

Embodied AI-Enhanced Vehicular Networks: An Integrated Vision Language Models and Reinforcement Learning Method

Ruichen Zhang¹, Member, IEEE, Changyuan Zhao², Hongyang Du³, Member, IEEE, Dusit Niyato⁴, Fellow, IEEE, Jiacheng Wang⁵, Member, IEEE, Suttinee Sawadsitang⁶, Xuemin Shen⁷, Fellow, IEEE, and Dong In Kim⁸, Life Fellow, IEEE

Abstract—This paper investigates adaptive transmission strategies in embodied AI-enhanced vehicular networks by integrating vision language models (VLMs) for semantic information extraction and deep reinforcement learning (DRL) for decision-making. The proposed framework aims to optimize both data transmission efficiency and decision accuracy by formulating an optimization problem that incorporates the Weber-Fechner law, serving as a metric for balancing bandwidth utilization and quality of experience (QoE). Specifically, we employ the large language and vision assistant (LLAVA) model to extract critical semantic information from raw image data captured by embodied AI agents (i.e., vehicles), reducing transmission data size by approximately more than 90% while retaining essential content for vehicular communication and decision-making. In the dynamic vehicular environment, we employ a generalized advantage estimation-based proximal policy optimization (GAE-PPO) method to stabilize decision-making under uncertainty. Simulation results show that attention maps from

LLAVA highlight the model's focus on relevant image regions, enhancing semantic representation accuracy. Additionally, our proposed transmission strategy improves QoE by up to 36% compared to DDPG and accelerates convergence by reducing required steps by up to 47% compared to pure PPO. Further analysis indicates that adapting semantic symbol length provides an effective trade-off between transmission quality and bandwidth, achieving up to a 61.4% improvement in QoE when scaling from 4 to 8 vehicles.

Index Terms—Embodied AI, vehicular networks, PPO, LLM, VLM, LLAVA, QoE.

I. INTRODUCTION

WITH the advent of the 6G era, the Internet of Vehicles (IoV) is expected to achieve unprecedented advancements, with traffic densities exceeding 0.1–10 Gbps/m² and connection densities reaching up to 10 million devices/km² [1]. These improvements will significantly enhance data rates, connectivity, and network capacity, fundamentally transforming IoV services such as real-time navigation, environmental perception, and autonomous decision-making [2]. As a key component of IoV, the vehicular network plays a pivotal role in enabling communication between vehicles and infrastructure, facilitating these advanced services. With the growing number of connected vehicles, the demand for these services is increasing, necessitating the deployment of numerous sensors in vehicles and along roadsides to collect and process large amounts of real-time data [3].

Artificial intelligence (AI) has emerged as a significant technology for enabling vehicles to process this data autonomously and efficiently. Traditional discriminative AI models have been applied to tasks such as object detection [4] and path planning [5] in vehicular networks. However, as vehicular networks grow more complex, these models struggle to maintain high performance under dynamic conditions and when managing multiple tasks simultaneously [6]. As a result, robust AI methods are required to address the increasing complexity of these environments.

Embodied AI has been proposed as a promising solution, which emphasizes the interaction between intelligent systems and their physical surroundings [7]. In vehicular networks, embodied AI systems embedded within vehicles enable real-time adaptation to dynamic environments [8]. Equipped with

Received 24 December 2024; revised 14 May 2025; accepted 16 June 2025. Date of publication 24 June 2025; date of current version 3 October 2025. This work was supported in part by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme under Grant FCP-NTU-RG-2022-010 and Grant FCP-ASTAR-TG-2022-003, in part by the Singapore Ministry of Education (MOE) Tier 1 under Grant RG87/22 and Grant RG24/24, in part by the NTU Centre for Computational Technologies in Finance (NTU-CCTF), in part by the RIE2025 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) under Grant I2301E0026, administered by A*STAR, in part by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), in part by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program under Grant IITP-2020-0-01821 and the ITRC support program under Grant IITP-2023-RS-2023-00258639 supervised by the IITP (Institute for ICT Planning & Evaluation), and in part by Chiang Mai University. Recommended for acceptance by G. Zhu. (Corresponding author: Suttinee Sawadsitang.)

Ruichen Zhang, Changyuan Zhao, Dusit Niyato, and Jiacheng Wang are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: ruichen.zhang@ntu.edu.sg; zhao0441@e.ntu.edu.sg; dnyato@ntu.edu.sg; jiacheng.wang@ntu.edu.sg).

Hongyang Du is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Pok Fu Lam Hong Kong (e-mail: duhy@eee.hku.hk).

Suttinee Sawadsitang is with the College of Arts, Media and Technology, Chiang Mai University, Suthep 50200, Thailand (e-mail: suttinee.s@cmu.ac.th).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Dong In Kim is with the Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea (e-mail: don-gin@skku.edu).

Digital Object Identifier 10.1109/TMC.2025.3582864

advanced sensors and actuators, these vehicles can perceive their surroundings, make intelligent decisions, and autonomously perform tasks such as obstacle avoidance, traffic management, and collaborative driving [9]. In particular, embodied AI-enhanced vehicular networks rely on two core components, i.e., efficient data processing and adaptive decision-making.

- For *data processing*, large language models (LLMs) can be employed to extract critical information via their advanced understanding and generation capabilities [10]. Specifically, multimodal LLMs analyze multimodal data such as images or videos captured by vehicles, converting them into text-based information that can be more efficiently transmitted across the network [11]. This method leverages multimodal technologies to reduce the amount of raw data that needs to be sent, thereby optimizing network efficiency and bandwidth usage. For example, in [12], an LLM-based data transmission framework was proposed to compress large volumes of sensory data in autonomous driving systems, significantly reducing transmission bandwidth while preserving the quality of information.
- For *decision-making*, deep reinforcement learning (DRL) can be utilized to develop adaptive strategies in dynamic environments by continuously adjusting actions based on real-time feedback [13]. Specifically, DRL trains agents to make optimal decisions by interacting with their environment and receiving feedback in the form of rewards or penalties. As part of an embodied AI system, this enables vehicles to adjust driving behaviors for tasks such as obstacle avoidance and cooperative driving, ensuring responsiveness and improved performance in complex scenarios [14]. For example, in [15], a DRL-based semantic-aware spectrum-sharing method was proposed to enhance vehicular decision-making, optimizing communication efficiency and safety outcomes in high-speed environments.

In addition to their individual strengths, LLMs and DRL complement each other in embodied AI systems. In particular, decision-making often requires continuous interaction between the system and human users. In an embodied AI framework, LLMs act as an optimal interface, enhancing the understanding of contextual and semantic data from multimodal inputs [16]. This processed information informs DRL-based decision-making, where the system uses the LLM-generated semantic insights to refine strategies in real-time. By leveraging LLMs, vehicles can convert complex multimodal data into actionable information, which in turn allows DRL to optimize decisions in dynamic vehicular environments.

Motivated by these insights, we propose a framework that integrates LLMs and DRL within embodied AI-enhanced vehicular networks, as illustrated in Fig. 1. *To the best of the authors' knowledge, this is the first work to explore the joint application of LLMs for data processing and DRL for decision-making in embodied AI-enhanced vehicular networks.* In this work, each vehicle operates as an embodied AI agent, utilizing LLMs to extract and communicate semantic information and employing DRL to optimize decision-making processes. The contributions of this work are summarized as follows:

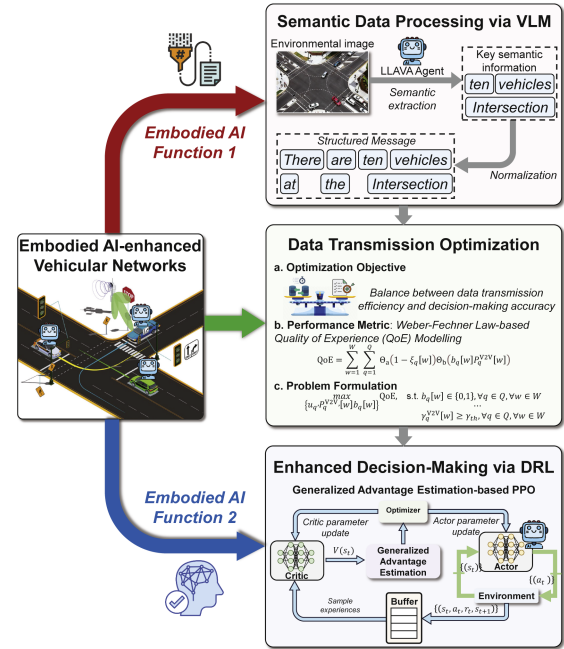


Fig. 1. The workflow of the proposed embodied AI framework for vehicular networks. The framework comprises two key functions: semantic data processing using LLaVA for efficient information extraction and enhanced decision-making via GAE-PPO to optimize transmission and decision strategies.

- *Data Transmission Formulation*: We formulate an optimization problem aimed at balancing data transmission efficiency and decision-making accuracy. To ensure the quality of transmitted information while managing limited bandwidth, we introduce the Weber-Fechner law as a metric for quantifying quality of experience (QoE). This law provides an objective function to ensure that the perceived quality of transmitted data remains high, even as bandwidth constraints increase.
- *Semantic Data Processing via VLM*: We leverage the large language and vision assistant (LLaVA) to extract semantic information from raw image data. Specifically, LLaVA converts visual inputs into structured text representations using its cross-modal alignment capability, which integrates a visual encoder, a projection matrix, and a BERT-based language model. This process significantly reduces the transmission bandwidth while preserving essential contextual details needed for vehicular communication and decision-making.
- *Enhanced Decision-Making via DRL*: We propose a generalized advantage estimation-based proximal policy optimization (GAE-PPO) to improve decision-making in dynamic vehicular environments. The GAE-PPO method incorporates generalized advantage estimation to reduce variance in policy gradient updates, stabilizing the training process. It also employs policy clipping and replay buffers to enhance convergence and robustness, ensuring that the decisions made by our embodied AI agents align with QoE optimization goals.

Simulation results show that the proposed embodied AI framework significantly outperforms the baselines in multiple aspects. The attention maps generated by LLAVA indicate that the model accurately focuses on relevant image regions during semantic extraction, effectively identifying key features such as cars and parking spaces. Compared to other VLM models (e.g., LLAVA-1.5-13b-hf, Qwen-VL-Chat, Deepseek-vl-7b-base, and Moondream2), LLAVA-1.5-7b-hf achieves superior accuracy while maintaining lower inference time, outperforming larger models like LLAVA-1.5-13b-hf by reducing inference time by 40% on average and demonstrating a smaller parameter size, making it computationally efficient. The analysis of achievable QoE as a function of the number of embodied AI agents demonstrates that our method maintains superior QoE under increasing network load compared to conventional methods, with a 36% improvement over DDPG and a 25.2% gain when scaling from 12 to 16 vehicles. Additionally, the convergence analysis reveals that the GAE-PPO method ensures faster and more stable convergence, reducing convergence steps by up to 54 compared to pure PPO. Finally, an investigation into the relationship between semantic symbol length and transmission quality confirms that adapting the symbol length provides an effective trade-off between communication quality and bandwidth utilization, achieving up to a 61.4% improvement in QoE when scaling from 4 to 8 vehicles.

The rest of this paper is organized as follows: Section II reviews the related work. Section III presents the system model, including the details of LLAVA-based semantic extraction and GAE-PPO for optimization. Section IV discusses the proposed embodied AI framework for vehicular network communication. Section V provides the simulation results and performance analysis, demonstrating the advantages of our method over conventional methods. Finally, Section VI concludes the work. Important notations used throughout the paper are summarized in Table I.

II. RELATED WORK

In this section, we review the relevant literature across two key domains, i.e., vehicular Networks and embodied AI. We highlight significant advancements and identify the gaps that our work aims to address.

