scientific reports

Check for updates

OPEN

A cognitive internet of things resource allocation method based on multi-agent reinforcement learning algorithm

Rong Wang, Yanjin Shen, Dongtao Wang✉ & Wan Li✉

This paper addresses the challenges of inter-vehicle communication, taking into consideration the stochastic nature of primary user spectrum occupancy, the highly dynamic fluctuation of channel states, and the timeliness requirements for communication among vehicles. The study investigates the joint channel selection and power control resource allocation problem in cognitive Internet of Things (CIoT) under high-speed mobility, with the aim of minimizing the system's Age of Information (AoI). The presented problem is modeled as a Markov Decision Process (MDP) and incorporates a meticulously designed reward function. Furthermore, to meet the timeliness demands, a multi-agent reinforcement learning approach is employed, with vehicles serving as intelligent agents that gather localized observational information and directly determine their transmission strategies. An improved Multi-agent Proximal Policy Optimization (IMAPPO) algorithm is proposed, which is based on a centralized training and distributed execution framework. Enhancements to the Actor network within the algorithm enable it to address the challenges presented by the discrete-continuous hybrid action space. Finally, the feasibility and effectiveness of the enhanced multi-agent proximal policy optimization algorithm are verified through simulations. The results demonstrate that compared to alternative approaches, the CIoT resource allocation scheme based on the improved multi-agent proximal policy optimization algorithm significantly reduces the AoI for vehicle users.

Keywords Cognitive vehicular networking, Age of information, Resource allocation, Multi-agent reinforcement learning

With the rapid advancement of the sixth-generation (6G) wireless communication technology,the Internet of Things (IoT) has increasingly emerged as an innovative paradigm for connecting numerous smart devices to provide various services[1]. Currently,6G mobile networks,such as those supporting 6G-enabled IoT,have been developed to realize high-quality communication for massive wireless devices[2]. Vehicle communication,commonly referred to as Vehicle-to-Everything(V2X)communication,is an integral part of the IoT[3]. It serves as an enabling technology for autonomous driving and intelligent vehicles,providing various vehicular data services. Spectrum access design in vehicular networks typically includes Vehicle-to-Infrastructure(V2I)and Vehicle-to-Vehicle(V2V)connections. It is a key technology that enhances transportation through collaborative support among nearby vehicles,offering satisfactory Quality of Service (QoS). V2X communication has been deemed indispensable for improving road safety,traffic efficiency,and the overall vehicular entertainment experience,providing wireless connectivity between vehicles and road infrastructure[4,5]. Yang et al.[6] introduced a technique based on IEEE 802.11p inter-vehicle cooperative channel estimation to acquire accurate Channel State Information (CSI) in V2X networks, aiming to improve the transmission of safety-critical data. This issue has been addressed by the 3rd Generation Partnership Project(3GPP)standard,which supports various QoS requirements for V2X networks and utilizes Device-to-Device (D2D) communication in Long-Term Evolution (LTE) and 5G cellular networks.

Therefore, our focus is on resource allocation(RA)in vehicular networks based on the 3GPP standard,including shared spectrum and the use of both PC5 and Uu radio interfaces for V2V and V2I links,respectively[7]. Anticipated future applications and entertainment related to traffic are expected to be facilitated by vehicles[8]. However,this

Hunan Automotive Engineering Vocational University, Zhuzhou 412000, China. ✉email: wang1047651955@163.com; 230116002@fzu.edu.cn

necessitates unrestricted internet access through high-capacity V2I links and the instantaneous transmission of safety-critical messages to neighboring vehicles via V2V communication in a reliable manner.

Traditionally developed mathematical approaches for vehicle communication systems primarily rely on the assumption of low mobility or static environments[9]. Therefore,it is imperative to develop new solutions for RA that can interact with rapidly changing environments and achieve optimal decisions for highly mobile vehicle systems.Although a considerable amount of literature has applied traditional optimization methods to address similar V2X resource allocation problems, they have encountered significant challenges in fully resolving these issues in two aspects.On the one hand, the rapidly changing channel conditions in the vehicular environment lead to significant uncertainty in the acquired Channel State Information(CSI), resulting in resource allocation uncertainty.On the other hand,it is challenging to mathematically express the maximization of throughput and reliability for the combination of V2X flows, let alone to find a systematic approach to the optimal solution.

Fortunately, the recent success of Deep Reinforcement Learning (DRL) in achieving human-level performance in video games[10] and AlphaGo[11] has sparked significant interest in applying reinforcement learning techniques to address problems in various domains, yielding notable progress since then[12–14]. DRL provides a robust and principled approach to deal with dynamic environments and execute sequential decisions under uncertainty, representing a unique and challenging approach to dynamic resource allocation in V2X. Additionally, by designing training rewards relevant to the ultimate objective, challenging optimization problems can be effectively addressed within the reinforcement learning framework[15]. The learning algorithm can autonomously devise an intelligent strategy to approach the ultimate goal. Another potential advantage of using DRL for resource allocation is that it enables the possibility of distributed algorithms. In this work, we investigate the use of multi-agent reinforcement learning tools to address the V2X spectrum access problem, treating each V2V link as an agent, learning to refine its resource-sharing strategy by interacting with the unknown vehicular environment, thus obtaining the optimal decision for resource allocation.

Related works

Addressing the challenges posed by V2V spectrum allocation, data allocation, and task assignment in the high-speed mobile Cognitive Internet of Things (CIoT). Reference[16] combined CR with reactive diffusion biological mechanisms, proposing a cluster-based distributed spectrum allocation algorithm for CIoT to determine the optimal cluster size, maximizing cluster throughput and minimizing communication latency. In[17], the authors presented a multi-band cooperative spectrum sensing and resource allocation framework for IoT based on cognitive 5G networks, significantly reducing energy consumption for spectrum sensing. Most of the research on Cognitive IoT mentioned above focuses on static application scenarios, typically assuming that IoT devices in these scenarios are stationary or exhibit slight mobility. Addressing the time-varying characteristics of V2V communication channels, Ahsan et al.[18] developed an intelligent resource allocation scheme for non-orthogonal multiple access (NOMA) IoT communication in the uplink, designing SARSA-Learning and deep reinforcement learning algorithms. Allahham et al.[19] proposed a Multi-Agent Deep Reinforcement Learning (MADRL) algorithm based on a distributed framework for dynamic network selection and resource allocation in the edge layer, aiming to enhance QoS for edge nodes and extend node battery life.

However, due to dynamic wireless environment variations and the rapid movements of network users, these methods face challenges such as non-convexity and non-global optimal solutions. Fortunately, reinforcement learning algorithms have garnered significant attention in the research on resource allocation in wireless communication systems. In[20], the authors explored the social-aware networking model of Cognitive IoT networks and introduced a novel coordinated resource management method based on Multi-Agent Deep Reinforcement Learning (MADRL) to optimize joint radio block allocation and transmission power control strategies. Huang et al.[21] applied reinforcement learning to joint relay selection and power allocation in secure cognitive radio relay networks, using Double Deep Q-Network (DDQN) to address the problems of maximizing throughput and maximizing confidentiality. Liu et al.[22] proposed a dynamic spectrum access scheme based on a collaborative large-scale data perception and Q-Learning algorithm to improve spectrum utilization in Cognitive Vehicular Networks and prevent harmful interference to Primary Users (PUs). Considering the timeliness of information transmission among vehicles, traditional performance metrics such as transmission delay do not effectively characterize the freshness of system information. To measure information freshness, Sanjit Kaul et al. described a Carrier Sense Multiple Access (CSMA) mechanism-based vehicular network in Kaul et al.[23], introducing the concept of Age of Information (AoI) for the first time, analyzing the performance of the vehicular network using AoI as a novel metric. Additionally, Sanjit Kaul utilized the average AoI value as an indicator to evaluate the real-time state update system performance in Kaul et al.[24]. They derived a general expression for the average AoI in a first-come, first-served queue and optimized the AoI values for three queuing systems individually. Yu et al.[25] proposed a novel prediction-based CIoT state update scheme, studying the trade-off between power consumption and state update performance, designing an optimization algorithm to reduce the Average Age of Information (AoI) by jointly adjusting transmission power and prediction range. Lin et al.[26] presented a dynamic beamforming mode selection and bandwidth allocation scheme based on a cooperative Multi-Agent Deep Reinforcement Learning (MADRL) algorithm, where each intelligent agent is responsible for either beam illumination or bandwidth allocation.

In recent years, leveraging multi-agent reinforcement learning (MARL) to optimize the Age of Information (AoI) has become a research hotspot in the field of wireless communications. Researchers have successfully applied MARL methods to address information freshness issues in various network environments. For instance, Wang et al.[27] investigated an AoI-oriented link scheduling strategy in Device-to-Device (D2D) communications. Liu et al.[28] utilized MARL to solve a priority-aware AoI resource allocation problem in Wi-Fi networks. These efforts have been extended to scenarios with higher mobility, such as the work by Shi et al.[29] , which combined AoI-aware data collection with energy replenishment in UAV-enabled IoT (UAV-IoT) systems. These studies

collectively demonstrate the significant potential of MARL in managing information freshness across diverse network settings. Regarding resource allocation, the work by Shoaib et al.[30] successfully achieved decentralized resource allocation in UAV communication networks through reward-based multi-agent learning, which aligns with the distributed decision-making philosophy in our study. Furthermore, the research by Mohiuddin et al.[31] employed RL for end-to-end UAV navigation and obstacle avoidance, further showcasing the effectiveness of RL methods in handling complex dynamic environments and high-mobility agent control problems.

