

Addressing Out-of-Distribution Challenges in Image Semantic Communication Systems with Multi-modal Large Language Models

Feifan Zhang^{a†}, Yuyang Du^{a†}, Kexin Chen^{a†}, Yulin Shao^b, Soung Chang Liew^{a*}

^a The Chinese University of Hong Kong, Hong Kong SAR, China

^b University of Macau, Macau SAR, China

Abstract—Semantic communication is a promising technology for next-generation wireless networks. However, the out-of-distribution (OOD) problem, where a pre-trained machine learning (ML) model is applied to unseen tasks that are outside the distribution of its training data, may compromise the integrity of semantic compression. This paper explores the use of multi-modal large language models (MLLMs) to address the OOD issue in image semantic communication. We propose a novel “Plan A - Plan B” framework that leverages the broad knowledge and strong generalization ability of an MLLM to assist a conventional ML model when the latter encounters an OOD input in the semantic encoding process. The novel framework integrates the anti-OOD ability of MLLMs with the domain expertise of ML models in tasks they have been trained for, thus enhancing the accuracy in the semantic encoding process. Further, at the receiver side of the communication system, we put forth a “generate-criticize” framework that allows one MLLM to challenge the image generated by another MLLM, which then revises the generated image in the next iteration. The joint effort of the two MLLMs significantly enhances the reliability of image reconstruction.

Index Terms—Semantic communication, multi-modal foundation model, generative AIs, out-of-distribution problem.

I. INTRODUCTION

From the first generation (1G) to the fifth generation (5G), advanced communication techniques have significantly improved the transmission rate of wireless networks. The upcoming 6G network is designed to support even higher transmission rates [1]–[4], facilitating diverse communication-reliant applications, such as augmented reality (AR), virtual reality (VR), and autonomous driving. However, bit-based transmission technologies are currently approaching the Shannon Capacity [5]. This motivates researchers to develop goal-oriented communication systems beyond bit transmission.

Semantic communication transcends the traditional bit-wise transmission by prioritizing the meaning and intent conveyed through information exchange. It effectively compresses data by filtering out irrelevant information while preserving semantic fidelity, significantly reducing the bandwidth requirement for information transmission [6]. As such, semantic communication is expected to become a key technology in next-generation wireless communication systems.

The work was partially supported by the Shen Zhen-Hong Kong-Macao technical program (Type C) under Grant No. SGD20230821094359004.

*S. C. Liew is the corresponding author (email: soung@ie.cuhk.edu.hk).

[†]These authors contribute equally to this work.

A semantic communication system consists of a semantic transmitter and a semantic receiver. At the transmitter side, the transmitter first performs semantic encoding, in which the sender extracts semantic features from the source data and compresses the intended message in a way that causes no semantic information loss. Subsequently, channel coding is applied to the semantic features to generate signals suitable for physical-channel transmission. The two encoding modules can be implemented separately, as in a conventional communication system, or they can be implemented as a joint source-channel coding (JSCC) integrated system. At the receiver side, reverse operations are carried out to reconstruct the semantic contents.

Many existing works realized semantic communication systems with machine learning (ML) models, such as convolutional neural network (CNNs) or transformers [7]–[10]. In [7], for example, the authors proposed a Generative Adversarial Network (GAN)-based model for image semantic coding and decoding. In [8], the authors put forth a generative semantic communication framework based on scene graphs, which uses a graph neural network (GNN) for semantic encoding at the transmitter and a diffusion model for decoding at the receiver. In [9], the authors introduced a JSCC framework for wireless image transmission. It employs convolutional neural networks to directly map image pixel values to complex transmission signals, demonstrating robust performance in environments with low signal-to-noise ratios (SNRs). A later work [10] proposed a receiver-feedback JSCC scheme that utilizes the transformer’s self-attention mechanism to boost the discriminative representation of semantic features.

A major issue not addressed in the above papers is the out-of-distribution (OOD) problem prevalent in conventional ML models [11]. An ML model in a semantic communication system is typically trained on a specific dataset assumed to be representative of the data distribution for the target task. The model learns to extract important semantic features based on the training data. However, when the pre-trained model is applied to a new, unseen task that is out of the distribution of the training data, the ML model may fail to correctly extract the true semantic information. Fig. 1 provides an example in which the training dataset of the semantic encoder does not contain knowledge about cats due to its limited scale. Although the semantic encoder can identify the backpack in

the image, it has no idea about the object besides the backpack. The semantic encoder might misclassify the cat as another type of animal it knows from its training data, such as a dog, thus introducing semantic distortion at the transmitter. Subsequently, the receiver reconstructs the image with the distorted semantic information and obtains a significantly different image in Fig. 1b.



Fig. 1: (a) The original image. (b) A reconstructed image with the OOD problem.

