
IMPLEMENTATION OF ZERO-SHOT SEMANTIC COMMUNICATION ON SOFTWARE DEFINED RADIO

Thanh Le Arif Dataesatu Homare Murakami Takeshi Matsumura
Wireless Systems Laboratory, Wireless Networks Research Center
National Institute of Information and Communications Technology (NICT)
Yokosuka, Kanagawa, Japan

ABSTRACT

Semantic communication has recently gained traction for its ability to reduce the amount of data transmitted over a communication link by transmitting a task-oriented representation instead of the raw source. Zero-shot semantic communication sends a general embedding from a vision-language model (VLM), so the same transmitter can serve new classification tasks without retraining. Most evidence for this advantage, however, comes from numerical simulation. We implement zero-shot semantic communication on a software-defined radio platform: a Raspberry Pi drives a pair of Analog Devices Active Learning Module (ADALM)-Pluto transceivers, with an image encoder at the transmitter and a text encoder at the receiver, and determines the zero-shot classification results via cosine similarity. We compare two VLMs, CLIP and MobileCLIP, across various channel conditions, *i.e.*, different signal-to-noise ratios (SNRs). We validate that the semantic link spends $9\times$ fewer channel uses per image than a JPEG plus 16-ary quadrature amplitude modulation baseline and still reaches 82% accuracy on CIFAR-10 at 22.3 dB, where the baseline scores 0%. On the traffic sign recognition dataset (TSRD), MobileCLIP correctly classifies 98.3% of unseen images at the same SNR. Offloading the image encoder to a neural processing unit reduces encoding to 13.4 ms per image, $49\times$ faster than a Raspberry Pi 4 CPU, placing the transmitter within a real-time budget. Our implementation is publicly available at <https://github.com/thanhlexyz/zsscsdr>.

Keywords orthogonal frequency-division multiplexing (OFDM) · software-defined radio · vision-language model · zero-shot classification

1 Introduction

Semantic communication transmits a compact representation to the receiver, optimized for the specific task (Le et al., 2026a). For instance, a receiver tasked with recognizing a traffic sign requires data that conveys the sign’s meaning rather than the original high-resolution image. Conventional wireless communication systems use bandwidth to reconstruct the entire image, even though the receiver ultimately discards most of it. Task-oriented semantic communication implements semantic communication by encoding only the information necessary for a given task (Xie et al., 2022; Liu et al., 2025). However, this approach is limited because the encoder is trained for a fixed task and dataset, necessitating retraining and redeployment when encountering previously unseen data.

Foundation models, including vision-language models (VLMs) (Radford et al., 2021; Vasu et al., 2024), address this challenge by having general embedding spaces that are inherently task-agnostic. Recent studies have transmitted such embeddings for zero-shot inference, which involves classifying into any user-defined labels without task-specific training (Hu et al., 2026). However, these results are primarily based on numerical simulations. Simulations do not account for nonidealities such as imperfect channel estimates. Therefore, this work seeks to determine whether the advantages of zero-shot semantic communication persist when implemented on a software-defined radio (SDR). Previous research has validated learned deep joint source-channel coding (Bourtsoulatz et al., 2019), a form of semantic communication that still attempts to recover the full data, on SDR (Ding et al., 2026), but has not addressed the zero-shot variant.

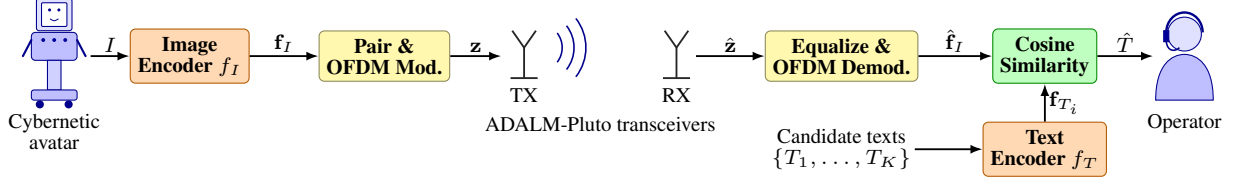


Fig. 1: Zero-shot semantic communication system model for transmitting cybernetic avatars’ observations to teleoperators.

Cybernetic avatar systems are a natural application for zero-shot semantic communication. A cybernetic avatar is a robotic system that acts as a teleoperator’s surrogate in a remote environment, over a Local 5G network shared with other avatars (Murakami et al., 2025, 2026). Future deployments envision many teleoperators collaboratively controlling many avatars (Hagita et al., 2024), so perception streams, control commands, and safety traffic compete for a finite radio pool (Dataesatu et al., 2024, 2025). Perception streams are the bulk of that load and the part least sensitive to reconstruction, so compressing them semantically frees bandwidth for the traffic that cannot wait (Ekolle et al., 2025, 2026). Because avatars handle general tasks that cannot all be specified at deployment, a zero-shot link that needs no retraining for new categories fits the operational model.