While the aforementioned studies have made significant progress, they have not fully considered the unique challenges posed by the specific scenario of Cognitive Internet of Things (CIoT) for vehicular networks. Compared to UAV or general D2D networks, the CIoT environment is characterized by: 1) stricter spectrum sharing constraints, which necessitate the constant avoidance of harmful interference to Primary Users (PUs); 2) more complex and rapid channel fading patterns caused by ground traffic and urban building obstructions; and 3) more frequent and unpredictable changes in the interaction topology among vehicles. Therefore, this research aims to fill this gap by proposing a MARL-based resource allocation method specifically tailored for the CIoT environment, capable of handling a hybrid action space while optimizing AoI.

In this paper, we employ a multi-agent reinforcement learning approach, with the vehicle users acting as intelligent agents, grouping all V2V decisions of vehicles into a single decision to address the scalability issue, collecting local observational information, and directly determining their transmission strategies. We propose an Improved Multi-Agent Proximal Policy Optimization (IMAPPO) algorithm based on a centralized training and distributed execution structure, enhancing the Actor network within the algorithm to handle the mixed discrete and continuous action space problem.

Contribution

In this paper, we propose an Improved Multi-Agent Proximal Policy Optimization (IMAPPO) algorithm based on a centralized training and distributed execution structure to address the challenges of joint channel selection and power control resource allocation in high-speed mobile Cognitive Internet of Things (CIoT).

The main contributions of this paper are as follows:

- Research conducted on inter-vehicle communication and resource allocation, where we model the formulated problem as a Markov Decision Process (MDP), a mathematical framework for modeling stochastic decision problems. We design a reward function incorporating effective interrupt thresholds to assist in making decisions in highly dynamic environments, ensuring reliability requirements.
- To meet the timeliness requirements of communication, we employ a multi-agent reinforcement learning approach, where vehicle users are considered agents capable of autonomously determining transmission strategies based on local observational information. To optimize this multi-agent system, we propose an Improved Multi-Agent Proximal Policy Optimization (IMAPPO) algorithm, which utilizes a centralized training and distributed execution structure. The enhanced Actor network within the algorithm addresses the mixed discrete and continuous action space problem.
- The effectiveness of the improved Multi-Agent Proximal Policy Optimization (IMAPPO) algorithm is demonstrated through simulation experiments. The simulation results reveal that, compared to other comparative schemes, this CIoT resource allocation solution can significantly reduce the Age of Information (AoI) for vehicle users, contributing to improved communication efficiency and timeliness.

System model and problem description

Network model

As shown in Fig 1, this paper considers a CIoT network model with a grid-like urban road layout, which includes a Primary Base Station (PBS), a Cognitive Base Station (CBS), K PUs, and L Vehicular User Pairs (VUPs). In this model, each PU corresponds to occupying an orthogonal channel (i.e., there are K orthogonal channels in total) and communicates wirelessly with the PBS.

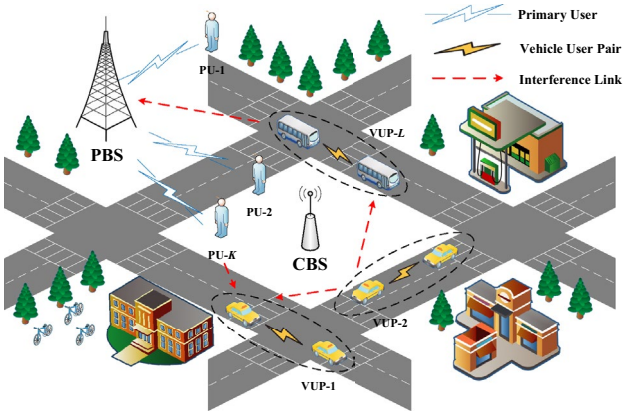

Fig. 1. CIoT network model with grid-like urban roads.

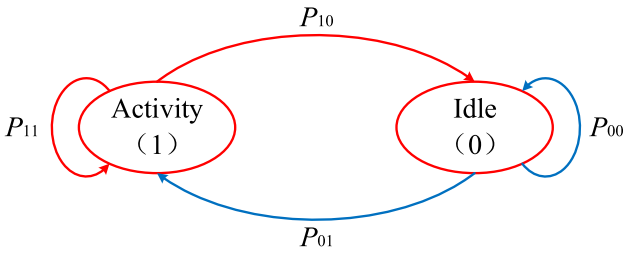

Fig. 2. Main user state transition diagram.

Symbol	Value
K	Number of primary users/Number of orthogonal channels
L	Number of vehicle user pairs
B	Spectral bandwidth
δ^2	Noise power
$PU-k$	The k-th primary user
$VUP-l$	The -th vehicle user pair
P_P	Transmission power of the primary user
$\bar{P}_l(t)$	Transmission power of the transmitter of the -th vehicle user pair in the t-th time slot
$P_{\max}$	Maximum transmission power of the vehicle user
$\bar{g}_{VUP-l}(t)$	Channel gain between the transmitter and receiver of $VUP-l$ in the t-th time slot
$\bar{g}_{VUP-l,PBS}(t)$	Gain from the transmitter of $VUP-l$ to PBS in the t-th time slot
$\bar{g}_{PU-k,VUP-l}(t)$	Gain from $PU-k$ to the receiver of $VUP-l$ in the t-th time slot
$\bar{g}_{VUP-l',VUP-l}(t)$	Gain from the transmitter of $VUP-l'$ to the receiver of $VUP-l$ in the t-th time slot

Table 1. System symbols.

As depicted in Fig 2, the state of the PU is described by a two-state discrete-time Markov chain process[26]. In a given time slot, the PU may be in one of the following two states: active (state 1) and idle (state 0). Here, P_{10} represents the probability of the PU transitioning from the active state to the idle state, while P_{11} denotes the probability of the PU remaining in the active state, where P_{11} =1-P_{10}. Similarly, P_{01} represents the probability of the PU transitioning from the idle state to the active state, while P_{00} denotes the probability of the PU remaining in the idle state, where P_{00} =1-P_{01}.

Each VUP is composed of a vehicle transmitter and a corresponding vehicle receiver for mutual communication, with each vehicle traveling in a straight line at a certain speed on the respective lane, immediately reversing direction when reaching the end of the road. VUPs access orthogonal channels in an Underlay spectrum sharing mode, with each VUP accessing a maximum of one channel in a time slot to complete its data transmission. To ensure the Quality of Service (QoS) for PUs, VUPs need to limit their transmission power when accessing channels occupied by PUs to avoid causing harmful interference. When accessing unoccupied channels, VUPs do not need to consider interference to PUs. Table 1 summarizes the main symbols in the system model.

As shown in Fig 3, the network model is divided into three stages in each time slot: the sensing stage, the decision stage, and the transmission stage.

During the sensing stage, the CBS performs spectrum sensing to acquire the status of each orthogonal channel. For simplicity, perfect spectrum sensing is assumed. A binary indicator function is defined $w_k(t)$ to represent the channel occupancy status:

$$\overline{w}_k(t) = \begin{cases} 0, & \text{Unoccupied} \\ 1, & \text{Occupied} \end{cases} \tag{1}$$

Here, $w_k(t) = 0$ represents that the k-th channel is unoccupied by the primary user (i.e., the primary user is idle) at the t-th time slot, and $w_k(t) = 1$ represents that the k-th channel is occupied by the primary user (i.e., the primary user is active) at the t-th time slot.

In the decision phase, VUPs obtain the channel gains of their relevant links and the status of each orthogonal channel broadcasted by the CBS through specific control channels. Subsequently, VUPs determine the access channels and transmission power of their transmitters based on the status of the orthogonal channels and the channel gains of their links.

$\overline{C}_l(t) \in \{0, 1, ..., k, ..., K\}$ denotes the selected channel accessed by the transmitter of VUP-l at the t-th time slot. Here, $\overline{C}_l(t) = 0$ indicates that the transmitter of VUP-l does not access any channel at the t-th time slot, while $\overline{C}_l(t) = K$ indicates that the transmitter of VUP-l accesses the k-th channel at the t-th time slot.

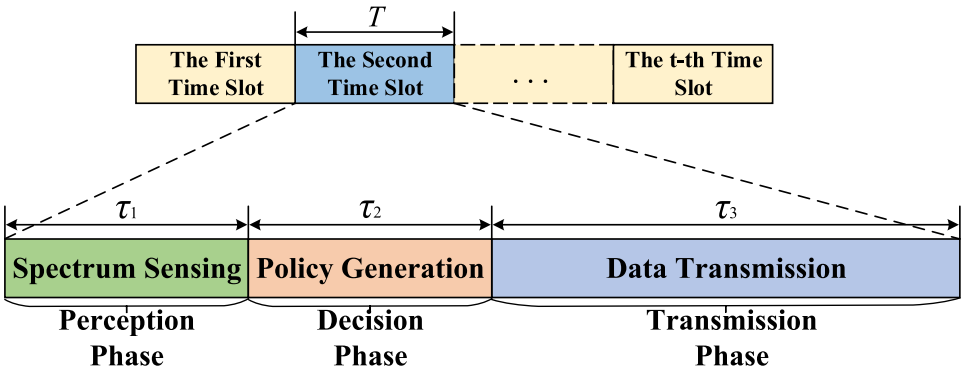

Fig. 3. Time slot model.

