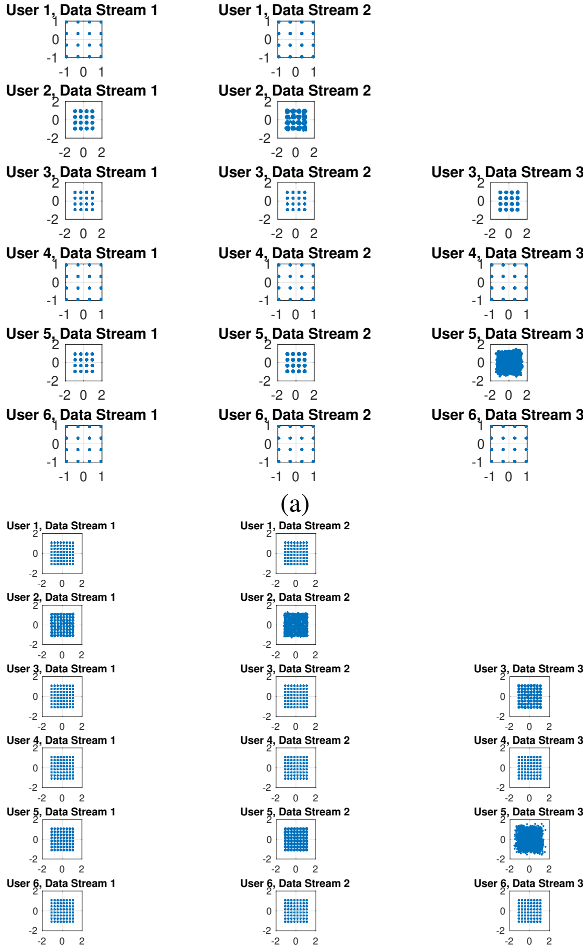
5.2 Massive MIMO

We conducted simulations to evaluate the performance of a VLM for massive MIMO semantic communication. The system comprises six users sharing the massive MIMO channel, with each user allocated a different number of data streams. Specifically, the number of data streams per user is 2, 2, 3, 3, 3, and 3, respectively. The massive MIMO configuration includes 128 transmit antennas and 64 receive antennas, resulting in a system size of 128×64. We evaluated the system using different modulation schemes, namely 16-QAM and 64-QAM, for comparative analysis.

The beamforming approach utilized joint spatial division multiplexing (JSDM) (Li, Han, and Molisch 2016; Adhikary et al. 2013). For data transmission, orthogonal frequency-division multiplexing (OFDM) was employed, with the number of OFDM subcarriers set to 256. A single-bounce ray tracing approximation with a parametrized number of scatterers was used as the scattering model (Shiu et al. 2000). The number of scatterers was set to 100. The channel model was configured for non-line-of-sight (NLoS) communication, operating at a radio frequency (RF) of 28 GHz.

In Figure 2, we illustrate the system’s output performance for 16-QAM and 64-QAM modulation schemes after processing through massive MIMO. Although some signal constellations appear slightly mixed, all configurations achieve a bit error rate (BER) of 0 after channel coding. The simulations were conducted at an SNR value of 5 dB.

To further explore system performance, we extended the simulations to higher-order QAM modulations, including 1024-QAM and 4096-QAM. The results, presented in Figure 3, depict the BER for these modulation schemes.



(a)

(b)

Figure 2: QAM signal output after massive MIMO processing. (a) 16-QAM, (b) 64-QAM.

5.3 Image Reconstruction Quality

Image reconstruction quality measures how accurately the reconstructed image $\hat{\mathbf{I}}$ matches the original image $\mathbf{I}$. Two widely used metrics are employed, Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

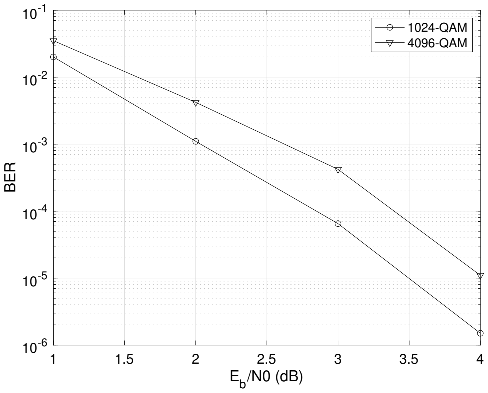


Figure 3: BER performance for 1024-QAM and 4096-QAM modulation schemes in massive MIMO.

