This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

SIMAC: A Semantic-Driven Integrated Multimodal Sensing And Communication Framework

Yubo Peng, *Graduate Student Member, IEEE*, Luping Xiang, *Senior Member, IEEE*, Kun Yang, *Fellow, IEEE*, Feibo Jiang, *Senior Member, IEEE*, Kezhi Wang, *Senior Member, IEEE*, and Dapeng Oliver Wu, *Fellow, IEEE*

***Abstract*—Traditional unimodal sensing faces limitations in accuracy and capability, and its decoupled implementation with communication systems increases latency in bandwidth-constrained environments. Additionally, single-task-oriented sensing systems fail to address users' diverse demands. To overcome these challenges, we propose a semantic-driven integrated multimodal sensing and communication (SIMAC) framework. This framework leverages a joint source-channel coding architecture to achieve simultaneous sensing, decoding, and transmission of sensing results. Specifically, SIMAC first introduces a multimodal semantic fusion (MSF) network, which employs two extractors to extract semantic information from radar signals and images, respectively. MSF then applies cross-attention mechanisms to fuse these unimodal features and generate multimodal semantic representations. Secondly, we present a large language model (LLM)-based semantic encoder (LSE), where relevant communication parameters and multimodal semantics are mapped into a unified latent space and input to the LLM, enabling channel-adaptive semantic encoding. Thirdly, a task-oriented sensing semantic decoder (SSD) is proposed, in which different decoded heads are designed according to the specific needs of tasks. Simultaneously, a multi-task learning strategy is introduced to train the SIMAC framework, achieving diverse sensing services. Finally, experimental simulations demonstrate that the proposed framework achieves diverse and higher-accuracy sensing services.**

***Index Terms*—Integrated multimodal sensing and communications; semantic communication; large language model; multi-task learning**

I. Introduction

A. Backgrounds

Unimodal sensing technologies, such as radar and visual sensing, have been extensively studied and widely deployed across various domains due to their respective strengths. In autonomous driving, radar enables precise measurement of object distance, velocity, and relative motion, making it essential for advanced vehicular systems [1]. In military applications, radar serves as a critical component for reconnaissance, surveillance, early warning, and missile defense systems [2]. Visual sensing, by contrast, captures rich image content and supports detailed object recognition and classification. Leading autonomous driving platforms, including those developed by Tesla and Waymo, leverage advanced visual perception algorithms to improve decision-making and situational awareness [3]. In intelligent surveillance systems, visual sensing enables automated video analytics for intruder detection, anomaly recognition, and fire monitoring, thus enhancing security and operational efficiency [4].

Despite their advantages, each modality suffers from inherent limitations. Radar cannot capture visual attributes such as color, texture, and shape, while visual sensing struggles with accurate spatial localization, particularly under challenging conditions involving low light or occlusion [5]. To overcome these shortcomings, multimodal sensing has emerged as a compelling approach, combining radar and visual inputs to exploit their complementary strengths [6]. By integrating radar's spatial awareness with the rich contextual information from visual data, multimodal systems offer a more comprehensive environmental understanding.

However, current sensing presents two major challenges: (1) The traditional decoupled architecture, where sensing is completed at the transmitter and the results are subsequently forwarded to the receiver [7]. This design poses sequential processing increases service latency, thereby limiting real-time responsiveness; and (2) sensing devices without multimodal capabilities must rely on other devices to access multimodal sensing services, which often entails transmitting large volumes of data and consequently imposes heavy communication overhead. These limitations not only undermine the efficiency of multimodal sensing but also hinder its deployment in scenarios with stringent latency and bandwidth constraints.

B. Challenges

Given this background, several challenges associated with traditional sensing technology are summarized as follows:

1) *Insufficient Information Sensing:* Unimodal sensing has inherent limitations; for example, radar lacks visual semantics, while vision struggles with precise spatial localization and is vulnerable to lighting and occlusion. These constraints limit the comprehensiveness of scene understanding.

The paper was partly funded by Jiangsu Major Project on Fundamental Research (Grant No.: BK20243059), Gusu Innovation Project (Grant No.: ZXL2024360), High-Tech District of Suzhou City (Grant No.: RC2025001) and Natural Science Foundation of China (Grant No. 62132004 and 62301122), the Major Program Project of Xiangjiang Laboratory (Grant No. XJ2023001 and XJ2022001), and Qiyuan Lab Innovation Fund (Grant No. 2022-JCJQ-LA-001-088). (Corresponding author: Luping Xiang.)

Yubo Peng (ybpeng@smail.nju.edu.cn), Luping Xiang (luping.xiang@nju.edu.cn), and Kun Yang (kunyang@nju.edu.cn) are with the State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China, and the School of Intelligent Software and Engineering, Nanjing University (Suzhou Campus), Suzhou, China.

Feibo Jiang (jiangfb@hunnu.edu.cn) is with the School of Information Science and Engineering, Hunan Normal University, Changsha, China.

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

Dapeng Oliver Wu (dpwu@ieee.org) is with the Department of Computer Science, City University of Hong Kong, Hong Kong.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

2) *High Communication Overhead:* To support devices without multimodal sensing capabilities, sensing results may be transmitted from the transmitter to the receiver. This process generates substantial data traffic, particularly in multimodal systems, which not only increases latency but also limits deployment in bandwidth-constrained scenarios.
3) *Limited Sensing Services:* Most sensing algorithms are designed for specific tasks (e.g., distance estimation or object recognition) and lack the flexibility to address diverse or user-specific requirements.

C. Contributions

Semantic communication (SC) is an innovative approach based on deep joint source-channel coding (JSCC), with the potential to transform communication system design and development [8], [9]. Unlike conventional systems, SC focuses on understanding and conveying the message's core meaning or intent, rather than transmitting all bits [10]. This paradigm shift allows for reduced redundancy and irrelevant data, improving transmission efficiency.

Integrated sensing and communication (ISAC) unifies sensing and data transmission, overcoming the limitations of traditional decoupled systems [11]. By enabling simultaneous sensing and communication, ISAC reduces latency and communication overhead, making it particularly effective for multimodal systems. This approach optimizes resource use and enhances efficiency in real-time applications.

Based on SC and ISAC, a semantic-driven integrated multimodal sensing and communication (SIMAC) framework is proposed to address the identified challenges. The main contributions are as follows:

1) We introduce a multimodal semantic fusion (MSF) network that employs two extractors to extract unimodal semantics from images and radar signals, respectively. One is based on the vision transformer (ViT) and the other is based on complex convolutional neural networks (CNNs). Then, a cross-attention mechanism is used to fuse these unimodal semantics, obtaining a comprehensive multimodal semantic representation. This approach fully leverages both physical position and visual information, addressing the first challenge.
2) We present an LLM-based semantic encoder (LSE), where a specialized embedding network maps both multimodal semantics and relevant communication parameters into a unified latent space and obtains an embedding. Then, an LLM is applied to perform semantic encoding on the generated embeddings. Due to the inclusion of communication parameters, LSE can flexibly adapt to various communication environments without retraining. Compared to traditional methods, only the semantic encoding needs transmission, reducing communication overheads and addressing the second challenge.
3) We design a multi-task-oriented sensing semantic decoder (SSD) with distinct decoding heads tailored to specific tasks, such as distance and angle prediction, velocity estimation, and image reconstruction. Additionally, a multi-task learning strategy is implemented to train these heads simultaneously, enhancing training efficiency. This approach enables users to access diverse sensing services, addressing the final challenge.
4) Based on the VIRAT Video Dataset [12], we construct a specific dataset to train and evaluate the proposed framework. The results demonstrate that our framework provides more diverse sensing services and higher accuracy with low communication costs.

D. Organization

The rest of the paper has the following structure: Section II introduces the related works, and Section III provides a detailed description of the system model. Section IV presents the proposed SIMAC framework, including the implementation of the MSF, LSE, and SSD modules. Section V employs experimental simulations to evaluate the performance of the proposed methods. Lastly, Section VI concludes this paper.

II. Related Works

This section reviews the related works about unimodal and multimodal sensing and ISAC. We also summarize the differences between our work and the existing works in Table I.

A. Unimodal Sensing

Sensing technologies have found extensive applications across various fields, due to their effectiveness in target detection and tracking, particularly in complex environments. For instance, in autonomous driving, Sun *et al.* [13] developed a high-resolution imaging radar system that delivers high-fidelity four-dimensional (4D) sensing through joint sparsity optimization in the frequency spectrum and array configurations. Sohail *et al.* [14] proposed a radar-based method for relative vehicle positioning, utilizing the dynamic range and azimuth of frequency-modulated continuous wave radar to achieve precise vehicle positioning. Additionally, Luo *et al.* [15] applied computer vision (CV)-based surface defect detection to monitor the status and integrity of bridge structures, ensuring their safety and reliability.

While these studies exploit the advantages of sensing technologies, they rely on unimodal data, which inherently limits the diversity and scope of their sensing capabilities. *Therefore, we propose an approach that integrates radar signals with visual data. In this method, the radar signals assist in locating the key target in rough visual information, while the visual modality improves the accuracy of the motion parameters estimation.*

B. Multimodal Sensing

The limitations of unimodal sensing, such as challenges in maintaining robustness and accuracy in complex environments, have driven interest in multimodal sensing technologies. Liu and Lin [16] introduced a multimodal dynamic hand gesture recognition method using a two-branch deformable network with Gram matching, ensuring reliable recognition and improving generalization across varying field-of-view

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

TABLE I: Comparison of Our Contributions with Related Literature

Contributions	Ours	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]
Unimodal sensing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Multimodal sensing	✓				✓	✓	✓				✓	✓	✓
ISAC	✓							✓	✓	✓	✓	✓	✓
Semantic communication	✓												
Channel self-adaption	✓				✓				✓			✓	✓
Multi-task learning	✓												

scenes. Deliali *et al.* [17] developed a framework to classify radar-based trajectories in multimodal traffic environments, enhancing performance under adverse lighting and weather conditions. Kim *et al.* [18] proposed an early fusion method that combines spatial and contextual properties from cameras and radar for improved 3D object detection.

Although these works demonstrate advancements in multimodal fusion and sensing accuracy, they largely overlook the communication cost of transmitting sensing results. Furthermore, their designs are often tailored to specific tasks, such as detection or imaging. *In this paper, we will address these gaps by introducing LSE to mitigate dynamic communication environment challenges and employing SSD to efficiently deliver diverse sensing services to users.*

C. Integrated Sensing and Communication

ISAC represents a paradigm shift by seamlessly unifying sensing and communication functionalities within a single framework, thereby significantly enhancing overall system efficiency. For unimodal ISAC systems, Zhang *et al.* [19] investigated an IRS-assisted, WPT-enabled ISAC architecture in which a base station (BS) simultaneously performs radar sensing and data reception from IoT devices. To extend ISAC capabilities to cellular-connected UAV systems, Wang *et al.* [20] proposed an extended Kalman filter-based data fusion algorithm that enables beyond-line-of-sight (LoS) sensing by providing accurate environmental information. Xiang *et al.* [21] introduced a green beamforming design for ISAC that utilizes beam-matching error to assess radar performance. In the context of multimodal ISAC, Xu *et al.* [22] proposed FVMNet, a radar-camera fusion network that enables real-time volumetric perception, demonstrating zero-shot generalization and robustness under diverse weather conditions. Jiang *et al.* [23] designed a vision-guided multiple-input multiple-output (MIMO) radar system capable of multi-subject vital sign monitoring in cluttered environments through adaptive beamforming. Yang *et al.* [24] developed a deep multimodal learning framework for wireless communications, introducing novel architectures for effective multi-source sensing data fusion and achieving improved performance in massive MIMO channel prediction tasks.