Therefore, the set of channel selection strategies for all transmitters of VUPs at the *t*-th time slot is denoted as $\overline{C}(t) = \{\overline{C}_l(t), l = 1, 2, ..., L\}$. $\overline{P}_l(t) \in [0, P_{\max}]$ represents the transmission power of the transmitter of VUP l at the *t*-th time slot, where $P_{\max}$ is the maximum transmission power for these transmitters. Thus, the set of power control strategies for all transmitters of VUPs at the th time slot is represented as $\overline{P}(t) = \{\overline{P}_l(t), l = 1, 2, ..., L\}$ We define an indicator function $\overline{z}_{k,l}(t)$ to represent the access status of the channels by the transmitter of VUP:

$$\overline{z}_{k,l}(t) = \begin{cases} 0, & \text{Not access} \\ 1, & \text{Access} \end{cases} \tag{2}$$

where $\overline{z}_{k,l}(t) = 0$ represents that the transmitter of VUP-l does not access the k-th channel in the t-th time slot, and $\overline{z}_{k,l}(t) = 1$ represents that the transmitter of VUP-l accesses the k-th channel in the t-th time slot. During the transmission phase, the transmitters of each VUP perform data transmission according to their own transmission policies.

Wireless transmission model

Assuming that the channel follows quasi-static block fading, with one block fading coinciding with a single time slot. Furthermore, to make the channel model more realistic, the proposed model considers both large-scale fading and small-scale fading. Large-scale fading is caused by path loss and shadow fading. For example, $\overline{g}_{VUP-l}(t)$ can be represented as:

$$\overline{g}_{VUP-l}(t) = \left|\overline{h}_{VUP-l}(t)\right|^2 \overline{L}_{VUP-l}(t) \tag{3}$$

where $\overline{L}_{VUP-l}(t) = \overline{G}\overline{\xi}_{VUP-l}(t)/\overline{d}_{VUP-l}(t)^{\alpha}$ represents the large-scale fading channel gain between the transmitter and receiver of VUP-l in the t-th time slot, $\bar{G}$ is the path loss constant, $\xi_{VUP-l}(t)$ is the log-normal shadow fading gain with a logarithmic normal distribution, α is the path loss exponent, $\bar{d}_{VUP-l}(t)$ is the distance between the transmitter and receiver of VUP-l in the t-th time slot, and $\bar{h}_{VUP-l}(t)$ represents the small-scale Rayleigh fading between the transmitter and receiver of VUP-l in the t-th time slot.

Therefore, the Signal-to-Interference-plus-Noise Ratio (SINR) between the transmitter and receiver of VUP-l in the t-th time slot can be expressed as:

$$SINR_{k,l}(t) = \frac{\overline{z}_{k,l}(t)\overline{P}_l(t)\overline{g}_{VUP-l}(t)}{\overline{I}_{PU-k}(t) + \overline{I}_{VUP-l'}(t) + \delta^2} \tag{4}$$

where:

$$\overline{I}_{PU-k}(t) = \overline{w}_k(t)\overline{P}_P(t)\overline{g}_{PU-k,VUP-l}(t) \tag{5}$$

$$\overline{I}_{VUP-l'}(t) = \sum_{l' \neq l}^{L} \overline{z}_{k,l'}(t)\overline{P}_{l'}(t)\overline{g}_{VUP-l',VUP-l}(t) \tag{6}$$

$\bar{I}_{PU-k}(t)$ represents the interference to VUP-l from the PUs sharing the same channel, and $\bar{I}_{VUP-l'}(t)$ represents the interference to VUP-l from other VUPs sharing the same channel. According to Shannon's capacity theorem, the transmission rate of VUP-l in the t-th slot can be expressed as:

$$\overline{R}_l(t) = B\sum_{k=1}^{K} \overline{z}_{k,l}(t)\log_2\left(1 + SINR_{k,l}(t)\right) \tag{7}$$

where B represents the bandwidth.

Age of information model

Let $E_l(t)$ represent the residual load at the end of the t-th slot. The update formula for the residual load from the end of slot t to the end of slot t-1 is given as follows:

$$E_l(t) = \begin{cases} E, & \frac{E_l(t-1)}{\bar{R}_l(t)} \leq \tau_3 \\ E_l(t-1) - \bar{R}_l(t)\tau_3, & \frac{E_l(t-1)}{\bar{R}_l(t)} > \tau_3 \end{cases} \tag{8}$$

where E represents the initial load of the VUP, and τ_3 represents the duration of data transmission in each slot. If the current residual load is successfully transmitted in slot VUP-l, $E_l(t)$ will be updated to E. Otherwise, the successfully transmitted load $E_l(t)$ will be subtracted from it.

Let $A_l(t)$ represent the Age of Information (AoI) at the end of slot VUP-l. According to the aforementioned update rule for the residual load, the update formula for the AoI metric from the end of slot $\mathrm{t}-1$ to the end of slot is as follows:

$$A_l(t) = \begin{cases} T, & \frac{E_l(t-1)}{\bar{R}_l(t)} \leq \tau_3 \\ A_l(t-1) + T, & \frac{E_l(t-1)}{\bar{R}_l(t)} > \tau_3 \end{cases} \tag{9}$$

It can be observed that if VUP-l successfully transmits the current residual load in slot t, it will be updated to T at the end of slot t. Otherwise, $A_l(t)$ will increase by T.

Problem description

As stated above, the primary objective of this paper is to select the optimal access channel and transmission power for VUP in high-speed mobile CIoT, minimizing the AoI of VUP to ensure timely communication between vehicles. Additionally, considering the requirements for energy conservation and cost reduction, the transmission power of VUP should be controlled to improve energy efficiency and system reliability. Therefore, a utility function is constructed that simultaneously considers the AoI and transmission power. The utility function represents the weighted sum of the system's AoI and transmission power, i.e., the sum of the weighted sum of AoI and transmission power for all VUP, as follows:

$$U(t) = \sum_{l=1}^{L} \left(\lambda_A A_l(t) + \lambda_P \overline{P_l}(t) \right) \tag{10}$$

where λ_A and λ_P represent the weighting coefficients for the AoI and power consumption parts, respectively $\lambda_A + \lambda_P = 1$, with denoting a trade-off parameter. The optimization problem is given by the following equation:

$$\min_{\bar{C}_l(t), \bar{P}_l(t)} U(t) \tag{11}$$

$$s.t. \sum_{k=1}^{K} \overline{z}_{k,l}(t) \leq 1, \forall l = 1, 2, ..., L \tag{12}$$

$$\sum_{l=1}^{L} \overline{z}_{k,l}(t) \leq X, \forall k = 1, 2, ..., K \tag{13}$$

$$\overline{C}_l(t) \in \{0, 1, ..., K\}, \forall l = 1, 2, ..., L \tag{14}$$

$$\overline{P}_l(t) \in [0, P_{\max}], \forall l = 1, 2, ..., L \tag{15}$$

$$\sum_{l=1}^{L} \overline{z}_{k,l}(t) \overline{P}_l(t) \overline{g}(t) \leq \overline{I}_k^{\mathrm{th}}(t), \forall k = 1, 2, \ldots, K \tag{16}$$

where the constraints (12) ensure that each VUP accesses at most one channel in each time slot. Constraints (13) ensure that each channel is accessed by at most X VUPs simultaneously. Constraints (14) and (15) define the channel and power selection ranges for the VUPs, respectively. Constraint (16) ensures that the interference caused by VUPs to the primary user (PU) does not exceed the interference threshold $\overline{I}_m^{\mathrm{th}}(t)$ when the PU occupies the channel.

The optimization problem in (11) to (16) is a non-linear, non-convex problem that is difficult to solve in polynomial time. Moreover, due to the presence of interference, the resource allocation strategies of individual VUPs (including channel selection and power control) are interdependent and interconnected. Additionally, each VUP can only access relevant information, making it challenging for traditional optimization methods to find the optimal transmission channel and power in the absence of global state information. Therefore, a multi-

agent reinforcement learning algorithm is proposed to enable each VUP to make decisions based on its local observations, aiming to minimize the weighted sum of the system's AoI and power consumption.

Resource allocation method based on IMAPPO algorithm

In the CIoT network model depicted in Fig 1, each VUP selects the optimal access channel and transmission power to minimize the weighted sum of system AoI and transmission power consumption. Considering the interdependence among decisions made by different VUPs, we first model the problem of joint channel selection and power control resource allocation as a fully cooperative multi-agent task, where all agents (i.e., VUPs) share the same objective. Subsequently, to address the multi-agent task and obtain decentralized policies for each VUP, we introduce a centralized training and distributed execution structure, developing a Multi-Agent Proximal Policy Optimization (MAPPO) algorithm. Due to centralized training, the algorithm learns decentralized policies that can work more collaboratively. Additionally, to tackle the joint optimization problem with mixed discrete-continuous action spaces, corresponding improvements have been made to the MAPPO algorithm.

Markov decision process

The resource allocation problem of joint channel selection and power control described in this paper is modeled as a Markov Decision Process (MDP), S represents the global state space of the environment, $\mathcal{O} = \{o_1, o_2, ..., o_L\}$ indicating the set of local observations for all VUPs. $\mathcal{A} = \{a_1, a_2, ..., a_L\}$ represents the joint action space for all VUPs, with a_l being the action of VUP-l. In each time slot, VUP-l selects actions based on policies $\pi_l(a_l \mid o_l)$. The collective actions of all VUPs cause the environment to transition from the current state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ according to the state transition probability $\mathcal{P}(s'|s, a)$.

