A Cross-Layer Value of Information-Based Communication–Control Joint Optimization Algorithm for Connected Vehicle Platoon

Yaqi Xu[1], Tony Q. S. Quek[2]
[1]State Key Laboratory of Wireless and Network Technology,
Beijing University of Posts and Telecommunications, Beijing, China
[2]Wireless Networks and Decision Systems,
Singapore University of Technology and Design, Singapore
Email: xuyq19hiyan@bupt.edu.cn

***Abstract*—Connected vehicle platooning is a promising traffic model that has recently attracted attention from academia and industry. The tight coupling between the control and communication systems involved in platoon necessitates their collaborative design. In this paper, a hierarchical joint optimization algorithm based on the Cross-layer Value of Information (C-VoI), named Hierarchical Dual-Clip Proximal Policy Optimization (HDC-PPO) is proposed. The proposed C-VoI metric provides a unified measure that quantifies the value of information by integrating both control and communication performance. Building upon this metric, the HDC-PPO algorithm performs multi-timescale joint optimization: the control layer enhances the platoon's string stability by generating optimal acceleration commands, while the communication layer maximizes the C-VoI by performing subchannel selection and transmission power allocation, thereby realizing value-driven communication resource management. Simulation results demonstrate that compared with state-of-the-art algorithms, the proposed method significantly improves both string stability and packet reception ratio, achieving superior overall communication efficiency and control performance for the platoon system.**

***Index Terms*—Connected Vehicle Platoon, Cross-layer Value of Information, Joint Optimization, Hierarchical Reinforcement Learning, V2X Resource Allocation, Platoon Control.**

I. Introduction

With the rapid development of vehicular communication and autonomous driving technologies, intelligent connected vehicles have become a key component of intelligent transportation system. In accordance with the 3GPP communication standard, Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications exhibit enhanced reliability and flexibility [1]. Consequently, this enables the support of more complex cooperative driving tasks. Connected vehicle platooning (CVP) is a notable application that facilitates coordinated vehicle movement with minimal inter-vehicle distances while preserving speed synchronisation. This enhances road capacity, reduces energy consumption, and improves traffic safety, consequently garnering heightened attention in both theoretical research and engineering practice.

The efficient operation of CVP fundamentally relies on the tight coupling between control and communication systems. On the one hand, dynamic variations in control system states directly influence the resource demands and scheduling strategies of the communication system. On the other hand, time-varying communication channel conditions significantly affect control strategies, particularly in determining inter-vehicle spacing and following speed. This bidirectional coupling characteristic makes it difficult for traditional independent optimization approaches to achieve global system optimum. Consequently, establishing an effective cooperative optimization mechanism has become crucial for enhancing platoon performance.

Preliminary progress has already been made in the study of joint optimization of communication and control. For instance, Liu et al. [2], [3] decomposed the problem into communication-aware DRL-based platoon control and control-aware DRL-based radio resource allocation subproblems, addressing them through a multi-timescale control and communication joint optimization framework. Building on this approach, Lei et al. [4] introduced the concept of Value of Information (VoI) to quantify the gain of communication on control performance, and realized value-driven resource allocation through iterative optimization of separate DRL modules for control and communication. The aforementioned approaches provide a partial revelation of the relationship between communication information value and control effectiveness. However, existing VoI definitions primarily focus on communication's contribution to control performance, failing to establish a unified quantitative metric that incorporates both communication and control performance.

To address the research gap, this paper introduces a Cross-layer Value of Information (C-VoI) metric that integrates control and communication quality as a bridge between platoon control and communication resource allocation. When a vehicle experiences significant control errors and operates in an unstable state, its C-VoI becomes substantially higher. Simultaneously, if the transmission link quality for this information is excellent, the ultimate utility of this information is further amplified. Our optimization objective is to maximize the sum of C-VoI across all links in the entire platoon, enabling

resources to adaptively allocate to the vehicles and links that most need them and can most effectively utilize them. To our knowledge, no existing research has achieved coordinated optimization of communication and control performance through a unified metric in the platoon scenario.

Achieving joint communication-control optimization based on C-VoI necessitates the coordination of decision making processes between the two systems. In CVP scenario, there exists a significant difference in time scales between the control and communication systems: communication processes operate at a millisecond-level frequency, whereas control strategies are typically updated over longer time scales [2] . This multi-time scale characteristic leads to gradient conflicts and decision misalignment during joint optimization, presenting challenges that conventional single-layer reinforcement learning frameworks struggle to address effectively.