To our knowledge, zero-shot semantic communication has not yet been demonstrated using radio hardware. We address this gap through the following contributions:

- **Hardware implementation.** A Raspberry Pi operates a pair of Analog Devices Active Learning Module (ADALM)-Pluto transceivers. A VLM encoder generates the payload, and the receiver selects outputs by cosine similarity against text prompts that can be updated at any time without retraining.
- **VLM comparison.** We compare two VLM encoders, Contrastive Language-Image Pre-training (CLIP) and the mobile-scale MobileCLIP, quantifying the accuracy each achieves relative to its encoding time on the Raspberry Pi. Offloading the encoder to a neural processing unit (NPU) reduces encoding to 13.4 ms per image, $49\times$ faster than a Raspberry Pi 4 CPU.
- **Measured link performance.** Under varying transmit power, the semantic link needs $9\times$ fewer channel uses than the conventional baseline and holds higher accuracy under noise: 98.3% on the traffic sign recognition dataset (TSRD) at 22.3 dB with MobileCLIP, where the baseline scores 0%.

2 Overview

2.1 Vision-Language Model

Conventional vision models are limited to a predefined set of labels. Classifying new data requires collecting new labels and retraining the model. CLIP (Radford et al., 2021) and its variants (Vasu et al., 2024) utilize natural language captions for supervision, pairing an image encoder $f_I(\cdot)$ with a text encoder $f_T(\cdot)$ in a shared d -dimensional space. The model is trained on millions of image-text pairs to handle a wide variety of topics. During training on a batch of M pairs, the model learns which of the $M \times M$ pairings are correct by increasing the cosine similarity of the M true pairs and decreasing that of the remaining $M^2 - M$ incorrect pairs. Let $l \triangleq [0, 1, \dots, M - 1]$ denote the correct indices. With $\mathbf{f}_I, \mathbf{f}_T$ stacking the M batch embeddings, the loss is a symmetric cross-entropy over their similarity matrix.

$$\mathcal{L}_{\text{CLIP}} = \left[\mathcal{L}_{ce}(\cos(\mathbf{f}_I, \mathbf{f}_T^\top), l) + \mathcal{L}_{ce}(\cos(\mathbf{f}_T, \mathbf{f}_I^\top), l) \right] / 2. \quad (1)$$

This loss function consists of two cross-entropy losses: the first asks which text matches each image, while the second asks which image matches each text.

A trained CLIP model can recognize previously unseen class sets by representing each class as a text caption, encoding it with the text encoder $f_T(\cdot)$, and assigning images to the nearest text embedding. Task adaptation is achieved by modifying the caption rather than retraining the classification model, a design choice that supports general data transmission.

2.2 System Model

We consider a remote monitoring task within a cyber-physical system, *e.g.*, a cybernetic avatar transmitting image observations to a teleoperator via a wireless link. Fig. 1 illustrates the complete end-to-end pipeline. The cybernetic

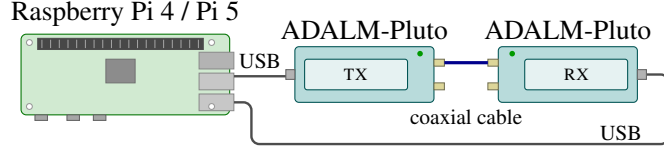


Fig. 2: The SDR hardware and its connections.

avatar captures an image I , which is processed by a pretrained image encoder to produce the embedding $\mathbf{f}_I = f_I(I)$. The transmitter pairs the d real components of $\mathbf{f}_I$ into $L = d/2$ complex symbols and modulates them onto orthogonal frequency-division multiplexing (OFDM) subcarriers as $\mathbf{z} \in \mathbb{C}^L$, subject to the average power constraint $\|\mathbf{z}\|^2/L \leq P$.

The symbols are transmitted to the receiver over a block Rayleigh fading channel.

$$\hat{\mathbf{z}} = h \mathbf{z} + \mathbf{n}. \quad (2)$$

In this model, the gain h is assumed to be distributed as $\mathcal{CN}(0, 1)$, remaining constant during the transmission of a single image and varying independently between images. The noise $\mathbf{n} \in \mathbb{C}^L$ is modeled as white Gaussian noise, with each entry distributed as $n_i \sim \mathcal{CN}(0, \sigma^2)$. Equation (2) represents the standard assumption for the wireless link. However, in this implementation, neither h nor $\mathbf{n}$ is sampled, as both are generated by the physical link between the two SDR transceivers. The signal-to-noise ratio (SNR) of the wireless link is defined as $\mu = 10 \log_{10}(P/\sigma^2)$ dB.