State and observation space

As mentioned above, the observational information for each VUP, apart from its own information, includes partial environmental information obtained through communication. Thus, the state observation for VUP-l at time slot t is defined as:

$$O_l(t) = \{\overline{W}(t), \overline{G}_l(t), E_l(t-1), A_l(t-1)\} \tag{17}$$

where,

$$\overline{W}(t) = \{\overline{w}_k(t), k = 1, 2, \ldots, K\} \tag{18}$$

$$\begin{aligned} \overline{G}_l(t) = \{&\overline{g}_{VUP-l}(t), \overline{g}_{VUP-l,PBS}(t), \\ &\overline{g}_{PU,VUP-l}(t), \overline{g}_{VUP,VUP-l}(t)\} \end{aligned} \tag{19}$$

$$\begin{cases} \overline{g}_{PU,VUP-l}(t) = \{\overline{g}_{PU-k,VUP-l}(t), k = 1, 2, ..., K\} \\ \overline{g}_{VUP',VUP-l}(t) = \{\overline{g}_{VUP-l',VUP-l}(t), (l' \neq l) \bigcap \\ (l' = 1, 2, ..., L)\} \end{cases} \tag{20}$$

$\bar{W}(t)$ represents the set of orthogonal channel states perceived by CBS. $\bar{G}_l(t)$ represents the channel gain set of VUP-l, including its own transmission link channel gain and interference link channel gain. $E_l(t-1)$ represents the remaining load at the end of the time slot for VUP-l. $A_l(t-1)$ represents the AoI of VUP-l at the end of the time slot t-1.

Due to the random nature of PU occupancy, $\bar{W}(t)$ reflects the randomness of PU occupancy for both CBS and VUPs. The dynamic variations in path loss due to the rapid movement of vehicles, along with the relationship between channel gain and path loss, $\bar{G}_l(t)$ reflect the highly dynamic nature of the channel state. According to Formula (9), $A_l(t-1)$ reflects the timeliness of communication between vehicles.

Action space

As stated above, as an agent, VUP needs to perform channel selection and power control. Thus, the action of VUP-l at time slot t is defined as:

$$a_l(t) = \{\overline{C}_l(t), \overline{P}_l(t)\} \tag{21}$$

where, $\bar{C}_l(t)$ represents the discrete action, and $\bar{P}_l(t)$ represents the continuous action.

Reward function

As mentioned earlier, the primary objective of this paper is to minimize the AoI of VUPs during CIoT communication while reducing transmission power consumption, thus the design of the reward function needs to consider the negative of the utility function in formula (10). As VUPs in the network use the Underlay spectrum sharing mode to access orthogonal channels, it is necessary to control the interference of VUPs to PUs to ensure the QoS of the PUs. Therefore, the reward function is designed as:

$$r(t) = -\left(\lambda_A \sum_{l=1}^{L} A_l(t) + \lambda_P \sum_{l=1}^{L} \overline{P}_l(t)\right) - \overline{\chi} \sum_{k=1}^{K} \overline{w}_k(t) \max\left\{0, \overline{I}_k(t) - \overline{I}_k^{\text{th}}(t)\right\} \tag{22}$$

where,

$$\overline{I}_k(t) = \sum_{l=1}^{L} \overline{z}_{k,l}(t) \overline{P}_l(t) \overline{g}_{VUP-l,PBS}(t) \tag{23}$$

The first term in formula (22) represents the negative of the system utility function, while the second term represents the interference penalty with
$\overline{X}$ as the interference penalty adjuster. When the PU does not occupy the channel, it is assumed that $\overline{I}_k^{\text{th}}(t) \gg \overline{I}_k(t)$.

Improved multi-agent proximal policy optimization algorithm

This section provides an initial introduction to the Proximal Policy Optimization (PPO) algorithm. Subsequently, we design a centralized training distributed execution MAPPO algorithm and enhance the Actor network of the algorithm to address the joint channel selection and power control resource allocation problem in a mixed discrete-continuous action space.

PPO algorithm

The PPO algorithm is a type of Policy Gradient (PG) algorithm that directly learns action probabilities. In traditional PG algorithms, a policy gradient estimator is used to update the policy.:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_\theta \log \pi_\theta\big(\, a(t)|\, s(t)\big)\, \hat{A}_t\right] \tag{24}$$

$$\hat{A}_t = \mathcal{R}(t) + \gamma V_\vartheta\left(s(t+1)\right) - V_\vartheta\left(s(t)\right) \tag{25}$$

Here, π_θ represents a stochastic policy, and $\hat{A}_t$ denotes an estimate of the true advantage at time step t. The estimator $\hat{g}$ is derived through the objective function :

$$J(\theta) = \hat{\mathbb{E}}_t \left[\log \pi_\theta\big(\, a(t)|s(t)\big)\, \hat{A}_t\right] \tag{26}$$

By adjusting parameters μ to train the Critic network, the aim is to minimize the loss:

$$J(\vartheta) = \hat{\mathbb{E}}_t \left[\left(y(t) - V_9\left(s(t)\right)\right)^2\right] \tag{27}$$

Herein,

$$y(t) = \mathcal{R}(t) + \gamma V_{\bar{\vartheta}}\left(s(t+1)\right) \tag{28}$$

$V_{\bar{\vartheta}}$ is achieved by using the latest ϑ regularly updated target value function. The target function is solved using the stochastic gradient descent method. However, because the magnitude of the gradient step in the parameter space often does not directly correspond to its magnitude in the policy space, executing stochastic policy gradients is often unstable and can result in excessively large steps in the policy space. Additionally, the original policy-based algorithm has a lower training efficiency since the collected trajectories are used to train the agent only once.

To attain reliable performance and learning efficiency, the PPO algorithm introduces a clipped surrogate objective function to avoid excessive modifications to the objective value. The PPO algorithm consists of two networks: the Actor network and the Critic network. The Actor network determines actions, while the Critic network evaluates the value of actions. The agent utilizes the Actor network to perform learning tasks in specific observation states of the environment and obtain the actions to be taken. The agent sends actions determined by the Actor network, observes the next state of the environment, and receives a positive or negative reward. The Critic network treats the obtained reward as a network parameter, evaluating whether the actions determined by the Actor network have led the environment into a more positive state and providing feedback to the Actor network. The alternative objective function can be represented as follows:

$$
\begin{aligned}
J'(\theta) &= \hat{\mathbb{E}}_t \left[\log \pi_\theta \left(a(t)|s(t) \right) \hat{A}_t \right] \\
&= \hat{\mathbb{E}}_t \left[\frac{\pi_\theta \left(a(t)|s(t) \right)}{\pi_{\theta_{old}} \left(a(t)|s(t) \right)} \hat{A}_t \right] \\
&= \hat{\mathbb{E}}_t \left[\rho_t(\theta) \hat{A}_t \right]
\end{aligned}
\tag{29}
$$

Here, $\pi_\theta \left(a(t)|s(t) \right)$ represents the probability of selecting actions $a(t)$ based on the new policy state $s(t)$, $\pi_{\theta_{old}} \left(a(t)|s(t) \right)$ represents the probability of selecting actions $a(t)$ based on the old policy state $s(t)$, and $\rho_t(\theta)$ denotes the ratio of the new and old policy distributions. Subsequently, the alternative objective function is clipped by imposing constraints on the difference between the new and old policies to avoid excessively large policy updates.

$$
\begin{aligned}
\bar{J}(\theta) &= \hat{\mathbb{E}}_t \left[\min \left(\rho_t(\theta) \hat{A}_t, \mathrm{clip} \left(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon \right) \hat{A}_t \right) \right] \\
&= \hat{\mathbb{E}}_t \left[F \left(\rho_t(\theta), \hat{A}_t \right) \right]
\end{aligned}
\tag{30}
$$

Here, ε is the clipping parameter, and the clip operation is used to constrain the ratio of the new and old policy distributions $\rho_t(\theta)$ within $[1-\varepsilon, 1+\varepsilon]$. When $\hat{A}_t > 0$, it indicates that the current action is better than the average, increasing the probability of that action, but the update step is limited to $1+\varepsilon$. When $\hat{A}_t < 0$ it indicates that the current action is worse than the average action, reducing the probability of that action, with the update step truncated at $1-\varepsilon$.