The dynamic and ever-changing nature of the real world inevitably introduces OOD data to realistic semantic communication systems, as novel situations, events, and entities continuously emerge. Therefore, building a practical semantic encoder with a strong ability to handle OOD data is crucial. This paper explores the use of multi-modal large language models (MLLMs), which converts a given input image into a textual description, to address the challenge. A conventional ML model’s training dataset may focus on a specific domain, and the model can perform quite well given tasks within its knowledge distribution. On the other hand, MLLMs are trained on extremely large-scale and general datasets [12]. Compared with a conventional ML model’s expertise in a domain-specific task, MLLMs have a broad knowledge base that can be applied to various fields and tasks, although they may not perform as well as an expert ML model in domain-specific tasks for which the ML model is specially trained. This paper demonstrates that the vast knowledge and generalization ability of a MLLM can be a valuable augmentation to a conventional ML model to overcome the OOD problem encountered in semantic communication. Given the general knowledge of the MLLM and the domain specific expertise of the conventional ML model, the rest of paper refers to these two models as “general AI model” and “expert AI model”, respectively.

This paper focuses mainly on the semantic compression process happening within the semantic encoder at the transmitter side. For system-building and demonstration purposes, we also deploy off-the-shelf generative AI models for image reconstruction at the receiver side. Furthermore, we investigate the reliability performance of the communication system in the face of strong channel noise. Our contributions are summarized as follows:

A “Plan A – Plan B” Framework: At the transmitter side, we put forth a novel “Plan A – Plan B” framework

that overcome the OOD problem in the semantic coding process with the help of MLLMs. For non-OOD cases, the expert AI expresses confidence in its judgment. Thus, we take the classification result of the expert model as “Plan A”. On the other hand, for OOD cases, the expert AI expresses extremely low confidence in its judgment (which, in essence, is just a random guess). Therefore, when the confidence of the expert AI’s judgment is below a pre-defined threshold, we turn to the general AI and treat its classification result as an alternative “Plan B”. When OOD occurs, our scheme, with the general AI’s augmentation, significantly outperforms the conventional expert-AI-empowered semantic communication system in semantic extraction. Our experimental results indicate a 13.3% classification accuracy improvement on a comprehensive dataset with 8.2% OOD and 91.8% non-OOD cases. For pure OOD cases, the classification accuracy of the semantic encoding process increases from close to 0 to 71.9%.

A “Generate-Criticize” Framework for Image Reconstruction: At the receiver side, we propose a “generate-criticize” framework that utilizes multi-MLLM cooperation to enhance the reliability of image reconstruction. Specifically, the receiver side deploys a text-to-image MLLM and an image-to-text MLLM, which are referred to as the image generation model and the criticizing model, respectively. As their names suggest, the first MLLM generates an image according to the text description of the desired image (which is recovered from the wireless signals received), while the second MLLM compares the generated image with the text description and points out the flaws of the generated image so that the first MLLM can revise its output accordingly.

The generate-criticize framework is inspired by an interesting observation during our experiments. We found that existing text-to-image MLLMs like Dall-E-3 [13] and Stable Diffusion [14] are poor at handling numbers (similar observations have also been reported in [15], [16]). However, such mistakes can be discovered by image-to-text MLLMs like GPT-4V [17].

Therefore, to mitigate the semantic distortion caused by the image generation MLLM’s poor handling of numbers (e.g., number of similar objects in an image), we deploy an image-to-text MLLM to challenge the generated image. Then in the next iteration, the image generation MLLM can revise the image according to the critic’s feedback. Our experimental results indicate that the “generate-criticize” framework significantly enhances the quality of image generation, boosting the probability of correct image generation by around 20% with up-to-four-round iterations.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Many previous investigations in semantic communication focus on pixel-wise image fidelity [7]–[10]. This paper, on the other hand, focuses on accurate element-wise semantic representation of the image. When interpreting an image, humans seldom focus on the RGB values of the pixels at the outset. Instead, they are more likely to first concentrate on identifying the objects or elements in the figure and interpreting what the figure tries to convey by analyzing the juxtaposition of these

objects. In other words, when reading an image, humans focus more on the element-wise semantic information rather than the pixel-wise image information. Therefore, as long as the element-wise semantic meaning of a figure can be successfully recovered at the receiver, the transmission can be considered as largely correct, regardless of specific pixel values. For example, Fig. 2a presents a plate with some fruits, while Fig. 2b renders the same scenario as an oil painting. Although the two figures are completely different from the perspective of a pixel-based analysis, they convey essentially the same meaning from a semantic perspective: a banana, an apple, and an orange are placed on a white plate.

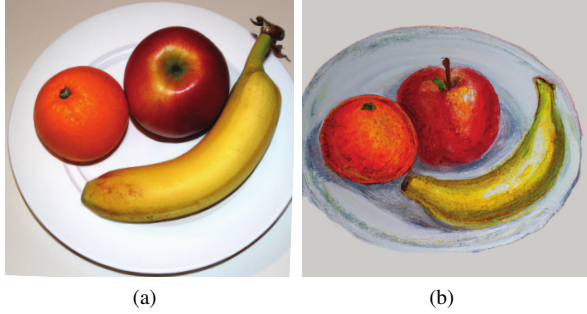


Fig. 2: (a) The photo of a plate and some fruits. (b) The oil painting of the same fruit plate.

In this section, we will first introduce our system framework and define the associated notations. Then, we present the loss function assumed in this paper that considers the integrity of semantic information from two angles: 1) the correct recovery of the transmitted element-wise information, and 2) the perceptual similarity between the transmitted image and the recovered image.