While existing ISAC studies have demonstrated the feasibility of performing sensing and communication simultaneously, they face two key limitations. First, conventional unimodal ISAC systems primarily rely on radio frequency (RF) signals and cannot perceive the environment through multiple modalities, making them inadequate in complex scenarios. Second, traditional multimodal ISAC approaches focus on enhancing the sensing capabilities of the BS, such as enabling more efficient beamforming, but do not offer sensing support to other users. *In contrast, unlike unimodal ISAC, SIMAC utilizes MSF to effectively integrate heterogeneous sensing data, thereby enabling robust multimodal perception. Compared to conventional multimodal ISAC, SIMAC is designed to provide multimodal sensing services to users who lack such capabilities, empowering them to perform downstream tasks, such as drone tracking and environmental modeling, locally and more efficiently. More importantly, SIMAC incorporates SC to jointly optimize the transmission and decoding of multimodal perception information between the BS and users. By transmitting only task-relevant semantic representations, SIMAC significantly reduces communication overhead while enhancing resource efficiency.*

III. Wireless Sensing and Communication System Model

A. System Model

As illustrated in Fig. 1, we consider a system consisting of N sensing targets (STs), a BS as the transmitter, and a user as the receiver. The primary objective of the BS is to transmit the image of the nth ST (i.e., communicated data) and its motion parameters (i.e., sensing results) to the user. Specifically, the BS utilizes its radar and camera to sense the nth ST from both digital signal and visual perspectives. The BS then uses an integrated SC and multimodal sensing system to communicate with the user, where the sensing decoding and data transmission are processed in parallelly. The detailed process is outlined as follows:

1) Sensing Data Acquisition: Due to its favorable properties, the linear frequency modulation (LFM) waveform is extensively employed in radar sensing. The frequency of this waveform varies linearly over time, making it inherently robust against doppler frequency shifts, which enhances signal processing gain. Assuming that the motion parameters of the ST n include the angle θ_n, distance d_n, and radial velocity v_n, as shown in Fig. 2, we adopt a single-input-multiple-output (SIMO) radar model to transmit the LFM waveform and capture the corresponding echo signal. The echo signal can be expressed as:

$$\mathbf{A}_n = \lambda \mathbf{a}(\theta_n) \mathrm{e}^{j2\pi\mu_n \mathbf{T}} s(\mathbf{T} - \tau_n), \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

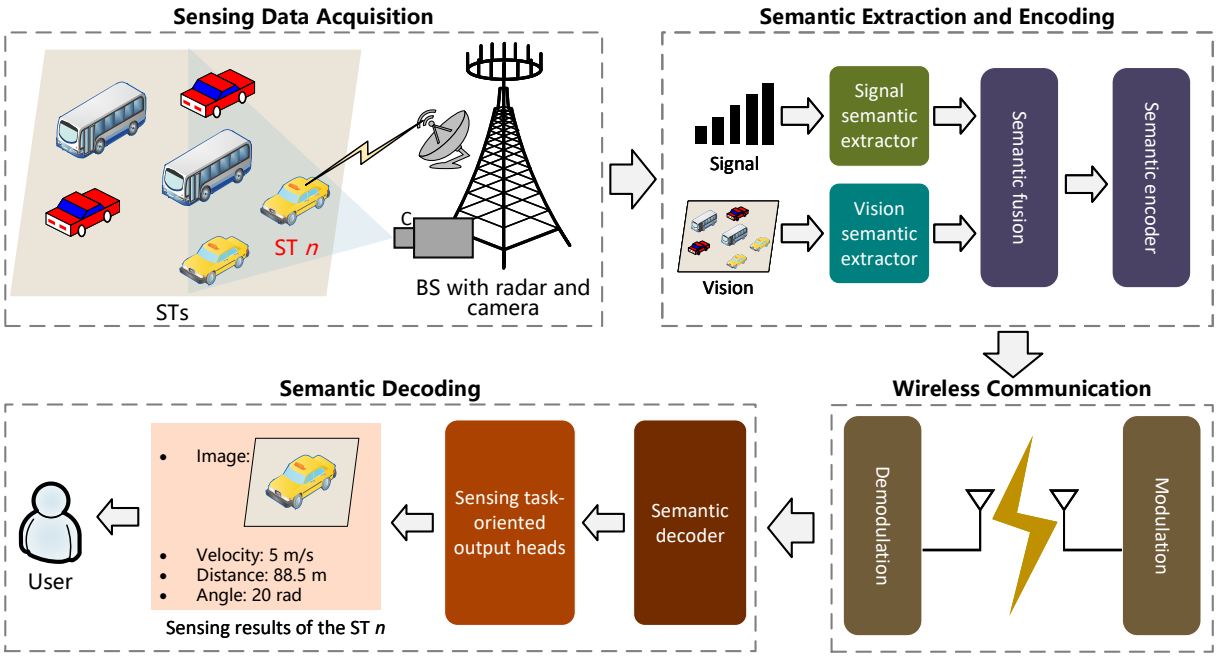


Fig. 1: The illustration of the integrated SC and multimodal sensing system model.

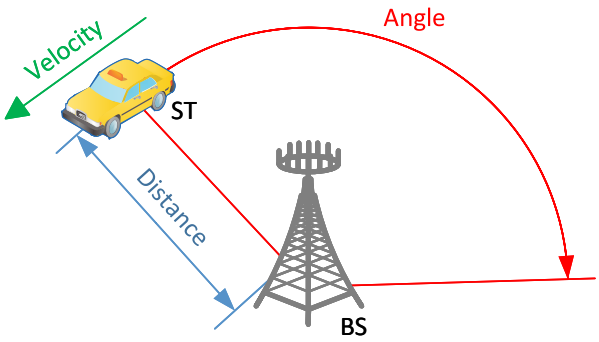


Fig. 2: The illustration of the motion parameters of the ST.

where,

$$\lambda = \frac{\xi \cdot \rho_n}{(4\pi)^{3/2} \cdot d_n^2}, \tag{2}$$

and

$$\xi = \frac{c}{f_c + K_t}, \tag{3}$$

while ρ_n is the radar cross-section (RCS) of the nth ST, which represents the ST's ability to reflect radar signals. $\mathbf{a}(\theta_n) = [1, \mathrm{e}^{-j\pi\cos\theta_n}, \ldots, \mathrm{e}^{-j(K-1)\pi\cos\theta_n}]$ is the steering vector, and K represents the number of receive antennas. $\mathbf{T} = [k\cdot\Delta t \mid k \in \mathbb{Z}, 0 \le k\Delta t \le T_r]$ denotes the time sequence of the sampling process, where T_r is the pulse repetition interval (PRI), $\Delta t = \frac{1}{F_s}$ is the sampling interval, and F_s is the sampling frequency. $\tau_n = \frac{2d_n}{c}$ and $\mu_n = \frac{2(f_c+K_t/2)v_n}{c}$ represent the time delay and doppler frequency shift, respectively, where f_c is the radar's central frequency, K_t is the frequency modulation slope, and $c = 3\times10^8$ m/s is the velocity of light.

BGR mode represents image pixels using blue, green, and red values, allowing direct display of sensor data without interpolation for optimal quality. Thus, it is widely used in image sensors [25]. We assume that the BS is equipped with a camera that utilizes the BGR mode to acquire high-quality images $\mathbf{m} \in \mathbb{R}^{W\times H\times 3}$, where W and H denote the width and height of the image in terms of the number of pixels, respectively. Note the captured image $\mathbf{m}$ may contain multiple STs, hence we aim to isolate the portion of the image corresponding to the nth ST, denoted as $\mathbf{m}_n$, using the latent information of the echo signal $\mathbf{A}_n$.

2) Semantic Extraction: Given that the echo signal $\mathbf{A}_n$ and the captured image $\mathbf{m}$ have distinct data dimensions and characteristics, we adopt two separate semantic extractors for each modality. The process of semantic extraction is described as follows:

$$\mathbf{s}_n^{\text{sig}} = S_{\text{sig}}(\mathbf{A}_n, \boldsymbol{\alpha}), \tag{4}$$

$$\mathbf{s}_n^{\text{vis}} = S_{\text{vis}}(\mathbf{m}, \boldsymbol{\beta}), \tag{5}$$

where $\mathbf{s}_n^{\text{sig}}$ and $\mathbf{s}_n^{\text{vis}}$ represent the semantic features of length L_s extracted from $\mathbf{A}_n$ and $\mathbf{m}$, respectively. $S_{\text{sig}}(\cdot)$ denotes the signal semantic extractor with parameters $\boldsymbol{\alpha}$, and $S_{\text{vis}}(\cdot)$ is the image semantic extractor with parameters $\boldsymbol{\beta}$.

To efficiently capture key information and explore the latent relationships between the two semantic features $\mathbf{s}_n^{\text{sig}}$ and $\mathbf{s}_n^{\text{vis}}$, a semantic fusion module is employed to combine these features and generate a comprehensive multimodal semantic representation $\mathbf{s}_n^{\text{mul}}$. This process can be expressed as:

$$\mathbf{s}_n^{\text{mul}} = S_{\text{mul}}(\mathbf{s}_n^{\text{sig}}, \mathbf{s}_n^{\text{vis}}, \boldsymbol{\gamma}), \tag{6}$$

where $S_{\text{mul}}(\cdot)$ is the semantic fusion module with parameters $\boldsymbol{\gamma}$.

3) Semantic Encoding: To minimize semantic distortion during wireless transmission, a JSCC encoder is employed to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

perform semantic encoding, taking $\mathbf{s}_n^{\text{mul}}$ as input. The encoding process is given by:

$$\mathbf{e}_n = F_{\text{se}}(\mathbf{s}_n^{\text{mul}}, \boldsymbol{\delta}), \tag{7}$$

where $\mathbf{e}_n$ represents the semantic encoding and $F_{\text{se}}(\cdot)$ is the JSCC encoder with parameters $\boldsymbol{\delta}$.

4) Wireless Communication: To ensure that the semantic encoding can be transmitted over the wireless channel, signal modulation techniques, such as QPSK and 16QAM, are employed to convert $\mathbf{e}_n$ into complex-valued symbols $\mathbf{c}_n$.

The complex-valued symbols $\mathbf{c}_n$ are transmitted over the channel, which is modeled as:

$$\mathbf{y}_n = \mathbf{H} \cdot \mathbf{c}_n + \mathbf{N}, \tag{8}$$

where $\mathbf{y}_n$ represents the received complex-valued symbols, $\mathbf{H}$ denotes the channel gain between the user and the BS. $\mathbf{N} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ represents Additive White Gaussian Noise (AWGN), where σ^2 is the noise power. Given that we consider a deep JSCC architecture, the channel model must be compatible with backpropagation to facilitate end-to-end training of both the encoder and decoder. Consequently, the wireless channel is simulated using neural network-based approaches [26]. It is notable that this semantic-level communication inherently mitigates privacy risks, as original sensor data is never exposed. Even if intercepted, the semantic features are difficult to decode without access to the task-specific decoder. Moreover, this framework is compatible with existing privacy-preserving techniques, such as homomorphic encryption [27], differential privacy [28], and secure multi-party computation can be directly applied to further enhance data security.

During wireless communication between the BS and the user, the transmission rate v can be expressed as:

$$v = B \log_2 \left(1 + \frac{P\|\mathbf{H}\|^2}{\sigma^2}\right), \tag{9}$$

where B and P represent the bandwidth and transmission power, respectively. Thus, the transmission delay is given by:

$$t^{\text{com}} = \frac{Z(\mathbf{c}_n)}{v}, \tag{10}$$

where $Z(\mathbf{c}_n)$ denotes the number of bits required to transmit the complex-valued symbols $\mathbf{c}_n$ to the user.