Hierarchical Reinforcement Learning (HRL) serves as an effective tool for addressing such multi-time scale coupling problems. By decomposing the policy structure across different levels, HRL significantly reduces the training complexity of traditional RL in high-dimensional state spaces and long-sequence decision tasks. Leveraging its hierarchical task abstraction and timescale separation capabilities, HRL has demonstrated remarkable advantages in large-scale system optimization. Previous research has explored HRL applications in multi-task and cross-layer optimization scenarios. For instance, Fu et al. [5] proposed a hierarchical resource decision framework for vehicle platooning that optimizes resource efficiency while reducing communication costs; Additionally, Zhang et al. [6] developed a hierarchical deep reinforcement learning approach that achieved faster convergence and superior performance in large-scale reconfigurable intelligent surface allocation. These studies demonstrate the vast potential of HRL applications in multi-timescale decision-making problems, especially in complex dynamic environments such as CVP.

In summary, this paper proposes a C-VoI-based Hierarchical Dual-Clip Proximal Policy Optimization (HDC-PPO). The proposed framework adopts a dual-layer structure: the control layer enhances platoon string stability by generating vehicle acceleration commands, while the communication layer maximizes C-VoI through subchannel selection and transmission power allocation, thereby realizing value-driven communication resource management. Notably, in the control layer, the larger timescale and gradual system state variations permit the use of standard PPO algorithm for policy updates, effectively ensuring training convergence and stability. In contrast, the communication module's smaller timescale, combined with time-varying channel conditions and complex action space, may lead to training instability caused by negative advantage samples. To address this, this paper introduce the dual-clip proximal policy optimization algorithm to suppress unstable updates, further enhancing the algorithm's training stability and robustness. Simulation results demonstrate that compared to existing advanced algorithms, the proposed method shows significant advantages in both platoon string stability and packet reception rate, effectively improving the system's overall control performance and communication efficiency. The main contributions of this paper are summarized as follows:

- The HDC-PPO framework is proposed to address the multi-timescale characteristics of control and communication, incorporating a two-layer policy structure and a dual-clip PPO mechanism to enhance training stability and convergence efficiency.
- The C-VoI metric is introduced, establishing a unified quantification of control and communication performance to support communication-control joint optimization.
- Through systematic experimental design and multi-dimensional performance evaluation, the proposed algorithm demonstrates significant advantages in key metrics such as platoon stability and communication reliability, providing a reliable solution for CVP control in connected vehicle networks.

This paper is organized as follows. Section 2 establishes the system model and problem formulation; Section 3 presents the design of the HDC-PPO algorithm; Section 4 provides simulation validation and result analysis; and Section 5 concludes the paper and discusses future research directions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Platoon Modeling

Consider a connected vehicle platoon travelling along a flat urban straight road, maintaining constant spacing and identical speeds. The vehicle set is denoted as $\mathcal{V}$, with each vehicle following the same dynamic model. The lead vehicle is numbered 0. The discrete-time dynamics model for vehicle i at time step k is defined as:

$$\begin{aligned} \dot{p}_{i,k} &= v_{i,k}, \\ \dot{v}_{i,k} &= u_{i,k}, \end{aligned} \tag{1}$$

where $p_{i,k}$, $v_{i,k}$, and $u_{i,k}$ denote the longitudinal position, velocity, and acceleration of vehicle i at time step k, respectively.

The platoon adopts a Predecessor-Leader Following (PLF) control strategy, in which each following vehicle receives the position, velocity, and acceleration information of both its immediate predecessor and the platoon leader through V2V communication.

The state of the leading vehicle is denoted by $p_{0,k}$, $v_{0,k}$, and $u_{0,k}$, where the acceleration $u_{0,k}$ is treated as an exogenous input. The desired inter-vehicle distance is determined based on the Constant Time Headway (CTH) policy:

$$d_{i,k}^{\text{des}} = \delta + h \cdot v_{i,k}, \tag{2}$$

where δ represents the minimum safety distance and h denotes the time headway constant. The tracking error for vehicle i at time step k comprises two components: The position error is defined as the deviation between the actual relative distance and the desired safety distance:

$$e_{p_{i,k}} = (p_{i-1,k} - p_{i,k}) - d_{i,k}^{\text{des}}. \tag{3}$$

The velocity error represents the deviation between the velocity of the current vehicle and a weighted combination of its preceding and leading vehicles' velocities:

$$e_{v_{i,k}} = \alpha(v_{0,k} - v_{i,k}) + (1-\alpha)(v_{i-1,k} - v_{i,k}), \quad (4)$$

where $\alpha \in [0,1]$ is a weighting factor indicating the relative importance of leader and predecessor information in the control process.