We define the element-wise semantic information of an image by $\mathbf{S} = [S_1, S_2, \dots, S_n]$, where n is the total number of elements contained in the image, and S_i represents the i^{th} element among them. For element S_i , it could be the background of the image or an object in the image.

Fig. 3 presents the overall framework of our semantic communication system. At the beginning of the transmission process, semantic encoding is conducted to extract the semantic information from image M . The encoding process first attempts to identify the elements within the image using a conventional ML model trained with domain-specific data. For element S_i , let us denote the expert model's confidence in its judgment by C_i . As discussed in the introduction, the expert model cannot precisely identify an element that never appears in its training data, and it may have a relatively low C_i in an OOD case. If C_i is below a pre-defined threshold ρ , we extract a sub-image that solely contains S_i by cropping the whole image (to avoid confusing the general AI in the identification task of other objects) and provide the sub-image to the general AI for further processing.

This paper trains YOLOv8 [18], a lightweight open-source CNN model, as the expert AI, and InstructBLIP [19], a large

open-source visual-language MLLM, as the general AI. We refer readers to Section III for details on how we train the expert AI with a domain-specific dataset, and how the image is jointly processed by the expert AI and the general AI.

After extracting the element-wise semantic information $\mathbf{S} = [S_1, S_2, \dots, S_n]$ with the joint effort of the expert AI and the general AI, the transmitter inputs $\mathbf{S}$ to a text-to-text joint source-channel coding (T2T-JSCC) encoder.¹ As in [20], [21], the T2T-JSCC encoder combines the conventional source coding, channel coding, and modulation into a single encoder and parameterizes the module by $F_{TX}(\cdot)$. After the semantic element S_i is encoded by the T2T-JSCC encoder, its corresponding output X_i becomes a normalized representation in a d -dimensional vector space, where d is a hyperparameter of the T2T-JSCC encoder. We write X_i as:

$$X_i = F_{TX}(S_i) \in \mathbb{C}^d. \quad (1)$$

With the above definition, the output of the T2T-JSCC encoder can be written as:

$$\mathbf{X} = F_{TX}(\mathbf{S}) = [F_{TX}(S_1), F_{TX}(S_2), \dots, F_{TX}(S_n)]. \quad (2)$$

For more information about the implementation and analysis of the T2T-JSCC encoder, we refer readers to Section II and Section III of [20] for details.

Vector $\mathbf{X}$ is the signal to be transmitted over the noisy channel. For X_i , its corresponding channel output Y_i can be written as:

$$Y_i = H(X_i) = X_i + N_i, \quad (3)$$

where N_i denotes the Gaussian noise over the transmitted signal X_i .

At the receiver side, the system applies a T2T-JSCC decoder to decode the received signal $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]$. The T2T-JSCC decoder is symmetrical to the encoder. We parameterize the T2T-JSCC decoder by $F_{RX}(\cdot)$ and write the decoded message as:

$$\hat{\mathbf{S}} = F_{RX}(\mathbf{Y}) = [\hat{S}_1, \hat{S}_2, \dots, \hat{S}_n]. \quad (4)$$

After obtaining $\hat{\mathbf{S}}$ through T2T-JSCC decoding, the receiver composes the text information into a prompt that describes the image. The prompt is then given to the image reconstruction module.

The image reconstruction module aims to build an image $\hat{M}$ that is consistent with the semantic information described in the given prompt. The image reconstruction module contains two MLLMs playing complementary roles. The first MLLM is realized with DALL-E-3 [13]. It takes text as input and generates images according to the given description. The second MLLM is realized with GPT-4V [17]. It analyzes the generated image and determines whether the image is consistent with the description in the prompt. If the generated

¹After the element-wise semantic compression, the textual description of the image can be transmitted via a conventional communication system with source coding (e.g., Huffman coding), channel coding (e.g., convolution coding), and modulation. Alternatively, we could apply the T2T-JSCC module for the same purpose. This paper chooses the later for system building.

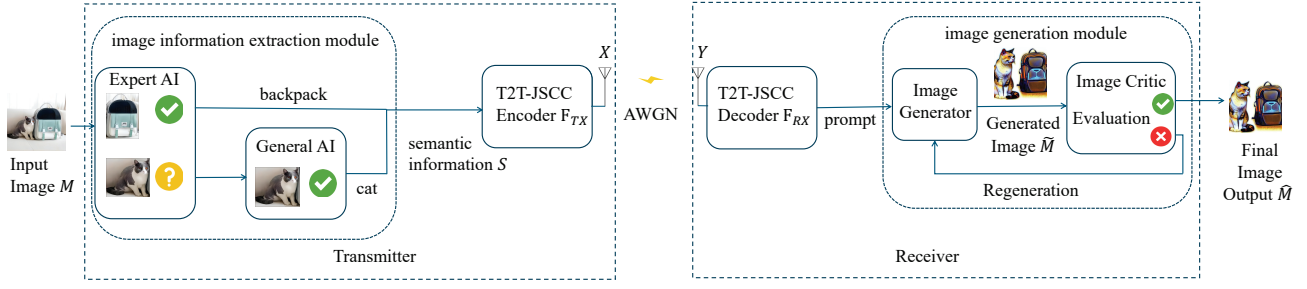


Fig. 3: The framework of the proposed image semantic transmission system.

image perfectly reflects the prompt, we take the image as the final output. Otherwise, the critic MLLM modifies the original prompt by adding suggestions on how the image should be revised to better reflect the semantic information. The new prompt is returned to DALL-E-3 for image generation in the next round. The above iteration continues until the critic MLLM is satisfied with the image (i.e., the image is consistent with the semantic information provided in the prompt) or the number of iterations reaches an upper limit.