5) Semantic Decoding: Upon receiving the symbols $\mathbf{y}_n$, signal demodulation is applied to convert them back into the received semantic encoding $\hat{\mathbf{e}}_n$. To support diverse sensing services for the user, a JSCC decoder is employed to obtain sensing results tailored for multiple tasks. Specifically, we consider a variety of sensing tasks, including distance, angle, velocity prediction, and ST reconstruction. The decoding process is therefore formulated as:

$$\mathbf{o}_n = F_{\text{sd}}(\hat{\mathbf{e}}_n, \boldsymbol{\epsilon}), \quad \mathbf{o}_n \in \{\hat{\theta}_n, \hat{d}_n, \hat{v}_n, \hat{\mathbf{m}}_n\}, \tag{11}$$

where $F_{\text{sd}}(\cdot)$ represents the semantic decoder parameterized by $\boldsymbol{\epsilon}$. The decoded results, $\mathbf{o}_n$, may include a selection of the reconstructed image $\hat{\mathbf{m}}_n$ of the nth ST, predicted distance $\hat{d}_n$, predicted velocity $\hat{v}_n$, and estimated angle $\hat{\theta}_n$.

B. Problem Formulation

To realize the data transmission and sensing decoding from the BS to the user via a wireless channel, the total execution time T^{exe} comprises the computation time for semantic extraction t^{st} and encoding t^{se} at the transmitter, the communication time for transmission t^{com}, and the computation time for semantic decoding at the receiver t^{sd}. Thus, the total execution time can be expressed as:

$$T^{\text{exe}} = t^{\text{st}} + t^{\text{se}} + t^{\text{com}} + t^{\text{sd}}. \tag{12}$$

To provide diversified services for the user, we consider multiple sensing tasks, including distance, velocity, angle prediction, and image reconstruction of the nth ST. The corresponding task losses are defined as follows:

$$\mathcal{L}_{\text{dp}} = ||d_n - \hat{d}_n||^2, \tag{13}$$

$$\mathcal{L}_{\text{ap}} = ||\theta_n - \hat{\theta}_n||^2, \tag{14}$$

$$\mathcal{L}_{\text{vp}} = ||v_n - \hat{v}_n||^2, \tag{15}$$

$$\mathcal{L}_{\text{sr}} = ||\mathbf{m}_n - \hat{\mathbf{m}}_n||. \tag{16}$$

The primary goal of the SIMAC framework is to minimize semantic distortion during wireless transmission while maximizing the accuracy of the decoded sensing results. Additionally, transmission delays must be accounted for to ensure the quality of service. Accordingly, the objective function of the proposed SIMAC framework can be expressed as:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\epsilon}} \; l_1\mathcal{L}_{\text{dp}} + l_2\mathcal{L}_{\text{ap}} + l_3\mathcal{L}_{\text{vp}} + l_4\mathcal{L}_{\text{sr}}, \tag{17a}$$

$$\text{s.t. } T^{\text{exe}} \leq T_{\max}, \tag{17b}$$

where $T_{\max}$ denotes the latency requirement for completing the sensing task. $l1$, $l2$, $l3$, and $l4$ are adjustment factors.

To address the optimization problem described in Eq. (17a), there are three key issues. First, images and radar signals have different modalities and are difficult to process using a single neural network. Second, fixed-strategy semantic encoding is difficult to adapt to dynamic changes in communication parameters, resulting in low semantic fidelity. Finally, traditional single-task learning has difficulty optimizing multiple objectives simultaneously. Therefore, we have meticulously designed specialized neural networks for the various modules within the SIMAC framework, which will be described in the next section.

IV. Semantic-Driven Integrated Multimodal Sensing and Communication

A. Overview

As illustrated in Fig. 3, the SIMAC framework first employs the MSF module to extract semantic features from radar and visual inputs and fuse them into a compact multimodal representation. To enable channel adaptivity, the LSE module incorporates real-time communication parameters, such as SNR and modulation scheme, expressed in natural language (e.g., "the SNR is 5 dB and the signal modulation is QPSK") into the semantic encoding process. At the receiver side, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

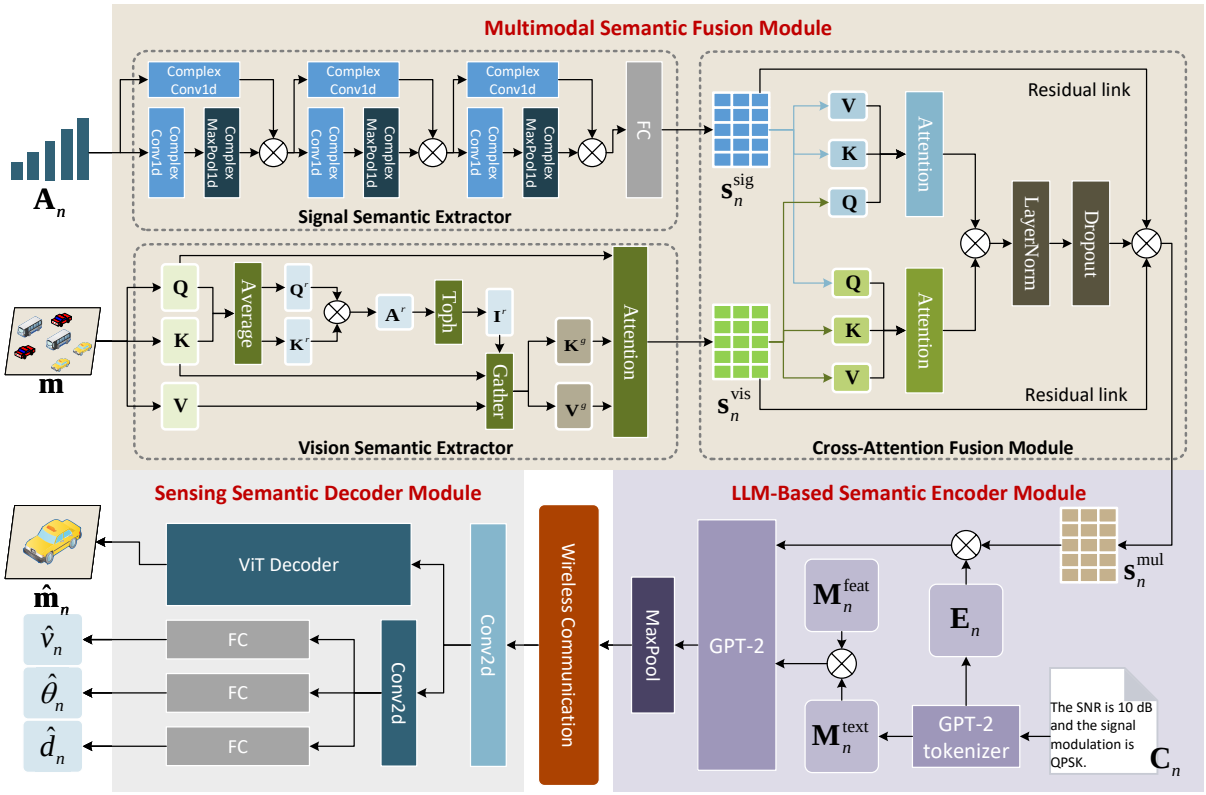


Fig. 3: The network design of the proposed SIMAC framework.

SSD module decodes the transmitted features and generates multi-task outputs, including distance, velocity, angle, and reconstructed images. Assuming the training dataset is denoted by $\mathcal{D}$, the overall workflow of the proposed SIMAC framework is summarized in **Algorithm 1**.

Algorithm 1 SIMAC Framework Workflow

Input: $\mathbf{m}, \mathbf{A}_n, \mathcal{D}$.
Output: $\mathbf{o}_n, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}$

Inference Phase:

1: Obtain the semantic representation $\mathbf{s}_n^{\text{mul}}$ from multimodal data $\mathbf{m}$ and $\mathbf{A}_n$ using **Algorithm 2**.
2: Obtain the channel-adaptive semantic encoding $\mathbf{e}_n$ using **Algorithm 3**.
3: Modulate $\mathbf{e}_n$ into $\mathbf{c}_n$ and perform wireless transmission according to Eq. (8).
4: Perform task-oriented semantic decoding using **Algorithm 4** and obtain the sensing result $\mathbf{o}_n$.

Training Phase:

5: Obtain the trained parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}$ by jointly training all the modules according to **Algorithm 5**, using $\mathcal{D}$.

B. Multimodal Semantic Fusion Module

MSF integrates radar signal and visual image modalities through a carefully designed framework that combines signal processing, transformer-based image feature extraction, and cross-attention fusion, as shown in Fig. 3. The detailed description of key modules is as follows:

1) Signal Semantic Extractor: We first process radar signal $\mathbf{A}_n$, which is represented as complex-valued tensors of shape (B, K, L_{sig}), where B is the batch size, K is the number of receiving antennas used by the BS for sensing, and L_{sig} is the input length. Then, each complex-valued convolutional layer performs operations on both the real and imaginary components of the signal $\mathbf{A}_n$, denoted by $\mathbf{x}_n^{\text{real}}$ and $\mathbf{x}_n^{\text{imag}}$, respectively. This process can be expressed as:

$$\mathbf{z}_n^{\text{real}} = \mathbf{W}_{\text{real}} * \mathbf{x}_n^{\text{real}} - \mathbf{W}_{\text{imag}} * \mathbf{x}_n^{\text{imag}}, \tag{18}$$

$$\mathbf{z}_n^{\text{imag}} = \mathbf{W}_{\text{real}} * \mathbf{x}_n^{\text{imag}} + \mathbf{W}_{\text{imag}} * \mathbf{x}_n^{\text{real}}, \tag{19}$$

$$\mathbf{z}_n^{\text{out}} = \text{ReLU}(\mathbf{z}_n^{\text{real}} + j \cdot \mathbf{z}_n^{\text{imag}}), \tag{20}$$

where $*$ denotes convolution, $\mathbf{W}_{\text{real}}$ and $\mathbf{W}_{\text{imag}}$ are the real and imaginary parts of the convolutional kernel, and j represents the imaginary unit. Next, complex max-pooling is applied, reducing the sequence length while retaining the semantic features. After three convolutional layers, each followed by a pooling operation, the real and imaginary parts are concatenated along the channel dimension:

$$\mathbf{z}_{n,3} = \text{Concat}(\mathbf{z}_{n,3}^{\text{real}}, \mathbf{z}_{n,3}^{\text{imag}}). \tag{21}$$

Finally, the signal is sent through by a fully connected layer that maps the concatenated features to a reduced dimensionality:

$$\mathbf{s}_n^{\text{sig}} = \text{Linear}(\mathbf{z}_{n,3}), \mathbf{s}_n^{\text{sig}} \in \mathbb{R}^{B \times L_s \times d}, \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

where $\mathrm{Linear}(\cdot)$ represents a fully connected layer, L_s is the sequence length of the signal semantic and d is the feature dimensionality.

2) Vision Semantic Extractor: To ensure high extraction accuracy and inference velocity, we utilize a lightweight ViT network——bilateral transformer (BiFormer) [29], as the backbone to extract visual features from input images $\mathbf{m}$. The key advantage of BiFormer over traditional ViT lies in its bilevel routing attention (BRA) mechanism.

First, $\mathbf{m}$ is divided into $S \times S$ non-overlapping regions using a patch embedding layer, with each region containing HW/S^2 feature vectors. This process reshapes $\mathbf{m}$ into $\mathbf{m}^r \in \mathbb{R}^{S^2 \times HW/S^2 \times C}$. Linear projections are then applied to obtain query, key, and value tensors, denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{S^2 \times HW/S^2 \times C}$, respectively:

$$\mathbf{Q} = \mathbf{m}^r \mathbf{W}^\mathrm{q}, \quad \mathbf{K} = \mathbf{m}^r \mathbf{W}^\mathrm{k}, \quad \mathbf{V} = \mathbf{m}^r \mathbf{W}^\mathrm{v}, \tag{23}$$

where $\mathbf{W}^\mathrm{q}$, $\mathbf{W}^\mathrm{k}$, and $\mathbf{W}^\mathrm{v}$ are the projection weights for the query, key, and value, respectively.