A. Vehicular Networks

Recent advancements in vehicular networks have garnered significant attention due to their potential to improve road safety, optimize resource allocation, and enhance communication reliability. For instance, Liang et al. [17] proposed a multi-agent reinforcement learning (MARL) framework for spectrum sharing between V2V and V2I communications, where their method improved spectrum efficiency by allowing vehicles to manage spectrum resources in high-mobility environments cooperatively. Additionally, Liang et al. [18] proposed a resource-sharing strategy leveraging graph-based methods to optimize vehicular network performance in dynamic environments, where performance was measured through reduced interference and enhanced communication reliability. Xue et al. [19] introduced

TABLE I
SUMMARY OF NOTATION

Notation	Description
$b_q[w]$	Indicator: V2V link q on subband w
$\text{CE}_\beta(\cdot)$	Channel encoder function
$\text{Enc}_\nu^{-1}(\cdot)$	Channel decoder function
$h_{w,BS}^{(w)}, h_{q,BS}^{(w)}$	Channel gain of V2I to BS
$h_{q,q}^{(w)}, h_{w,q}^{(w)}, h_{q',q}^{(w)}$	Channel gain of V2V links/interference
$\text{MI}(\cdot \cdot)$	Mutual information
$\mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{MI}}$	Cross-entropy, MI loss functions
$p(\cdot)$	Predicted probability
$P_w^{\text{V2I}}, P_q^{\text{V2V}}[w]$	Transmission power of V2I and V2V links
$q(\cdot)$	Actual probability
$\text{SE}_\alpha(\cdot)$	Semantic encoder function
$\text{SE}_\mu^{-1}(\cdot)$	Semantic decoder function
$\mathbf{S}_i^{(w)}$	Semantic symbol vector (vehicle i , link w)
σ_a^2, σ_b^2	Gaussian noise power (V2I, V2V)
$\Theta_a(\cdot), \Theta_b(\cdot)$	Normalization functions for QoE
$\mathbf{V}_i^{(w)}$	Textual message for vehicle i on link w
$\mathbf{X}_i^{(w)}$	Transmitted signal (vehicle i , link w)
$\mathbf{Y}_i^{(w)}$	Received signal (vehicle i , link w)
$\hat{\mathbf{S}}_i^{(w)}$	Recovered semantic symbols
$\xi_q[w]$	Cosine similarity (semantic accuracy)
$\gamma_w^{\text{V2I}}, \gamma_q^{\text{V2V}}[w]$	SINR for V2I, V2V links

DRL to tackle power allocation in vehicular networks, focusing on improving ultra-reliable low-latency communication (URLLC) while reducing packet duplication, which directly impacted communication reliability and overall system latency. Similarly, Zhang et al. [20] developed a federated learning-based method for gradient quantization in vehicle edge computing networks. This method improved communication efficiency by minimizing bandwidth consumption during model updates and enhanced resource sharing across distributed vehicular nodes. Furthermore, Yao et al. [21] proposed a DRL-based secure transmission scheme for V2V communications using joint radar and communication systems. Their method protected V2V links from eavesdropping and interference while maintaining communication integrity, spectral efficiency, and signal-to-noise ratio (SNR) even in high-mobility environments.

Despite these significant contributions, most existing research primarily focused on traditional performance metrics like spectrum efficiency, latency, and security. However, new metrics such as semantic data transmission and decision-making efficiency become essential when considering embodied AI. Additionally, when incorporating LLMs for data extraction and processing, the QoE must be considered. Our work addresses this by introducing the Weber-Fechner law as a novel metric for balancing bandwidth and QoE in data transmission for vehicular networks.

B. Embodied AI

With the rapid evolution of AI, the concept of embodied AI has emerged, emphasizing the interaction between physical entities and intelligent systems. The term ‘‘embodied AI’’ refers to AI systems embedded within physical objects, enabling them to perceive, interpret, and respond to their environment [22]. The

related work on embodied AI can be broadly divided into two primary areas, i.e., data extraction and decision-making.

For data processing, most of the studies have focused on utilizing LLMs to efficiently process and transmit critical information, reducing the need for raw data transmission. For instance, Song et al. [23] proposed an embodied agents framework that used an LLM-based method for few-shot grounded planning in embodied agents. Their method allowed the system to extract key information from multimodal inputs, enabling agents to dynamically update plans based on minimal data, significantly reducing the volume of sensory data required for transmission. Similarly, in [24], Zhang et al. introduced a multi-embodied system that enhances LLMs' ability to extract relevant data for multi-agent collaboration. Their method improved task execution by refining the plan based on environmental inputs. This led to more efficient data extraction and reduced the amount of data exchanged between agents and their environment. Moreover, Mower et al. [25] developed a framework that integrates LLMs with embodied robotic systems, where data extraction is optimized through task feedback and structured reasoning. By leveraging LLMs to extract essential task information from multimodal data, the framework reduced the need to transmit large volumes of raw sensory input, enhancing the efficiency of robotic decision-making and communication. Additionally, in [26], Zhang et al. explored the vulnerabilities of multimodal LLMs in embodied AI, focusing on extracting relevant actions from complex, multimodal data and demonstrating how these systems can reduce unnecessary transmissions by filtering unsafe actions, underscoring the importance of secure data extraction for efficient operation in embodied AI systems. Furthermore, Jiang et al. [27] proposed a large AI model-based multimodal semantic communication framework, which employs latent diffusion and personalized LLM-based extraction to effectively reduce multimodal data transmission while preserving semantic integrity. Zhang et al. [28] introduced an LLM-driven semantic correction mechanism that explicitly rectifies semantic distortions, significantly enhancing transmission reliability and efficiency in lossy wireless communication environments.

For decision-making, most of the research in embodied AI has focused on using DRL to develop strategies in dynamic environments. For instance, Ying et al. [30] introduced the PEAC algorithm, which used RL to pre-train embodied agents across multiple embodiments. By learning task-agnostic strategies, PEAC improved decision-making efficiency and adaptability in handling dynamic environments. Similarly, in [31], Long *et al.* developed a human-in-the-loop simulation embodied platform that integrated RL to learn control policies from human demonstrations, enhancing decision-making efficiency in surgical tasks. Moreover, Liu et al. [32] explored language learning as a byproduct of meta-reinforcement learning, showing how agents indirectly learned language cues through decision-making tasks, further highlighting RL's role in extracting and utilizing environmental data for task performance improvement. Additionally, in [33], Tan et al. proposed the TWOSOME framework, combining LLMs and DRL for decision-making in embodied environments. Using PPO to interact with tasks, they

optimized decision-making and reduced the need for predefined datasets by enabling real-time data extraction and optimization.

Despite these significant contributions, most existing works on embodied AI primarily focus on improving algorithmic efficiency and decision-making without fully exploring its potential for resource optimization in vehicular networks. Additionally, while LLMs and DRL have been successfully applied in various embodied AI systems, their integration in optimizing both communication and decision-making processes in vehicular networks remains underexplored. Therefore, this work aims to fill this gap by leveraging the strengths of embodied AI, LLMs, and DRL to enhance both resource optimization and decision-making within the context of vehicular networks.

III. SYSTEM MODEL

As illustrated in Fig. 2, we consider a cellular-based vehicular communication network in an urban environment, where I vehicles equipped with embodied AI systems drive within the communication range of a base station (BS). To associate a vehicle with a BS, the system selects the link based on the strongest received signal strength indicator (RSSI) value at the BS [34]. The network comprises W vehicle-to-infrastructure (V2I) links and Q vehicle-to-vehicle (V2V) links, which support both direct communication between vehicles and communication with the BS. The V2I links are used for transmitting high-priority or aggregated data to the BS, while V2V links facilitate direct communication between nearby vehicles, reducing network load and latency in localized areas. For clarity, we use $i \in \{1, 2, \dots, I\}$ to denote the index of vehicle, $w \in \{1, 2, \dots, W\}$ to denote the index of V2I link, and $q \in \{1, 2, \dots, Q\}$ to denote the index of V2V link, respectively.

A. Image Capture and Semantic Extraction

In this model, each vehicle is equipped with onboard sensors that capture real-time environmental images, denoted by I_i . These images are processed by a vision-LLM model, i.e., LLAVA, to extract relevant semantic information and convert it into textual descriptions. Each vehicle uses its onboard embodied AI system to process and extract semantic information locally, performing that computational tasks are offloaded from the BS and minimizing the volume of raw data that needs to be transmitted. The extracted semantic information \mathbf{M}_i contains details about other vehicles, pedestrians, road conditions, and traffic signals.

The semantic processing system within each vehicle utilizes the textual information \mathbf{M}_i to generate a structured message, denoted by $\mathbf{V}_i^{(w)}$, where w indicates the specific V2I link over which the information will be transmitted. The structured message $\mathbf{V}_i^{(w)}$ consists of l_{\max} words, i.e.,

$$\mathbf{V}_i^{(w)} = [v_{i,1}^{(w)}, v_{i,2}^{(w)}, v_{i,l}^{(w)}, \dots, v_{i,l_{\max}}^{(w)}], \quad (1)$$

where $v_{i,l}^{(w)}$ represents the l -th word in the structured message for the i -th vehicle on link w . This message is then passed through the semantic encoder of the vehicle's transmitter, which extracts

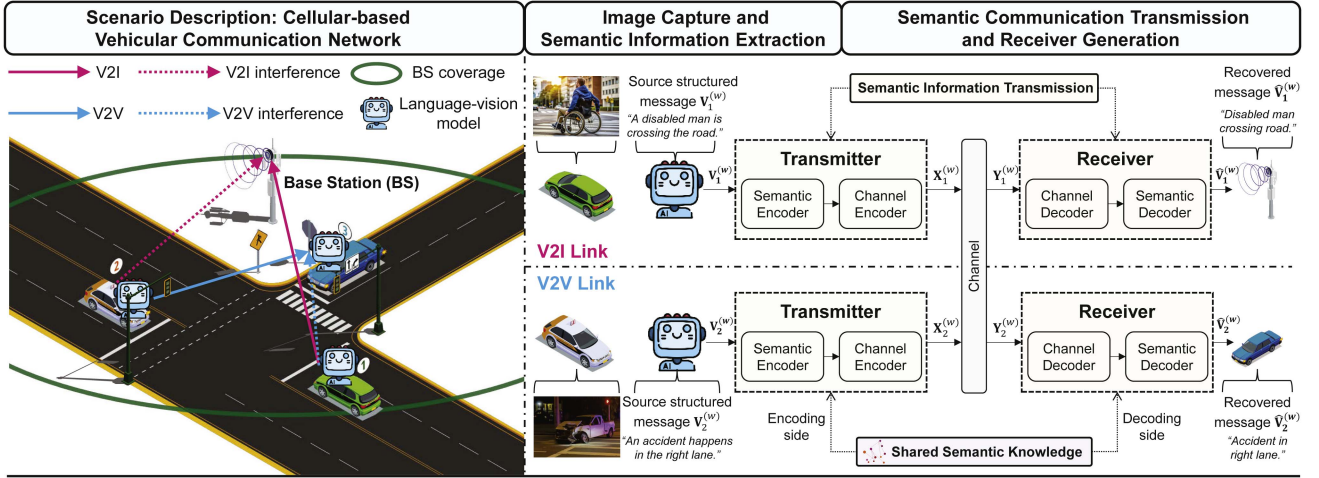


Fig. 2. System model illustrates a cellular-based vehicular communication network, where embodied AI vehicles utilize semantic communication to encode and decode structured messages for efficient and reliable data exchange [29].

and encodes the underlying semantic information into a symbol vector $\mathbf{S}_i^{(w)}$. The encoding process can be formalized as

$$\mathbf{S}_i^{(w)} = \text{SE}_\alpha \left(\mathbf{V}_i^{(w)} \right), \quad (2)$$

where $\text{SE}_\alpha(\cdot)$ represents the semantic encoder function with a set of hyper-parameters α . The resulting semantic symbol vector $\mathbf{S}_i^{(w)}$ is then represented as

$$\mathbf{S}_i^{(w)} = \left[s_{i,1}^{(w)}, s_{i,2}^{(w)}, s_{i,u}^{(w)}, \dots, s_{i,u_{\max}}^{(w)} \right], \quad (3)$$

where u denotes the average number of semantic symbols used per word with the unit of *semantic unit*.

B. SemCom Transmission and Reconstruction

After the semantic information is encoded into a symbol vector $\mathbf{S}_i^{(w)}$, it is transmitted over the wireless channel. The transmitted signal for each V2I or V2V link is denoted as:

$$\mathbf{X}_i^{(w)} = \text{CE}_\beta \left(\mathbf{S}_i^{(w)} \right), \quad (4)$$

where $\text{CE}_\beta(\cdot)$ denotes the channel encoder function parameterized by β . The expression in (4) represents the process of preparing the semantic symbols for transmission over the corresponding w -th link.