B. Communication Modeling for CVP

On the basis of 3GPP cellular V2X architecture [1], this paper investigates a vehicular platoon communication system that incorporates both Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) communications. V2I communication connects each vehicle to a Roadside Unit (RSU) for transmitting information such as traffic control and entertainment data. In contrast, V2V communication is primarily responsible to ensure reliable driving information sharing for safety driving.

Under the PLF control strategy adopted in this study, each platoon member (PM) communicates with both the leading vehicle (PL) and its immediate predecessor. Consequently, intra-platoon V2V communication can be categorized into two modes: Leader-to-Member (L2M) communication, where the PL broadcasts status information to all PMs via multicast, and Member-to-Member (M2M) communication, which involves unicast transmission between adjacent following vehicles.

The system assumes $\mathcal{M} = \{1, \ldots, M\}$ subchannels. Similar to [7], this paper simplify the V2I communications in that V2I links have preoccupied the spectrum sub-bands in an orthogonal manner, and their transmission powers are fixed, i.e., the mth V2I link preoccupies the mth sub-band. Therefore, the interference experienced by a V2V link originates from the corresponding V2I link in the same sub-band as well as other V2V links that reuse the same spectrum. Since the spectrum and power configuration for V2I links are fixed, the focus of system optimization shifts to the spectrum allocation and power control of V2V links. Orthogonal Frequency Division Multiplexing (OFDM) is employed, and channel fading is assumed to be independent across different sub-channels. The time domain is divided into equal-length time slots, and indexed by $t \in T$.

To model various fading environments in vehicular networks, the V2V channel is characterized as a Rayleigh fading channel. The channel gain from node i to node j can be expressed as:

$$\boldsymbol{G}_{i,j} = |\boldsymbol{h}_{i,j}|^2 \, \boldsymbol{d}_{i,j}^{-\varepsilon}, \quad (5)$$

where, $d_{i,j}^{-\varepsilon}$ is path loss.

The transmit power constraints for different links are defined as follows:

$$\begin{aligned} \mathbf{0} \leq \mathbf{P}_{\mathrm{L2M}} \leq \mathbf{P}_{\mathrm{L2M}'}^{\max} \\ \mathbf{0} \leq \mathbf{P}_{\mathrm{M2M}} \leq \mathbf{P}_{\mathrm{M2M}}^{\max}, \end{aligned} \quad (6)$$

where, $P_{L2M}^{\max}$ and $P_{M2M}^{\max}$ denote the maximum transmit power for $L2M$ and $M2M$ links, respectively.

The spectrum allocation is modeled as $\{a_i[m], b_i[m]\}$, with $a_i[m] = 1$ indicating that the L2M link for vehicle i reuses the m-th subchannel, and $b_i[m] = 1$ indicating that the M2M link for vehicle i reuses the m-th subchannel.

The interference model of CVP communication networks can be elaborated as follows. The interference in the L2M multicast link includes components from the V2I link and other M2M links. The interference received by receiver i on the m-th subchannel is given by:

$$I_i^{L2M}[m] = P_i^{V2I} G_i^{V2I}[m] + \sum_{j=1}^{N} b_j[m] P_j^{M2M} G_{i,j}^{M2M}[m]. \quad (7)$$

The interference for the M2M link of vehicle i over the m-th subchannel includes components from the V2I link, the L2M link, and other M2M links:

$$\begin{aligned} I_i^{M2M}[m] =& P_i^{V2I} G_i^{V2I}[m] + a_i[m] P_i^{L2M} G_i^{L2M}[m] \\ &+ \sum_{j \neq i} b_j[m] P_j^{M2M} G_{i,j}^{M2M}[m]. \end{aligned} \quad (8)$$

Here, $G_i^{V2I}[m]$ and $G_i^{L2M}[m]$ represent the channel gains from the m-th V2I link and the L2M transmitter to vehicle i, respectively. $G_{i,j}^{M2M}[m]$ denotes the channel gain from vehicle j to vehicle i on the m-th subchannel.