Second, region-level queries and keys, $\mathbf{Q}^\mathrm{r}$ and $\mathbf{K}^\mathrm{r} \in \mathbb{R}^{S^2 \times C}$, are computed by averaging the query and key tensors over each region. Using these, we calculate the adjacency matrix $\mathbf{A}^\mathrm{r} \in \mathbb{R}^{S^2 \times S^2}$ to quantify semantic relationships between regions:

$$\mathbf{A}^\mathrm{r} = \mathbf{Q}^\mathrm{r} \left(\mathbf{K}^\mathrm{r}\right)^\top, \tag{24}$$

where $\top$ represents the transpose operation.

The adjacency matrix is then pruned by retaining the top-h semantic connections for each region, yielding the routing index matrix $\mathbf{I}^\mathrm{r} \in \mathbb{N}^{S^2 \times h}$:

$$\mathbf{I}^\mathrm{r} = \mathrm{toph}(\mathbf{A}^\mathrm{r}), \tag{25}$$

where $\mathrm{toph}(\cdot)$ is the row-wise top-h selection operator. Hence, the i-th row of $\mathbf{I}^\mathrm{r}$ contains the indices of the h most semantically relevant regions for the i-th region.

Next, with the region-to-region routing index matrix $\mathbf{I}^\mathrm{r}$, fine-grained token-to-token attention is performed by gathering the corresponding key and value tensors:

$$\mathbf{K}^\mathrm{g} = \mathrm{gather}\left(\mathbf{K}, \mathbf{I}^\mathrm{r}\right), \quad \mathbf{V}^\mathrm{g} = \mathrm{gather}\left(\mathbf{V}, \mathbf{I}^\mathrm{r}\right), \tag{26}$$

where $\mathbf{K}^\mathrm{g}$ and $\mathbf{V}^\mathrm{g}$ are the gathered key and value tensors. Attention is then applied to the gathered key-value pairs, producing the output tensor:

$$\mathbf{O} = \mathrm{Attention}\left(\mathbf{Q}, \mathbf{K}^\mathrm{g}, \mathbf{V}^\mathrm{g}\right) + \mathrm{LCE}(\mathbf{V}), \tag{27}$$

$$\mathrm{Attention}\left(\mathbf{Q}, \mathbf{K}^\mathrm{g}, \mathbf{V}^\mathrm{g}\right) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{g}\top}}{\sqrt{C}}\right)\mathbf{V}^\mathrm{g}, \tag{28}$$

where $\sqrt{C}$ is the scaling factor, and $\mathrm{LCE}(\mathbf{V})$ refers to a local context enhancement term [30].

Finally, a linear projection layer F_Proj is applied to the output tensor to obtain the vision semantic while ensuring it has the same shape with the signal semantic $\mathbf{s}_n^\mathrm{sig}$. This can be expressed as:

$$\mathbf{s}_n^\mathrm{vis} = F_\mathrm{Proj}(\mathbf{O}), \mathbf{s}_n^\mathrm{vis} \in \mathbb{R}^{B \times L_\mathrm{s} \times d}. \tag{29}$$

3) Cross-Attention Fusion Module: To achieve deep multimodal fusion, we integrate the signal semantics $\mathbf{s}_n^\mathrm{sig}$ and vision semantics $\mathbf{s}_n^\mathrm{vis}$ through a bidirectional cross-attention mechanism [31]. This mechanism enables each modality to dynamically attend to the other, thereby capturing complementary semantic cues across domains. In the MSF network, when $\mathbf{s}_n^\mathrm{sig}$ acts as the query, $\mathbf{s}_n^\mathrm{vis}$ serves as the key and value, and vice versa. The two cross-attention branches are computed as:

$$\mathbf{z}_n^\mathrm{vis} = \mathrm{Attention}(\mathbf{W}_q^1 \mathbf{s}_n^\mathrm{sig}, \mathbf{W}_k^2 \mathbf{s}_n^\mathrm{vis}, \mathbf{W}_v^2 \mathbf{s}_n^\mathrm{vis}), \tag{30}$$

$$\mathbf{z}_n^\mathrm{sig} = \mathrm{Attention}(\mathbf{W}_q^2 \mathbf{s}_n^\mathrm{vis}, \mathbf{W}_k^1 \mathbf{s}_n^\mathrm{sig}, \mathbf{W}_v^1 \mathbf{s}_n^\mathrm{sig}), \tag{31}$$

where $\mathbf{W}_q^1, \mathbf{W}_k^1, \mathbf{W}_v^1$ are the learnable linear projections for the radar signal modality, and $\mathbf{W}_q^2, \mathbf{W}_k^2, \mathbf{W}_v^2$ are for the vision modality. Notably, the two cross-attention directions serve different fusion purposes: In Eq. (30), the signal features query vision features, enabling the model to capture rich spatial and visual context to complement the signal representation. In Eq. (31), the vision features query signal features, allowing the model to enhance visual understanding with physical or geometric cues from the signal domain. Therefore, the bidirectional design allows mutual enhancement between the modalities, rather than a one-way dependency.

The attention outputs from both branches are fused as:

$$\mathbf{z}_n^\mathrm{fusion} = \mathbf{z}_n^\mathrm{vis} + \mathbf{z}_n^\mathrm{sig}. \tag{32}$$

To avoid vanishing gradients and to preserve modality-specific cues, we further apply residual connections and normalization:

$$\mathbf{s}_n^\mathrm{mul} = \mathrm{LayerNorm}(\mathbf{z}_n^\mathrm{fusion}) + \mathbf{s}_n^\mathrm{sig} + \mathbf{s}_n^\mathrm{vis}, \mathbf{s}_n^\mathrm{mul} \in \mathbb{R}^{B \times L_\mathrm{s} \times d}. \tag{33}$$

As a result, this fused representation $\mathbf{s}_n^\mathrm{mul}$ serves as a semantically aligned and contextually enriched embedding for downstream tasks. The inference process of MSF is summarized in **Algorithm 2**. Overall, the advantage of MSF is the combination of complex operations, ViT-based feature extraction, and attention-driven fusion, effectively fusing the latent information between radar signal processing and image. Importantly, compared to traditional concatenation- or addition-based fusion, cross-attention adaptively weighs the contribution of each modality based on learned semantic correlations, leading to more robust and discriminative semantic representations. This is particularly beneficial under partial observation or modality degradation (e.g., occlusion in vision or noise in radar), where complementary information can be leveraged to recover accurate perception.

Algorithm 2 Inference of MSF

Input: $\mathbf{m}$, $\mathbf{A}_n$.
Output: $\mathbf{s}_n^\mathrm{mul}$.
1: Extract signal semantic $\mathbf{s}_n^\mathrm{sig}$ using Eqs. (20)-(22).
2: Extract vision semantic $\mathbf{s}_n^\mathrm{vis}$ using Eqs. (23)-(29).
3: Obtain the fused multimodal semantic $\mathbf{s}_n^\mathrm{mul}$ via cross-attention using Eqs. (30)-(33).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

C. LLM-Based Semantic Encoder Module

LSE is designed to encode multimodal semantic information by leveraging the representational capabilities of LLMs [32]. Specifically, the encoder integrates textual descriptions of communication parameters with multimodal features to generate channel-adaptive semantic encoding $\mathbf{e}_n$.

First, the input to the LSE consists of two components: the multimodal semantic representation $\mathbf{s}_n^{\text{mul}}$ derived from the MSF network and textual communication parameters C_n. The textual input C_n undergoes tokenization using the GPT-2 tokenizer [33], producing input token IDs $\mathbf{C}_n^{\text{text}}$ and the corresponding attention mask $\mathbf{M}_n^{\text{text}}$. The end-of-sequence, "`eos_token`" symbol, replaces padding tokens to ensure compatibility with the pre-trained GPT-2 model [33]. The tokenized input is then mapped to an embedding space using the GPT-2 word embedding layer $\text{Embed}(\cdot)$, which can be expressed as:

$$\mathbf{E}_n = \text{Embed}(\mathbf{C}_n^{\text{text}}), \mathbf{E}_n \in \mathbb{R}^{B\times L_{\text{t}}\times d}, \tag{34}$$

where L_{t} is the length of the text embedding.

Second, the multimodal semantic representation $\mathbf{s}_n^{\text{mul}}$ and textual embeddings $\mathbf{E}_n$ are concatenated along the sequence dimension to form a fused input:

$$\mathbf{F}_n^{\text{input}} = \text{Concat}(\mathbf{s}_n^{\text{mul}}, \mathbf{E}_n), \mathbf{F}_n^{\text{input}} \in \mathbb{R}^{B\times L_{\text{fusion}}\times d}, \tag{35}$$

where $L_{\text{fusion}} = L_{\text{s}} + L_{\text{t}}$. Simultaneously, a fused attention mask is constructed:

$$\mathbf{F}_n^{\text{mask}} = \text{Concat}(\mathbf{M}_n^{\text{feat}}, \mathbf{M}_n^{\text{text}}), \mathbf{F}_n^{\text{mask}} \in \mathbb{R}^{B\times L_{\text{fusion}}}, \tag{36}$$

where $\mathbf{M}_n^{\text{feat}} \in \mathbb{R}^{B\times L_{\text{s}}}$ is initialized as a matrix of ones.

Third, the fused input $\mathbf{F}_n^{\text{input}}$ and the fused attention mask $\mathbf{F}_n^{\text{mask}}$ are passed into the pre-trained GPT-2 model, which generates contextually enriched representations from its final hidden state:

$$\mathbf{S}_n^{\text{fusion}} = \text{GPT2}(\mathbf{F}_n^{\text{input}}, \mathbf{F}_n^{\text{mask}}), \mathbf{S}_n^{\text{fusion}} \in \mathbb{R}^{B\times L_{\text{fusion}}\times d}. \tag{37}$$

To reduce the computational burden and extract the most salient features, a max-pooling operation is applied along the sequence dimension, reducing the dimension by half:

$$\mathbf{S}_n^{\text{pool}} = \text{MaxPool}(\mathbf{S}_n^{\text{fusion}}), \mathbf{S}_n^{\text{pool}} \in \mathbb{R}^{B\times L_{\text{fusion}}\times d/2}. \tag{38}$$

Finally, the pooled output is subsequently activated using a hyperbolic tangent function:

$$\mathbf{e}_n = \tanh(\mathbf{S}_n^{\text{pool}}), \mathbf{e}_n \in \mathbb{R}^{B\times L_{\text{fusion}}\times d/2}. \tag{39}$$

We summarize the inference process of LSE in **Algorithm 3**. Overall, this enriched semantic encoding $\mathbf{e}_n$ captures multimodal contextual dependencies and textual semantics, making it suitable for downstream tasks such as semantic reconstruction and communication. By integrating GPT-2 into the pipeline, the LSE effectively bridges the gap between textual and non-textual modalities, achieving a comprehensive representation of multimodal inputs.

Algorithm 3 Inference of LSE

Input: $\mathbf{s}_n^{\text{mul}}$, C_n.
Output: $\mathbf{e}_n$.
1: Tokenize C_n to obtain $\mathbf{C}_n^{\text{text}}$ and $\mathbf{M}_n^{\text{text}}$.
2: Map $\mathbf{C}_n^{\text{text}}$ to embedding space $\mathbf{E}_n$ using Eq. (34).
3: Concatenate $\mathbf{s}_n^{\text{mul}}$ and $\mathbf{E}_n$ to obtain $\mathbf{F}_n^{\text{input}}$ using Eq. (35).
4: Initiate $\mathbf{M}_n^{\text{feat}}$.
5: Construct fused attention mask $\mathbf{F}_n^{\text{mask}}$ according to Eq. (36).
6: Generate fused semantic representation $\mathbf{S}_n^{\text{fusion}}$ using GPT-2 according to Eq. (37).
7: Apply max-pooling to reduce the dimensions using Eq. (38).
8: Activate pooled representation to obtain $\mathbf{e}_n$ by Eq. (39).