Each V2I link is allocated a specific frequency subband, while the Q V2V links may share these subbands, leading to potential interference. The signal-to-interference-plus-noise ratios (SINRs) for the w -th V2I link and the q -th V2V link on the w -th subband are respectively denoted as

$$\gamma_w^{\text{V2I}} = \frac{P_w^{\text{V2I}} h_{w,\text{BS}}^{(w)}}{\sigma_a^2 + \sum_{q=1}^Q b_q[w] P_q^{\text{V2V}}[w] h_{q,\text{BS}}^{(w)}}, \quad (5)$$

and

$$\gamma_q^{\text{V2V}}[w] = \frac{P_q^{\text{V2V}}[w] h_{q,q}^{(w)}}{\sigma_b^2 + P_w^{\text{V2I}} h_{w,q}^{(w)} + \sum_{q' \neq q}^Q b_{q'}[w] P_{q'}^{\text{V2V}}[w] h_{q',q}^{(w)}}, \quad (6)$$

where P_w^{V2I} and $P_q^{\text{V2V}}[w]$ represent the transmission powers for the w -th V2I and q -th V2V links, respectively. σ_a^2 and σ_b^2 denote the corresponding Gaussian noise, respectively. The binary variable $b_q[w] = 1$ indicates that the q -th V2V link shares the w -th V2I link.¹ The channel gains $h_{w,\text{BS}}^{(w)}$, $h_{q,\text{BS}}^{(w)}$, $h_{q,q}^{(w)}$, $h_{w,q}^{(w)}$, and $h_{q',q}^{(w)}$ represent both small-scale fading and large-scale path loss effects, affecting the quality of the signal for both V2I and V2V links. The received semantic signal at the receiver, managed by the embodied AI agents, is expressed as

$$\mathbf{Y}_i^{(w)} = \mathbf{H}_i^{(w)} \mathbf{X}_i^{(w)} + \mathbf{N}, \quad (7)$$

where $\mathbf{H}_i^{(w)}$ represents the overall channel gain matrix for the w -th V2I and V2V links, and \mathbf{N} denotes the noise vector.

Upon receiving the transmitted semantic symbols, the vehicle receiver, controlled by the embodied AI system, decodes and reconstructs the original semantic information using channel and semantic decoding techniques. The decoded signal is represented as

$$\hat{\mathbf{S}}_i^{(w)} = \text{SE}_\mu^{-1} \left(\text{Enc}_\nu^{-1} \left(\mathbf{Y}_i^{(w)} \right) \right), \quad (8)$$

where $\hat{\mathbf{S}}_i^{(w)}$ denotes the recovered semantic symbol vector for the i -th vehicle on the w -th link, $\text{SE}_\mu^{-1}(\cdot)$ is the semantic decoder with parameter μ , and $\text{Enc}_\nu^{-1}(\cdot)$ is the channel decoder with parameter ν . The μ and ν consist of network settings of the semantic decoder and channel decoder, respectively.

To assess the accuracy of the reconstructed information, the cross-entropy (CE) loss function is adopted [37], as shown in (9) shown at the bottom of next page, where $q(v_{i,l}^{(w)})$ is the actual

¹ The interaction between V2V and V2I links is crucial for supporting dynamic data exchange in vehicular networks. Specifically, the V2V links facilitate direct, localized communication between vehicles, reducing latency. In contrast, V2I links are used for relaying high-priority or aggregated data from vehicles to the BS, providing centralized control and data management. The V2V and V2I communications are distinct. That is, V2V transmissions are not forwarded to the BS via V2I links. Such a system setting has been adopted in existing works, e.g., [35] and [36].

probability of the l -th word in the original message $\mathbf{V}_i^{(w)}$, and $p(v_{i,l}^{(w)})$ is the predicted probability in the recovered message $\hat{\mathbf{V}}_i^{(w)}$.

Additionally, mutual information (MI) between the transmitted and received symbols is evaluated as in (10) shown at the bottom of this page, where $p(\mathbf{X}_i^{(w)})$ and $p(\mathbf{Y}_i^{(w)})$ are the marginal probabilities of the transmitted and received symbols, respectively. This metric provides insight into the efficiency of semantic communication, helping to evaluate how effectively the transmitted information is preserved through the communication channel.

Using the dual representation of the Kullback–Leibler (KL) divergence [38], we can derive a lower bound for the MI as follows:

$$\text{MI}(\mathbf{X}_i^{(w)}; \mathbf{Y}_i^{(w)}) \geq \mathcal{L}_{\text{MI}}(\mathbf{X}_i^{(w)}, \mathbf{Y}_i^{(w)}; \alpha, \beta, T), \quad (11)$$

where \mathcal{L}_{MI} denotes the loss function used to train the neural network parameters α, β , and T for mutual information estimation. It is formulated as a negative mutual information objective based on the contrastive learning principle, which facilitates the network to retain as much semantic information as possible when transmitting through noisy communication channels [37]. T represents the temperature parameter that T controls the sharpness of the probability distribution, effectively adjusting the confidence level of the model's predictions during training.

C. Performance Metrics and Formulated Problem

To evaluate the effectiveness of the reconstructed semantic information transmitted through V2I and V2V links, we compute the cosine similarity between the bidirectional encoder representation of Transformers (BERT) embeddings of the original and reconstructed messages [39]. The cosine similarity describes the sentence semantic similarity between the recovered $\hat{\mathbf{V}}_i^{(w)}$ and actual $\mathbf{V}_i^{(w)}$, which reflects whether the semantic of the sentences are correctly transmitted. Specifically, the cosine similarity is defined as

$$\xi_q[w] = \frac{|B(\mathbf{V}_i^{(w)}) \cdot B(\hat{\mathbf{V}}_i^{(w)})|}{\|B(\mathbf{V}_i^{(w)})\| \|B(\hat{\mathbf{V}}_i^{(w)})\|}, \quad (12)$$

where $B(\cdot)$ represents the BERT model used to generate feature vectors for the sentences. The value ξ ranges from 0 to 1, with $\xi = 1$ indicating perfect similarity and $\xi = 0$ indicating no similarity.

To optimize the performance of our vehicular communication system, we introduce a QoE metric inspired by the

Weber-Fechner law,² which models the relationship between the magnitude of a stimulus and its perceived intensity [42]. The QoE is influenced by the semantic similarity (i.e., $\xi_q[w]$) between transmitted and received images, which depends on both the number of semantic symbols and the SINR of the communication link. The QoE can be expressed as

$$\text{QoE} = \sum_{w=1}^W \sum_{q=1}^Q \Theta_a(1 - \xi_q[w]) \Theta_b(b_q[w] P_q^{\text{V2V}}[w]), \quad (13)$$

where $\Theta_a(\cdot)$ and $\Theta_b(\cdot)$ are functions that normalize the impact of semantic similarity $\xi_q[w]$ and the transmission power $P_q^{\text{V2V}}[w]$ on the overall QoE. To ensure clarity, the QoE metric integrates both the semantic accuracy of reconstructed messages and the efficiency of resource utilization. This dual consideration helps strike a balance between the quality of the transmitted information and the optimal use of network resources.

The optimization problem focuses on achieving high QoE while ensuring efficient resource utilization. In this context, the embodied AI agents (i.e., vehicles) play a crucial role in generating and adapting communication strategies. These agents dynamically optimize these parameters, i.e., transmission power, semantic symbol allocation, and channel usage, based on the real-time analysis of the environment and communication needs. Thus, the optimization problem is formulated as follows:

$$\max_{\{P_q^{\text{V2V}}[w], b_q[w], u_q\}} \text{QoE}, \quad (14a)$$

$$\text{s.t. } b_q[w] \in \{0, 1\}, \quad \forall q \in Q, \forall w \in W, \quad (14b)$$

$$\sum_{w=1}^W b_q[w] \leq 1, \quad \forall q \in Q, \quad (14c)$$

$$\sum_{q=1}^Q b_q[w] \leq 1, \quad \forall w \in W, \quad (14d)$$

$$u_q \in \{1, \dots, u_{q_{\max}}\}, \quad \forall q \in Q, \quad (14e)$$

$$\xi_q[w] \geq \xi_{\text{th}}, \quad \forall q \in Q, \forall w \in W, \quad (14f)$$

² Weber-Fechner law is a classical psychophysical principle, which models the nonlinear relationship between stimulus and perception, suggesting that human sensitivity to changes in stimuli is proportional to their relative, not absolute, magnitude. In our context, semantic distortion and transmission effort jointly affect perceived communication quality. To reflect this perceptual principle, we adopt a multiplicative formulation rather than a linear weighted sum. This design ensures that degradation in either semantic accuracy or transmission reliability leads to a sharp drop in overall QoE, emphasizing the coupled impact of the two dimensions. Such multiplicative QoE formulations based on Weber-Fechner law have also been used in other studies, such as [40] and [41].

$$\mathcal{L}_{\text{CE}}(\mathbf{V}_i^{(w)}, \hat{\mathbf{V}}_i^{(w)}; \alpha, \beta, \mu, \nu) = - \sum_{l=1}^{l_{\max}} q(v_{i,l}^{(w)}) \log p(v_{i,l}^{(w)}) + (1 - q(v_{i,l}^{(w)})) \log(1 - p(v_{i,l}^{(w)})). \quad (9)$$

$$\text{MI}(\mathbf{X}_i^{(w)}; \mathbf{Y}_i^{(w)}) = \sum_{\mathbf{X}_i^{(w)}, \mathbf{Y}_i^{(w)}} p(\mathbf{X}_i^{(w)}, \mathbf{Y}_i^{(w)}) \log \frac{p(\mathbf{X}_i^{(w)}, \mathbf{Y}_i^{(w)})}{p(\mathbf{X}_i^{(w)}) p(\mathbf{Y}_i^{(w)})} = \mathbb{E}_{p(x,y)} \left[\log \frac{p(\mathbf{X}_i^{(w)}, \mathbf{Y}_i^{(w)})}{p(\mathbf{X}_i^{(w)}) p(\mathbf{Y}_i^{(w)})} \right]. \quad (10)$$

$$P_q^{V2V}[w] \in [P_{\min}, P_{\max}], \forall q \in Q, \forall w \in W, \quad (14g)$$

$$\gamma_q^{V2V}[w] \geq \gamma_{th}, \quad \forall q \in Q, \forall w \in W. \quad (14h)$$

Here, constraint (14b) ensures that each V2V link either shares or does not share a V2I link. Constraints (14c) and (14d) ensure that each vehicle and each subband are allocated appropriately, i.e., no overlapping resource usage occurs. Constraint (14e) specifies the range of the semantic symbol allocation u_q . Constraint (14f) ensures that the semantic similarity $\xi_q[w]$ remains higher than or equal to the minimum threshold ξ_{th} to maintain communication quality. Constraints (14g) and (14h) limit the transmission power and require the SINR to be higher than or equal to the threshold γ_{th} to ensure efficient communication.

It is observed that Problem (14) is non-convex and involves combinatorial aspects due to constraints such as (14b), (14c), and (14d). This complexity, combined with the highly dynamic nature of vehicular networks, presents a significant challenge. Specifically, traditional optimization algorithms often fall short in adapting to rapid changes in channel state or network state, requiring repeated re-execution of computationally expensive procedures. To address these issues, we propose an embodied AI framework that integrates LLAVA and PPO to solve it in the next section.

IV. PROPOSED EMBODIED AI FRAMEWORK

In this section, we present an embodied AI framework in which LLAVA is utilized for semantic information extraction from environmental images, while PPO is employed to dynamically optimize communication parameters in response to environmental changes.

A. LLAVA-Based Semantic Information Extraction

To enable the extraction of semantic information from images captured by the embodied AI agents (i.e., vehicles), we employ the LLAVA model. First, we provide an overview of LLAVA, followed by details on its design for semantic data generation.