Improved MAPPO algorithm

This section proposes an improved Multi-Agent Proximal Policy Optimization (IMAPPO) algorithm, the algorithmic framework of which is illustrated in Fig.4. The algorithm adopts a centralized training and distributed execution structure. The centralized Critic network estimates the joint action-value function based on global information, including $o(t)$ and $a(t)$. On the other hand, the distributed Actor network makes decisions solely

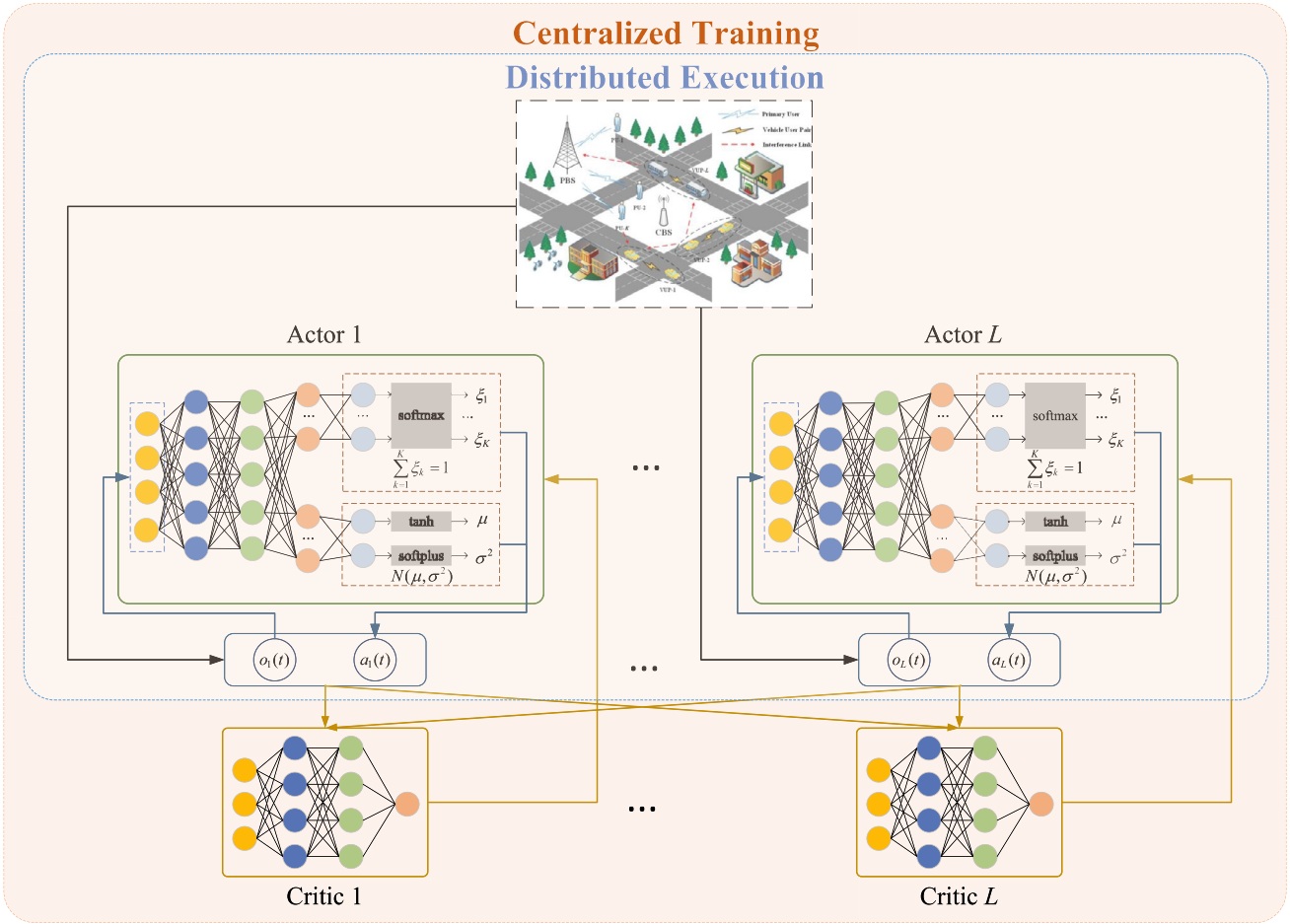

Fig. 4. IMAPPO algorithm framework diagram.

based on local observation information from VUP. After the training process, there is no requirement for global information, and VUP can independently select its actions in a decentralized manner.

To address the issue of the mixed discrete-continuous action space, the Actor network was correspondingly improved, as depicted in Fig.5. Given that there are K+1 channel selection strategies for VUP, the output layer is configured with K+1 neurons that utilize the softmax activation function to generate action probabilities $\xi_k, k \in \{1, ..., K+1\}$ and $\sum_{k=1}^{K+1} \xi_k = 1$ for hese K+1 strategiesand. Furthermore, actions are sampled randomly from the probability distribution composed of these K+1 probabilities . The remaining two neurons in the output layer respectively output the mean μ and variance σ^2 of continuous actions. Subsequently, continuous actions are obtained by sampling from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

Require: Initialize the parameters θ of the Actor network and the parameters g of the Critic network
Initialize the experience memory pool D
Initialize the number of training episodes $E_{\max}$, and the number of iterations per episode T_{sum}
Ensure: Action $A(t)$ at each step
Initialize state $S(1)$
for episode $= 1$ to $E_{\max}$ **do**
  for $t = 1$ to T_{sum} **do**
    for agent $l = 1$ to L **do**
      Agent l obtains local observation $o_l(t)$ from the environment
      Actor network selects action $a_l(t)$ based on policy $\pi_{\theta_{old}}{}^l(a_l(t)|o_l(t))$
    end for
    Obtain reward $r(t)$, transition to next state $S(t+1)$
    Store all agents' data $\{o(t), a(t), r(t), o(t+1))\}$ into experience memory pool D
    if $t \mod \tau = 0$ **then**
      for $u = 1$ to U **do**
        Sample a small batch randomly from experience pool D
        Calculate $\hat{A}_t^l$ based on Equation (32)
        Calculate $\overline{J}(\theta^l)$ based on Equation (31)
        Calculate $J(g^l)$ based on Equation (34)
        Update Actor network parameters: $\theta^l \leftarrow \theta^l + \zeta \nabla_{\theta^l} \overline{J}(\theta^l)$
        Update Critic network parameters: $g^l \leftarrow g^l - \zeta \nabla_{g^l} J(g^l)$
      end for
      Update old parameters: $\theta_{\text{old}} \leftarrow \theta^l$
      Reset experience pool D
    end if
  end for
end for

Algorithm 1. Resource Allocation Algorithm Based on IMAPPO for Cognitive Internet of Things

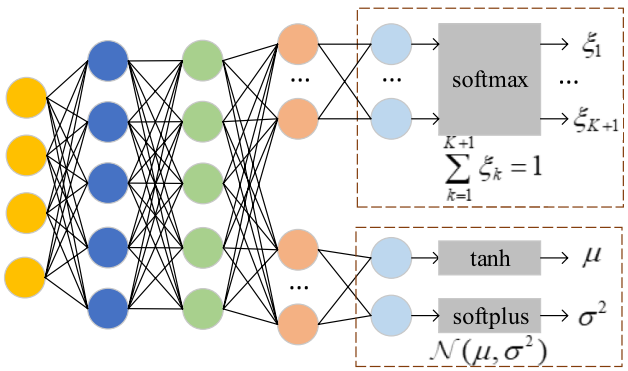

Fig. 5. IMAPPO algorithm framework diagram.

First, utilize a clipped loss function to update the policy of the Actor network:

$$\overline{J}(\theta^l) = \hat{\mathbb{E}}_t \left[F\left(\rho_t(\theta^l), \hat{A}_t^l\right)\right] \tag{31}$$

where $\hat{A}_t^l$ is the joint advantage function, representing the contribution of the collective actions of all agents, including the given agent, to the collective reward.

$$\hat{A}_t^l = \delta_t^l + (\gamma\vartheta)\delta_{t+1}^l + \cdots + (\gamma\vartheta)^{\tau-t+1}\delta_{\tau-1}^l \tag{32}$$

$$\delta_t^l = r(t) + \gamma Q_{\bar{g}^l}\left(o(t+1), a(t+1)\right) - Q_{\bar{g}^l}\left(o(t), a(t+1)\right) \tag{33}$$

The Critic network is updated via the following loss function:

$$J(g^l) = \hat{\mathbb{E}}_t \left[\left(y_t^l - Q_{g^l}\left(o(t), a(t)\right)\right)^2\right] \tag{34}$$

The Critic network of each agent is updated separately but shares the same joint action-value function. Algorithm 1 provides a detailed training process for the cognitive IoT resource allocation algorithm based on IMAPPO.

Experimental simulation and result analysis

Simulation settings

In this section, a series of experimental simulations are conducted to evaluate the performance of the cognitive IoT resource allocation method proposed in this paper based on the IMAPPO algorithm. A coordinate system is established with the cognitive base station as the origin, and four bidirectional roads, each 2000m long, are established with endpoints at (1000, 1000), (-1000, 1000), (-1000, -1000), and (1000, -1000). VUPs are randomly distributed along these four roads. It is assumed that the primary base station (PBS) is located at coordinates (3000, 3000), and primary users (PUs) are randomly distributed within a circle centered at PBS with a radius of 3000m. In the proposed IMAPPO algorithm, the neural network consists of three fully connected hidden layers containing 512, 256, and 128 neurons, respectively.ReLU is used as the activation function, and Adam is used as the optimizer. Unless otherwise specified, the parameters for the IMAPPO algorithm and the cognitive IoT system are provided in Table 2.

The experiments were conducted using Python 3.9 to develop the DRL environment, with deep neural networks constructed and executed via Pytorch 2.0.1, and the Adam optimizer used for gradient descent. NVIDIA GeForce RTX 3060 GPU was utilized to accelerate the training of all models. To evaluate the performance of the proposed IMAPPO algorithm in this paper, it is compared with the following several contrast algorithms.

P-DDPG Algorithm: The Parameterized Deep D-Policy Gradient (P-DDPG) algorithm[32], a method designed for hybrid action spaces, is employed as a state-of-the-art baseline. This algorithm extends the DDPG framework where the actor network simultaneously outputs a discrete action (channel selection) and its corresponding continuous action parameter (transmission power). This allows the agent to directly select a transmission power P from the continuous range $[0, P_{max}]$ for the chosen channel, based on the policy learned to maximize the Q-value estimated by the critic network.