D. Sensing Semantic Decoder Module

SSD, used as the JSCC decoder, integrates ViT-based and CNN-based modules to extract and decode semantic information and it is capable of reconstructing target images and estimating angles, velocities, and distances. The specific design of SSD is described as follows:

First, SSD reshapes the received semantic encoding $\hat{\mathbf{e}}_n$ into a spatial representation compatible with convolutional operations. To achieve this, the $\text{Conv2d}(\cdot)$ operation rearranges $\hat{\mathbf{e}}_n$ into a size of $(B \times W_d \times H_d \times d/2)$, where $H_{\text{d}} = W_{\text{d}} = \sqrt{L_{\text{fusion}}}$, assuming L_{fusion} is a perfect square. This process can be expressed as:

$$\mathbf{Z}_n^{\text{grid}} = \text{ReLU}(\text{Conv2d}(\mathbf{F}_n^{\text{grid}})), \mathbf{Z}_n^{\text{grid}} \in \mathbb{R}^{B\times C_d'\times H_d\times W_d}, \tag{40}$$

where C_d' is the channel dimensionality of the backbone's output. These enriched features are shared among four task-specific decoding heads, each optimized for a distinct semantic task of image reconstruction, distance prediction, angle estimation, and velocity estimation.

Second, for the task of image reconstruction, a ViT decoder [34] serves as the primary decoder. $\mathbf{Z}_n^{\text{grid}}$ undergoes up-sampling to double its spatial resolution, followed by the addition of a positional embedding, $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{H_d'\times W_d'\times D}$. This embedding retains spatial information during transformer-based processing:

$$\mathbf{Z}_n^{\text{embed}} = \text{Upsample}(\mathbf{Z}_n^{\text{grid}}) + \mathbf{E}_{\text{pos}}, \mathbf{Z}_n^{\text{embed}} \in \mathbb{R}^{B\times H_d'\times W_d'\times D}, \tag{41}$$

where $\text{Upsample}(\cdot)$ is the upsampling operation. The transformer blocks within the ViT decoder perform global feature aggregation, and the final prediction layer maps the transformed features back to pixel space. The reconstructed image is derived via:

$$\hat{\mathbf{m}}_n = \text{Unpatchify}(\text{Transformer}(\mathbf{Z}_n^{\text{embed}})), \tag{42}$$

where $\text{Transformer}(\cdot)$ represents multiple transformer blocks and $\text{Unpatchify}(\cdot)$ converts a patch to an image.

Third, the other sensing results, including angle, velocity, and distance, are extracted from the shared backbone features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

$\mathbf{Z}_n^{\text{grid}}$ through their specialized heads, respectively. This can be expressed as:

$$\hat{\theta}_n = \text{sigmoid}(\mathbf{W}_{\text{angle}} \cdot \text{Conv2d}(\mathbf{Z}_n^{\text{grid}}) + \mathbf{b}_{\text{angle}}), \quad (43)$$

$$\hat{v}_n = \text{sigmoid}(\mathbf{W}_{\text{rate}} \cdot \text{Conv2d}(\mathbf{Z}_n^{\text{grid}}) + \mathbf{b}_{\text{rate}}), \quad (44)$$

$$\hat{d}_n = \text{sigmoid}(\mathbf{W}_{\text{distance}} \cdot \text{Conv2d}(\mathbf{Z}_n^{\text{grid}}) + \mathbf{b}_{\text{distance}}), \quad (45)$$

where $\text{sigmoid}(\cdot)$ is the sigmoid activation function, $\mathbf{W}_{\text{angle}}$, $\mathbf{W}_{\text{rate}}$, and $\mathbf{W}_{\text{distance}}$ are the weights of the three task output heads. $\mathbf{b}_{\text{angle}}$, $\mathbf{b}_{\text{rate}}$, and $\mathbf{b}_{\text{distance}}$ are the bias.

Finally, we summarize the inference process of SSD in **Algorithm 4**. The SSD efficiently integrates ViT-based image reconstruction and auxiliary sensing tasks. By optimizing a unified multitask objective, SSD can provide diversified sensing results using the same semantic information. Moreover, users can select suitable output heads to deploy locally, according to their personalized requirements.

Algorithm 4 Inference of SSD

Input: $\hat{\mathbf{e}}_n$.
Output: $\hat{\mathbf{m}}_n, \hat{\theta}_n, \hat{v}_n, \hat{d}_n$.
1: Obtain the refined feature $\mathbf{Z}_n^{\text{gird}}$ using Eq. (40).
2: Reconstruct the image of the ST using Eqs. (41)-(42).
3: Predict the distance of the ST using Eq. (43).
4: Estimate the velocity of the ST using Eq. (44).
5: Predict the angle of the ST using Eq. (45).

E. Multi-Task Learning-Based Training Process

To ensure the SIMAC framework can achieve diversified sensing services, multi-task learning is used to jointly train the modules in this framework. Multi-task learning leverages a unified framework to address the simultaneous optimization of multiple objectives, targeting image reconstruction, angle estimation, and distance prediction. Specifically, the image reconstruction task is guided by the L1 loss, promoting pixel-level accuracy. Simultaneously, the angle, velocity, and distance prediction tasks are supervised using the mean squared error (MSE) loss, which minimizes deviations from the ground truth. The total multi-task loss is formulated as a weighted sum:

$$\mathcal{L}_{\text{MTL}} = l_1\mathcal{L}_{\text{sr}} + l_2\mathcal{L}_{\text{ap}} + l_3\mathcal{L}_{\text{vp}} + l_4\mathcal{L}_{\text{dp}}, \quad (46)$$

where each task loss is given in Eqs. (13)-(16).

During training, input data comprising the captured image $\mathbf{m}$ and echo signal $\mathbf{A}_n$. Moreover, dynamic channel characteristics are introduced based on the SNR and modulation schemes (e.g., BPSK, QPSK, 8PSK, 16QAM). These parameters, along with contextual information about the channel, are used to condition the model predictions. The forward pass of the model generates reconstructed images, estimated angles, and predicted distances. Losses for each task are computed and backpropagated to update the model parameters.

Assuming the training dataset is $\mathcal{D}$, the training process of the SIMAC framework is summarized in **Algorithm 5**.

Algorithm 5 Training process based on multi-task learning

Input: $\mathcal{D}$.
Output: $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}$.
1: **for** each training epoch **do**
2: **for** each batch $(\mathbf{m}_n, \mathbf{A}_n)$ from $\mathcal{D}$ **do**
3: Generate communication parameters C_n using dynamic SNR and modulation scheme.
4: Predict the sensing results $\hat{\mathbf{m}}_n, \hat{d}_n, \hat{v}_n, \hat{\theta}_n$ according to **Algorithms 2-4**.
5: Compute total loss using Eq. (46).
6: Backpropagate and update model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}$ with the optimizer.
7: **end for**
8: **end for**

V. Experimental Setup and Numerical Results

This section presents the simulation dataset, parameter configurations, and evaluation results. The simulations are conducted on a server equipped with an Intel Xeon CPU (2.3 GHz, 256 GB RAM) and two NVIDIA RTX 4090 GPUs (24 GB SGRAM each), leveraging the PyTorch framework to implement the proposed schemes.

A. Experimental Settings

1) Dataset Setup: Based on the VIRAT Video Dataset [12], we perform a series of operations to construct a specialized dataset for training and testing our proposed methods. The detailed procedure is as follows:

First, we select videos representing three specific scenes from [12] as the raw data. Each video is sampled at a velocity of one frame per second, resulting in approximately 10,000 RGB images (i.e., $\mathbf{m}$).

Second, for each extracted frame, we assume that all STs are cars. A YOLOv10 [35] model detects the bounding box coordinates of cars in the scenes. Based on these bounding box coordinates, we employ the segment anything model (SAM) [36] to isolate the car images from the raw frames, which serve as labels for image reconstruction (i.e., $\mathbf{m}_n$). This process produces approximately 800,000 images of STs.

Third, we assume the BS is positioned at the lower-right corner of each image. Accordingly, we calculate the distance of each ST to the BS as follows:

$$d_n = \sqrt{(x^{\text{BS}} - x_n^{\text{center}})^2 + (y^{\text{BS}} - y_n^{\text{center}})^2}, \quad (47)$$

where $(x^{\text{BS}}, y^{\text{BS}})$ represents the coordinates of the BS, and $(x_n^{\text{center}}, y_n^{\text{center}})$ denotes the center coordinates of the nth ST's bounding box.

Similarly, we estimate the angle of each ST relative to the BS as follows:

$$\theta_n = \arctan2(y^{\text{BS}} - y_n^{\text{center}}, x_n^{\text{center}} - x^{\text{BS}}), \quad (48)$$

where $\arctan2(\cdot)$ computes the arc tangent of the input coordinates, returning a radian value within the range $[-\pi, \pi]$.

Additionally, the tracking capability of the YOLO model is employed to track the same object across consecutive frames. The velocity of the object is calculated as the ratio

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

of the displacement of the center point of the object detection bounding box to the frame duration. This can be expressed as:

$$v_n = \frac{\|\mathbf{p}_{n,t+1} - \mathbf{p}_{n,t}\|}{\Delta t}, \tag{49}$$

where $\mathbf{p}_{n,t} = (x_{n,t}^{\text{center}}, y_{n,t}^{\text{center}})$ and $\mathbf{p}_{n,t+1} = (x_{n,t+1}^{\text{center}}, y_{n,t+1}^{\text{center}})$ represent the center points of the nth ST at frame t and $t+1$, respectively, and Δt is the time interval between the two frames.

Finally, we generate the echo signal $\mathbf{A}_n$ for each ST according to Eq. (1).

2) Parameter Settings: Considering that the task of image reconstruction is more difficult than the other tasks, the adjustment factors of task losses are set to $l_1 = 100$, $l_2 = 1$, $l_3 = 1$, and $l_4 = 1$, respectively. The bandwidth is set to $B = 1$ kHz, the power is set to $P = 1$ W, and the SNR is varied between 0 dB and 25 dB. When the SNR is low, the transmitted symbols are easy to be impacted. Hence, to achieve more accurate transmission, we apply the following modulation scheme:

$$\text{Modulator} = \begin{cases} \text{BPSK}, & 0 \text{ dB} \leq \text{SNR} \leq 10 \text{ dB}, \\ \text{QPSK}, & 10 \text{ dB} < \text{SNR} \leq 18 \text{ dB}, \\ \text{8PSK}, & 18 \text{ dB} < \text{SNR} \leq 22 \text{ dB}, \\ \text{16QAM}, & \text{otherwise}. \end{cases} \tag{50}$$

The radar's central frequency is $f_c = 10\,\text{GHz}$. The PRI is configured as $T_r = 1 \times 10^{-4}\,\text{s}$, with a sampling frequency of $F_s = 60\,\text{MHz}$, corresponding to a sampling interval of $\Delta t = \frac{1}{F_s} = 1.67\,\text{ns}$. Since only cars are considered as the ST, all RCSs of the STs are set to $\rho_n = 100$. We assume the SIMO radar model utilizes $K = 10$ antennas to transmit the LFM waveform and capture the echo signal.

The AWGN and Rayleigh channels are considered, and the SNR is randomly chosen for each forward pass to enhance the robustness of the SIMAC framework to channel noise. In the inference phase, we evaluate the SC model under fixed SNR conditions at $[5, 10, 15, 20, 25]$ dBs.