The transmission model of CVP communication networks can be elaborated as follows. The channel quality for vehicle i in the L2M multicast link on the m-th subchannel is given by:

$$\beta_i^{L2M}[m] = \frac{G_i^{L2M}[m]}{N_0 + I_i^{L2M}[m]}, \quad (9)$$

where $G_i^{L2M}[m]$ represents the channel gain from the L2M transmitter to vehicle i on the m-th subchannel. For the L2M multicast scenario where the PL broadcasts to all platoon members, the transmission condition on the m-th subchannel is determined by the receiver with the worst channel condition:

$$\beta^{L2M}[m] = \min_{i \in \mathcal{V}} \beta_i^{L2M}[m]. \quad (10)$$

According to Shannon's formula, the normalized transmission rate of the multicast link is given by:

$$R^{L2M}[m] = \log_2\left(1 + P^{L2M} \beta^{L2M}[m]\right). \quad (11)$$

The transmission rate of the M2M link for vehicle i over the m-th subchannel can be expressed as:

$$R_i^{M2M}[m] = \log_2\left(1 + \frac{P_i^{M2M} G_i^{M2M}[m]}{N_0 + I_i^{M2M}[m]}\right), \quad (12)$$

where $G_i^{M2M}[m]$ denotes the channel gain for vehicle i's M2M link on the m-th subchannel.

C. Cross-layer Value of Information

To jointly optimize communication and control performance, this paper defines the C-VoI metric. The C-VoI for vehicle i is defined as:

$$\text{C-VoI}_i = C_i \cdot \tilde{I}_i, \quad (13)$$

where $C_i \in [0,1]$ quantifies the vehicle's control performance, and $\tilde{I}_i \in [0,1]$ represents the normalized communication performance for vehicle i.

The control term C_i measures the deviation of the vehicle's current state from its desired state:

$$C_i = 1 - \min\left(1, \frac{|e_{p_i}|}{e_{p_{\max}}} + \frac{|e_{v_i}|}{e_{v_{\max}}}\right), \tag{14}$$

where e_{p_i} and e_{v_i} denote the position and velocity errors of vehicle i.

The communication term $\tilde{I}_i$ captures the vehicle's transmission capability over a recent time window:

$$\tilde{I}_i = \frac{1}{T}\sum_{t\in\mathcal{T}}\left(\omega_{\text{L2M}}\frac{R_i^{\text{L2M}}(t)}{R_{\max}} + \omega_{\text{M2M}}\frac{R_i^{\text{M2M}}(t)}{R_{\max}}\right), \tag{15}$$

where $R_i^{\text{L2M}}(t)$ and $R_i^{\text{M2M}}(t)$ are the instantaneous rates of L2M and M2M links for vehicle i.

By multiplying C_i and $\tilde{I}_i$, the C-VoI metric prioritizes links that are both control-critical and capable of reliable transmission. Links with poor control performance but acceptable communication are assigned higher C-VoI to receive more resources, while links with severely degraded channels are deprioritized, ensuring efficient and practical resource allocation.

D. Problem Formulation

The objective of this paper is to joint optimize the control and communication performance of the platoon system, formulated as the following maximization problem:

$$\max_{\substack{a_i[m],b_i[m]\\ P_i^{L2M},P_i^{M2M}}} \quad \sum_{i\in\mathcal{V}} \text{C-VoI}_i \tag{16a}$$

$$\text{s.t.} \quad \sum_{m=1}^{M} a_i[m] \le 1, \tag{16b}$$

$$\sum_{m=1}^{M} b_i[m] \le 1, \quad labeleq:16c \tag{16c}$$

$$0 \le P_i^{L2M} \le P_{\max}^{L2M}, \tag{16d}$$

$$0 \le P_i^{M2M} \le P_{\max}^{M2M}, \tag{16e}$$

$$d_i \ge d_{\min}, \tag{16f}$$

$$a_{\min} \le a_i \le a_{\max}, \tag{16g}$$

(16b) ensures that each resource block is assigned to at most one vehicle at a given time, preventing resource conflicts; (**??**) limits the transmission power of each communication link within the physical maximum; (16d) maintains a minimum inter-vehicle distance to guarantee safety; and (16e) constrains vehicle acceleration within feasible bounds to ensure driving comfort and controllability.