Parameter	Value
Bandwidth B	10 MHz
Number of primary users M	6
Number of vehicle user pairs N	12
Vehicle speed v	50 km/h
Transmit power of primary user P_p	2 W
Maximum transmit power of vehicle user pair transmitter $P_{\max}$	1 W
Time slot T	100 ms
Sensing phase duration τ_1	20 ms
Decision phase duration τ_2	20 ms
Transmission phase duration τ_3	60 ms
Noise power δ^2	1×10^{-8} W
Initial learning rate α	0.001
Target network decay ϵ	0.01
Discount factor γ	0.99
Number of training episodes $E_{\max}$	500
Iterations per episode T_{num}	1000

Table 2. simulation parameters.

MADDQN Algorithm: The Multi-Agent Double Deep Q-Network (MADDQN) algorithm presented in[7] is employed. The core architecture of each agent is based on the DDQN algorithm. Each agent selects the access channel from the channel selection set, discretizes the transmission power, and chooses the transmission power $\tilde{P}$ from the discrete power set $\tilde{P} = \left\{0, \frac{1}{4}P_{\max}, \frac{1}{2}P_{\max}, \frac{3}{4}P_{\max}, P_{\max}\right\}$.

MADQN Algorithm: The Multi-Agent Deep Q-Network (MADQN) algorithm presented in[15] is employed. The core architecture of each agent is based on the DQN algorithm. The action space of each agent is consistent with that in the aforementioned MADDQN algorithm.

Figure 6 shows the variation of the average reward values of the proposed IMAPPO algorithm at each episode during the training phase. It can be seen from the figure that between Episode 0 and Episode 1000, the average reward value of the IMAPPO algorithm increases with the increase of episodes, indicating that the IMAPPO algorithm is learning the action policy in the direction of maximizing the average reward value. After Episode 200, the average reward value of the IMAPPO algorithm stabilizes around a specific value, which demonstrates the convergence of the IMAPPO algorithm. It is noted that there are significant fluctuations in the average reward value curve around Episode 220 and Episode 430. This is because there is exploration noise in the IMAPPO algorithm, which is intended to prevent the algorithm from converging to a local optimum.

Figure 7 illustrates the variation of the average information age under different VUP (Vehicle User Pair) quantities for the proposed IMAPPO algorithm and four other comparative algorithms. The IMAPPO algorithm consistently demonstrates superior performance. Compared to the P-DDPG algorithm, IMAPPO shows a performance improvement of approximately 8.7%. The performance improvement is 16.95% when compared to the MADDQN algorithm, 40.73% against the MADQN algorithm, and 45.95% versus the RANDOM algorithm. From the graph, it is evident that regardless of the change in the number of VUPs, the average age of information obtained using the IMAPPO algorithm is significantly lower than that of the other four comparative algorithms, thus confirming its superiority. Additionally, the average age of information for all five algorithms increases with the growth in VUP quantity. This is attributed to the fact that as the number of VUPs increases, their data transmission requires more spectrum resources. However, the limited spectrum resources result in a decrease in the transmission rate for each VUP. According to formulas (8) and (9), a lower transmission rate means the remaining data load of VUPs cannot be offloaded in a timely manner, consequently leading to an increase in the age of information.

Figure 8 depicts the variation of average power consumption under different VUP quantities for the proposed IMAPPO algorithm and four other comparative algorithms. It is evident that the average power consumption for all algorithms increases with the growth in VUP quantity. The proposed IMAPPO algorithm consistently achieves the lowest power consumption. Compared to the P-DDPG algorithm, the IMAPPO algorithm demonstrates a performance improvement of approximately 10.3%. The performance improvement is 26.55% when compared to the MADDQN algorithm, 41.24% against the MADQN algorithm, and 48.69% versus the RANDOM algorithm.This increase in power consumption can be attributed to the fact that the growth in the number of VUPs leads to an increase in the age of information for all VUPs. To ensure a lower age of information, VUPs need to consume more energy to transmit information more quickly. Additionally, across various VUP quantities, the proposed IMAPPO algorithm achieves lower power consumption compared to the four other comparative algorithms, highlighting its efficiency.

Figure 9 demonstrates the variation in average information age under different initial loads for the proposed IMAPPO algorithm and four other comparative algorithms. It is evident that the average information age for all

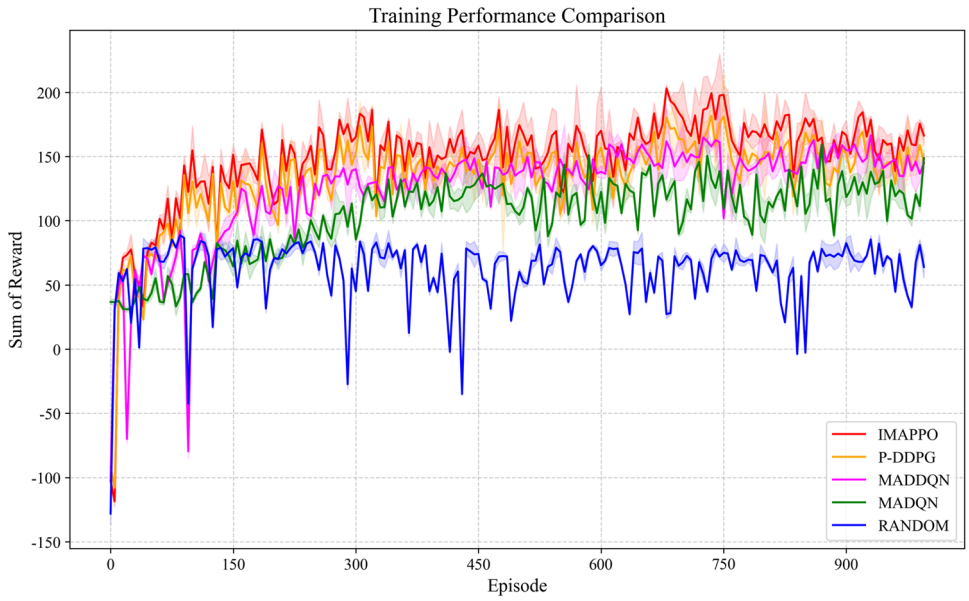

Fig. 6. IMAPPO algorithm in training phase average reward per episode.

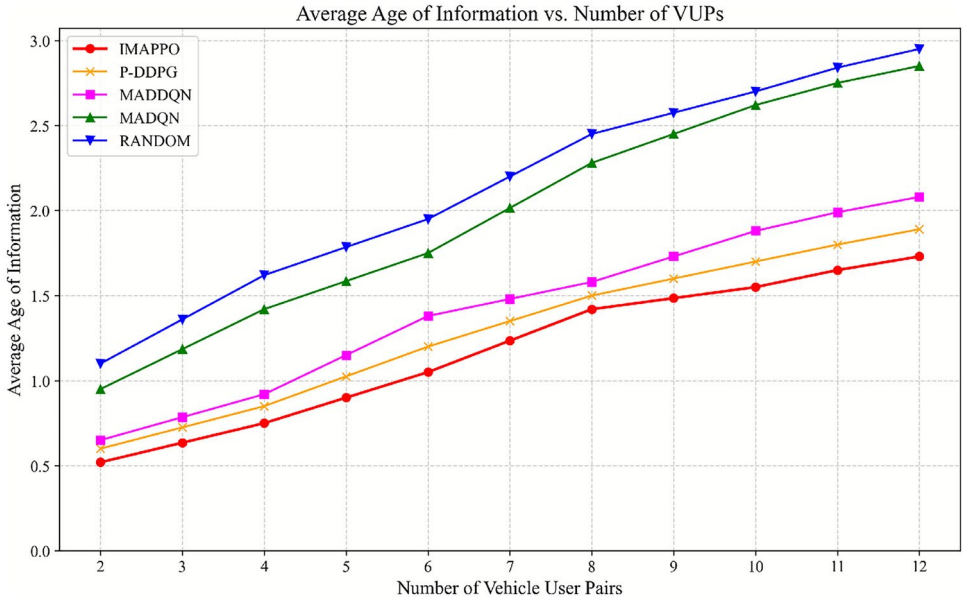

Fig. 7. Different algorithms' average information age under different VUP quantities.