1) *Overview of LLAVA*: LLAVA is a large multimodal model designed to serve as a general-purpose visual assistant by combining VLMs with powerful visual encoders based on contrastive language-image pre-training [43]. The model is trained to align image features with language prompts, enabling capabilities such as image understanding, visual reasoning, and natural language interaction with visual data. The architecture of LLAVA consists of a visual encoder, a linear projection layer, and a language model. Specifically, given an input image \mathbf{I}_i captured by the i -th vehicle's onboard sensor, LLAVA applies a visual encoder $g(\cdot)$ to extract the corresponding visual feature $\mathbf{Z}_i = g(\mathbf{I}_i)$. These features are then linearly projected to match the word embedding space of the LLM, enabling coherent integration with the language model. This allows the model to interpret the visual content and respond to user-defined semantic queries in natural language. The detailed structure of this pipeline is illustrated in Fig. 3, which highlights how the visual features are transformed and grounded in the language model's representation

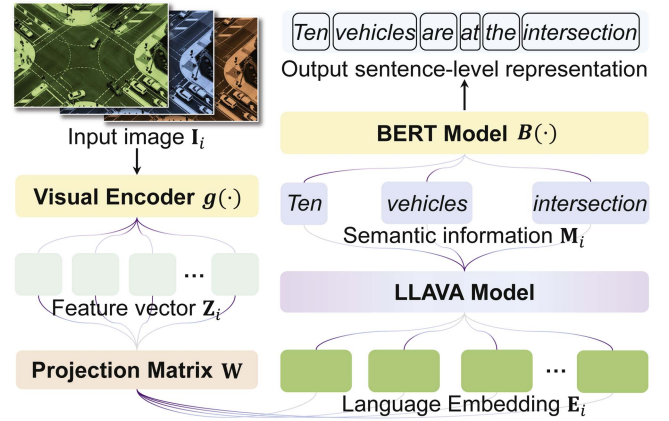


Fig. 3. The architecture of the LLAVA model for semantic extraction and language embedding. The input image \mathbf{I}_i is processed by the visual encoder $g(\cdot)$ to generate feature vectors, which are then transformed through a projection matrix W and processed by the LLAVA model to extract semantic information \mathbf{M}_i . The BERT model $B(\cdot)$ is used to generate the final sentence-level representation, capturing key information from the visual input.

space. Through this process, LLAVA effectively maps complex visual inputs into interpretable semantic representations, making it well-suited for embodied AI systems that require real-time understanding and communication of environmental data.

2) *LLAVA Design for Extraction*: The proposed LLAVA design in this work aims to support real-time vehicular communication scenarios. The model processes captured images to generate semantic information used for subsequent decision-making. The LLAVA model is employed in the following steps:

Data Generation Using LLAVA: Given input images \mathbf{I}_i , LLAVA extracts semantic information \mathbf{M}_i . First, a contrastive language-image pretraining (CLIP-based) visual encoder is employed to obtain image-level features, i.e.,

$$\mathbf{Z}_i = g(\mathbf{I}_i), \quad (15)$$

where $g(\cdot)$ denotes the CLIP-based visual encoder, and \mathbf{Z}_i represents the extracted features from the vehicle's surrounding environment.

To bridge the visual and language modalities, a trainable linear projection is applied to align the extracted visual features with the language model's word embedding space, which is expressed as

$$\mathbf{E}_i = \mathbf{W} \cdot \mathbf{Z}_i, \quad (16)$$

where \mathbf{E}_i is the resulting embedding vector projected onto the language space, and \mathbf{W} is a learnable matrix.

The language model then processes this embedding to generate the semantic output, i.e.,

$$\mathbf{M}_i = \text{LLAVA}(\mathbf{I}_i; \theta_i), \quad (17)$$

where θ_i denotes the set of LLAVA model parameters for the i -th vehicle.

LLAVA Model Fine-Tuning: To improve performance in the vehicular network domain, LLAVA is fine-tuned using training data collected from realistic driving scenarios. The objective is to minimize the semantic deviation between the ground-truth

Algorithm 1: LLAVA-Based Semantic Information Extraction for Vehicular Networks.

- 1: **Input:** Input image I_i from vehicle i ; contrastive language–image pre-training encoder $g(\cdot)$;
LLAVA model parameters θ ; projection matrix \mathbf{W} ;
BERT model for representation;
 - 2: Use pretrained contrastive language–image pre-training encoder to obtain visual features from the input image via (15); # *Extract Visual Features*
 - 3: Project the extracted visual features into the word embedding space using a trainable linear projection via (16); # *Project Features to Embedding Space*
 - 4: Utilize LLAVA to generate semantic information from the projected features via (17); # *Derive Semantic Information Using LLAVA*
 - 5: Fine-tune LLAVA with real-world vehicular data to minimize semantic mismatch via (18); # *Fine-Tune LLAVA for Scenario Adaptation*
 - 6: Encode the generated semantic information using BERT to obtain a sentence-level representation $B(\mathbf{M}_i)$; # *Semantic Representation with BERT*
 - 7: **Output:** Extracted semantic information \mathbf{M}_i ;
-

semantic label \mathbf{M}_i and the generated output $\hat{\mathbf{M}}_i$, defined as

$$\min_{\theta} \sum_{i=1}^I \mathcal{L}(\mathbf{M}_i, \hat{\mathbf{M}}_i) = \min_{\theta} \sum_{i=1}^I \|\mathbf{M}_i - \hat{\mathbf{M}}_i\|^2, \quad (18)$$

Remark 1: This MSE-based loss formulation ensures that the extracted semantic representation \mathbf{M}_i remains close to the actual context of the physical environment, enhancing the communication fidelity of the embodied AI agent.

In addition, we use the cross-entropy (CE) loss to strengthen alignment between the predicted and target words in semantic responses, which is expressed as

$$\mathcal{L}_{\text{CE}}(\mathbf{M}_i, \hat{\mathbf{M}}_i) = \sum_{l=1}^{l_{\max}} q(v_{i,l}) \log p(v_{i,l}), \quad (19)$$

where $q(v_{i,l})$ and $p(v_{i,l})$ denote the ground-truth and predicted probabilities of the l -th token $v_{i,l}$.

Semantic Information Representation: To provide a compact and meaningful representation, the output \mathbf{M}_i is further encoded via BERT to produce a sentence-level embedding $B(\mathbf{M}_i)$. This vectorized representation is later used in downstream modules such as transmission quality evaluation and decision-making.

In such a manner, key information captured by vehicles can be processed as semantic information through LLAVA. The overall algorithm for semantic information extraction and representation using LLAVA is summarized in Algorithm 1.

B. PPO-Based Transmission Strategy

The proposed PPO-based transmission strategy begins with a service request processed by LLAVA. LLAVA interprets the request and extracts relevant semantic information from images captured by vehicle-mounted cameras. After extracting

the semantic information via LLAVA, an efficient transmission strategy is required for vehicular networks. To address this, we propose a PPO-based method to optimize the transmission policy.

1) *Overview of PPO:* PPO is a policy gradient method within the actor-critic framework, where the actor network defines the transmission policy (parameterized by θ_A), and the critic network estimates the value function (parameterized by θ_C). PPO aims to optimize the policy to maximize expected rewards while ensuring smooth policy transitions to prevent drastic and destabilizing updates [44]. The core of PPO lies in its use of a surrogate objective function that encourages updates to the policy without large shifts from the previous policy $\pi_{\theta_A^{\text{old}}}$. The surrogate objective function at each time step t is defined as

$$J(\theta_A) = \mathbb{E}_t \left[\rho_t(\theta_A) A_t^{\pi_{\theta_A^{\text{old}}}} \right], \quad (20)$$

where $\rho_t(\theta_A) = \frac{\pi_{\theta_A}(a_t|s_t)}{\pi_{\theta_A^{\text{old}}}(a_t|s_t)}$ represents the probability ratio between the new and old policies, and $A_t^{\pi_{\theta_A^{\text{old}}}}$ is the advantage function at time t , which is calculated as

$$A_t^{\pi_{\theta_A^{\text{old}}}} = r(s_t, a_t) + \gamma V_{\theta_C}(s_{t+1}) - V_{\theta_C}(s_t), \quad (21)$$

where $r(s_t, a_t)$ is the reward for taking action a_t in state s_t , and γ is the discount factor. To prevent overly large updates that may lead to unstable learning, PPO applies a clipped version of the objective function, i.e.,

$$J_{\text{clip}}(\theta_A) = \mathbb{E}_t \left[\min \left(\rho_t(\theta_A) A_t^{\pi_{\theta_A^{\text{old}}}}, \text{clip}(\rho_t(\theta_A), 1 - \epsilon, 1 + \epsilon) A_t^{\pi_{\theta_A^{\text{old}}}} \right) \right], \quad (22)$$

where ϵ is a hyper-parameter that limits the update magnitude, ensuring the probability ratio $\rho_t(\theta_A)$ remains within $[1 - \epsilon, 1 + \epsilon]$.

Remark 2: To ensure stability during training and improve the sample efficiency, we incorporate generalized advantage estimation into the critic network's value function, i.e., GAE-PPO. This technique is used to optimize the advantage function, preventing the advantage variance from growing too large during training [45]. The generalized advantage estimation is computed as $A_t^{\pi_{\theta_A^{\text{old}}}} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$, where λ is the hyperparameter introduced from GAE to adjust the number of steps of time differences considered during training. This advantage estimation ensures that the variance estimation does not diverge during training, allowing both actor and critic networks to maintain numerical stability.

Next, the actor network is updated by performing gradient descent on $J_{\text{clip}}(\theta_A)$, while the critic network is updated by minimizing the mean squared error between the predicted and target value functions. The loss function is

$$\mathcal{L}_C(\theta_C) = \mathbb{E}_t \left[(V_{\theta_C}(s_t) - V_{\text{tar}}(s_t))^2 \right], \quad (23)$$

where $V_{\theta_C}(s_t)$ is the value calculated by the value network with hyper-parameters set θ_C , and $V_{\text{tar}}(s_t)$ is the target value

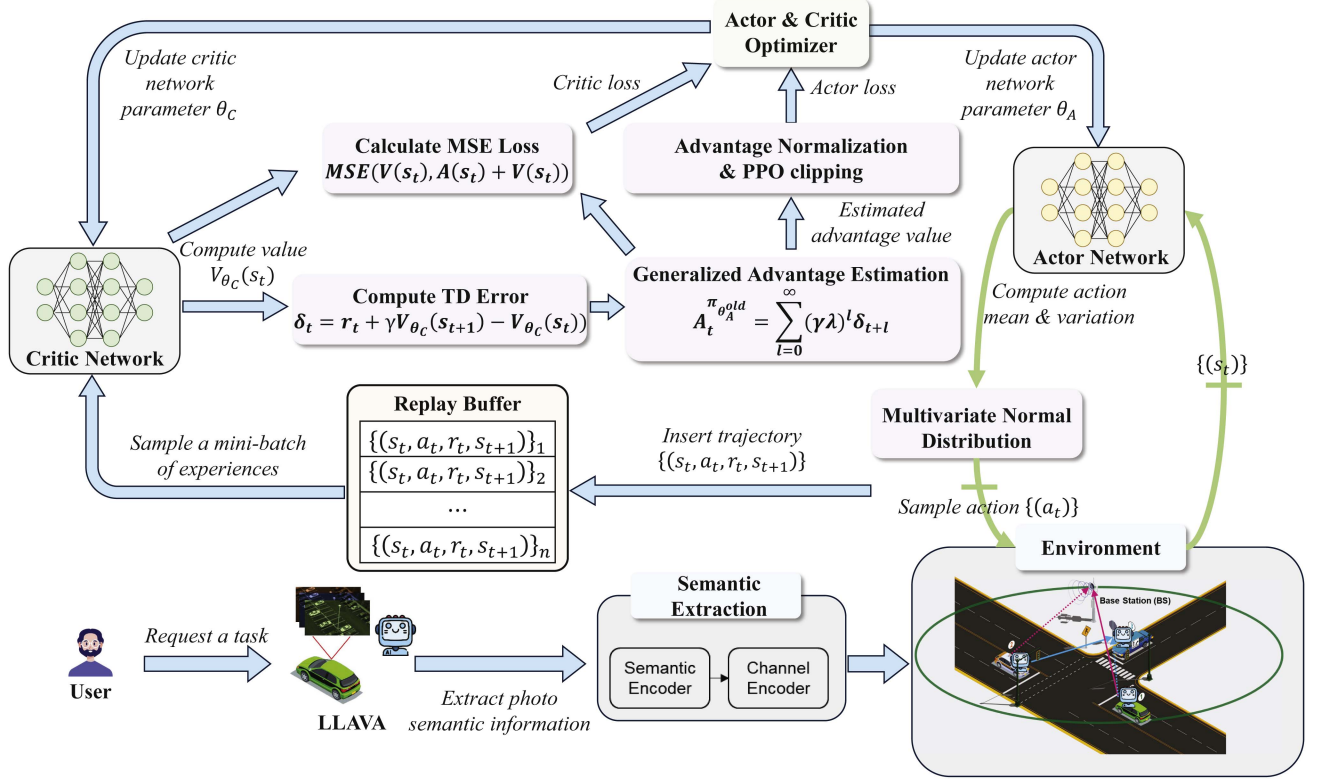


Fig. 4. The workflow of the GAE-PPO method for optimizing transmission strategies. The actor and critic networks are updated iteratively through a combination of replay buffer sampling, temporal-difference (TD) error calculation, advantage estimation, and policy clipping. The actor network generates actions using a multivariate normal distribution, while the critic network evaluates state values, contributing to the stability and convergence of the learning process.

calculated using n -step returns, i.e.,

$$V_{\text{tar}}(s_t) = r(s_t, a_t) + \gamma V_{\theta_C}(s_{t+1}). \quad (24)$$

For a clear representation of the proposed GAE-PPO method, we provide the detailed workflow in Fig. 4. Specifically, our GAE-PPO method provides the interaction between the actor and critic networks during the optimization process. The actor network generates actions, while the critic network evaluates the corresponding state values. Through iterative updates, the replay buffer is used to store experience samples, and the TD error is calculated to refine the value estimates. By combining advantage normalization, PPO clipping, and generalized advantage estimation, the GAE-PPO method can achieve stable and efficient training for optimizing vehicular network transmission strategies.