The receiver equalizes and demodulates the channel output $\hat{\mathbf{z}}$ to obtain a noisy embedding $\hat{\mathbf{f}}_I$. The receiver maintains a set of candidate texts $\{T_1, \dots, T_K\}$ corresponding to the classes to be distinguished. Each candidate text is encoded using the pretrained text encoder as $\mathbf{f}_{T_i} = f_T(T_i)$. The receiver then selects the label whose text embedding is closest to the received image embedding.

$$\hat{T} = \arg \max_{T_i} \cos(\hat{\mathbf{f}}_I, \mathbf{f}_{T_i}). \quad (3)$$

The selected label $\hat{T}$ is then communicated to the operator. Both encoders are pretrained and operate in inference mode only, without retraining; hence, serving a new class list requires only updating the candidate texts at the receiver.

3 Proposed Implementation

3.1 Hardware and Software Architecture

A Raspberry Pi 4 or Pi 5 hosts the demonstration and controls two ADALM-Pluto transceivers via USB (Fig. 2). The two SubMiniature version A (SMA) ports on the two ADALM-Pluto transceivers are directly connected by a coaxial cable. The image encoder f_I and the OFDM modulator are implemented on the transmitter, whereas equalization, demodulation, the text encoder f_T , and cosine similarity computation are performed on the receiver. The link operates at a carrier frequency of 915 MHz, well inside the ADALM-Pluto tuning range of 325 MHz to 3.8 GHz. The 915 MHz band carries LoRa and other industrial Internet-of-Things traffic, so it is representative of the unlicensed sub-GHz spectrum available to drones, autonomous robots, or cybernetic avatars.

We deliberately designed this architecture to maintain a modest hardware footprint. The Raspberry Pi consumes only a few watts, *e.g.*, less than 7 W for the Pi 4, which is within the operational limits of drones and mobile robots, so it serves as a representative computing platform for these applications. All other components are commodity and open-source, including off-the-shelf ADALM-Pluto transceivers and the `pyadi-iiio`, `CLIP`, and `OpenCLIP` Python libraries. The Python implementation is available at https://github.com/thanhlexyz/zssc_sdr. This availability enables cost-effective reproduction and facilitates further development based on the proposed implementation.

3.2 OFDM Implementation

OFDM serves as the data transmission scheme responsible for carrying the embedding over the radio link (Wu et al., 2022). OFDM also forms the foundation of the physical layer in 4G, 5G, and recent Wi-Fi standards. Consequently, a demonstration based on OFDM can be readily transferred to deployed systems. The structural elements common to these standards, such as the cyclic prefix and comb-type pilots, are retained. The SDR demonstration is implemented from scratch because full protocol stacks are large and do not build efficiently on the Raspberry Pi and ADALM-Pluto devices. Our minimal implementation enabled a fast and compatible prototype on the chosen hardware.

Fig. 3 illustrates the frame structure, which consists of a 64-sample cyclic prefix followed by a single 512-point OFDM symbol, totaling 576 samples or 288 μ s at 2 MHz. Each embedding is transmitted within a single OFDM frame, with

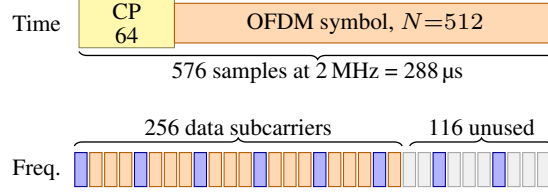


Fig. 3: OFDM frame structure; blue cells mark the comb pilots.

Table 1: OFDM and transceiver parameters

Parameter	Value
Transceivers	$2 \times$ ADALM-Pluto
Host	Pi 4 (link), Pi 5 (processing time)
Carrier frequency	915 MHz
Sample rate	2 MHz
FFT size N	512
Subcarrier spacing	3.91 kHz
Cyclic prefix	64 samples (32 μ s)
Pilot spacing	every 4th subcarrier (128 pilots)
Data subcarriers	256 of 372 usable subcarriers
Frame length	576 samples (288 μ s)
Nominal TX power	-73 to -13 dBm (swept)
RX gain	30 dB, manual
Measured SNR range	-2.6 to 43.7 dB

the cyclic prefix configured to mitigate multipath propagation. In the absence of a guard interval, multipath propagation causes symbol overlap, resulting in inter-symbol interference and a frequency-selective channel. Prepending a copy of the symbol tail to the output of the inverse Fourier transform renders the channel convolution circular. The forward transform subsequently diagonalizes the channel into orthogonal subchannels, so that (2) holds on each subcarrier with its own coefficient h . The receiver recovers these coefficients per frame from the pilots, after removing the carrier frequency offset estimated from the cyclic prefix and pilots. The 64-sample cyclic prefix, one eighth of the symbol or 32 μ s, exceeds the typical indoor delay spread at 915 MHz.