III. Algorithm Description

To address the multi-timescale decision-making problem between communication and control in CVP system, the HDC-PPO framework is adopted to jointly optimize both layers. Strategies are executed and updated at different timescales, thereby achieving joint optimization of communication and control. The architecture of the algorithm framework is illustrated in Fig. 1.

A. Semi-Markov Decision Process Formulation

The joint optimization problem is modeled as a Semi-Markov Decision Process (SMDP) with a two-layer structure: a control layer with a longer timescale and a communication layer with a shorter timescale. This decomposition enables coordinated optimization of control performance and communication efficiency.

State Space: *Control layer:* For agent i at interval k, the state includes delayed observations, past actions, and current communication delay:

$$S_{i,k}^{PL} = x_{i,k-\tau}, \tau, U_{i,k-1}^{\text{hist}}, \tag{17}$$

where $x_{i,k-\tau}$ contains speed, acceleration, and position; $U_{i,k-1}^{\text{hist}}$ is historical control actions; τ is the current observation delay.

Communication layer: For agent i at interval (k,t), the state includes channel information and the latest control action:

$$S_{i,(k,t)}^{CM} = G_{i,(k,t)}, I_{i,(k,t)}, U_{i,k}^{\text{hist}}, \tag{18}$$

where $G_{i,(k,t)}$ and $I_{i,(k,t)}$ are channel gain and interference, and $U_{i,k}^{\text{hist}}$ is used in computing C-VoI.

Action Space: *Control agent:* Continuous longitudinal acceleration

$$u_{i,k}^{PL} \in [u_{\min}^{PL}, u_{\max}^{PL}], \tag{19}$$

which guides communication decisions.

Communication agent: Discrete subchannel selection and continuous transmit power

$$a_{i,(k,t)}^{CM} = [m_i, P_{i,m,(k,t)}^{V}], \tag{20}$$

with $m_i \in \text{M} \cup -1$ and $P_{i,m,(k,t)}^{V} \in A_p$. If $P_{i,m,(k,t)}^{V} = 0$, power is set to -100 dBm.

Reward Function: *Control layer:* Minimize tracking errors while ensuring smoothness:

$$R_{i,k}^{PC} = -\left(\alpha_1\frac{|err_{p_{i,k}}|}{err_{p,\max}} + \alpha_2\frac{|err_{v_{i,k}}|}{err_{v,\max}} + \alpha_3\frac{|u_{i,k}|}{u_{\max}}\right), \tag{21}$$

Communication layer: Maximize cumulative C-VoI:

$$R_{i,(k,t)}^{CM} = -\sum C\text{-VoI}_{i,(k,t)}. \tag{22}$$

Through the above definitions of state, action, and reward, a complete SMDP model for hierarchical joint optimization is established.

B. Hierarchical Dual-Clip Proximal Policy Optimization Algorithm

Building on the hierarchical reinforcement learning structure, this study adopts the HDC-PPO algorithm to update the neural networks of both the control layer and the communication layer.

The Proximal Policy Optimization (PPO) algorithm [10] serves as the core learning framework due to its robustness