Structural Similarity Index (SSIM) SSIM measures the perceptual similarity between the original and reconstructed images by comparing luminance, contrast, and structural details. It is defined as:

$$\text{SSIM}(\mathbf{I}, \hat{\mathbf{I}}) = \frac{(2\mu_{\mathbf{I}}\mu_{\hat{\mathbf{I}}} + C_1)(2\sigma_{\mathbf{I}\hat{\mathbf{I}}} + C_2)}{(\mu_{\mathbf{I}}^2 + \mu_{\hat{\mathbf{I}}}^2 + C_1)(\sigma_{\mathbf{I}}^2 + \sigma_{\hat{\mathbf{I}}}^2 + C_2)} \tag{17}$$

Where:

- $\mu_{\mathbf{I}}$ and $\mu_{\hat{\mathbf{I}}}$: Mean pixel values of the original and reconstructed images.
- $\sigma_{\mathbf{I}}^2$ and $\sigma_{\hat{\mathbf{I}}}^2$: Variances of the original and reconstructed images.
- $\sigma_{\mathbf{I}\hat{\mathbf{I}}}$: Covariance between the original and reconstructed images.
- C_1 and C_2: Stabilization constants to avoid division by zero.

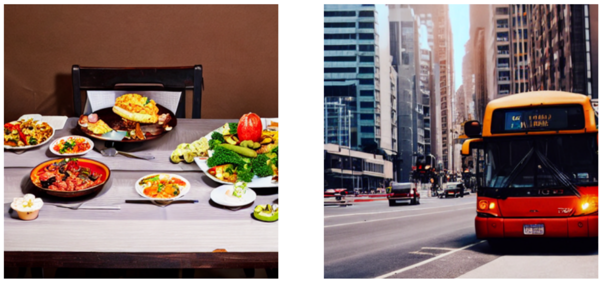
SSIM values range from 0 to 1, where 1 indicates perfect similarity. This metric is particularly useful for evaluating perceptual quality, as it aligns with human visual perception.

In Fig. 4, we plotted the two generated images at the receiver for the two transmitted images in Fig 1 based on stable diffusion, then we applied BLIP captioning, and obtained their captions as “a table with a variety of food on it” and “a bus is driving down the street in the city”, respectively. Observe that one caption is exactly the same as the caption in the transmit side.

We computed the SSIM between the two images of Fig. 1a and Fig. 4a, and their SSIM is 0.167; and the SSIM between the two images of Fig. 1b and Fig. 4b, is 0.146. This indicates a low degree of perceptual similarity, suggesting significant differences in the visual content or structure of the two images. However, semantic communication cares more about meaning instead of image similarity.

Peak Signal-to-Noise Ratio (PSNR) PSNR evaluates the fidelity of the reconstructed image by comparing the mean squared error (MSE) between the original and reconstructed images. It is defined as:

$$\text{PSNR} = 10 \log_{10}\left(\frac{L^2}{\text{MSE}}\right) \tag{18}$$



(a) (b)

Figure 4: Two generated images at the receiver and their BLIP captioning. (a) “a table with a variety of food on it”. (b) “a bus is driving down the street in the city”.

Where:

- L: The maximum possible pixel value (e.g., 255 for 8-bit images).
- MSE: The mean squared error between the pixel values of the original and reconstructed images:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{I}_i - \hat{\mathbf{I}}_i)^2 \tag{19}$$

  Here, N is the total number of pixels in the image, and $\mathbf{I}_i$ and $\hat{\mathbf{I}}_i$ are the pixel intensities of the original and reconstructed images, respectively.

Higher PSNR values indicate better reconstruction quality, with values above 30 dB generally considered good for visual tasks. The PSNR between the two food images (Fig. 1a and Fig. 4a) is approximately 8.78 dB, and PSNR between the two bus images (Fig. 1b and Fig. 4b) is approximately 8.28 dB. The relatively low PSNR values suggests a significant level of difference or distortion between the original images and reconstructed images.

5.4 Semantic Similarity

Semantic similarity measures the alignment between the semantic meaning of the original and reconstructed images. This is particularly important in semantic communication, where preserving meaning is more critical than achieving pixel-perfect reconstruction.

The semantic similarity is evaluated using the cosine similarity between the text embedding $\mathbf{z}_T$ of the transmitted description S and the image embedding $\mathbf{z}_I$ of the reconstructed image $\hat{\mathbf{I}}$.