2) *MDP Design*: In our proposed method, the Markov decision process (MDP) framework is used to model the problem of optimizing transmission strategies for the vehicular network [46]. The key components are designed as follows.

Action Space: The action space represents the decisions made by the embodied AI agents to control the transmission in the vehicular network. Specifically, the actions include channel selection, transmission power control, and the number of semantic symbols transmitted in each V2V link. The action taken by the

agent at time step t can be represented as

$$a_t = \left(\underbrace{\{b_q[w]\}}_Q, \underbrace{\{P_q^{V2V}[w]\}}_Q, \underbrace{\{u_q\}}_Q \right), \quad (25)$$

Thus, the dimension of the action space is $3Q$.

State Space: The state space captures the environment's observable parameters at each time step t . For our framework, the state s_t includes the channel gain of the V2V links and V2I links, and the SINR for both V2V and V2I links. Thus, the state at time t can be expressed as

$$s_t = \left(\underbrace{\{\mathbf{H}_i^{(w)}\}}_W, \underbrace{\{\gamma_q^{V2V}(t)\}}_Q, \underbrace{\{\gamma_w^{V2I}(t)\}}_W \right), \quad (26)$$

Thus, the dimension of the state space is $2W + Q$.

Reward Function: The reward function is designed to align with the optimization objective of the system. Specifically, our reward function encourages the maximization of the proposed QoE while considering communication constraints. The reward at time step t is given by

$$r_t = \sum_{w=1}^W \sum_{q=1}^Q \Theta_a(1 - \xi_q[w](t)) \Theta_b(b_q[w] P_q^{V2V}[w](t)). \quad (27)$$

To address the challenge of ensuring that the constraints are satisfied, we employ a penalty-based reward function. Specifically, we add penalty terms to the reward function to discourage actions that violate the constraints. The designed reward function can be expressed as

$$\begin{aligned} \tilde{r}_t = & r_t - \lambda_1 \sum_{q=1}^Q \max(0, \xi_{\text{th}} - \xi_q[w](t)) \\ & - \lambda_2 \sum_{q=1}^Q \max(0, \gamma_{\text{th}} - \gamma_q^{\text{V2V}}[w](t)), \end{aligned} \quad (28)$$

where λ_1 and λ_2 are penalty coefficients that control the weight of the penalty terms. The first penalty term ensures that the semantic similarity $\xi_q[w](t)$ remains higher than or equal to the minimum threshold ξ_{th} , while the second penalty term ensures that the SINR $\gamma_q^{\text{V2V}}[w](t)$ remains higher than or equal to the required threshold γ_{th} . By incorporating these penalty terms, we effectively guide the embodied AI agents to make decisions that satisfy the system constraints.

Remark 3: The proposed GAE-PPO-based method inherently provides resilience against unexpected scenarios such as sensor failures or severe channel interference. In particular, the agent continuously adjusts its transmission policy based on real-time environmental feedback, and any degradation in sensing or communication quality is immediately reflected through the observed rewards. Moreover, the penalty-based reward design discourages actions that violate semantic similarity and SINR constraints, guiding the embodied AI agents to adopt more robust behaviors. This adaptability ensures stable and efficient communication even under dynamic and adverse conditions, as similarly observed in [47] and [48].

Based on this MDP formulation, we summarize the proposed GAE-PPO-based transmission strategy optimization algorithm for vehicular networks in Algorithm 2. Specifically, the proposed algorithm utilizes the GAE to reduce variance in the policy gradient estimation, which improves the stability and efficiency of the learning process. The algorithm begins by initializing the actor and critic networks, followed by the exploration phase, where actions are selected based on the actor's policy. The observed rewards and transitions are stored, and the actor and critic networks are updated using gradient descent to optimize the transmission policy.

Remark 4: While our proposed framework is originally designed for embodied AI in vehicular networks, it exhibits adaptability and generalizability to a variety of other application domains and input modalities. On the application level, the framework can be extended to robotics and UAV networks, as long as the agent is equipped with visual sensors. The LLAVA-based semantic extraction module interprets environmental inputs into compact and meaningful representations, which are then processed by the GAE-PPO algorithm for decision-making. This modular design, separating perception from policy learning, facilitates straightforward adaptation to diverse embodied AI tasks. Similar designs have been used in robotic manipulation and UAV path planning, as shown in [47]. Moreover, although our current implementation relies on LLAVA, a vision-language

Algorithm 2: GAE-PPO-Based Transmission Strategy Optimization for Vehicular Networks.

```

1 Input : Episodes, discount factor  $\gamma$ , PPO
           hyper-parameter  $\epsilon$ , learning rate, number of
           embodied AI agents;
2 Initialize actor network  $\theta_A$ , critic network  $\theta_C$ , and
   environment;
3 Observe initial state  $s_0$  from environment;
4 for each iteration do
5   for each exploration step  $t$  do
6     Select action  $a_t$  using the actor network based
       on policy  $\pi_{\theta_A}$ ;
7     Execute action  $a_t$  and observe reward  $r_t$  and
       next state  $s_{t+1}$ ;
8     Store transition  $(s_t, a_t, r_t, s_{t+1})$  in replay
       buffer  $\mathcal{D}$ ;
9   for each update step  $t$  do
10    Sample mini-batch from  $\mathcal{D}$ ;
11    Update actor network  $\theta_A$  via gradient
       descent on  $J_{\text{clip}}(\theta_A)$ ;
12    Update critic network  $\theta_C$  by minimizing
        $\mathcal{L}_C(\theta_C)$ ;
13    Update state  $s_t \leftarrow s_{t+1}$ ;
14 Output : Optimized transmission policy  $\pi_{\theta_A}$ ;

```

model tailored to image-based inputs, the architecture of the framework allows replacement of the visual encoder with other multimodal models (e.g., ImageBind [49] and Flamingo [50]) to accommodate additional modalities such as audio, text, or radar data. This substitution would not require modification of the downstream DRL module, making our approach extensible to future embodied AI systems with diverse sensory inputs.

C. Complexity Analysis

Computational Complexity Analysis: The computational complexity of the proposed embodied AI framework is influenced by LLAVA and GAE-PPO. For LLAVA, it primarily utilizes a Transformer-based architecture to process cross-modal data, combining visual and textual information [51]. The complexity of its self-attention mechanism is $\mathcal{O}(L^2 \cdot d)$, where L is the input sequence length (e.g., the number of tokens or visual patches) and d is the embedding dimension. Additionally, the feature projection step incurs a computational cost proportional to the projection matrix size, resulting in an overall complexity of $\mathcal{O}(L^2 \cdot d + L \cdot d^2)$. For the GAE-PPO method, the computational complexity depends on the structure of the DNNs used in the actor and critic networks [52]. Specifically, the complexity of the DNNs is $\mathcal{O}(\sum_{p=1}^P n_{p-1} \cdot n_p)$, where n_p represents the number of neurons in the p -th layer, and P is the total number of layers. The size of the input layer (i.e., n_0) corresponds to the dimension of the state space, while the size of the output layer (i.e., n_P) corresponds to the dimensionality of the action space. Moreover, GAE introduces an additional computational overhead of $\mathcal{O}(T)$ per iteration, where T is the number of timesteps

used in advantage computation. Consequently, the overall complexity of the GAE-PPO method is as $\mathcal{O}(T \cdot \sum_{p=1}^P n_{p-1} \cdot n_p)$. Thus, the overall computational complexity of the embodied AI framework is $\mathcal{O}(L^2 \cdot d + L \cdot d^2) + \mathcal{O}(T \cdot \sum_{p=1}^P n_{p-1} \cdot n_p)$.

Space Complexity Analysis: Each vehicle in our framework maintains a lightweight neural policy network, with a parameter count denoted by P . Thus, the space complexity for each vehicle is $\mathcal{O}(P)$. As the number of vehicles increases to I , and given that each vehicle operates independently with its own local policy, the total model storage requirement across the system scales linearly as $\mathcal{O}(I \cdot P)$. In addition, a shared buffer is maintained to store historical interaction data across time. The size of this buffer is determined by the data stored per time slot D_t and the total number of time slots T , leading to an overall space complexity of $\mathcal{O}(T \cdot D_t)$ [53]. Therefore, the combined system space complexity is $\mathcal{O}(I \cdot P + T \cdot D_t)$.

Communication Complexity Analysis: In our framework, each vehicle communicates over a fixed resource block and only performs power control and semantic-aware transmission, without involving dynamic spectrum allocation or spectrum sharing [54]. Therefore, the wireless signal processing complexity stems primarily from the evaluation of interference across vehicles. For each vehicle, interference computation requires estimating the channel gains and received interference from its K neighbors, considering both V2V and V2I transmissions. This includes path loss, shadowing, and fast fading computations. As such, the communication channel complexity for the whole system is $\mathcal{O}(I \cdot K)$.

V. NUMERICAL RESULTS

A. Parameter Settings

The parameters in our framework are categorized into *Scenario Settings* and *Embodied AI Framework Settings*.

1) *Scenario Settings:* The text dataset employed for semantic extraction in this work is the European Parliament dataset, comprising approximately 2 million sentences and 53 million words [55]. We simulate a vehicular network with 4 embodied AI vehicles, each supporting both V2V and V2I communication links. Following [15], the height of the base BS is set to 25 m, and the height of vehicles is set to 1.5 m. The V2I transmit power is fixed at 23 dBm, while the V2V transmit power varies between -100 dBm and 23 dBm. The noise powers of σ_a^2 and σ_b^2 are both set at -114 dBm, and the carrier frequency is 1GHz. The width of road lanes is configured as 3.5 m, with shadowing deviations of 3 dB and 8 dB for V2I and V2V links, respectively. For the system objectives, $\Theta_a(x)$ is set to x and Θ_b is expressed as $\Theta_b(x) = \frac{1}{1+e^{-x}}$. The minimum thresholds for SSIM and SINR constraints are set to $\xi_{th} = 0.3$ and $\gamma_{th} = 10$ dB, respectively.

2) *Embodied AI Framework Settings:* The framework consists of two key components, i.e., LLAVA and GAE-PPO. Specifically, the LLAVA model used in our simulations is LLAVA-v1.5-7B³, containing 7 billion adjustable parameters. It processes a dataset comprising 30 driving scenario images [56], efficiently extracting semantic information to support vehicular

TABLE II
CRITICAL PARAMETER SETTINGS

System Parameter	Value
V2I transmit power	23 dBm
V2V transmit power	[-100, 5, 15, 23] dBm
V2I deviation of shadowing	3 dB
V2V deviation of shadowing	8 dB
Number of NN layers	3
Clipping parameter ϵ	0.5
Initial learning rate α	1×10^{-4}
Final learning rate α	1×10^{-8}
Discount factor γ	0.99

communication. The transmitter network includes a semantic encoder and a channel encoder, with hyper-parameters α and β . The architecture consists of three Transformer layers, each with 8 attention heads, and the Rectified Linear Unit (ReLU) as the activation function. Two dense layers, with sizes of 256 and 16, are connected to the last Transformer layer. The receiver network, comprising a semantic decoder and a channel decoder, is similarly structured but with the layers applied in reverse order. For the GAE-PPO algorithm, we use the *Adam* optimizer and *Tanh* as the activation function. The neural network includes three layers, each with 512 hidden units [57]. The replay buffer has a maximum size of 1×10^6 , and the network is updated at the end of each episode. The penalty coefficients λ_1 and λ_2 are both set to 1. Other critical parameters are listed in Table II [58].