We define the loss function for our semantic communication system as

$$L = (1 - \alpha)L_1\langle S, \hat{S} \rangle + \alpha L_2\langle M, \hat{M} \rangle, \quad (5)$$

where $L_1\langle S, \hat{S} \rangle$ denotes the textual dissimilarity between the lists of elements at the transmitter side and the receiver side, falling within $[0, 1]$; $L_2\langle M, \hat{M} \rangle$ reflects the visual distance between the original image and the recovered image, falling within $[0, 1]$; and $1 - \alpha$ and α , $0 < \alpha < 1$, denote the weights of L_1 and L_2 , respectively.

It is important to note that typical measurements of textual dissimilarity/similarity must be conducted within a high-dimensional semantic space generated by a natural language processing (NLP) network. Therefore, to calculate $L_1\langle S, \hat{S} \rangle$, we need to transform S and $\hat{S}$ into corresponding vectors within the high-dimensional semantic space. For that purpose, we first concatenate the elements within S and $\hat{S}$, and then we leverage an NLP embedding network (this paper uses SentenceBERT [22] as an example) to transform the element concatenations into two corresponding vectors within the semantic space of the NLP embedding network. Let us denote the dimension of the semantic space by G_1 , and denote the two output vectors of the NLP embedding network by T and $\hat{T}$, respectively.

With T and $\hat{T}$, we calculate the cosine angle between the two vectors, which is denoted by $CA\langle T, \hat{T} \rangle$ and could be written as

$$CA\langle T, \hat{T} \rangle = \frac{\sum_{n=1}^{G_1} T_n \hat{T}_n}{\sqrt{\sum_{n=1}^{G_1} (T_n)^2 \cdot \sum_{n=1}^{G_1} (\hat{T}_n)^2}}. \quad (6)$$

We note that a large $CA\langle T, \hat{T} \rangle$ indicates a small angle between T and $\hat{T}$, corresponding to high similarity between S and $\hat{S}$. Thus, we define the dissimilarity between T and $\hat{T}$ as

$$L_1\langle S, \hat{S} \rangle = 1 - CA\langle T, \hat{T} \rangle. \quad (7)$$

For $L_2\langle M, \hat{M} \rangle$, the perceptual image loss during the communication process, this paper applies Learned Perceptual Image Patch Similarity (LPIPS) [23] as a measurement of distance². Compared with the traditional pixel-wise metrics, LPIPS aims to mimic the perceptual results of human vision, thereby better conforming to the way human judgment measures the similarity between two images. LPIPS utilizes deep neural networks (this paper uses AlexNet [24] as an example) to compute the distance between the transmitted image M and the reconstructed image $\hat{M}$ in the feature space of the network to measure the dissimilarity of the two images. Specifically, $L_2\langle M, \hat{M} \rangle$ is given by

$$L_2\langle M, \hat{M} \rangle = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (y_{hw}^l - \hat{y}_{hw}^l)\|_2^2, \quad (8)$$

where H_l, W_l, C_l represent the height, width, and channel dimension of the feature generated by AlexNet at the l^{th} layer; $y_{hw}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ is the feature that AlexNet generates at the l^{th} layer according to the transmitted image, while $\hat{y}_{hw}^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ is the feature that AlexNet generates at the l^{th} layer according to the reconstructed image; and $\odot$ represents the channel-wise feature multiplication operation. For more explanations on (8) given above, we refer readers to Section III of [23] for more details.

III. TRANSMITTER DESIGN

This section focuses on the design of our semantic transmitter. We first give implementation details of our novel “Plan A – Plan B” framework in subsection III-A. In subsection III-B, we present the adaptation we made on the general AI model for fulfilling the “Plan A – Plan B” framework. Finally, in subsection III-C, we provide a brief introduction to how we adapt the T2T-JSCC encoder reported in [20] for our task.

A. The Plan A – Plan B Framework

We leverage the YOLOv8 project for the realization of the expert AI. The original YOLOv8 model reported in [18] was built upon the COCO dataset [25], which contains

²In [23], the authors referred to their metric as “similarity”. However, as we can see from equation (1) in [23], the metric increases when the differences between the two images become more distinct. In other words, LPIPS is, in essence, a *distance* or *dissimilarity* rather than a *similarity*. Although this paper follows the traditional name developed in [23] for consistency, we point out this naming inconsistency to prevent readers’ potential confusion.

over 330,000 images and 80 object categories. For concept-proving purposes, this paper reorganizes the training dataset of YOLOv8 to manually construct an OOD scenario. Specifically, among the 80 classes within the COCO dataset, there are ten labels related to animals, such as “cat”, “bird”, and “elephant”. We removed all images containing animal label(s) from the training dataset but kept the testing dataset unchanged. Consequently, the model learns nothing about animals from the training data but faces animal images in the testing dataset. For the realization of the general AI, we use InstructBLIP’s pre-trained model weights reported in [19] and adapt its output layer to make it fulfill the Plan A – Plan B framework (see implementation details presented in Section III-B).