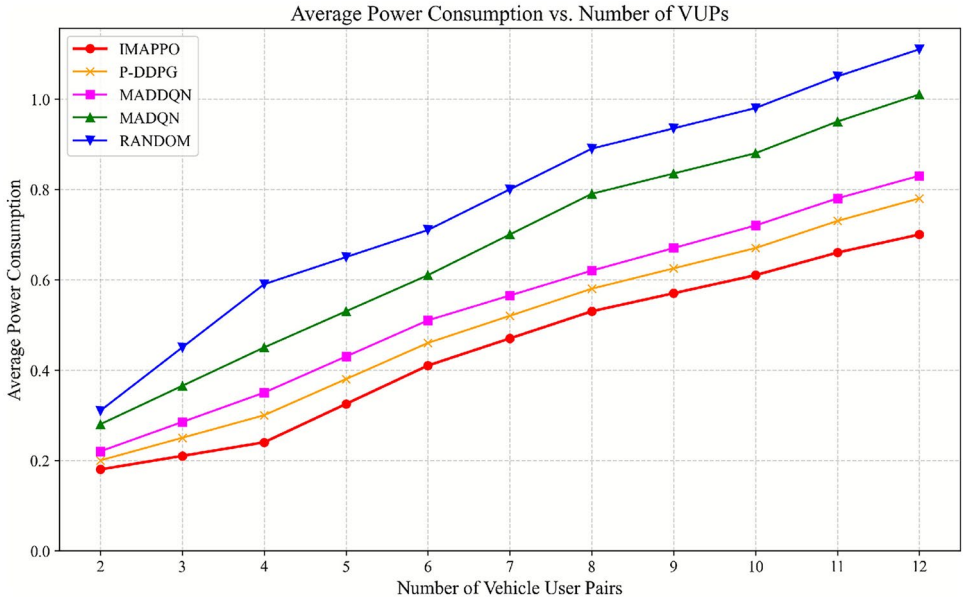

Fig. 8. Different algorithms' average power consumption under different VUP quantities.

algorithms increases with the rise in the initial load. The IMAPPO algorithm shows a performance improvement of approximately 7.4% compared to the P-DDPG algorithm, 12.69% compared to the MADDQN algorithm, 43.29% compared to the MADQN algorithm, and 49.54% compared to the RANDOM algorithm.According to formulas (8) and (9), it is understood that the remaining load of VUPs increases with the growth in the initial load, consequently leading to an increase in the average information age of VUPs. Additionally, from the graph, it is observable that across different initial loads, the proposed IMAPPO algorithm consistently achieves a lower average information age compared to the four other comparative algorithms.

Figure 10 exhibits the variation in average power consumption under different initial loads for the proposed IMAPPO algorithm and four other comparative algorithms. It is observable that the average power consumption increases with the rise in the initial load. The IMAPPO algorithm demonstrates a performance improvement of approximately 17.5% compared to the P-DDPG algorithm, 22.31% compared to the MADDQN algorithm, 44.81% compared to the MADQN algorithm, and 53.45% compared to the RANDOM algorithm.The increase in the initial load results in an increase in the average information age, prompting VUPs to consume more energy to lower the average information age. Consequently, the average power consumption also increases. Additionally,

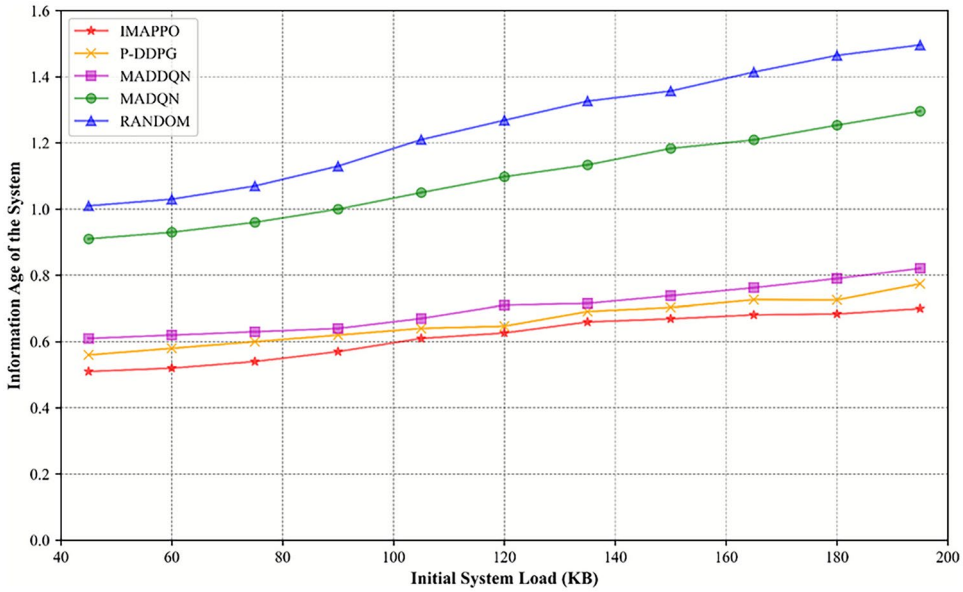

Fig. 9. Different algorithms' average information age under different initial loads.

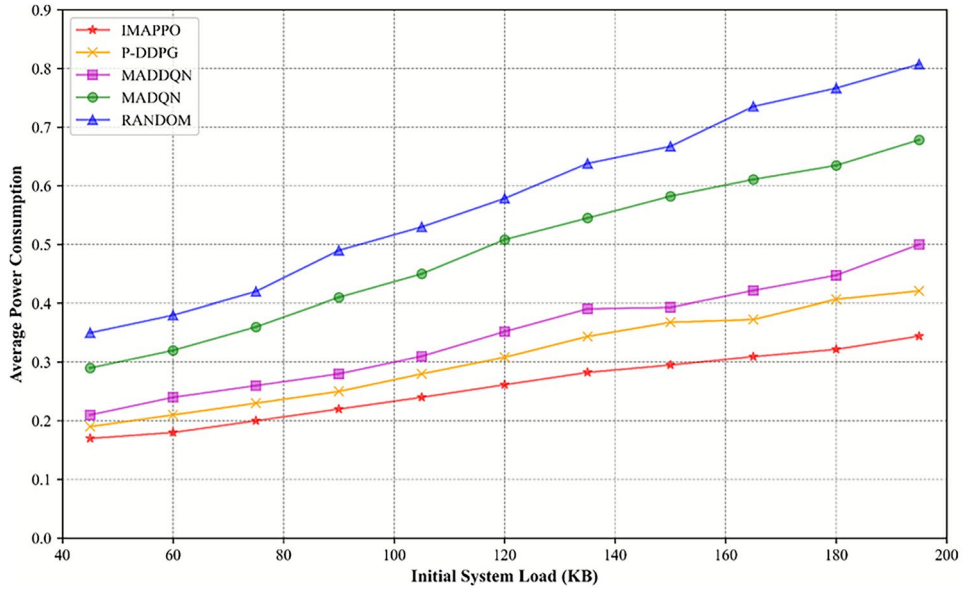

Fig. 10. Different algorithms' average power consumption under different initial loads.

from the graph, it is apparent that across different initial loads, the proposed IMAPPO algorithm achieves lower average power consumption compared to the four other comparative algorithms.

Complexity analysis

This section is dedicated to discussing and comparing the computational costs of our proposed IMAPPO algorithm with all baseline algorithms.In this subsection, we first analyze the complexity of the algorithms from a theoretical perspective. During the inference (decision-making) phase, a single-step decision for all learning-based algorithms (IMAPPO, P-DDPG, MADDQN, MADQN) only requires a single forward pass of their Actor/policy network. Since we have configured neural network architectures of similar scales for all algorithms, their theoretical complexity and FLOPs (Floating Point Operations) during inference are of the same order of magnitude. The RANDOM algorithm makes decisions through simple random selection, and its computational cost is negligible. During the training phase, the computational overhead varies among algorithms: the PPO-based IMAPPO algorithm typically requires multiple mini-batch update iterations on each data batch to optimize the policy; P-DDPG needs to update both the Actor and Critic networks simultaneously; whereas DQN-based

Algorithm	Avg. Training Time	Inference FLOPs	Avg. Inference Latency
IMAPPO (ours)	5.2 GPU-hours	1.5 GFLOPs	2.1 ms/decision
P-DDPG	4.5 GPU-hours	1.4 GFLOPs	2.0 ms/decision
MADDQN	3.8 GPU-hours	1.2 GFLOPs	1.9 ms/decision
MADQN	3.5 GPU-hours	1.2 GFLOPs	1.9 ms/decision
RANDOM	N/A	Negligible	0.1 ms (Negligible)

Table 3. Comparison of computational complexity and performance.

algorithms (MADDQN, MADQN) usually update only once per sample. Therefore, theoretically, the training process of IMAPPO requires more computational resources, followed by P-DDPG, while the training overhead for MADDQN/MADQN is relatively lower.

As can be seen from the table 3, IMAPPO indeed has the longest training time, which is consistent with our theoretical analysis. The RANDOM algorithm requires no training, and its decision latency and computational load are essentially negligible. However, the inference FLOPs for all learning-based algorithms are of the same order of magnitude, and their millisecond-level average inference latency demonstrates that they are fully capable of meeting the real-time decision-making requirements of IoV scenarios. We believe that the significant performance improvement in terms of Age of Information and power consumption justifies the increased training cost.

Limitations and future work

Although our proposed IMAPPO algorithm has demonstrated superior performance in simulations, we acknowledge that this study has several limitations, which also point to directions for future research.

Spectrum Sensing Assumption: This study assumes perfect sensing by the Cognitive Base Station (CBS), meaning it can determine the occupancy status of Primary User (PU) channels without error. This is an idealized assumption. In practice, spectrum sensing is subject to errors like false alarms (mistaking an idle channel for occupied) and missed detections (mistaking an occupied channel for idle). These errors can lead to wasted spectrum resources or even harmful interference to PUs. Future work should incorporate imperfect sensing models (e.g., probability models based on energy detection) and formulate the problem as a Partially Observable Markov Decision Process (POMDP) to develop more robust resource allocation strategies under sensing uncertainty.

Vehicle Mobility Model: To focus on the resource allocation algorithm, we employed a simplified straight-line vehicle mobility model. However, real-world urban traffic is far more complex, involving acceleration, deceleration, lane changes, turns, and stops. These complex mobility patterns lead to faster and more drastic variations in Channel State Information (CSI). Future research plans to integrate our algorithmic framework with mainstream traffic simulators like Simulation of Urban MObility (SUMO) or Verkehr In Städten - SIMulationsmodell (VISSIM). This will allow us to validate and refine our algorithm's performance and adaptability in more realistic and dynamic environments.