The ADALM-Pluto is a direct-conversion transceiver, so local-oscillator leakage and offsets in the in-phase and quadrature signal paths manifest as a spurious tone. This spurious tone appears on the subcarrier located at the center of the frequency band. Conversely, the analog filter exhibits roll-off near half the sample rate, attenuating the subcarriers at the edges of the frequency band. As a result, nine subcarriers centered on each of these two points are reserved and not used for data transmission. Six of these 18 coincide with the 128 comb pilots on every fourth subcarrier, so twelve are nulled and $512 - 128 - 12 = 372$ remain available. One embedding occupies the first 256 of the 372, and the remaining 116 are left unused. The ℓ_2 -normalized embedding meets the constraint $\|\mathbf{z}\|^2/L \leq P$ of Sec. 2.2, and the transmitter scales each frame to the same peak amplitude.

The receiver does not perform closed-loop tracking; consequently, each frame arrives at an unknown offset within the receive buffer. Frame timing is the peak of the cyclic-prefix correlation at a 512-sample lag, accumulated over the cyclically repeated frames in the receive buffer. Table 1 summarizes the resulting parameters and the transceiver settings.

4 Experiment

Our evaluation answers the following research questions: **RQ1**: How does zero-shot accuracy vary across channel conditions and unseen datasets?

RQ2: How much bandwidth does zero-shot semantic transmission save compared to a traditional baseline?

RQ3: What is the processing time on the Raspberry Pi?

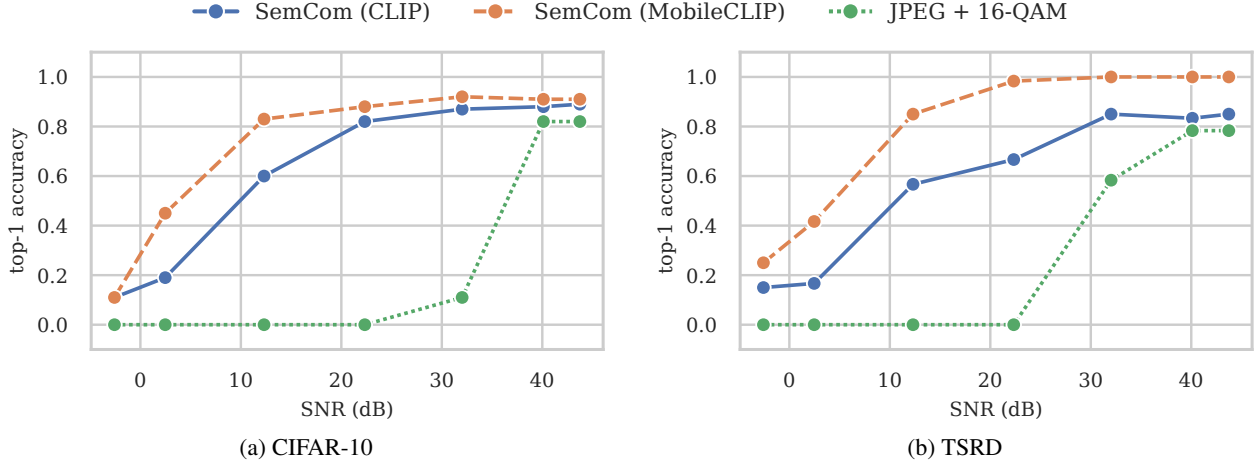


Fig. 4: Accuracy of zero-shot semantic communication (SemCom) and the JPEG + 16-QAM baseline versus measured SNR.



Fig. 5: Three example classes from each dataset (first evaluated image of each class): TSRD signs at native resolution and CIFAR-10 thumbnails at 32×32 .

4.1 Experimental Setup

Platform. Link measurements use a Raspberry Pi 4; processing time is additionally reported on a Raspberry Pi 5, on its CPU and with a Hailo-10H NPU over PCIe.