3) Evaluation Metrics: We use the root mean squared error (RMSE) to evaluate the performance of the proposed method in distance, velocity, and angle prediction tasks. RMSE quantifies the absolute average deviation between predictions and ground truth. These metrics are defined as follows:

$$\text{RMSE}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \sqrt{\frac{1}{I}\sum_{i=1}^{I} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2}, \tag{51}$$

where $\hat{\mathbf{x}}_i \in (0, 1)$ denotes the predicted value, $\mathbf{x}_i$ represents the ground truth, and I is the total number of samples. Particularly, $\mathbf{x}_i$ is the normalized value of speed, distance, or angle.

To evaluate image reconstruction performance, we adopt peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) as metrics. PSNR measures the quality of reconstructed images and is expressed in decibels (dB), with higher values indicating better quality:

$$\text{PSNR}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2}\right), \tag{52}$$

where MAX_I is the maximum possible pixel value, typically 255 for 8-bit images. Similarly, SSIM is a metric that gauges the perceived similarity between two images, factoring in three key components - luminance, contrast, and structure. The definition of SSIM is outlined as follows:

$$\text{SSIM}(\mathbf{x}_i, \hat{\mathbf{x}}_i) = \frac{(2\varphi_{\mathbf{x}_i}\varphi_{\hat{\mathbf{x}}_i} + c_1)(2\phi_{\mathbf{x}_i\hat{\mathbf{x}}_i} + c_2)}{(\varphi_{\mathbf{x}_i}^2 + \varphi_{\hat{\mathbf{x}}_i}^2 + c_1)(\phi_{\mathbf{x}_i}^2 + \phi_{\hat{\mathbf{x}}_i}^2 + c_2)}, \tag{53}$$

where $\varphi_{\mathbf{x}_i}$ and $\varphi_{\hat{\mathbf{x}}_i}$ are mean values, $\phi_{\mathbf{x}_i}^2$ and $\phi_{\hat{\mathbf{x}}_i}^2$ are variances, $\phi_{\mathbf{x}_i\hat{\mathbf{x}}_i}$ is their covariance, and c_1 and c_2 are two small constants to prevent division by zero.

B. Visualization of Sensing Results

As illustrated in Fig. 4, the SIMAC framework processes the raw image $\mathbf{m}$, the radar signal $\mathbf{A}_n$, and the sensing outputs, including the reconstructed image $\hat{\mathbf{m}}_n$ and the predicted motion attributes. A distinctive feature of the SIMAC framework is its capability to produce diversified sensing outputs for the same image by integrating various radar signals. Since the training data is derived from simulations, there may be discrepancies between the predicted results and real-world scenes. Fig. 5 presents a representative comparison between raw sensing images and their reconstructed counterparts. The reconstructed images demonstrate significantly enhanced visual quality, as evidenced by the SNR gains over the raw inputs. Remarkably, the model is even able to recover missing regions in the original images, highlighting the strong generative capacity of the ViT-based decoder.

Overall, the proposed MSF module enables the SIMAC framework to leverage the semantic information of radar signals, which serve as queries for localizing corresponding spatiotemporal features during training. These findings demonstrate the framework's capacity to fuse multimodal information effectively for target localization and motion attribute estimation.

C. Evaluation for Multimodal Sensing

1) Benchmark Schemes: To evaluate the advantages of the multimodal sensing in the proposed SIMAC framework compared to the unimodal sensing, we consider the following benchmark schemes:

- CV-based Sensing (CV-S): This variant only uses the image modality to perform sensing.
- Radar Signal-based Sensing (RS-S): This variant only uses the radar signal to perform sensing.

2) Evaluation Results: As illustrated in Figs. 6–8, SIMAC consistently outperforms both baselines across all three prediction tasks, distance, angle, and velocity, under AWGN and Rayleigh channels. Specifically, in distance prediction under the Rayleigh channel and Scene 1, SIMAC achieves an RMSE of 0.009, compared to 0.054 for CV-S and 0.020 for RS-S, corresponding to performance gains of 83.3% and 55.0%, respectively. In angle prediction, SIMAC maintains a near-constant RMSE of around 0.010 across all SNR levels and scenes, offering over 85% reduction compared to CV-S and RS-S in highly noisy conditions (Fig. 7(b), Rayleigh, Scene 3). Similarly, for velocity prediction in Scene 2 under AWGN,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

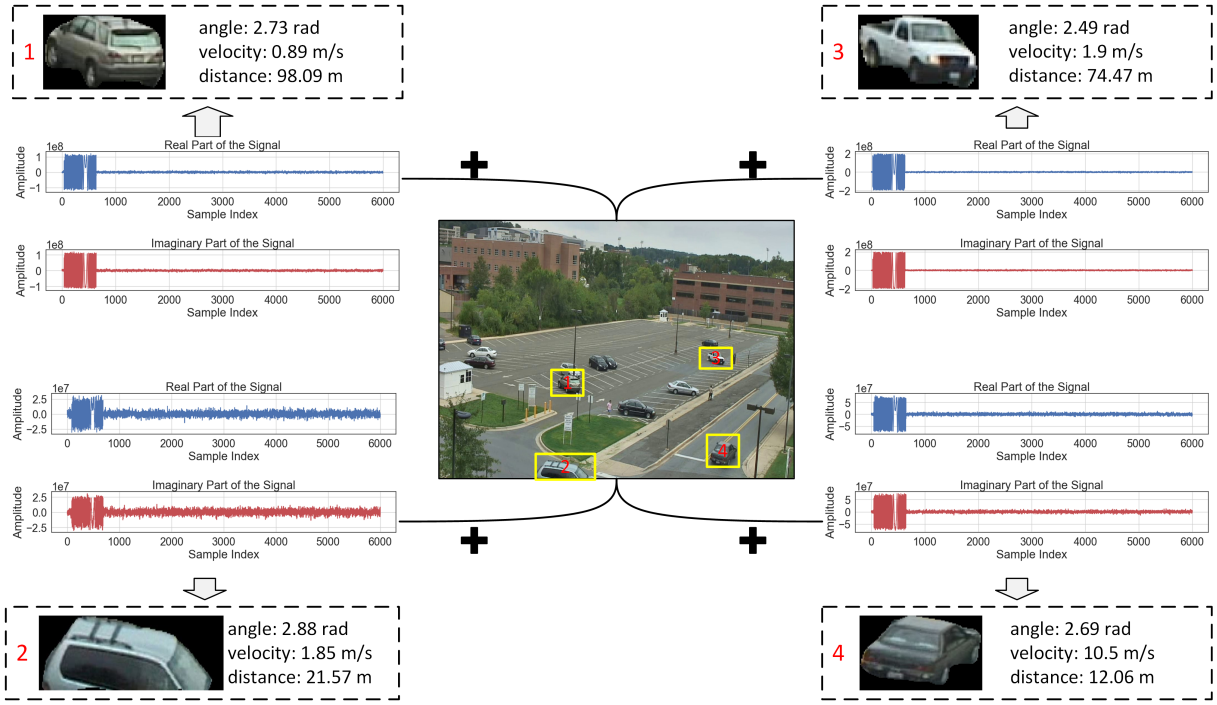


Fig. 4: Visualization of the SIMAC framework's running process.

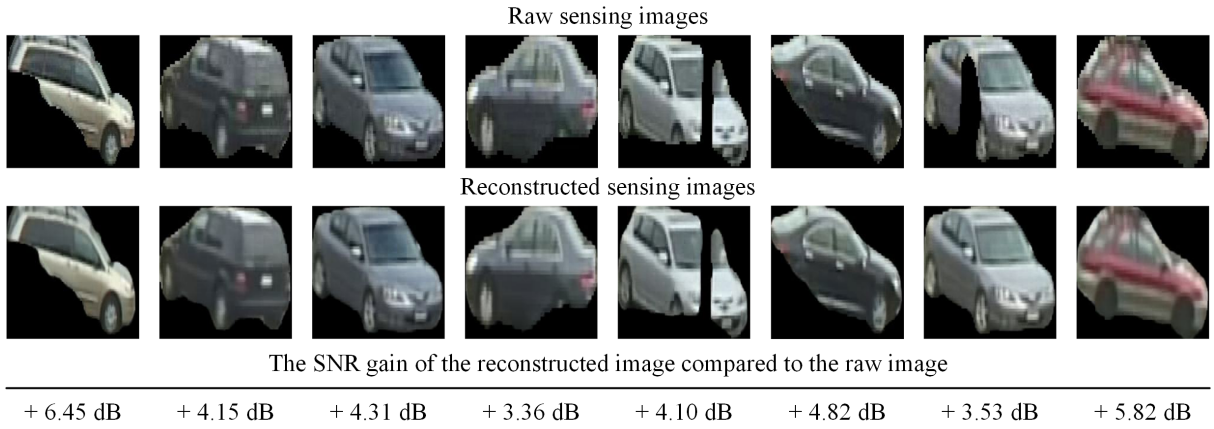


Fig. 5: Comparison results of raw sensing images and reconstructed images.

SIMAC yields an RMSE of 0.0038, improving by 55.8% over CV-S and 25.5% over RS-S.

Overall, these results validate the effectiveness of the proposed multimodal sensing framework. By integrating complementary features from both visual and radar modalities, SIMAC achieves superior distance prediction accuracy and enhanced robustness across diverse environments and channel conditions.

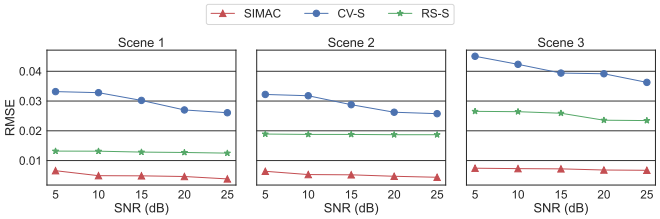
D. Evaluation for LLM

1) Benchmark Schemes: To evaluate the advantages of the GPT-2 in the proposed SIMAC framework compared to the traditional models, we consider the following benchmark schemes:

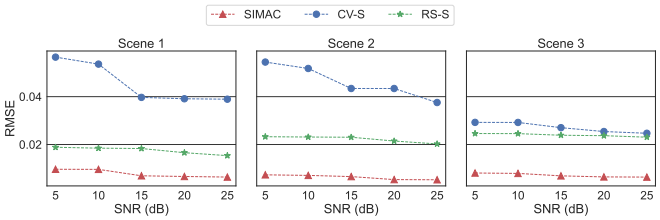
- SIMAC (with LSTM): This variant uses the LSTM [37] to replace the GPT-2 as the semantic encoder.
- SIMAC (with GRU): This variant uses the GRU [38] to replace the GPT-2 as the semantic encoder.

2) Evaluation Results: As demonstrated in Figs. 9–13, the GPT-2-based SIMAC consistently outperforms its RNN-based counterparts, including LSTM and GRU, across all SNR levels and evaluation metrics. Specifically, under the AWGN channel at an SNR of 25 dB (Figs. 9–11(a)), it achieves substantial reductions in RMSE for angle, distance, and velocity estimation, up to 40.2%, 12.5%, and 18.1% compared to LSTM, and as high as 90%, 102%, and 80% relative to GRU, respectively. For image reconstruction (Figs. 12–13(a)), GPT-2 provides significant improvements, with PSNR gains of 1.2 dB over LSTM and 3.1 dB over GRU, while also enhancing SSIM by 3.2% and 11.5%, achieving values exceeding 0.85. Under the more challenging Rayleigh fading conditions (Figs. 10-13(b)), GPT-2 demonstrates superior robustness, maintaining

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

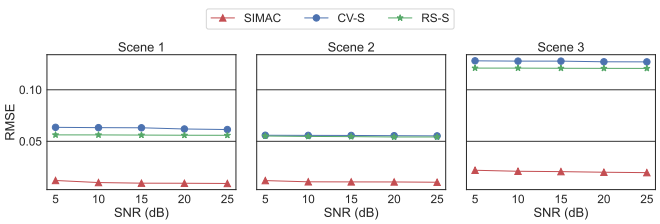


(a) AWGN channel

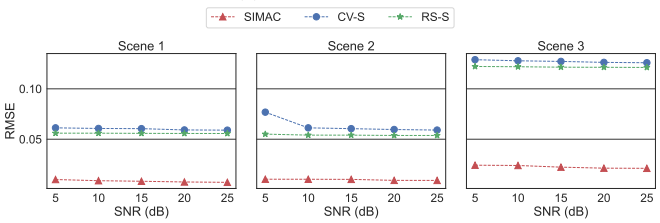


(b) Rayleigh channel

Fig. 6: Distance prediction comparisons across different schemes under different channels.