B. Simulation Results

Fig. 5 demonstrates an example of semantic information extraction when an embodied AI vehicle captures images of its surroundings. As shown, thanks to LLAVA's powerful cross-modal capabilities, the system can effectively extract key details such as the number of parking spaces, which ones are occupied, and where available parking spots are located. Compared to transmitting the image directly (i.e., 614 Kilobyte in our simulations), sending the extracted text (i.e., 12.1 Kilobyte in our simulations) significantly reduces the required bandwidth. This highlights the efficiency of using text-based semantic information over raw image data for communication purposes in vehicular networks. Moreover, the right part of Fig. 5 presents a detailed analysis of the semantic extraction process, focusing on LLAVA's attention mechanism and how the textual output is associated with visual elements. This analysis demonstrates how LLAVA combines input text with visual features to generate the final output, ensuring that the extracted information is well-grounded in the environmental context.

- *Input-Output Text Relevancy Analysis:* The bar charts show the relevancy scores between input tokens and output tokens, reflecting the contribution of each input word to generating specific output tokens. Tokens such as "spaces," "occupied," and "cars" are found to be highly relevant to responses like "four," "three," and "one." This indicates that LLAVA can effectively direct its attention to specific, relevant parts of the input to generate accurate output, ensuring the correct interpretation of semantic elements,

³ [Online]. Available: <https://huggingface.co/llava-hf/llava-1.5-7b-hf>

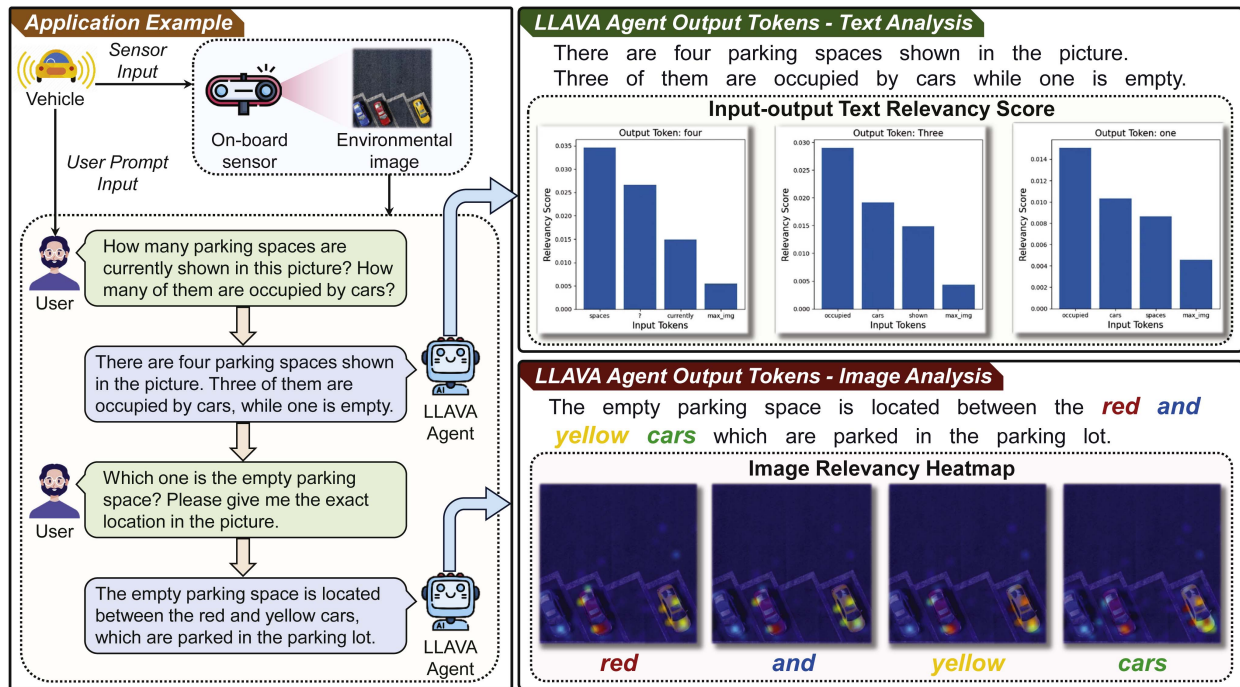


Fig. 5. The illustration of semantic information extraction. In the left part, the environmental image is obtained through the onboard sensor. Then, the semantic information is extracted according to the user's needs via the embodied AI. In the right part, we demonstrate the relevance between the extracted semantic information and user needs, as well as the environmental images.

such as the number of parking spaces and their occupancy status.

- **Relevancy Heatmaps:** The heatmaps at the bottom of Fig. 5 highlight the image regions most relevant to specific output tokens. For instance, the highlighted areas show that tokens "red" and "yellow" are directly associated with the respective cars in the image. This demonstrates LLAVA's capability to ground text in specific visual features, linking language tokens to corresponding visual objects. The heatmaps validate the system's ability to perform cross-modal grounding, which enhances the reliability of extracted semantic information by correctly associating textual descriptions with visual inputs.

Next, to evaluate the performance of our LLAVA-1.5-7b-hf model, we compare it with several state-of-the-art vision-LLM models, including:

- **LLAVA-1.5-13b-hf [59]:** A vision-LLM with about 13.4 billion parameters, designed for high-accuracy multimodal semantic extraction through advanced cross-modal alignment techniques.
- **Qwen-VL-Chat [60]:** A multimodal conversational model with about 9.66 billion parameters, specializing in visual-language reasoning tasks requiring both image and text understanding.
- **Deepseek-vl-7b-base [61]:** A vision-LLM with about 7.34 billion parameters optimized for extracting structured semantic information from complex visual and textual data.
- **Moondream2 [62]:** A lightweight vision-LLM with 1.8 billion parameters, targeting general-purpose semantic reasoning with the focus on computational efficiency.

Fig. 6 compares these vision-LLM models in terms of parameter size, inference time, and output accuracy for semantic queries, highlighting their trade-offs in computational efficiency and performance. The first bar chart shows the parameter sizes of the evaluated models. LLAVA-1.5-7b-hf, with 7 billion parameters, is significantly smaller than LLAVA-1.5-13b-hf, which has 13 billion parameters, approximately 46.2% fewer parameters. Despite this reduction in size, LLAVA-1.5-7b-hf achieves comparable or superior performance, making it particularly suitable for real-time decision-making in resource-constrained vehicular networks. The second and third bar charts evaluate inference time for two semantic queries, both requiring the identification of parking space availability and its exact location. LLAVA-1.5-7b-hf achieves inference times of 1.2 seconds and 1.3 seconds for the first and second queries, respectively, which are 40% and 38% faster compared to LLAVA-1.5-13b-hf. While LLAVA-1.5-13b-hf delivers accurate answers, its larger parameter size results in higher latency. Other models, such as Qwen-VL-Chat and deepseek-vl-7b-base, either fail to provide consistent, correct answers or exhibit even higher inference delays, further emphasizing the efficiency of LLAVA-1.5-7b-hf. On the left side of Fig. 6, LLAVA-1.5-7b-hf demonstrates robust semantic alignment between visual inputs and textual representations. For instance, it consistently produces correct answers for both semantic queries, while the other models either provide vague or incorrect responses. This performance advantage can be attributed to its efficient cross-modal alignment mechanism, which balances parameter efficiency and semantic understanding. These characteristics make it a strong candidate for embodied AI vehicular networks, where low-latency and high-reliability decision-making are essential.

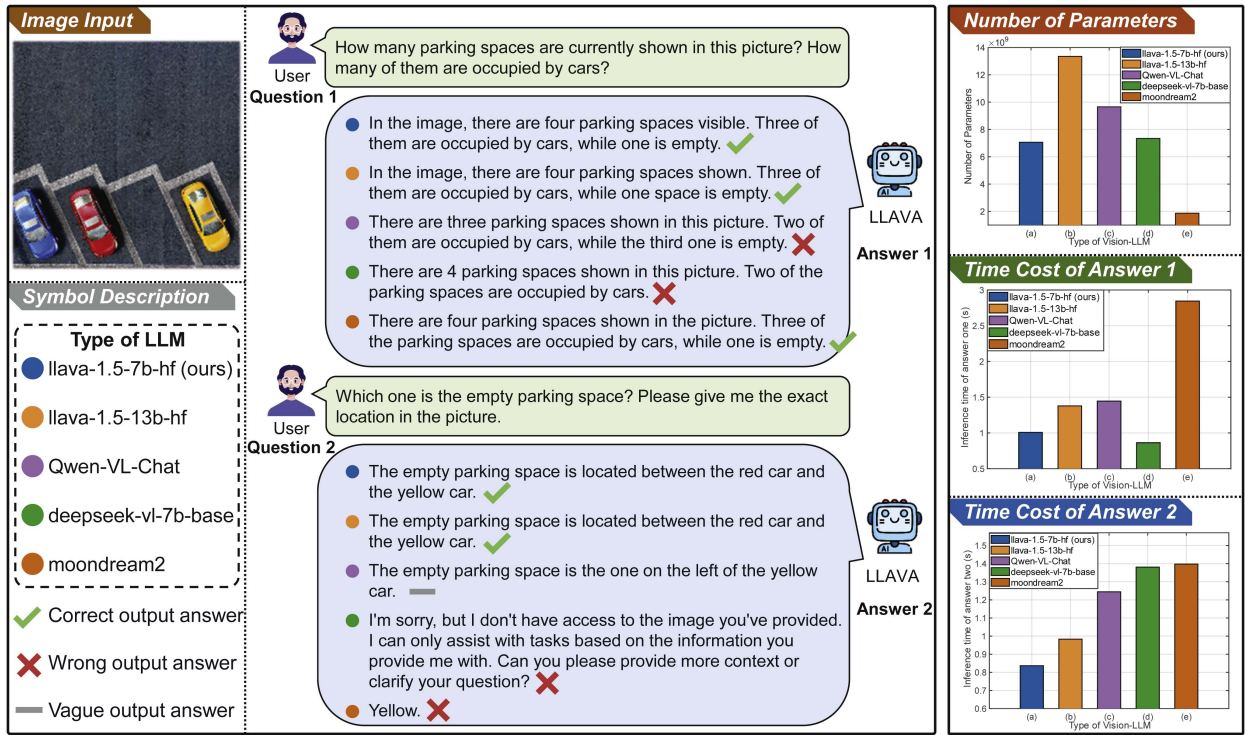


Fig. 6. Comparison of vision-LLM models in terms of parameter size, inference time, and output accuracy for semantic queries. The left side illustrates the semantic query results for two questions related to parking spaces, highlighting the correctness of answers provided by different models. The bar charts on the right compare the parameter sizes and inference times for each model, showing the balance between computational efficiency and output accuracy. All experiments are conducted on a server equipped with an Intel Xeon Platinum 8380 CPU and NVIDIA A100 80 GB GPU.

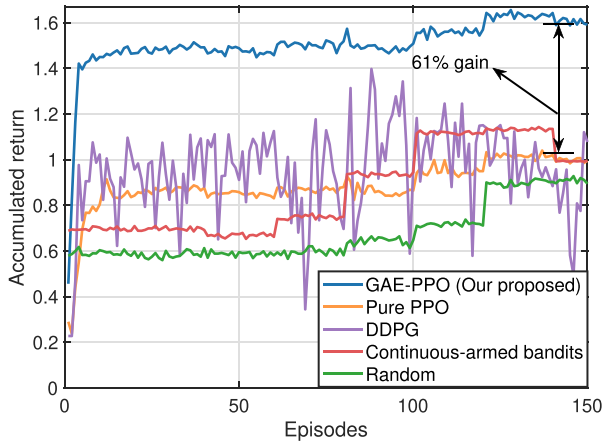


Fig. 7. Convergence behavior with different methods.