Model. This study evaluates two VLMs, both pretrained and used exclusively for inference in their released forms: CLIP ViT-B/32 (Radford et al., 2021) and MobileCLIP2-S0 from the MobileCLIP family (Vasu et al., 2024), hereafter CLIP and MobileCLIP. Both models use a $d=512$ embedding space, so each image feature requires $L=256$ complex symbols, and the OFDM configurations are identical. Image embeddings undergo ℓ_2 normalization before transmission.

Datasets and prompts. We evaluate on two datasets that differ in resolution, namely, CIFAR-10 and TSRD. CIFAR-10 provides 32×32 images across ten classes. We use one hundred test images and the prompt template “a photo of a {class}”. TSRD (Yang et al., 2016) provides traffic sign images of varying size. We use 60 images (median 112×106 pixels) from six of its classes, 10 per class. The candidate prompts for this dataset are: “speed limit is 30”, “speed limit is 50”, “speed limit is 60”, “speed limit is 80”, “pedestrian crossing”, and “do not enter”. For CIFAR-10, we use the first ten test indices of each class, resulting in a fixed image set rather than a randomly sampled one. Fig. 5 shows three example classes from each dataset.

Baseline. The conventional transmission pipeline transmits the original image data directly. JPEG compression is applied to each image, with the quality parameter determined by a binary search to meet a 1146-byte size constraint and optimized Huffman tables. If the lowest quality setting exceeds this limit, the system incrementally reduces image resolution. We tuned this budget to the smallest size at which the baseline still reaches roughly 80% accuracy on a clean channel (Sec. 4.2); a tighter budget would starve the baseline regardless of channel condition, thereby weakening the comparison.

The resulting bitstream includes a two-byte synchronization word, a two-byte length field, and a 16-bit cyclic redundancy check (CRC) for error detection. With 16-ary quadrature amplitude modulation (16-QAM), the compressed image fills nine OFDM frames. At the receiver, the image is decoded and classified using the same CLIP model (Radford et al., 2021).

Channel. All results are obtained using the link described in Sec. 3.1, employing the OFDM parameters specified in Table 1. Transmit power determines the operating point and is the sole parameter varied in these experiments. The ADALM-Pluto transceiver has a maximum transmission power of +7 dBm. The nominal transmit power is swept across $\{-73, -63, \dots, -13\}$ dBm. Receiver gain remains fixed in manual mode; therefore, any change in transmit power is reflected in the SNR of the channel.

At each transmit power setting, 200 probe frames consisting of unit-power complex Gaussian tones are used to determine the operating SNR, calculated from the residual error power after removal of the common phase. The seven transmit power settings correspond to measured SNRs of $-2.6, 2.5, 12.3, 22.3, 32.0, 40.1$, and 43.7 dB. We plot all subsequent accuracy results against these measured SNR values.

Metrics. We report accuracy as the proportion of images for which the nearest candidate prompt corresponds to the ground-truth class. We measure bandwidth in channel uses per image because the semantic path transmits analog symbols and does not generate a measurable bitstream. For RQ3, we report the median wall-clock time for both transmit-side and receive-side processing, excluding OFDM, 16-QAM, and the transceivers.

4.2 Accuracy Across Channel Conditions

When evaluated against the link with different measured SNRs, the semantic and conventional accuracies exhibit distinct trends (Fig. 4). As SNR decreases, both zero-shot semantic transmission paths gradually decline in performance, ultimately approaching chance level. For CLIP, CIFAR-10 classification accuracy decreases from 89% to 11% as the link SNR degrades from 43.7 to -2.6 dB.

On TSRD, zero-shot semantic communication demonstrates a similar trend, with CLIP accuracy decreasing from 85% to 15%, compared to a 16.7% accuracy of a random classifier for six classes. Notably, this behavior is independent of the specific model used to generate the embedding, *e.g.*, both CLIP and MobileCLIP have relatively similar trends.

In contrast, the baseline accuracy remains at exactly 0% for all SNR values at or below 22.3 dB across both datasets. The minimum accuracy is zero because frames that fail synchronization or the CRC produce no output image. In such cases, there is no data to decode or re-embed, and the trial is consequently recorded as incorrect. Above this threshold, the baseline accuracy increases sharply, reaching 11% at 32.0 dB and 82% at 40.1 dB on CIFAR-10.

Zero-shot semantic communication and the baseline have a substantial performance gap at low SNR. With CLIP, semantic transmission achieves 82% accuracy on CIFAR-10 at 22.3 dB. The baseline does not attain comparable accuracy until 40.1 dB. Across the four SNR settings from -2.6 to 22.3 dB, the baseline fails to classify any samples. Once the link quality improves, the gap narrows, with accuracies of 89% and 82% on CIFAR-10, indicating that the primary benefit of zero-shot semantic communication is concentrated in the low-SNR regime.