Scalability: Our centralized training, distributed execution framework is effective for a moderate number of agents. However, as the scale of the vehicular network grows very large (e.g., hundreds of vehicles), the central critic's need to process global information and joint actions from all agents can lead to a sharp increase in training complexity, posing a scalability challenge. A future research direction is to explore more scalable MARL paradigms, such as fully decentralized training methods or hierarchical reinforcement learning, to accommodate ultra-large-scale vehicular networks.

Conclusion

In this paper, we address the challenges of communication between vehicles, considering the randomness of primary user spectrum occupancy, highly dynamic channel states, and the timeliness requirements of inter-vehicle communication. We investigate the problem of joint channel selection and power control resource allocation for CIoT under high-speed mobility, aiming to minimize the Age of Information in the system. The formulated problem is modeled as a Markov Decision Process, and a carefully designed reward function is introduced. Furthermore, to meet the timeliness requirements, we employ a multi-agent reinforcement learning approach, where vehicle users act as agents, collecting local observation information and directly determining their own transmission strategies. We propose a multi-agent proximal policy optimization algorithm based on a centralized training and distributed execution structure, with improvements made to the Actor network to handle the issue of mixed discrete-continuous action spaces. Finally, through simulations, we verify the feasibility and effectiveness of the improved multi-agent proximal policy optimization algorithm. The results demonstrate that, compared to other benchmark schemes, the CIoT resource allocation scheme based on the improved multi-agent proximal policy optimization algorithm significantly reduces the Age of Information for vehicle users.

Future work will include a comprehensive analysis and comparison of the robustness of single-agent and multi-agent reinforcement learning algorithms to better understand when trained Q-networks need to be updated and how to efficiently execute such updates. Additionally, we will explore the development of other techniques to maintain computational accuracy while reducing computation time.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 8 March 2025; Accepted: 12 January 2026
Published online: 07 February 2026

References

1. Liu, C., Feng, W., Chen, Y., Wang, C.-X. & Ge, N. Cell-free satellite-UAV networks for 6g wide-area internet of things. *IEEE J. Sel. Areas Commun.* **39**(4), 1116–1131 (2020).
2. Do, D.-T., Van Nguyen, M.-S., Voznak, M., Kwasinski, A. & Souza, J. N. Performance analysis of clustering car-following v2x system with wireless power transfer and massive connections. *IEEE Internet Things J.* **9**(16), 14610–14628 (2021).
3. Noor-A-Rahim, M. et al. A survey on resource allocation in vehicular networks. *IEEE Trans. Intell. Transp. Syst.* **23**(2), 701–721 (2020).
4. Boban, M., Kousaridas, A., Manolakis, K., Eichinger, J. & Xu, W. Connected roads of the future: Use cases, requirements, and design considerations for vehicle-to-everything communications. *IEEE Veh. Technol. Mag.* **13**(3), 110–123 (2018).
5. Xiang, H., Zhou, W., Daneshmand, M. & Peng, M. Network slicing in fog radio access networks: Issues and challenges. *IEEE Commun. Mag.* **55**(12), 110–116 (2017).
6. Yang, Y., Fei, D. & Dang, S. Inter-vehicle cooperation channel estimation for IEEE 802.11 p v2i communications. *J. Commun. Netw.* **19**(3), 227–238 (2017).
7. Mafuta, A. D., Maharaj, B. T. & Alfa, A. S. Decentralized resource allocation-based multiagent deep learning in vehicular network. *IEEE Syst. J.* **17**(1), 87–98 (2022).
8. Lu, N., Cheng, N., Zhang, N., Shen, X. & Mark, J. W. Connected vehicles: Solutions and challenges. *IEEE Internet Things J.* **1**(4), 289–299 (2014).
9. Liang, L., Ye, H. & Li, G. Y. Toward intelligent vehicular networks: A machine learning framework. *IEEE Internet Things J.* **6**(1), 124–135 (2018).
10. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**(7540), 529–533 (2015).
11. Silver, D. et al. Mastering the game of go with deep neural networks and tree search. *Nature* **529**(7587), 484–489 (2016).
12. He, Y. et al. Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks. *IEEE Trans. Veh. Technol.* **66**(11), 10433–10445 (2017).
13. He, Y., Yu, F. R., Zhao, N. & Yin, H. Secure social networks in 5g systems with mobile edge computing, caching, and device-to-device communications. *IEEE Wirel. Commun.* **25**(3), 103–109 (2018).
14. Mao, H., Alizadeh, M., Menache, I., & Kandula, S. Resource management with deep reinforcement learning. In: Proceedings of the 15th ACM Workshop on Hot Topics in Networks, pp. 50–56 (2016)
15. Liang, L., Ye, H. & Li, G. Y. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE J. Sel. Areas Commun.* **37**(10), 2282–2292 (2019).
16. Li, J. et al. A bio-inspired solution to cluster-based distributed spectrum allocation in high-density cognitive internet of things. *IEEE Internet Things J.* **6**(6), 9294–9307 (2019).
17. Ejaz, W. & Ibnkahla, M. Multiband spectrum sensing and resource allocation for IoT in cognitive 5g networks. *IEEE Internet Things J.* **5**(1), 150–163 (2017).
18. Ahsan, W., Yi, W., Qin, Z., Liu, Y. & Nallanathan, A. Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach. *IEEE Trans. Wirel. Commun.* **20**(8), 5083–5098 (2021).
19. Allahham, M. S. et al. Multi-agent reinforcement learning for network selection and resource allocation in heterogeneous multi-rat networks. *IEEE Trans. Cognit. Commun. Netw.* **8**(2), 1287–1300 (2022).
20. Yang, H., Zhong, W.-D., Chen, C., Alphones, A. & Xie, X. Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive internet of things. *IEEE Internet Things J.* **7**(6), 5677–5689 (2020).
21. Huang, C., Chen, G., Gong, Y. & Han, Z. Joint buffer-aided hybrid-duplex relay selection and power allocation for secure cognitive networks with double deep q-network. *IEEE Trans. Cognit. Commun. Netw.* **7**(3), 834–844 (2021).
22. Liu, X., Sun, C., Yau, K.-L.A. & Wu, C. Joint collaborative big spectrum data sensing and reinforcement learning based dynamic spectrum access for cognitive internet of vehicles. *IEEE Trans. Intell. Transp. Syst.* **25**(1), 805–815 (2022).
23. Kaul, S., Gruteser, M., Rai, V., & Kenney, J. Minimizing age of information in vehicular networks. In: 2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 350–358 (2011). IEEE
24. Kaul, S., Yates, R., & Gruteser, M. Real-time status: How often should one update? In: 2012 Proceedings IEEE INFOCOM, pp. 2731–2735 (2012). IEEE
25. Yu, B., Cai, Y., Zou, Y., Li, B. & Chen, Y. Can we improve the information freshness with prediction for cognitive IoT?. *IEEE Internet Things J.* **9**(18), 17577–17591 (2022).
26. Lin, Z., Ni, Z., Kuang, L., Jiang, C. & Huang, Z. Dynamic beam pattern and bandwidth allocation based on multi-agent deep reinforcement learning for beam hopping satellite systems. *IEEE Trans. Veh. Technol.* **71**(4), 3917–3930 (2022).
27. Wang, L., Wang, Q., Chen, H., & Zhou, S. Age of information-oriented link scheduling in device-to-device networks. IEEE Transactions on Wireless Communications (2025)
28. Liu, P. et al. Priority-aware resource allocation in AoI-oriented UL-OFDMA Wi-Fi networks based on multi-agent reinforcement learning. *IEEE Internet of Things J.* https://doi.org/10.1109/JIOT.2025.3593060 (2025).
29. Shi, K. et al. AoI-aware data collection and energy replenishment for multi-UAV-enabled IoT systems. *IEEE Trans. Green Commun. Netw.* https://doi.org/10.1109/TGCN.2025.3542611 (2025).
30. Shoaib, M. et al. Decentralized resource allocation in UAV communication networks through reward based multi agent learning. *Sci. Rep.* **15**, 33122. https://doi.org/10.1038/s41598-025-18353-8 (2025).
31. Mohiuddin, M. B. et al. Reinforcement learning for end-to-end UAV slung-load navigation and obstacle avoidance. *Sci. Rep.* **15**, 34621. https://doi.org/10.1038/s41598-025-18220-6 (2025).
32. Jin, Y., Liu, Q. & Ji, Z. Application of priority deep deterministic strategy algorithm in autonomous driving. *J. Shanghai Univ. (Nat. Sci. Ed.)* **29**(1), 105–117 (2023).

Acknowledgements

This work was supported by Hunan Provincial Department of Education Scientific Research Outstanding Youth Project (25B1058,22B0994,24B1026) and in part by Hunan Provincial Natural Science Foundation of China (2023JJ60217,2023JJ60216,2023JJ50207).

Author contributions

R.W. conceptualization: research design; methodology: development of methods; software: software development; validation: validation; formal analysis: data analysis; investigation: experimental research; resources: pro-

vision of resources. D.W. writing—original draft. Y.S. visualization: preparation of figures and tables. W.L: data analysis. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This study does not involve human subjects, animal experiments, or sensitive data. We strictly adhere to academic rigor and standards to ensure the scientific nature and reliability of the research.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-36380-x.

Correspondence and requests for materials should be addressed to D.W. or W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.