After extracting semantic information, the proposed GAE-PPO method is used to transmit the data. Fig. 7 shows a convergence comparison of our GAE-PPO method against several baselines, i.e., pure PPO, DDPG, continuous-armed bandits, and a random policy. The results indicate that our GAE-PPO method achieves superior performance compared to pure PPO, providing approximately 61% gain in accumulated return. This improvement is attributed to the GAE mechanism, which stabilizes the learning process and results in more effective policy updates. Compared to DDPG, which demonstrates significant

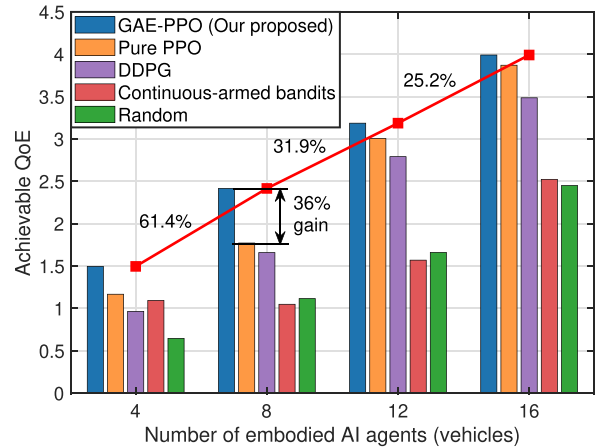


Fig. 8. Achievable QoE versus the number of embodied AI agents.

variability with fluctuations, the method struggles with consistent convergence, indicating its instability and unsuitability for reliable performance. Furthermore, the continuous-armed bandit baseline performs slightly better than random exploration but still suffers from slow and unstable convergence, confirming that it fails to capture temporal dependencies required for sequential decision-making. In contrast, our GAE-PPO method leverages generalized advantage estimation, which significantly contributes to stable convergence behavior, making it more reliable in vehicular network environments.

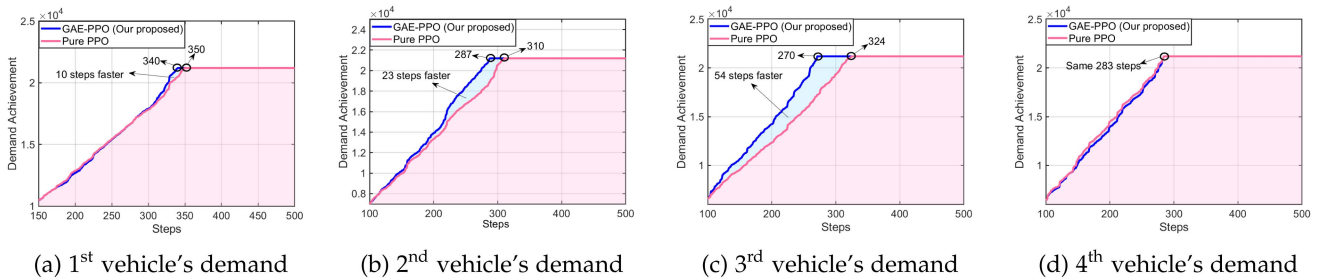


Fig. 9. Comparison of GAE-PPO and pure PPO in fulfilling communication demands.

Fig. 8 shows the achievable QoE performance of different transmission strategies as the number of embodied AI agents (i.e., vehicles) varies. The proposed GAE-PPO method consistently achieves superior performance compared to DDPG, pure PPO, continuous-armed bandits, and a random strategy across all scenarios. Specifically, our GAE-PPO method demonstrates a 36% gain in QoE over DDPG at a vehicle count of 8, highlighting its effectiveness in dynamic vehicular environments. The red line in Fig. 8 illustrates the QoE improvements as the number of vehicles increases. Notably, we observe a 61.4% increase in QoE when scaling from 4 to 8 vehicles, followed by a 31.9% improvement from 8 to 12 vehicles and a 25.2% increase when moving from 12 to 16 vehicles. The continuous-armed bandit baseline yields limited QoE improvements and shows diminished scalability, which further confirms the necessity of modeling the problem as a Markov Decision Process with sequential learning and adaptive policy updates. These results demonstrate the scalability of our GAE-PPO method, as it maintains consistent performance gains even as the network size grows. The observed improvements can be largely attributed to the use of GAE, which provides more accurate advantage estimation, leading to smoother learning and better resource utilization.

To understand the advantages of our proposed method, Fig. 9 illustrates the successfully delivered payloads for each vehicle of GAE-PPO and pure PPO methods across four embodied AI agents (i.e., vehicles), highlighting the number of steps each method requires to meet communication demands. It shows that GAE-PPO consistently outperforms pure PPO by achieving faster convergence for most vehicles. Specifically, GAE-PPO reduces the required steps by 10, 23, and 54 for vehicles 1, 2, and 3, respectively. For vehicle 4, both methods achieve the communication demand in the same number of steps, indicating comparable performance in this scenario. The efficiency of GAE-PPO is attributed to its generalized advantage estimation mechanism, which stabilizes the training process and enhances the balance between exploration and exploitation. This enables GAE-PPO to adapt effectively to dynamic vehicular communication environments, prioritizing optimal actions that lead to faster demand fulfillment.

To explore the impact of dataset size, Fig. 10 illustrates the relationship between the SSIM, the SINR, and the semantic sentence length. It shows that SSIM is significantly influenced by both SINR and sentence length, reflecting how these factors

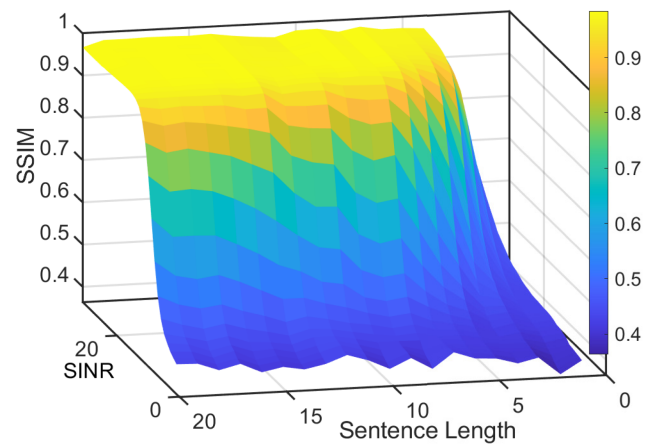


Fig. 10. The SSIM versus different words sentence length and SINR.

affect the quality of semantic communication in the vehicular network. At higher SINR values, the SSIM remains consistently high regardless of the sentence length, indicating that a reliable channel effectively preserves the quality of transmitted semantic information, even with longer sentences. However, as SINR decreases, we observe a notable drop in SSIM, particularly for longer sentences. This suggests that in low-SINR environments, transmitting longer semantic information leads to a greater degradation in quality due to increased noise interference. Moreover, it also indicates that shorter sentence lengths tend to maintain higher SSIM values across varying SINR conditions. This highlights the importance of optimizing sentence length to ensure reliable communication quality, especially in challenging network conditions where SINR is low. These observations underscore the trade-off between semantic content richness and communication quality and suggest that dynamically adapting sentence length based on SINR can be an effective strategy to enhance the performance of embodied AI-enhanced vehicular networks.

VI. CONCLUSION

In this paper, we have proposed a framework for optimizing transmission and decision-making in embodied AI-enhanced vehicular networks. By integrating LLAVA for semantic information extraction and GAE-PPO for stable decision-making, the proposed method can improve transmission efficiency and

adaptability in dynamic vehicular environments. We have also utilized attention maps from LLAVA to analyze how semantic information is extracted, demonstrating LLAVA's ability to focus on the most relevant visual regions for accurate semantic representation. The results have demonstrated that our method consistently outperforms traditional methods in terms of convergence speed, transmission efficiency, and overall system performance, validating the effectiveness of using embodied AI framework in vehicular networks.

REFERENCES

- [1] S. B. Prathiba, G. Raja, S. Anbalagan, A. K. S, S. Gurumoorthy, and K. Dev, "A hybrid deep sensor anomaly detection for autonomous vehicles in 6G-V2X environment," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 3, pp. 1246–1255, May/June 2023.
- [2] A. Nahar, K. K. Mondal, D. G. Gui, and R. Buyya, "Clouds on the road: A software-defined fog computing framework for intelligent resource management in vehicular ad-hoc networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12778–12792, Dec. 2024.
- [3] L. Wang et al., "Privacy-preserving and secure distributed data sharing scheme for VANETs," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 13882–13897, Dec. 2024.
- [4] S. Liu et al., "FedAGL: A communication-efficient federated vehicular network," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3704–3720, Feb. 2024.
- [5] Z. Wang et al., "A reliable physical layer key generation scheme based on RSS and LSTM network in VANET," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 692–707, Jan. 2024.
- [6] F. Tang, B. Mao, N. Kato, and G. Gui, "Comprehensive survey on machine learning in vehicular network: Technology, applications and challenges," *IEEE Commun. Survey Tut.*, vol. 23, no. 3, pp. 2027–2057, Third Quarter, 2021.
- [7] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied AI: From simulators to research tasks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 2, pp. 230–244, Apr. 2022.
- [8] M. Cunneen, M. Mullins, and F. Murphy, "Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions," *Appl. Artif. Intell.*, vol. 33, no. 8, pp. 706–731, 2019.
- [9] R. Zhang et al., "Toward democratized generative AI in next-generation mobile edge networks," *IEEE Netw.*, early access, Feb. 11, 2025, doi: [10.1109/MNET.2025.3541078](https://doi.org/10.1109/MNET.2025.3541078).
- [10] R. Zhang et al., "Generative AI agents with large language model for satellite networks via a mixture of experts transmission," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3581–3596, Dec. 2024.
- [11] R. Zhang et al., "Covert prompt transmission for secure large language model services," 2025, *arXiv:2504.21311*.
- [12] S. Wang et al., "OmniDrive: A holistic LLM-agent framework for autonomous driving with 3D perception, reasoning and planning," 2024, *arXiv:2405.01533*.
- [13] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, and K. B. Letaief, "Energy efficiency maximization in RIS-assisted SWIPT networks with RSMA: A PPO-based approach," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1413–1430, May 2023.
- [14] R. Zhang et al., "Generative AI-enabled vehicular networks: Fundamentals, framework, and case study," *IEEE Netw.*, vol. 38, no. 4, pp. 259–267, Jul. 2024.
- [15] Z. Shao et al., "Semantic-aware spectrum sharing in internet of vehicles based on deep reinforcement learning," 2024, *arXiv:2406.07213*.
- [16] R. Zhang et al., "Interactive AI with retrieval-augmented generation for next generation networking," *IEEE Netw.*, vol. 38, no. 6, pp. 414–424, Nov. 2024.
- [17] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.
- [18] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4579–4592, Jul. 2018.
- [19] J. Xue, K. Yu, T. Zhang, H. Zhou, L. Zhao, and X. Shen, "Cooperative deep reinforcement learning enabled power allocation for packet duplication URLLC in multi-connectivity vehicular networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 8, pp. 8143–8157, Aug. 2024.
- [20] C. Zhang et al., "Distributed deep reinforcement learning based gradient quantization for federated learning enabled vehicle edge computing," *IEEE Internet Things J.*, vol. 12, no. 5, pp. 4899–4913, Mar. 2025.
- [21] Y. Yao, F. Shu, Z. Li, X. Cheng, and L. Wu, "Secure transmission scheme based on joint radar and communication in mobile vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 10027–10037, Sep. 2023.
- [22] M. Savva et al., "Habitat: A platform for embodied AI research," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347.
- [23] C. H. Song, B. M. Sadler, J. Wu, W. -L. Chao, C. Washington, and Y. Su, "LLM-planner: Few-shot grounded planning for embodied agents with large language models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2998–3009.
- [24] Y. Zhang et al., "Towards efficient LLM grounding for embodied multi-agent collaboration," 2024, *arXiv:2405.14314*.
- [25] C. E. Mower et al., "Ros-LLM: A ROS framework for embodied AI with task feedback and structured reasoning," 2024, *arXiv:2406.19741*.
- [26] H. Zhang et al., "BadRobot: Jailbreaking LLM-based embodied AI in the physical world," 2024, *arXiv:2407.20242*.
- [27] F. Jiang et al., "Large AI model empowered multimodal semantic communications," *IEEE Commun. Mag.*, vol. 63, no. 1, pp. 76–82, Jan. 2025.
- [28] S. Guo, Y. Wang, J. Ye, A. Zhang, P. Zhang, and K. Xu, "Semantic importance-aware communications with semantic correction using large language models," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 232–245, 2025.
- [29] S. Tang, Q. Yang, L. Fan, X. Lei, A. Nallanathan, and G. K. Karagiannis, "Contrastive learning-based semantic communications," *IEEE Trans. Commun.*, vol. 72, no. 10, pp. 6328–6343, Oct. 2024.
- [30] C. Ying et al., "PEAC: Unsupervised pre-training for cross-embodiment reinforcement learning," 2024, *arXiv:2405.14073*.
- [31] Y. Long, W. Wei, T. Huang, Y. Wang, and Q. Dou, "Human-in-the-loop embodied intelligence with interactive simulation environment for surgical robot learning," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4441–4448, Aug. 2023.
- [32] E. Z. Liu et al., "Simple embodied language learning as a byproduct of meta-reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 21997–22008.
- [33] W. Tan et al., "True knowledge comes from practice: Aligning LLMs with embodied environments via reinforcement learning," 2024, *arXiv:2401.14151*.
- [34] B. Shin et al., "LTERSSI based vehicular localization system in long tunnel environment," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 11102–11114, Nov. 2023.
- [35] M. Huang, Z. Shen, and G. Zhang, "Joint spectrum sharing and V2V/V2I task offloading for vehicular edge computing networks based on coalition formation game," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 11918–11934, Sep. 2024.
- [36] F. Liu, J. Chen, Q. Zhang, and B. Li, "Online MEC offloading for V2V networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 6097–6109, Oct. 2023.
- [37] H. Xie, Z. Qin, G. Y. Li, and B. -H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.
- [38] K. Li et al., "Biasing federated learning with a new adversarial graph attention network," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 2407–2421, Mar. 2025.
- [39] R. Cheng, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "A wireless AI-generated content (AIGC) provisioning framework empowered by semantic communication," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 2137–2150, Mar. 2025.
- [40] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment," in *Proc. IEEE Int. Conf. Commun.*, 2010, pp. 1–5.
- [41] H. Du et al., "Attention-aware resource allocation and QoE analysis for metaverse xURLLC services," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2158–2175, Jul. 2023.
- [42] D. Sheng, Q. Qi, J. Wang, L. Li, W. Yu, and J. Liao, "PsyQoE: Improving quality-of-experience assessment with psychological effects in video streaming," *IEEE Trans. Serv. Comput.*, vol. 17, no. 6, pp. 4138–4150, Nov/Dec. 2024.
- [43] C. -W. Xie, S. Sun, X. Xiong, Y. Zheng, D. Zhao, and J. Zhou, "RA-CLIP: Retrieval augmented contrastive language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19265–19274.