We now introduce the implementation of the Plan A – Plan B framework. We first applied Plan A, the expert AI model, for object identification of the input image. Assume that there are n objects in the image. For object i , the expert model draws a rectangle to frame the object and generates the corresponding recognition result $S_{i,A}$, where the subscript A denotes that this is the recognition result of Plan A. Note that $S_{i,A}$ should be one of the labels within the revised COCO dataset, which has 70 classes only.

Let us denote the expert model’s confidence in $S_{i,A}$ by C_i . If C_i is below a pre-defined threshold ρ , we treat the identification result as unreliable. For such cases, we rely on Plan B, the general AI, for a reliable semantic encoding process. We crop the image within the rectangle frame to obtain a sub-image that contains the object solely. We then give the sub-image to InstructBLIP, which performs an object classification task with a simple prompt (i.e., “What is in the figure?”) and generates $S_{i,B}$. Note that $S_{i,B}$ is one of the labels within the original COCO dataset (with 80 classes that include animal-related ones). We refer readers to subsection III-B for how the classification task is conducted within the MLLM.

After InstructBLIP corrects the recognition results of all low-confidence objects, we finally obtain the semantic information vector $\mathbf{S} = [S_1, S_2, \dots, S_n]$, in which

$$S_i = \begin{cases} S_{i,A}, & C_i \geq \rho \\ S_{i,B}, & C_i < \rho \end{cases} \quad (9)$$

B. Adapt General Models for the Semantic Compression Task

The object identification within the semantic compression process expects answers that precisely describe the image with a set of nouns. However, the original output of a MLLM may contain many irrelevant words, including verbs, adjectives, articles, or even greetings like “hello” and “hi”, which is useless for describing an image. Our recent work [26] developed a novel cross-entropy transformation (CET) scheme that narrows the vocabulary of the output for precise object identification purposes. The CET scheme maps the complex self-attention matrix extracted from the last fully connected layer of the transformer network, which considers the full vocabulary list, into a probability distribution among target classes. This subsection gives a quick introduction of the CET

scheme at a conceptual level. We refer interested readers to Section II of [26] for illustrations and implementation details.

Let us start with the structure of the general AI model. InstructBLIP is an open-source MLLM with vision and language ability. There are three components in InstructBLIP: 1) an image encoder that deals with the image input, 2) a text-in-text-out language model to manage the output, and 3) an image-text transformer that bridges the input and output modules. Thanks to the modular architectural design, InstructBLIP is highly flexible, and we can quickly adapt a wide range of text-to-text language models for implementation. Without loss of generality, our implementation of the CET scheme utilizes FlanT5 [27], an instruction-tuned model based on Transformer T5 [28], as the text-in-text-out model.

Within the general AI model, given the sub-image of object i , let us denote the self-attention matrix at the last fully connected layer of FlanT5 by e . The size of e is $P \times P$, where P is the cardinality of InstructBLIP’s full vocabulary list after tokenization (let us denote the full vocabulary list by $\mathbf{l}^F$). Element e_{k_1, k_2} within e is the attention weight between word k_1 and word k_2 within $\mathbf{l}^F$. Let us denote the number of possible classes in our classification task by M , and denote the list of these classification labels (also after tokenization) by $\mathbf{l}^S$, in which the superscript S indicates that the list is essentially a subset of the full vocabulary list $\mathbf{l}^F$. Given $\mathbf{l}^S$, the CET scheme extracts the corresponding M rows from the original self-attention matrix e . Let us denote the new $M \times P$ matrix by e^E , where the superscript E notes that the matrix is a smaller one after extracting the M rows. Meanwhile, with the tokenized label list $\mathbf{l}^S$ (of size $M \times 1$) and the tokenized full vocabulary list $\mathbf{l}^F$ (of size $P \times 1$), we conduct one-hot encoding and obtain an $M \times P$ matrix $\mathbf{b}$, in which

$$b_{j,k} = \begin{cases} 1, & l_j^S = l_k^F \\ 0, & l_j^S \neq l_k^F \end{cases} \quad (10)$$

With e^E and $\mathbf{b}$, we calculate the cross-entropy loss between the two $M \times P$ matrices by,

$$L_j = - \sum_{x=1}^P e_{j,x}^E \cdot b_{j,x} \quad (11)$$

In the resulting $M \times 1$ vector $\mathbf{L}$, element L_j represents the expected loss of using classification label j as the identification result. Note that L_j is negatively related to the probability of label j being selected. Therefore, a reciprocal process is needed for transforming L_j to an intermediate variable D_j that is positively related to the probability of label j being selected. This paper follows the realization in [26] and lets $D_j = -L_j$. This is followed by the normalization process. Eventually, at the end of the CET scheme, we obtain a normalized $M \times 1$ probability vector $\mathbf{P}$, in which element P_j denotes the probability of label j being selected.