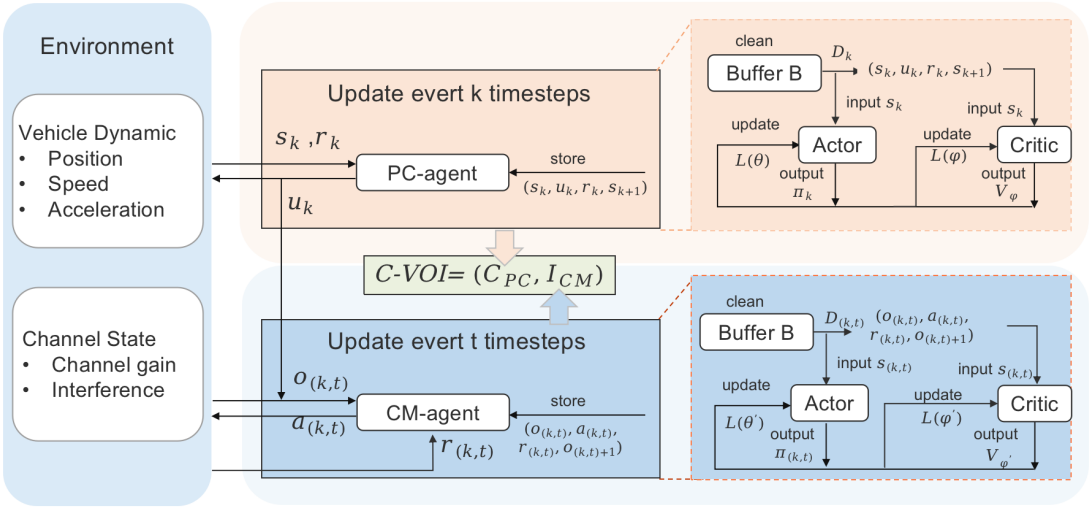

Fig. 1. Hierarchical Reinforcement Learning Framework for Communication Control Joint Optimization.

and sample efficiency. PPO is a policy gradient method that maximizes a clipped objective to prevent destructive large policy updates, ensuring stable convergence even in stochastic environments.

C. PPO for Control Layer

Given the smoother dynamics and slower timescale of the control layer, standard PPO is employed for stable policy updates. The actor network processes the input state s_k through a series of fully connected (FC) layers and outputs a control policy, from which the action u_k is sampled according to the target policy $\pi_\theta(u|s_k)$. By interacting with the environment, a set of transition tuples $(s_k g_k, r_k, s_{k+1})$ is collected and stored in the replay buffer $\mathcal{B}$. The policy is updated every k time steps using the sampled mini-batch.

The PC-agent aims to obtain the optimal policy by maximizing the expected cumulative reward:

$$\mathrm{J}\left(\mathrm{I}_\theta\right) = \mathrm{E}\left[\sum_{\mathrm{k}=0}^{\infty} \gamma_{\mathrm{k}} \mathrm{r}_{\mathrm{k}}\right] \approx \mathrm{E}\left[\ \mathrm{V}_\psi\left(\mathrm{s}_0\right)\right], \tag{23}$$

where γ is the discount factor and $V_\psi(s_k)$ denotes the state-value function estimated by the critic network.

To prevent excessive policy shifts that may lead to instability during training, PPO introduces a probability ratio constraint. For each update using batch samples $\mathcal{D}_E$, the new and old policies are required to satisfy:

$$J\left(\Pi_\theta^{\text{new}}\right) - J\left(\Pi_\theta^{\text{old}}\right) \geq \alpha E_{D_k}\left[r(\theta) A\left(s_k, g_k\right)\right] - \beta E_{D_k}[r(\theta) - 1] \tag{24}$$

where $r(\theta) = \frac{\Pi_\theta^{\text{new}}\left(g_k | s_k\right)}{\Pi_\theta^{\text{old}}\left(g_k | s_k\right)}$ is the probability ratio between the new and old policies, A_k is the advantage function, and ϵ is the clipping threshold (typically 0.1–0.2). The first term represents the Surrogate Objective, while the second term acts as a Penalty Term that restricts large deviations between successive policies. This clipping mechanism ensures stable and monotonic policy improvement by preventing over-updates.

The advantage function $A(s_k, g_k)$ evaluates the relative improvement of the selected action compared with the average policy and is estimated using Generalized Advantage Estimation (GAE) as:

$$A(s_k, g_k) = \delta_k + (\gamma\lambda)\delta_{k+1} + \cdots + (\gamma\lambda)^{K-k-1}\delta_{K-1}, \tag{25}$$

where λ is the GAE coefficient that balances bias and variance. This mechanism ensures stable and sample-efficient training by constraining the policy update magnitude within a reasonable range.

The loss function of the actor network is defined as:

$$L(\theta) = \mathbb{E}_D\left[\min\left(\frac{\pi_\theta^{\text{new}}\ (g_k \mid s_k)}{\pi_\theta^{\text{old}}\ (g_k \mid s_k)} A\left(s_k, g_k\right), u\left(A\left(s_k, g_k\right)\right)\right)\right], \tag{26}$$

where the clipping function $u(A_t)$ is used to enforce the constraint defined in (27).

$$u(A) = \begin{cases} (1+\varepsilon)A, & A \geq 0 \\ (1-\varepsilon)A, & A < 0 \end{cases}. \tag{27}$$

The critic network evaluates the state value $V_\psi(s_k)$ through a residual-based estimation to reduce the variance and provide a more accurate gradient signal for the actor. Its loss function is formulated as:

$$L(\varphi) = \mathbb{E}_{D_k}\left[r_k + \gamma V_\psi(s_{k+1}) - V_\psi(s_k)\right]. \tag{28}$$

where $\hat{V}_t$ is the target value derived from the sampled return.