In the above two examples of captioning, the size the two captions have different sizes. To make them size size, we used CLIP embeddings (Radford et al. 2021) CLIP embeddings are always of the same fixed size because the model is designed to produce consistent-dimensional representations for both text and images (Radford et al. 2021). The text encoder processes tokenized inputs, which are padded or truncated to a maximum sequence length (e.g., 77 tokens), and outputs embeddings that are reduced to a fixed size, typically by using the [CLS] token or mean pooling. Similarly, the image encoder maps images to the same fixed-dimensional vector space, regardless of their resolution. For instance, in the CLIP model clip-vit-base-patch32, both text and image embeddings are 512-dimensional. This design ensures compatibility between embeddings and facilitates tasks like similarity computation, retrieval, and classification within a shared semantic space.

The semantic similarity is computed as

$$\cos(\theta) = \frac{\mathbf{z}_T \cdot \mathbf{z}_I}{\|\mathbf{z}_T\| \|\mathbf{z}_I\|} \tag{20}$$

Where:

- $\mathbf{z}_T$: The embedding of the transmitted text description, obtained from CLIP's text encoder.
- $\mathbf{z}_I$: The embedding of the reconstructed image, obtained from CLIP's image encoder.
- $\| \cdot \|$: The Euclidean norm of the vector.
- $\mathbf{z}_T \cdot \mathbf{z}_I$: The dot product of the two embeddings.

Cosine similarity values range from -1 to 1:

- A value of 1 indicates perfect semantic alignment.
- A value of 0 indicates no correlation between the two embeddings.
- Negative values suggest opposing semantics.

A high cosine similarity indicates that the reconstructed image $\hat{\mathbf{I}}$ effectively retains the semantic meaning of the original image $\mathbf{I}$ as expressed through the transmitted text description S. The semantic similarity between the two food images (Fig. 1a and Fig. 4a) is 0.9076, while the semantic similarity between the two bus images (Fig. 1b and Fig. 4b) is exactly 1. This highlights the exceptional success of semantic communication.

We conducted the semantic communication process on the entire MSCOCO dataset with E_b/N_0 values ranging from 1 dB to 4 dB using the massive MIMO channel described in Section 5.2. The system consists of six users sharing a massive MIMO channel, with each user assigned a different number of data streams: 2, 2, 3, 3, 3, and 3 streams, respectively. The massive MIMO setup includes 128 transmit antennas and 64 receive antennas, resulting in a 128×64 system configuration. We utilized 1024-QAM modulation, with all six users transmitting images simultaneously. For each E_b/N_0 value, the semantic similarity scores were averaged across all images. Figure 5 illustrates the relationship between semantic similarity and E_b/N_0. Observe that with higher E_b/N_0, the semantic similarity increases.

5.5 Trade-offs and Discussion

The semantic communication has some trade-offs between bandwidth efficiency, reconstruction quality, and semantic fidelity:

- Bandwidth Efficiency vs. Quality: Reducing the size of the transmitted text improves bandwidth efficiency but may degrade reconstruction quality due to limited information.

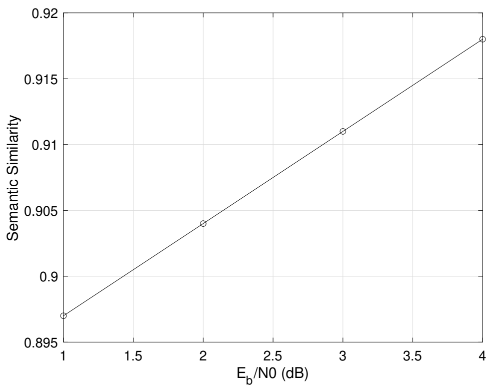


Figure 5: Semantic similarity versus E_b/N_0 in massive MIMO semantic communication.

- Semantic Similarity vs. Pixel Fidelity: High semantic similarity ensures meaningful content but does not guarantee pixel-perfect reconstructions.
- Noise Resilience: The use of semantic embeddings provides robustness against channel noise, as small perturbations in text embeddings or transmitted descriptions have minimal impact on reconstructed meaning. Observe Fig. 5, the semantic similarity just slightly increases with $E_b/N0$ increases from 1 dB to 4 dB.

By optimizing these trade-offs, the proposed system demonstrates its potential for efficient and meaningful image transmission in bandwidth-constrained scenarios.