C. The T2T-JSCC encoder

The T2T-JSCC encoder transforms semantic information vector $\mathbf{S}$ into an $N \times d$ complex vector $\mathbf{X}$ that is ready for

TABLE I: Performance comparisons for different models/scheme over D and D_{nOOD} .

Tested Model/Scheme	Performance on D_{nOOD}			Performance on D		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
YOLOv8(expert AI solely)	0.6095	0.6177	0.6136	0.5176	0.5670	0.5412
InstructBLIP (general AI solely)	0.7420	0.5252	0.6150	0.7333	0.5620	0.6363
Plan A – Plan B	0.7727	0.6869	0.7272	0.7468	0.6903	0.7174

transmission over the noisy wireless channel. The processing inside the T2T-JSCC encoder includes source coding, channel coding, and digital modulation.

Since we do not need to transmit a very long sentence as in [20], we re-developed the T2T-JSCC encoder with the following adjustments. To begin with, we follow the baseline method in [20] and utilize differentiable cross-entropy as the loss function. Furthermore, given that the simplified task does not require a very complex semantic space, we reduce the dimension of the output signal by setting a smaller hyperparameter d (specifically, from 128 to 50). Additionally, during the training process, we use object labels from the COCO dataset as new training materials to retrain the model.

With the above modifications, our new T2T-JSCC encoder is more lightweight than the original one reported in [20]. However, it is stable and reliable enough to meet our requirements and demonstrates strong anti-noise performance (as proven in the following experimental section).

IV. RECEIVER DESIGN

This section focuses on the design of our semantic receiver. We propose a “generate-criticize” image reconstruction framework for more reliable reconstruction based on semantic information recovered from the T2T-JSCC decoder.

At the receiver side, the recovered semantic information is a vector containing the labels of all identified objects within the transmitted image. For example, let’s say the transmitted image contains three oranges and one apple, and the successfully recovered semantic vector is: $\hat{S} = [\text{orange}, \text{apple}, \text{orange}, \text{orange}]$. We then asked DALL-E-3 to reconstruct this image based on this vector and obtained the image shown in Fig. 4a. However, the generated image exhibits inconsistencies: the image contains one apple and four oranges, which does not align with the original object counts. This shortcoming reflects the limited counting ability of the generation model, as observed in prior research [13].

To overcome this problem, we added a critic into the image reconstruction process to ensure reliable performance. GPT-4V has strong image analysis capabilities for tasks such as object counting and detection. During the critique process, GPT-4V will initially provide a “YES” or “NO” response regarding the accuracy of the types and quantities of all objects in the generated image. Subsequently, GPT-4V generates a detailed prompt to guide DALL-E-3 in producing a more accurate reconstruction. For example, given the example in Fig. 4a, GPT-4V will give a “NO” at the beginning of the response and then tell the image generation model (i.e., DALL-E-3) that the number of orange of should be three instead of four. With

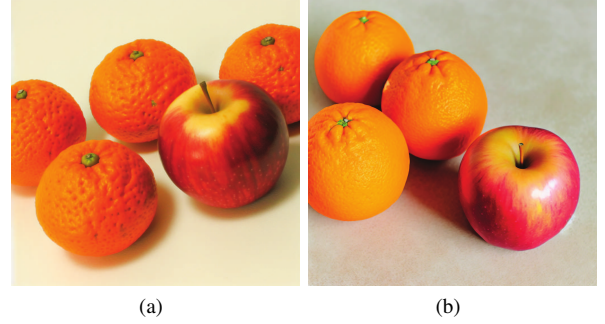


Fig. 4: An example of the “generate-criticize” image reconstruction framework.

the modification pointed out by GPT-4V, DALL-E-3 starts a new round of image generation. However, DALL-E-3 cannot always guarantee that the modified image after this iteration is correct. Therefore, the process will continue to iterate until GPT-4V outputs “YES” or the number of iterations reaches an iteration upper bound. Finally, DALL-E-3 can generate a correct image, as shown in Fig. 4b.

V. EXPERIMENTS AND VALIDATIONS

We first evaluate the anti-OOD performance of the proposed Plan A – Plan B framework. We crop the images in the original COCO validation dataset based on the ground truth annotations.³ And then we remove images that are too small to be identified even for human eyes. After the above operations, we obtain a new dataset D containing a total of 14,423 images, which has 1,183 OOD images (i.e., animal classes). Let us denote the set of OOD and non-OOD images by D_{OOD} and D_{nOOD} , respectively. To comprehensively evaluate the framework, we compared the object recognition accuracy of YOLOv8 (Plan A), InstructBLIP (Plan B), and Plan A – Plan B scheme on dataset D and D_{nOOD} , respectively. We did not test pure OOD cases, as this may not be very practical (more or less, an actual scenario should contain a few familiar cases).