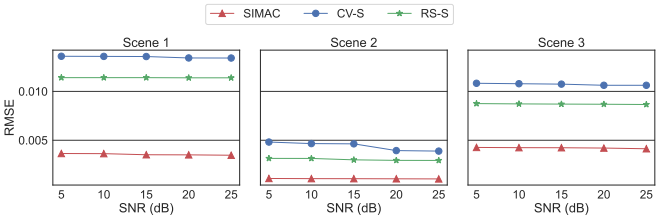


(a) AWGN channel

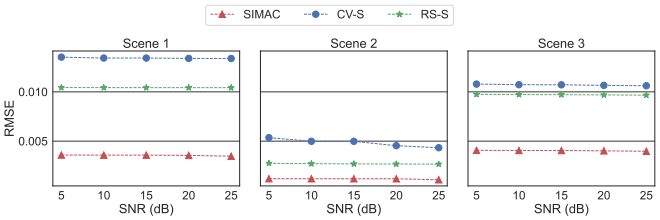


(b) Rayleigh channel

Fig. 7: Angle prediction comparisons across different schemes under different channels.

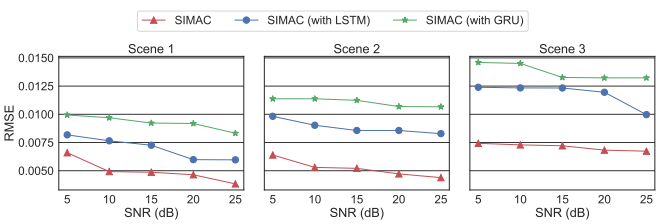


(a) AWGN channel

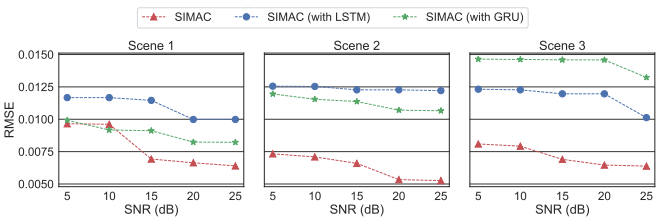


(b) Rayleigh channel

Fig. 8: Velocity prediction comparisons across different schemes under different channels.



(a) AWGN channel



(b) Rayleigh channel

Fig. 9: Distance prediction comparisons across different schemes under different channels.

30–40% lower RMSE in angle and velocity estimation at high SNRs and achieving up to 1.2 dB PSNR gains and over 20% improvements in SSIM compared to LSTM across all evaluated scenarios.

These results underscore GPT-2's strong capability in modeling semantic dependencies within multimodal sensing data. Its transformer-based architecture effectively captures long-range contextual information, thereby enhancing both accuracy and resilience under diverse channel conditions. Consequently, integrating GPT-2 as the semantic encoder significantly improves the overall performance of the SIMAC framework, affirming the potential of large language models in advancing semantic communications and multimodal joint sensing.

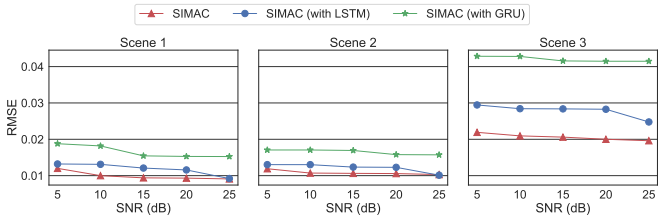
E. Evaluation for Ablation

1) Benchmark Schemes: To evaluate the effectiveness of each module in the proposed SIMAC framework, we consider the following benchmark schemes:

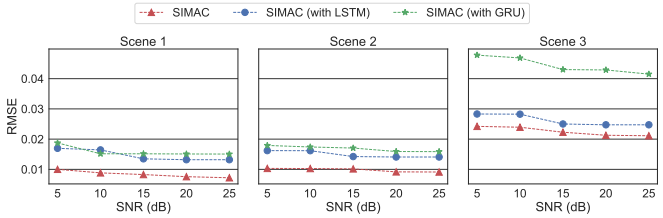
- SIMAC (w/o LSE): This variant excludes communication parameters during training and inference.
- SIMAC (w/o SSD): In this variant, the SIMAC framework operates without multiple output heads, and each sensing task is trained independently.

2) Evaluation Results: Figs. 14–18 present comprehensive comparisons under both AWGN and Rayleigh channels across three distinct scenes. The full SIMAC consistently outperforms its counterparts in all metrics. For instance, in distance prediction under the Rayleigh channel and Scene 2 (Fig. 14(b)), SIMAC achieves an RMSE of 0.005, compared to 0.009 for SIMAC (w/o LSE) and 0.007 for SIMAC (w/o SSD), reflecting a 44.4% and 57.1% reduction in error, respectively. The benefit of the LSE module is particularly evident under low SNR conditions, where the semantic representation enhances robustness against noise. For example, in angle prediction (Fig.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

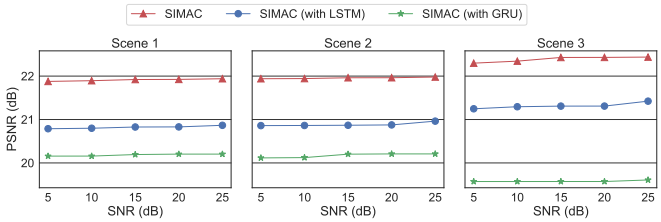


(a) AWGN channel

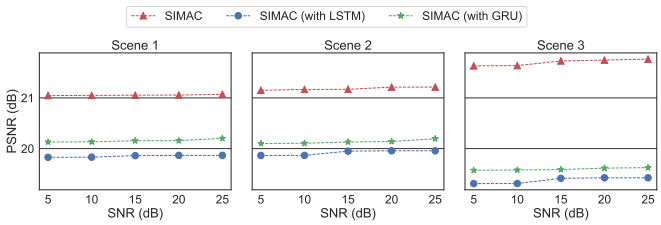


(b) Rayleigh channel

Fig. 10: Angle prediction comparisons across different schemes under different channels.

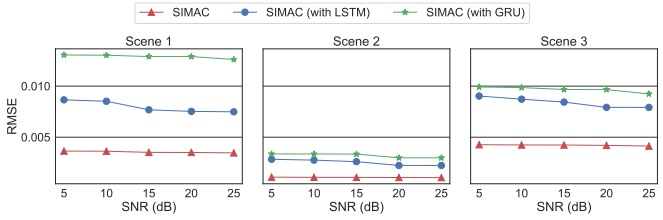


(a) AWGN channel

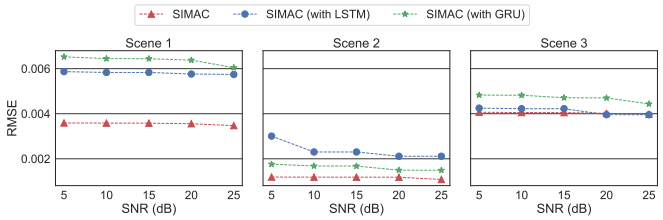


(b) Rayleigh channel

Fig. 12: PSNR comparisons across different schemes under different channels.

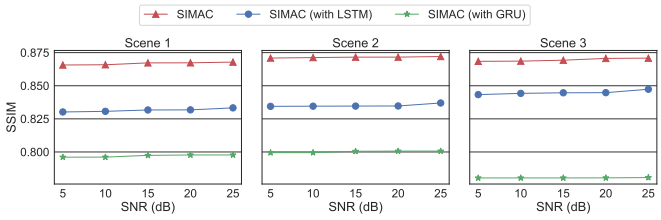


(a) AWGN channel

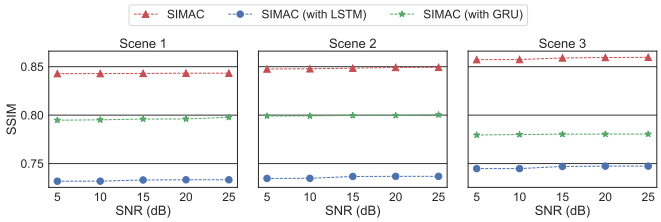


(b) Rayleigh channel

Fig. 11: Velocity prediction comparisons across different schemes under different channels.



(a) AWGN channel



(b) Rayleigh channel

Fig. 13: SSIM comparisons across different schemes under different channels.

15(a), AWGN, Scene 1), SIMAC reduces RMSE to 0.038, while SIMAC (w/o LSE) and SIMAC (w/o SSD) remain above 0.042 and 0.006, respectively, highlighting the LSE's critical role in extracting context-aware features. Meanwhile, the SSD module significantly improves multi-task synergy, especially in velocity prediction, where tasks are inherently coupled. In Fig. 16(b) (Rayleigh, Scene 1), SIMAC achieves an RMSE of 0.0024, compared to 0.0036 (w/o LSE) and 0.006 (w/o SSD). In terms of image reconstruction quality, SIMAC also demonstrates PSNR and SSIM improvements, respectively, across all scenes (Figs. 17 and 18).

These results validate the indispensability of both the LSE and SSD modules. Specifically, the LSE module enhances robustness across varying channel conditions, while the SSD module facilitates more effective feature learning. The synergy between the two modules enables SIMAC to generalize effectively across heterogeneous environments while maintaining superior sensing accuracy and perceptual fidelity.

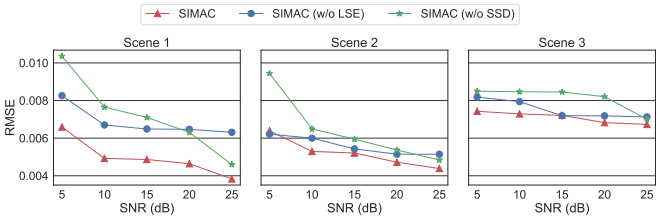
F. Evaluation for Complexity

TABLE II: Parameters, Complexity, and Inference Delay of the SIAMC Framework

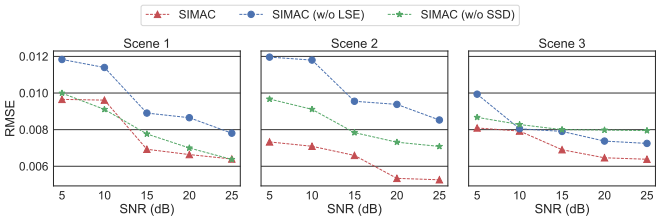
	MSF	LSE	SSD	SIMAC
Parameters	17.5M	124.4M	34.3M	176.3M
Complexity (FLOPs)	6.1G	8.3G	5.3G	19.8G
Inference delay (ms)	0.68	0.44	0.35	1.5

Table II summarizes the model size, computational complexity, and inference delay of SIMAC and its key components. The reported inference delay and computational complexity represent the average values per sample processed by SIMAC.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications

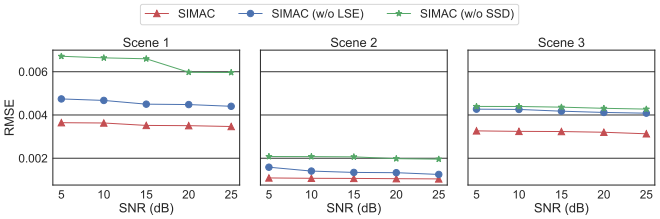


(a) AWGN channel

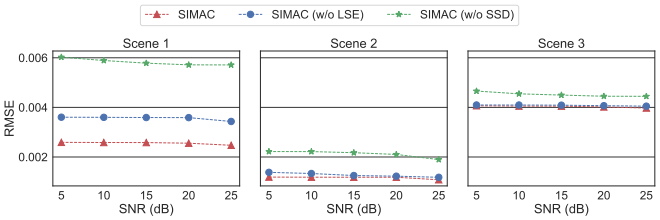


(b) Rayleigh channel

Fig. 14: Distance prediction comparisons across different schemes under different channels.