- [44] J. Ji, K. Zhu, and L. Cai, "Trajectory and communication design for cache-enabled UAVs in cellular networks: A deep reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 6190–6204, Oct. 2023.
- [45] L. Yang et al., "Constrained update projection approach to safe policy optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 9111–9124.
- [46] C. Li et al., "Dynamic offloading for multiuser multi-CAP MEC networks: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2922–2927, Mar. 2021.
- [47] B. Y. Yachour, T. Ahmed, and M. Mosbah, "Efficient DRL-based selection strategy in hybrid vehicular networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 20, no. 3, pp. 2400–2411, Sep. 2023.
- [48] B. Tian et al., "UAV-assisted wireless cooperative communication and coded caching: A multiagent two-timescale DRL approach," *IEEE Trans. Mobile Comp.*, vol. 23, no. 5, pp. 4389–4404, May 2024.
- [49] R. Girdhar et al., "ImageBind: One embedding space to bind them all," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15180–15190.
- [50] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.
- [51] A. Vaswani, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [52] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1274–1285, Jun. 2020.
- [53] D. Abode et al., "Unsupervised graph-based learning method for sub-band allocation in 6G subnetworks," in *Proc. IEEE 100th Veh. Technol. Conf.*, 2024, pp. 1–6.
- [54] S. Liu, F. Dai, and H. Cui, "Joint resource allocation and interference management for vehicular networks," in *Proc. Int. Conf. Power, Commun., Comput. Netw. Technol.*, 2023, pp. 1–5.
- [55] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proc. Mach. Transl. Summit X: Papers*, 2005, pp. 79–86.
- [56] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26296–26306.
- [57] X. He, Y. Yang, J. Lee, G. He, and Q. Yan, "Deep reinforcement learning based AoI minimization for NOMA-enabled integrated satellite-terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 74, no. 2, pp. 3567–3572, Feb. 2025.
- [58] T. Zhang, K. -Y. Lam, J. Zhao, and J. Feng, "Joint device scheduling and bandwidth allocation for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 3–18, Jan. 2025.
- [59] B. Lin et al., "Video-LLaVA: Learning united visual representation by alignment before projection," 2023, *arXiv:2311.10122*.
- [60] J. Bai et al., "Qwen-VL: A frontier large vision-language model with versatile abilities," 2023, *arXiv:2308.12966*.
- [61] H. Lu et al., "Deepseek-VL: Towards real-world vision-language understanding," 2024, *arXiv:2403.05525*.
- [62] R. Murthy et al., "Mobileaiibench: Benchmarking LLMs and LMMs for on-device use cases," 2024, *arXiv:2406.10290*.



Ruichen Zhang (Member, IEEE) received the BE degree from Henan University (HENU), China, in 2018, and the PhD degree from Beijing Jiaotong University (BJTU), China, in 2023. He is currently a post-doctoral research fellow with the College of Computing and Data Science, Nanyang Technological University (NTU), Singapore. In 2024, he was a visiting scholar with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. His research interests include the Internet of Agents, LLM-empowered network-

working, reinforcement learning-enabled wireless communication, generative AI models, and heterogeneous networks. He has received two Best Paper Awards from IEEE IWCNC and has two ESI Highly Cited Papers and one ESI hot papers. He also serves as a guest editor for several IEEE journals, including *IEEE Transactions on Cognitive Communications and Networking*, *IEEE Transactions on Artificial Intelligence*, and *IEEE Journal on Selected Areas in Sensors*. In addition, he serves as the managing editor of the *IEEE Transactions on Network Science and Engineering (TNSE)* starting from 2025.



Changyuan Zhao received the BSc degree in computing and information science from the University of Science and Technology of China, Hefei, China, in 2020, and the MAEng degree in computer science from the Institute of Software, CAS, Beijing, China, in 2023. He is currently working toward the PhD degree with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include generative AI, communication security, and resource allocation.



Hongyang Du (Member, IEEE) received the BEng degree from the Beijing Jiaotong University, China, and the PhD degree from the Nanyang Technological University, Singapore. He is an assistant professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong, where he directs the Network Intelligence and Computing Ecosystem (NICE) Laboratory. He serves as the editor-in-chief assistant (2022–2024) and editor (2025–Present) of *IEEE Communications Surveys & Tutorials*, the editor of *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Technology*, *IEEE Open Journal of the Communications Society*, and the guest editor for *IEEE Vehicular Technology Magazine*. He is the recipient of the IEEE ComSoc Young Professional Award for Best Early Career Researcher in 2024, IEEE Daniel E. Noble Fellowship Award from the IEEE Vehicular Technology Society in 2022, the IEEE Signal Processing Society Scholarship from the IEEE Signal Processing Society in 2023, the Singapore Data Science Consortium (SDSC) Dissertation Research Fellowship in 2023, and NTU Graduate College's Research Excellence Award in 2024. He was recognized as an exemplary reviewer of the *IEEE Transactions on Communications* and *IEEE Communications Letters*. His research interests include edge intelligence, generative AI, and network management.



Dusit Niyato (Fellow, IEEE) received the BEng degree from the King Mongkuts Institute of Technology Ladkrabang (KMUTL), Thailand and the PhD degree in electrical and computer engineering from the University of Manitoba, Canada. He is a professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests are in the areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.



Jiacheng Wang (Member, IEEE) received the MS and PhD degrees from the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, in 2018 and 2022, respectively. From 2021 to 2022, he was a visiting researcher with the College of Computing and Data Science, Nanyang Technological University, Singapore, where he is now a postdoctoral research fellow. His research interests include generative AI, integrated sensing and communications, network optimization, and edge intelligence. He has published more than 40 papers, including in *IEEE Journal on Selected Areas in Communications*, *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Cognitive Communications and Networking*, *IEEE Transactions on Vehicular Technology*, *IEEE CMOST*, *IEEE Wireless Communications*, *IEEE Network*, *IEEE Wireless Communications Letters*, *IEEE GLOBECOM*, *IEEE ICC*, and *IEEE WCNC*. He is the lead contributor to the IEEE ComSoc Best Readings on Generative AI and Large Language Models for Networking and has served as a guest editor for *IEEE WCM*, *IEEE Transactions on Cognitive Communications and Networking*, *IEEE Transactions on Network Science and Engineering*, *IEEE Open Journal of the Communications Society*, *IoT Magazine*, and *Networking Letters*.



Sutfinee Sawadsitang received the BEng degree in computer engineering from the King Mongkut's University of Technology Thonburi, Bangkok, Thailand, in 2012, the MEng degree from Shanghai Jiao Tong University, Shanghai, China, in 2015, and the PhD degree from the SIMTech-NTU Joint Lab on Complex Systems, Nanyang Technological University, Singapore, in 2020. She is currently a lecturer with College of Arts, Media and Technology, Chiang Mai University, Chiang Mai University, Thailand. Her research interests include transportation systems, operations research, optimization, and high-performance computing.



Xuemin (Sherman) Shen (Fellow, IEEE) received the PhD degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is a university professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. He is a registered professional engineer of Ontario, Canada, an Engineering Institute of Canada fellow, a Canadian Academy of Engineering fellow, a Royal Society of Canada fellow, a Chinese Academy of Engineering Foreign member, and an international fellow with the Engineering Academy of Japan. He received the "West Lake Friendship Award" from Zhejiang Province in 2023, President's Excellence in Research from the University of Waterloo in 2022, the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R.A. Fessenden Award in 2019 from IEEE Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from the Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He serves/served as the general chair for the 6G Global Conference 2023, and ACM Mobihoc 2015, Technical Program Committee chair/co-chair for IEEE Globecom 2024, 2016 and 2007, IEEE Infocom 2014, IEEE VTC 2010 Fall, and the chair for the IEEE ComSoc Technical Committee on Wireless Communications. He is the past president of the IEEE ComSoc, the vice president for Technical & Educational Activities, vice president for Publications, member-at-large on the Board of Governors, chair of the Distinguished Lecturer Selection Committee, and member of IEEE fellow Selection Committee of the ComSoc. He served as the editor-in-chief of the *IEEE Internet of Things Journal*, *IEEE Network*, and *IET Communications*.



Dong In Kim (Life Fellow, IEEE) received the PhD degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1990. He was a tenured professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a distinguished professor with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea. He was the first recipient of the NRF of Korea Engineering Research Center in Wireless Communications for RF Energy Harvesting from 2014 to 2021. He received several research awards, including the 2023 IEEE ComSoc Best Survey Paper Award and the 2022 IEEE Best Land Transportation Paper Award. He was selected the 2019 recipient of the IEEE ComSoc Joseph LoCicero Award for Exemplary Service to Publications. He has been listed as a Highly Cited Researcher by Clarivate Analytics in 2020 and 2022. He was the general chair of the IEEE ICC 2022, Seoul. Since 2001, he has been serving as an editor, an editor at large, and an area editor of Wireless Communications I for *IEEE Transactions on Communications*. From 2002 to 2011, he served as an editor and a founding area editor of Cross-Layer Design and Optimization for *IEEE Transactions on Wireless Communications*. From 2008 to 2011, he served as the co-editor-in-chief for the *IEEE/KICS Journal of Communications and Networks*. He served as the founding editor-in-chief for the *IEEE Wireless Communications Letters* from 2012 to 2015. He is a fellow of the Korean Academy of Science and Technology and a member of the National Academy of Engineering of Korea.