MobileCLIP matches or exceeds CLIP’s semantic transmission accuracy at every SNR on both datasets, with the largest gains between 2.5 and 22.3 dB. On CIFAR-10, MobileCLIP reads 45% against CLIP’s 19% at 2.5 dB and 83% against 60% at 12.3 dB. In the best channel condition, the two achieve 91% and 89%. TSRD shows the same pattern more sharply, with gains of 10.0 to 31.7 percentage points.

4.3 Bandwidth and Cross-Dataset Comparison

The semantic payload requires a single OFDM frame with 256 tones (Fig. 6). The baseline approach requires nine frames, totaling 2304 tones, because 1146 JPEG bytes encoded at four bits per tone cannot fit in fewer frames. Consequently, semantic transmission needs $9\times$ fewer channel uses per image.

Furthermore, semantic transmission achieves higher classification accuracy across both evaluated models. At the highest evaluated setting, CLIP achieves 89% accuracy on CIFAR-10 versus the baseline’s 82%, and 85.0% versus 78.3% on TSRD. MobileCLIP achieves 91% accuracy on CIFAR-10 and 100% on TSRD. The more efficient representation yields higher accuracy because the baseline discards task-relevant details to meet its bit budget, resulting in semantic recovery from a degraded reconstruction.

For CLIP, the performance margin remains consistent across both datasets, despite a median $12\times$ difference in pixel count, with improvements of 7.0 and 6.7 percentage points. Although a TSRD sign provides JPEG with substantially more data to compress than a 32×32 thumbnail, the semantic advantage remains nearly unchanged. Therefore, the

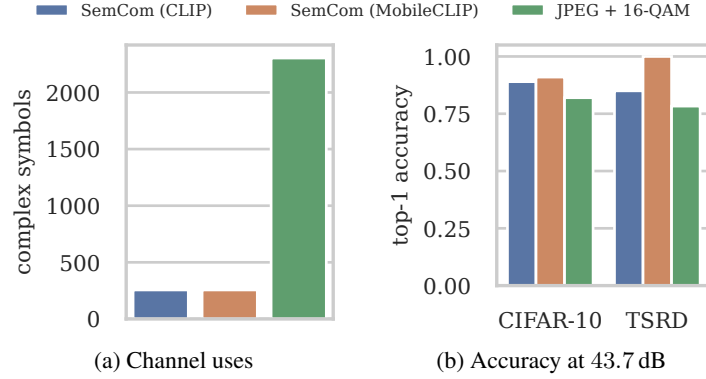


Fig. 6: Cost versus accuracy. Zero-shot semantic communication needs $9\times$ fewer channel uses and is more accurate.

Table 2: Median per-image encode and decode time across compute platforms, in milliseconds

Platform	Method	CIFAR-10		TSRD	
		Encode	Decode	Encode	Decode
Pi 4	SemCom (CLIP)	661.1	0.5	647.4	0.5
Pi 4	SemCom (MobileCLIP)	889.9	0.5	657.7	0.5
Pi 4	JPEG + 16-QAM	1.4	674.5	3.0	661.7
Pi 5	SemCom (CLIP)	262.6	0.2	264.7	0.2
Pi 5	SemCom (MobileCLIP)	200.3	0.2	206.9	0.2
Pi 5	JPEG + 16-QAM	0.5	265.0	1.4	273.8
Pi 5+NPU	SemCom (CLIP)	13.4	0.2	13.5	0.3

observed advantage is attributable to the representation itself. MobileCLIP increases the margin to 9.0 and 21.7 percentage points, indicating that model selection affects the magnitude of the advantage but not its direction.

4.4 Processing Time

Each scheme runs a single VLM forward pass per image, which dominates the computational cost, but the two schemes place it at opposite ends of the link (Table 2). In the semantic communication scheme the transmitter incurs it, and the receiver computes the cosine similarities as described in (3), requiring only a single matrix-vector product with the cached prompt embeddings, which takes 0.497 ms on CIFAR-10 on the Pi 4. In the baseline scheme, the receiver incurs this cost, as it must decode the JPEG and re-embed the reconstructed image before making a decision, requiring 674.5 ms. Consequently, receive-side processing of zero-shot semantic communication is less costly, whereas transmit-side processing becomes more expensive.