6 Conclusions and Future Work

In this work, we have presented a semantic communication scheme designed to efficiently transmit images by converting them into textual descriptions and transmitting these over Massive MIMO channels using VLMs. The proposed system introduces several notable advantages. By transmitting compact textual descriptions instead of raw image data, the approach achieves significant bandwidth efficiency, reducing the data requirements for image transmission. Furthermore, the semantic representation S demonstrates robustness to noise, as minor distortions in the transmitted text or its embeddings result in minimal degradation of the reconstructed image. The use of a Massive MIMO configuration enables the simultaneous transmission of multiple data streams, making the system scalable for multiple users or higher data rates. Additionally, by leveraging the capabilities of VLMs, the proposed model ensures that the reconstructed images retain the semantic essence of the original inputs, even when pixel-level details are not perfectly preserved.

Looking ahead, future work will focus on optimizing the image captioning and generation models to enhance the quality of the reconstructed images further. Moreover, exploring real-time implementation and deployment aspects, including latency minimization and hardware integration, will be critical to advancing the practical feasibility of the proposed system.

References

Adhikary, A.; Nam, J.; Ahn, J.-Y.; and Caire, G. 2013. Joint spatial division and multiplexing—The large-scale array regime. *IEEE transactions on information theory*, 59(10): 6441–6463.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bi, X.; Li, S.; Yu, C.; and Zhang, Y. 2022. A Novel Approach Using Convolutional Transformer for Massive MIMO CSI Feedback. *IEEE Wireless Communications Letters*, 11(5): 1017–1021.

Björnson, E.; Sanguinetti, L.; Hoydis, J.; and Debbah, M. 2015. Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer? *IEEE Transactions on wireless communications*, 14(6): 3059–3075.

Björnson, E.; Sanguinetti, L.; Wymeersch, H.; Hoydis, J.; and Marzetta, T. L. 2019. Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays. *Digital Signal Processing*, 94: 3–20.

Chen, D.; and Laneman, J. N. 2006. Modulation and demodulation for cooperative diversity in wireless systems. *IEEE Transactions on Wireless Communications*, 5(7): 1785–1794.

Cicchetti, G.; Grassucci, E.; Park, J.; Choi, J.; Barbarossa, S.; and Comminiello, D. 2024. Language-Oriented Semantic Latent Representation for Image Transmission. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

Ding, Y.; Davidson, T. N.; Luo, Z.-Q.; and Wong, K. M. 2003. Minimum BER block precoders for zero-forcing equalization. *IEEE Transactions on Signal Processing*, 51(9): 2410–2423.

Dzabraev, M.; Kalashnikov, M.; Komkov, S.; and Petiushko, A. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3354–3363.

Gao, X.; Edfors, O.; Rusek, F.; and Tufvesson, F. 2015. Massive MIMO performance evaluation based on measured propagation data. *IEEE Transactions on Wireless Communications*, 14(7): 3899–3911.

Huang, J.; Tao, J.; Liu, B.; Lian, Z.; and Niu, M. 2020. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3507–3511. IEEE.

Jiang, F.; Tang, C.; Dong, L.; Wang, K.; Yang, K.; and Pan, C. 2024. Visual Language Model based Cross-modal Semantic Communication Systems. ArXiv:2407.00020 [cs.CV], arXiv:2407.00020.

Jiang, Y.; Varanasi, M. K.; and Li, J. 2011. Performance analysis of ZF and MMSE equalizers for MIMO systems: An in-depth study of the high SNR regime. *IEEE Transactions on Information Theory*, 57(4): 2008–2026.

Larsson, E. G.; Edfors, O.; Tufvesson, F.; and Marzetta, T. L. 2014. Massive MIMO for next generation wireless systems. *IEEE communications magazine*, 52(2): 186–195.

Le, H.; Sahoo, D.; Chen, N. F.; and Hoi, S. C. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*.

Lee, S.; Yu, Y.; Kim, G.; Breuel, T.; Kautz, J.; and Song, Y. 2020. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Z.; Han, S.; and Molisch, A. F. 2016. Hybrid beamforming design for millimeter-wave multi-user massive MIMO downlink. In *2016 IEEE International Conference on Communications (ICC)*, 1–6. IEEE.

Liang, S. D.; and Mendel, J. M. 2022. Multimodal Transformer for Parallel Concatenated Variational Autoencoders. *arXiv preprint arXiv:2210.16174*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, C.; Guo, C.; Yang, Y.; and Jiang, N. 2022. Adaptable Semantic Compression and Resource Allocation for Task-Oriented Communications. *arXiv preprint arXiv:2204.08910*.