Table I presents the precision, recall, and F1 score of different models in D and D_{nOOD} . For the non-OOD dataset, YOLOv8 can achieve a precision of 60.95% and a recall of 61.77%. Its lower precision indicates that it overfits on

³This is because the testing set of the COCO dataset does not provide ground truth annotations. Instead, the dataset asks users to submit their identification results to a cloud serve for evaluation. In our experiment, we need to leverage the ground truth within an image to construct the experimental dataset. Hence, we choose the validation set. Note that the validation set was never used in our model training, i.e., no data leakage happens in our training process.

the identification of high-proportion classes, thus leading to misidentification and reduced accuracy. InstructBLIP achieves 74.2% and 52.52% for precision and recall in the same tested dataset. That shows InstructBLIP is more “cautious” than YOLOv8 in identifying samples of high-proportion classes, i.e., it prefers to sacrifice recall to ensure precision. For the model’s general performance (i.e., F1 score), YOLOv8 and InstructBLIP have similar performance, reaching 0.6136 and 0.6150, respectively. The Plan A – Plan B scheme combines two models’ advantages effectively and achieves the best performance in all aspects (i.e., both precision, recall, and F1 score).

For the comprehensive dataset D , the performance of YOLOv8 drops due to the involvement of OOD cases. InstructBLIP, on the other hand, demonstrates less obvious changes due to its anti-OOD ability. The Plan A-Plan B scheme shows higher performance than the above two single-model schemes, reaching the highest performance in terms of precision, recall, and F1 score. This experimental observation indicates the superiority of our Plan A-Plan B scheme in the object identification process within semantic encoding.

To further evaluate the anti-OOD performance of different models/schemes, we create a group of new testing data by selecting a total of 100 images from D_{nOOD} and D_{OOD} with different proportions. We adjust the proportion of OOD cases in the tested dataset and compare the recognition accuracy of YOLOv8 and the Plan A – Plan B scheme. Fig. 5 shows that YOLOv8’s performance continues to decrease as the proportion of OOD cases increases. In contrast, the Plan A – Plan B scheme can maintain relatively stable performance regardless of the proportion of OOD cases in the tested data. This further justifies the strong anti-OOD performance of the proposed framework.

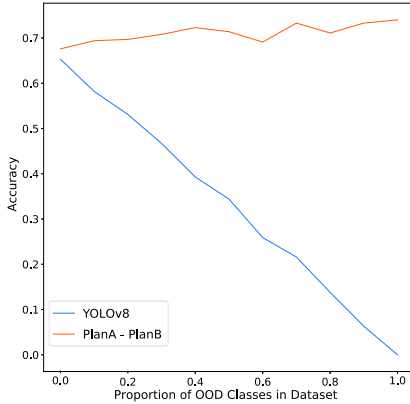


Fig. 5: The accuracies of YOLOv8 and the Plan A – Plan B scheme under different proportions of OOD cases in the dataset.

We then investigate the performance of the generate-criticize framework. We are interested in how the iteration limit of the image critic affects the image generation process. We first randomly select an object from the 80 possible classes and randomly generate quantity N that describes the

number of this object.⁴ We construct a prompt based on the selected object and the corresponding N . We then leverage the generate-criticize framework to build the corresponding image. After the image is constructed, we manually determine whether the generated images were consistent with the semantics in the prompt. For each iteration limit, we conduct the above process 100 times to calculate the average accuracy.

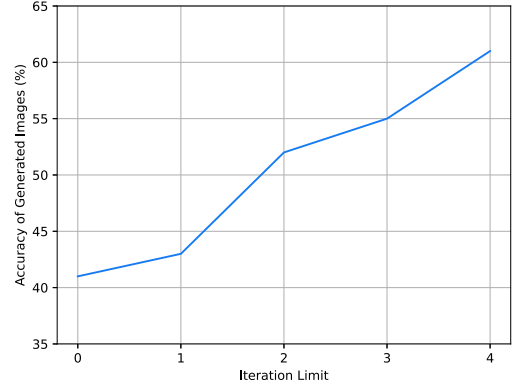


Fig. 6: The accuracy of generated images corresponding to different iteration limits.

Fig. 6 shows that as the iteration limit of the image critic increases, the accuracy of the generated images continues to improve. When the iteration limit is set to 4, the corresponding accuracy rate is nearly 20% higher than the scheme without the critic. These results indicate that the generate-criticize framework can significantly improve the reliability of image reconstruction by leveraging the image critic to provide feedback and guide the generation process.

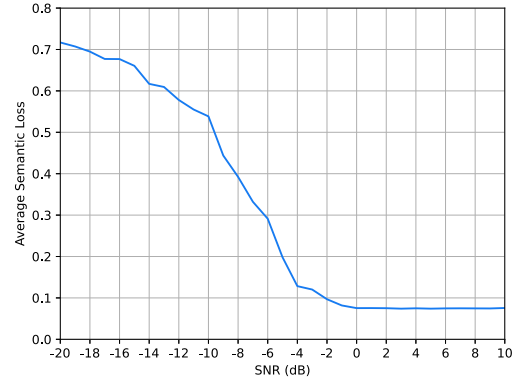


Fig. 7: The average loss performance of the whole system when transmitting 100 images (including 8 OOD cases) under different noise conditions.

Finally, we study the system’s anti-noise performance by investigating the semantic loss (see (II) for the definition of the system’s semantic loss) under different channel noise. Our detailed experimental setup is as follows: in the training process of the T2T JSCC encoder and decoder, we assume

⁴Here we assume $N \geq 3$, given that our prior observations suggest that a text-to-image model tends to generate incorrect object counts when the number of objects is larger than three.

an additive white Gaussian noise (AWGN) channel with an SNR of 10dB. While in the testing process, we consider noise conditions that are significantly larger than the expected one (with the SNR ranging from -20dB to 5dB).