The mobile-scale model does not make encoding faster on the Raspberry Pi 4: although MobileCLIP is smaller, it encodes slower than CLIP on both datasets, because all Pi 4 inference runs in PyTorch on an unaccelerated Cortex-A72, whereas these architectures are designed to leverage mobile accelerators, *e.g.*, the iPhone GPU. A Raspberry Pi 5 confirms this attribution: its Cortex-A76 cores encode $2.5\times$ faster for CLIP and up to $4.4\times$ faster for MobileCLIP, reversing the ordering so that MobileCLIP becomes 23.7% faster on CIFAR-10 and 21.8% faster on TSRD. A Hailo-10H NPU on the Pi 5 further reduces CLIP encoding to 13.4 ms, $19.6\times$ faster than its CPU and $49\times$ faster than the Pi 4, at the cost of 8-bit quantization, which leaves accuracy unchanged at 83.3% on TSRD and reduces it from 90.0% to 88.0% on CIFAR-10 without the radio link. Encoding therefore fits a real-time budget, so transmitter-side latency no longer limits the scheme.

This asymmetry matters for practical deployment. In the baseline scheme a receiver aggregating multiple transmitters incurs the encoder cost once per link, whereas in the semantic scheme that cost is eliminated, so the receiver’s load barely grows with the number of senders, *e.g.*, drones, autonomous robots, or cybernetic avatars.

5 Conclusion

This study demonstrates the feasibility of zero-shot semantic communication on physical radio hardware. A Raspberry Pi equipped with two ADALM-Pluto transceivers successfully classifies previously unseen images using only the transmitted embedding. The implemented scheme needs $9\times$ fewer channel uses while achieving higher accuracy than the JPEG + 16-QAM baseline. The computational cost profile differs: encoding moves to the transmitter, but that is not a barrier, as a Hailo-10H NPU reduces it to 13.4 ms per image, $49\times$ faster than an unaccelerated Raspberry Pi 4 CPU, for at most two percentage points of accuracy.

Several challenges remain. Because the present link uses a coaxial cable, repeating these measurements over the air in an anechoic chamber is the next step toward field deployment. Learned radio functions are also vulnerable to adversarial perturbations (Le et al., 2025). Formal verification has been applied to deep learning-based traffic prediction (Le and Matsumura, 2026), massive MIMO power control (Le et al., 2026b), and semantic communication under adversarial noise (Le et al., 2026a). Neural network verification (Duong et al., 2025a,b,c, 2026a,b) now also covers vision (Duong et al., 2026c,d) and language (Duong et al., 2026e) models, and extending it to the vision-language models used here could bound the receiver’s error within the noise range measured on this testbed. Finally, because zero-shot semantic communication is agnostic to embedding content, it can carry embeddings from a vision-language-action model, instructing remote robotic agents with minimal data.

Acknowledgments

This research is partially supported by JST Moonshot R&D (JPMJMS2216).

References

- Thanh Le, Hai Duong, ThanhVu Nguyen, and Takeshi Matsumura. Verifying DNN-Based Semantic Communication Against Generative Adversarial Noise. *arXiv preprint arXiv:2602.08801*, 2026a.
- Huiqiang Xie, Zhijin Qin, Xiaoming Tao, and Khaled B. Letaief. Task-Oriented Multi-User Semantic Communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597, 2022.
- Sige Liu, Nan Li, Yansha Deng, and Tony Q. S. Quek. Goal-Oriented Semantic Communication for Wireless Visual Question Answering. *IEEE Journal on Selected Areas in Communications*, 43(12):4247–4261, 2025.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15963–15974, 2024.
- Jiangjing Hu, Haotian Wu, Wenjing Zhang, Fengyu Wang, Wenjun Xu, Hui Gao, and Deniz Gündüz. Zero-Shot Semantic Communication With Multimodal Foundation Models. *IEEE Transactions on Vehicular Technology*, 75(5): 8759–8764, 2026.
- Eirina Bourtsoulatz, David Burth Kurka, and Deniz Gündüz. Deep Joint Source-Channel Coding for Wireless Image Transmission. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, 2019.
- Jiarun Ding, Peiwen Jiang, Chao-Kai Wen, and Shi Jin. Prototyping Real-Time Semantic Communications With SDRs and Edge GPUs. *IEEE Wireless Communications Letters*, 15:2045–2049, 2026.
- Homare Murakami, Kazuo Ibuka, Atsushi Wakayama, Shinichi Hama, Jun Amagai, and Takeshi Matsumura. Towards Stable Teleoperation of Cybernetic Avatars: Design and Evaluation of a Hybrid Traffic Management Platform. In *GLOBECOM 2025 - 2025 IEEE Global Communications Conference*, pages 5393–5398, 2025.
- Homare Murakami, Kazuo Ibuka, Atsushi Wakayama, Shinichi Hama, Jun Amagai, and Takeshi Matsumura. Toward Stable Cybernetic Avatar Teleoperation: A Reliability-Ensuring Platform with Communication Quality Monitoring and Jitter Control. *IEICE Transactions on Communications*, 2026. early access.
- Norihiro Hagita, Ryota Kanai, Hiroshi Ishiguro, Kouta Minamizawa, Fumihito Arai, Fumio Shimpo, Takeshi Matsumura, and Yoko Yamanishi. Cybernetic Avatars: Teleoperation Technologies from In-Body Monitoring to Social Interaction. *Science Robotics*, 9(96), 2024. Art. no. eadg1842.