Lu, L.; Li, G. Y.; Swindlehurst, A. L.; Ashikhmin, A.; and Zhang, R. 2014. An overview of massive MIMO: Benefits and challenges. *IEEE journal of selected topics in signal processing*, 8(5): 742–758.

Luo, X.; Chen, H.-H.; and Guo, Q. 2022. Semantic communications: Overview, open issues, and future research directions. *IEEE Wireless Communications*, 29(1): 210–219.

Luong, N. C.; Le Van, T.; Feng, S.; Du, H.; Niyato, D.; and Kim, D. I. 2024. Edge computing for metaverse: Incentive mechanism versus semantic communication. *IEEE Transactions on Mobile Computing*, 41(5): 6196–6211.

Marzetta, T. L. 2015. Massive MIMO: an introduction. *Bell Labs Technical Journal*, 20: 11–22.

Marzetta, T. L.; and Yang, H. 2016. *Fundamentals of massive MIMO*. Cambridge University Press.

Moffat, A. 2019. Huffman coding. *ACM Computing Surveys (CSUR)*, 52(4): 1–35.

Nam, H.; Park, J.; Choi, J.; Bennis, M.; and Kim, S.-L. 2024. Language-Oriented Communication with Semantic Coding and Knowledge Distillation for Text-to-Image Generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Nam, H.; Park, J.; Choi, J.; and Kim, S.-L. 2023. Sequential Semantic Generative Communication for Progressive Text-to-Image Generation. In *Proceedings of the IEEE International Conference on Sensing, Communication, and Networking (SECON)*.

Ngo, H. Q.; Ashikhmin, A.; Yang, H.; Larsson, E. G.; and Marzetta, T. L. 2017. Cell-free massive MIMO versus small cells. *IEEE Transactions on Wireless Communications*, 16(3): 1834–1850.

Parthasarathy, S.; and Sundaram, S. 2021. Detecting expressions with multimodal transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 636–643. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; and Rombach, R. 2024. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*.

Shiu, D.-S.; Foschini, G. J.; Gans, M. J.; and Kahn, J. M. 2000. Fading correlation and its effect on the capacity of multielement antenna systems. *IEEE Transactions on communications*, 48(3): 502–513.

Szott, S.; Kosek-Szott, K.; Gawłowicz, P.; Gómez, J. T.; Bellalta, B.; Zubow, A.; and Dressler, F. 2022. Wi-Fi Meets ML: A Survey on Improving IEEE 802.11 Performance With Machine Learning. *IEEE Communications Surveys & Tutorials*, 24(3): 1843–1893.

Wang, P.; Cheng, Y.; Dong, B.; Hu, R.; and Li, S. 2022. WIR-Transformer: Using Transformers for Wireless Interference Recognition. *IEEE Wireless Communications Letters*.

Weng, Z.; Qin, Z.; Tao, X.; Pan, C.; Liu, G.; and Li, G. Y. 2023. Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications*, 22(9): 6227–6240.

Wu, H.; Shao, Y.; Ozfatura, E.; Mikolajczyk, K.; and Gündüz, D. 2024. Transformer-aided wireless image transmission with channel feedback. *IEEE Transactions on Wireless Communications*.

Xie, B.; Sidulova, M.; and Park, C. H. 2021. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. *Sensors*, 21(14): 4913.

Xie, H.; Qin, Z.; and Li, G. Y. 2023. Semantic communication with memory. *IEEE Journal on Selected Areas in Communications*, 41(8): 2658–2669.

Xie, H.; Qin, Z.; Tao, X.; and Letaief, K. B. 2022. Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(9): 2584–2597.

Yoo, H.; Jung, T.; Dai, L.; Kim, S.; and Chae, C.-B. 2022. Real-Time Semantic Communications with a Vision Transformer. *arXiv preprint arXiv:2205.03886*.

Zhao, B.; Gong, M.; and Li, X. 2022. Hierarchical multimodal transformer to summarize videos. *Neurocomputing*, 468: 360–369.

Zheng, J.; Zhang, J.; Du, H.; Niyato, D.; Ai, B.; Debbah, M.; and Letaief, K. B. 2024. Mobile cell-free massive MIMO: Challenges, solutions, and future directions. *IEEE Wireless Communications*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.