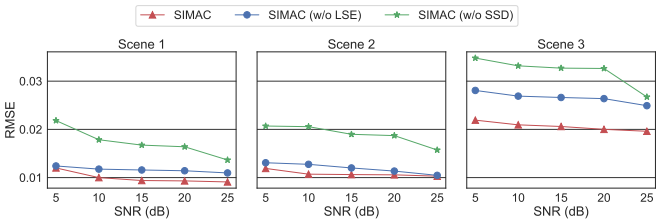


(a) AWGN channel

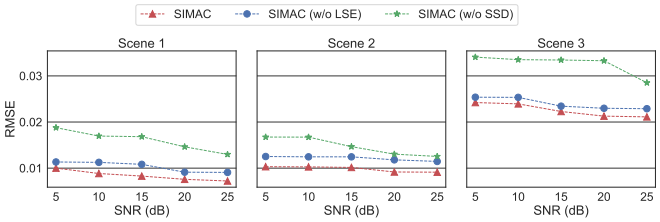


(b) Rayleigh channel

Fig. 16: Velocity prediction comparisons across different schemes under different channels.

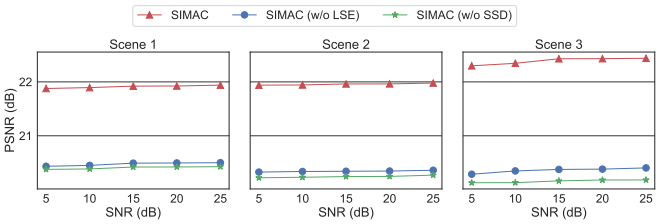


(a) AWGN channel

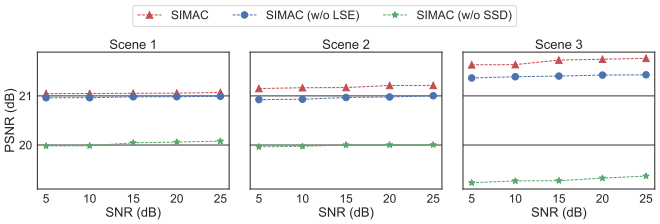


(b) Rayleigh channel

Fig. 15: Angle prediction comparisons across different schemes under different channels.



(a) AWGN channel



(b) Rayleigh channel

Fig. 17: PSNR comparisons across different schemes under different channels.

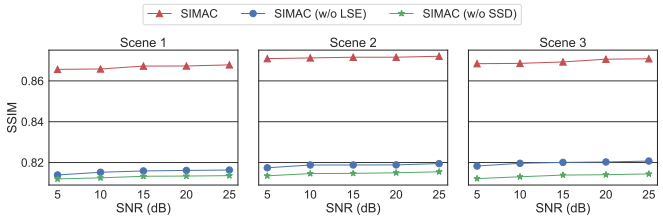
Notably, SIMAC achieves an inference delay of only 1.5 ms, which is sufficiently low to meet the demands of real-time applications. These results indicate that SIMAC imposes no significant computational burden on the communication system, striking a favorable balance between intelligent processing capabilities and resource efficiency. Consequently, SIMAC shows strong potential for latency-sensitive scenarios, such as autonomous driving, industrial automation, and UAV-based sensing.

G. Discussion

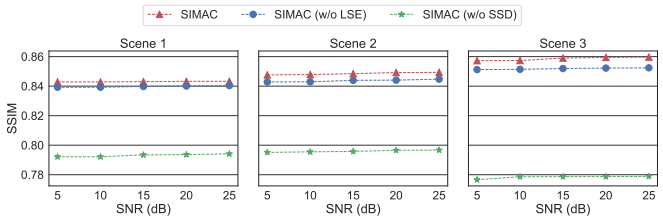
While the proposed SIMAC framework demonstrates promising performance in multimodal sensing and SC, several limitations remain, offering potential directions for future research:

1) The current design focuses on the fusion of only two sensing modalities, vision and radar signal. As a fact that real-world environments often involve richer and more diverse modalities (e.g., LiDAR, thermal, audio). Effectively extending SIMAC to handle more than two modalities poses challenges in terms of semantic space unification, modality weighting, and training efficiency.
2) The framework assumes that all predefined modalities are consistently available and equally necessary for all tasks. However, in practical scenarios, certain modalities may be redundant, unreliable, or irrelevant depending on the specific environment or task requirements. Therefore, it is crucial to explore adaptive modality selection and fusion strategies that can dynamically activate the most informative modalities while suppressing others, based on semantic relevance and system constraints.
3) The SIMAC framework currently operates in a single-device setting, where all sensing data is collected and fused locally. This does not reflect the distributed nature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JSAC.2025.3610398, IEEE Journal on Selected Areas in Communications



(a) AWGN channel



(b) Rayleigh channel

Fig. 18: SSIM comparisons across different schemes under different channels.

of many real-world systems, such as vehicle-to-vehicle perception or multi-camera surveillance networks. A promising extension would be to enable collaborative ISAC across multiple devices, where distributed agents share partial semantic representations to build a global understanding of the environment.

VI. Conclusions and Future Works

To address the challenges of limited accuracy and restricted capabilities in unimodal sensing, high communication overhead in decoupled sensing-communication systems, and the inability of single-task sensing to meet diversified user demands, we propose the SIMAC framework. This framework integrates multimodal sensing with SC to enable low-cost and high-accuracy sensing services. Specifically, the framework first employs the MSF network to extract and fuse semantic information from radar signals and images using cross-attention mechanisms, generating comprehensive multimodal representations. Then, it incorporates the LSE that maps communication parameters and multimodal semantics into a unified latent space, enabling channel-adaptive semantic encoding. Furthermore, it introduces the SSD to feature task-specific decoding heads and a multi-task learning strategy to deliver diversified sensing services. We conducted experimental simulations across four sensing tasks and benchmarks, demonstrating that the SIMAC framework substantially enhances sensing accuracy and service diversity.

In future work, as discussed in Section V-G, we aim to enhance the scalability and adaptability of SIMAC in the following directions. First, we will explore the integration of additional sensing modalities beyond radar and vision (e.g., LiDAR, cloud-point), and design efficient semantic fusion strategies that scale with modality number. Second, we plan to investigate adaptive modality selection mechanisms, enabling the system to dynamically determine which modalities are most relevant under specific environmental or task conditions, thereby reducing redundancy and improving efficiency. Third, we will extend SIMAC to multi-device cooperative sensing scenarios, such as multi-BS or vehicle-to-vehicle deployments.

References

[1] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 98–117, 2020.

[2] L. Aravinda, K. Sneha, and N. Suchitha, "IoT military security system for tracing of missile using ultrasonic radar," *Journal of Science & Technology (JST)*, vol. 9, no. 1, pp. 155–160, 2024.

[3] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li *et al.*, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2022.

[4] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

[5] S. Wang, L. Mei, R. Liu, W. Jiang, Z. Yin, X. Deng, and T. He, "Multi-modal fusion sensing: A comprehensive review of millimeter-wave radar and its integration with other modalities," *IEEE Communications Surveys & Tutorials*, 2024.

[6] H. Liu and Z. Liu, "A multimodal dynamic hand gesture recognition based on radar–vision fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[7] R. Zhang, L. Cheng, S. Wang, Y. Lou, Y. Gao, W. Wu, and D. W. K. Ng, "Integrated sensing and communication with massive MIMO: A unified tensor approach for channel and target parameter estimation," *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 8571–8587, 2024.

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

[9] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, "Large AI model empowered multimodal semantic communications," *IEEE Communications Magazine*, vol. 63, no. 1, pp. 76–82, 2025.

[10] L. Dong, Y. Peng, F. Jiang, K. Wang, and K. Yang, "Explainable semantic federated learning enabled industrial edge network for fire surveillance," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 12, pp. 14 053–14 061, 2024.

[11] C. Luo, J. Hu, L. Xiang, K. Yang, and B. Lei, "Optimizing placement and power allocation in reconfigurable intelligent sensing surfaces for enhanced sensing and communication performance," *IEEE Communications Letters*, vol. 29, no. 1, pp. 11–15, 2025.

[12] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, 2011, pp. 3153–3160.

[13] S. Sun and Y. D. Zhang, "4D automotive radar sensing for autonomous vehicles: A sparsity-oriented approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 879–891, 2021.

[14] M. Sohail, A. U. Khan, M. Sandhu, I. A. Shoukat, M. Jafri, and H. Shin, "Radar sensor based machine learning approach for precise vehicle position estimation," *Scientific Reports*, vol. 13, no. 1, p. 13837, 2023.

[15] K. Luo, X. Kong, J. Zhang, J. Hu, J. Li, and H. Tang, "Computer vision-based bridge inspection and monitoring: A review," *Sensors*, vol. 23, no. 18, p. 7863, 2023.

[16] H. Liu and Z. Liu, "A multimodal dynamic hand gesture recognition based on radar–vision fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[17] A. Deliali, F. Tainter, C. Ai, and E. Christofa, "A framework for mode classification in multimodal environments using radar-based sensors," *Journal of Intelligent Transportation Systems*, vol. 27, no. 4, pp. 441–458, 2023.

[18] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "Craft: Camera-radar 3D object detection with spatio-contextual fusion transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 1160–1168, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25198

[19] B. Zhang, K. Yang, and K. Wang, "Performance analysis of IRS-assisted and wireless power transfer enabled isac systems," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia*, 2023, pp. 1012–1017.

[20] Y. Wang, K. Zu, L. Xiang, Q. Zhang, Z. Feng, J. Hu, and K. Yang, "ISAC enabled cooperative detection for cellular-connected UAV network," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
[21] L. Xiang, K. Xu, J. Hu, and K. Yang, "Green beamforming design for integrated sensing and communication systems: A practical approach using beam-matching error metrics," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 5935–5940, 2024.
[22] H. Xu, X. Zhang, J. He, Y. Yu, and Y. Cheng, "Real-time volumetric perception for unmanned surface vehicles through fusion of radar and camera," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.
[23] S. Jiang, A. Alkhateeb, D. W. Bliss, and Y. Rong, "Vision-guided mimo radar beamforming for enhanced vital signs detection in crowds," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 4, pp. 4640–4649, 2024.
[24] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive mimo channel prediction," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 1885–1898, 2021.
[25] K. Tan, J. Wu, H. Zhou, Y. Wang, and J. Chen, "Integrating advanced computer vision and AI algorithms for autonomous driving systems," *Journal of Theory and Practice of Engineering Science*, vol. 4, no. 01, pp. 41–48, 2024.
[26] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large generative model assisted 3D semantic communication," *arXiv preprint arXiv:2403.05783*, 2024.
[27] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Computing Surveys (Csur)*, vol. 51, no. 4, pp. 1–35, 2018.
[28] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.
[29] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. W. Lau, "Biformer: Vision transformer with bi-level routing attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 10 323–10 333.
[30] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 853–10 862.
[31] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.
[32] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6G communications," *IEEE Wireless Communications*, pp. 1–8, 2024.
[33] A. P. B. Veyseh, V. Lai, F. Dernoncourt, and T. H. Nguyen, "Unleash GPT-2 power for event detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6271–6282.
[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
[35] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.
[36] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.
[37] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.
[38] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.