- Arif Dataesatu, Atsushi Wakayama, Kazuo Ibuka, Homare Murakami, and Takeshi Matsumura. Multi-Layer Resource Management Strategies for Cybernetic Avatars in Dynamic Local 5G Environments. In *2024 27th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 1–6, 2024.
- Arif Dataesatu, Atsushi Wakayama, Kazuo Ibuka, Homare Murakami, and Takeshi Matsumura. Fair Resource Allocation with Dynamic Multi-Layer Service Area Management in Local Networks Supporting Cybernetic Avatars. In *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, pages 1–7, 2025.
- Zie Eya Ekolle, Homare Murakami, and Takeshi Matsumura. A Goal-Oriented Semantic Communication for a Reliable Cybernetic Avatar Teleoperation System. In *2025 IEEE Future Networks World Forum (FNWF)*, pages 1–6, 2025.
- Zie Eya Ekolle, Homare Murakami, and Takeshi Matsumura. A Semantic Communication Model with Noise Reduction in a Cybernetic Avatar Teleoperation. In *2026 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 505–510, 2026.
- Haotian Wu, Yulin Shao, Krystian Mikolajczyk, and Deniz Gündüz. Channel-Adaptive Wireless Image Transmission With OFDM. *IEEE Wireless Communications Letters*, 11(11):2400–2404, 2022.
- Yi Yang, Hengliang Luo, Huarong Xu, and Fuchao Wu. Towards Real-Time Traffic Sign Detection and Classification. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2022–2031, 2016.
- Thanh Le, Hai Duong, Yusheng Ji, ThanhVu Nguyen, and John C. S. Lui. FGGM: Formal Grey-Box Gradient Method for Attacking DRL-Based MU-MIMO Scheduler. *arXiv preprint arXiv:2510.26075*, 2025.
- Thanh Le and Takeshi Matsumura. Formal Verification for Deep Learning-Based Mobile Network Traffic Prediction. In *2026 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, pages 246–250, 2026.
- Thanh Le, Takeshi Matsumura, Yusheng Ji, and John C. S. Lui. Formal Verification for Deep Learning-Based Power Control in Massive MIMO. In *2026 IEEE Vehicular Technology Conference (VTC2026-Fall), Recent Results Track*, 2026b. to be published.
- Hai Duong, ThanhVu Nguyen, and Matthew B. Dwyer. NeuralSAT: A High-Performance Verification Tool for Deep Neural Networks. In *International Conference on Computer Aided Verification (CAV)*, pages 409–423, 2025a.
- Hai Duong, David Shriver, ThanhVu Nguyen, and Matthew B. Dwyer. Compositional Neural Network Verification via Assume-Guarantee Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 64158–64182, 2025b.
- Hai Duong, ThanhVu Nguyen, and Matthew B. Dwyer. Generating and Checking DNN Verification Proofs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pages 65887–65909, 2025c.
- Hai Duong, Thanh Le, and ThanhVu Nguyen. CLAD: Constrained Abstract Domain for Neural Network Verification. *arXiv preprint arXiv:2609.34628*, 2026a.
- Hai Duong, Thanh Le, and ThanhVu Nguyen. Verifying Neural Networks with Reinforcement Learning. *arXiv preprint arXiv:2609.34553*, 2026b.
- Hai Duong, Thanh Le, Lam Nguyen, and ThanhVu Nguyen. Verifying Structural Robustness of Deep Neural Network. *Proceedings of the ACM on Software Engineering*, 3(FSE):1492–1514, 2026c.
- Hai Duong, Lam Nguyen, Thanh Le, and ThanhVu Nguyen. Verifying Neural Network Robustness with Dual Perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27916–27925, 2026d.
- Hai Duong, Thanh Le, and ThanhVu Nguyen. ZonoGPT: Towards an Abstract Domain for Verifying Large GPT Models. *arXiv preprint arXiv:2609.34457*, 2026e.