In the experiment, we assume $\alpha = 0.1$ and transmit images through the tested AWGN channel. Figure 7 illustrates the average semantic loss when different channel conditions are considered. As we can see from the figure, even if the SNR is as low as -2dB, which is far from the SNR setting in the training process, the average loss remains very close to that of 10dB-SNR case. This experimental result demonstrates the ultra-robustness of the proposed scheme in the face of the strong channel noise.

VI. CONCLUSION

This paper explores the application of MLLM in semantic communication to address the OOD problem in the semantic compression process. Our proposed “Plan A - Plan B” framework leverages the broad knowledge scope and strong generalization ability of MLLMs to assist conventional ML models in semantic encoding when encountering OOD cases. On the receiver side, our “generate-criticize” framework utilizes the cooperation of multiple MLLMs to improve the reliability of image reconstruction. Experimental results demonstrate the effectiveness of our proposed frameworks in terms of the accuracy of the semantic compression process, the probability of correct image generation, and the robustness in the face of strong channel noise. These findings highlight the potential of MLLMs in enabling reliable and efficient semantic communication for next-generation wireless networks. Future work could explore the application of our proposed frameworks to other types of data, such as video and audio, with the assistance of more powerful MLLMs. Additionally, the development of more advanced JSCC schemes, which could further enhance the practicality of MLLM-empowered semantic communication systems, might also be an interesting topic for the community.

REFERENCES

- [1] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, “What should 6G be?” *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.
- [2] Y. Du, S. C. Liew, and Y. Shao, “Efficient FFT computation in IFDMA transceivers,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6594–6607, 2023.
- [3] Y. Du, L. Hao, and Y. Lei, “SER optimization in OFDM-IM systems with nonlinear power amplifiers,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 15 046–15 051, 2023.
- [4] Y. Shao, D. Gündüz, and S. C. Liew, “Federated edge learning with misaligned over-the-air computation,” *IEEE Trans. Wireless Communications*, vol. 21, no. 6, pp. 3951–3964, 2021.
- [5] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [6] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” *IEEE J. Sele. Areas Commu.*, vol. 41, no. 1, pp. 5–41, 2022.
- [7] M. U. Lokumarambage, V. S. S. Gowrisetty, H. Rezaei, T. Sivalingam, N. Rajatheva, and A. Fernando, “Wireless end-to-end image transmission system using semantic communications,” *IEEE Access*, 2023.
- [8] M. Yang, D. Gao, F. Xie, J. Li, X. Song, and G. Shi, “SG2SC: A generative semantic communication framework for scene understanding-oriented image transmission,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 486–13 490.
- [9] E. Boursoulatz, D. B. Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Trans. Cognit. Commun. Networking*, vol. 5, no. 3, pp. 567–579, 2019.
- [10] H. Wu, Y. Shao, E. Ozfatura, K. Mikolajczyk, and D. Gündüz, “Transformer-aided wireless image transmission with channel feedback,” *IEEE Trans. Wireless Commun.*, 2024.
- [11] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, “Towards out-of-distribution generalization: A survey,” *arXiv preprint arXiv:2108.13624*, 2021.
- [12] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, “LLMind: Orchestrating AI and IoT with LLMs for complex task execution,” *arXiv preprint arXiv:2312.09007*, 2023.
- [13] G. G. James Betker, T. Li Jing, L. L. Jianfeng Wang, J. LongOuyang, Y. JoyceLee, P. WesamManassra, Y. CaseyChu, and A. Ramesh, “Improving image generation with better captions.” [Online]. Available: <https://cdn.openai.com/papers/dall-e-3.pdf>
- [14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 10 684–10 695.
- [15] A. Borji, “Qualitative failures of image generation models and their application in detecting deepfakes,” *Image and Vision Computing*, vol. 137, p. 104771, 2023.
- [16] E. Leivada, E. Murphy, and G. Marcus, “DALL·E 2 fails to reliably capture common syntactic processes,” *Social Sciences & Humanities Open*, vol. 8, no. 1, p. 100648, 2023.
- [17] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, “The dawn of LLMs: Preliminary explorations with GPT-4V (ision),” *arXiv preprint arXiv:2309.17421*, vol. 9, no. 1, p. 1, 2023.
- [18] J. Q. Glenn Jocher, Ayush Chaurasia, “YOLO by Ultralytics,” 2023, accessed: February 30, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [19] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [20] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, “Reinforcement learning-powered semantic communication via semantic similarity,” *arXiv preprint arXiv:2108.12121*, 2021.
- [21] N. Farsad, M. Rao, and A. Goldsmith, “Deep learning for joint source-channel coding of text,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 2326–2330.
- [22] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 586–595.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft CoCo: Common objects in context,” in *2014 European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.
- [26] K. Chen, Y. Du, T. You, M. Islam, Z. Guo, Y. Jin, G. Chen, and P.-A. Heng, “LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery,” *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [27] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma et al., “Scaling instruction-finetuned language models,” *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.