D. DC-PPO for Communication Layer

Due to the high-frequency dynamics and complex discrete–continuous action space in the communication layer, which involves both subchannel selection and power control, standard PPO often suffers from training instability caused by samples with negative advantages. To address this issue, a Dual-Clip PPO mechanism is introduced. This approach incorporates an additional lower bound constraint to suppress

excessive gradient amplification induced by negative advantage estimates:

$$L(\theta') = \mathbb{E}_{D_t}\left[\max\left(\min\left(r(\theta')A_{\theta'_{\text{old}}}(a_t,o_t),\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(r(\theta'),1-\epsilon,1+\epsilon\right)A_{\theta'_{\text{old}}}(a_t,o_t)\right),cA_{\theta'_{\text{old}}}(a_t,o_t)\right)\right], \tag{29}$$

where $c > 1$ is the lower clipping coefficient. When the $A_t < 0$ and r_t becomes excessively large, the update gradient is limited within cA_t, effectively preventing divergence and enhancing stability in large-scale multi-agent learning.

The critic network minimizes the mean-square value error to refine the advantage estimation:

$$\mathrm{L}\left(\varphi'\right) = \mathrm{E}_{\mathrm{D_t}}\left[\mathrm{r_t} + \gamma \mathrm{V}_{\varphi'}\left(\mathrm{o_{t+1}}\right) - \mathrm{V}_{\varphi'}\left(\mathrm{o_t}\right)\right]. \tag{30}$$

This dual-clipping mechanism ensures faster convergence and robust training performance, particularly suited for rapidly fluctuating communication environments.

IV. Simulation Results

This section presents a comprehensive evaluation of the proposed HDC-PPO algorithm against existing methods, along with ablation studies validating the dual-clip optimization mechanism.

A. Simulation Settings

This article establishes a connected vehicle platoon experimental environment based on a joint simulation platform integrating SUMO and Python. SUMO is used to generate vehicle trajectories and traffic scenarios, while Python implements the vehicular communication protocol modeling and channel characteristic simulation, thereby enabling closed-loop validation of communication-control joint optimization. Experiments were implemented using the PyTorch framework and conducted on a high-performance workstation with an NVIDIA RTX 4090 GPU.

The simulation scenario is set on a unidirectional straight highway. The platoon size is set to $N = 5$ vehicles. To enhance realism, the driving data for the leading vehicle is sourced from an open-source real-world driving dataset provided in [8]. Specifically, acceleration and velocity trajectories of real vehicles are extracted from the NGSIM dataset. At the communication level, the simulation employs a V2V/V2I hybrid communication mechanism. The parameter configurations for V2I and V2V communications follows the 3GPP specification [9]. Specific parameters are listed in Table I.

Regarding algorithm design, the proposed HDC-PPO algorithm adopts a hierarchical reinforcement learning architecture composed of two layers: a control layer and a communication layer, both utilizing an Actor–Critic framework with four fully connected (FC) layers. The control layer captures temporal dependencies and performs vehicle-level decision-making, while the communication layer is responsible for subchannel selection and transmit power allocation. Each control episode is comprised of 100 control intervals (i.e., $K = 100$), where each control interval is set to $T = 0.1s$. The width of all hidden layers in the networks is set to 500, using the ReLU activation function. The Adam optimizer is chosen with a learning rate of 1×10^{-4}. The discount factor γ is set to 0.999, the clipping parameter ϵ is set to 0.2, and the training process runs for 500 episodes.

TABLE I
Numerical Calculation Parameter Settings

Parameter	Value
Path Loss	LOS/NLOS
Number of Subchannels	25
Carrier Frequency	2 GHz
Total Bandwidth	10 MHz
V2I Transmit Power	23 dBm
L2M Maximum Transmit Power	23 dBm
M2M Maximum Transmit Power	23 dBm
Noise Power	-114 dBm

B. Results and Discussion

1) Hierarchical Coordination and C-VoI Effectiveness: The proposed hierarchical framework is evaluated against two baseline approaches: the CACC and Random Allocation method, which employs CACC for platoon control combined with random resource assignment, and the Double DRL Iterative Optimization strategy, where control and communication layers are optimized independently without inter-layer coordination.

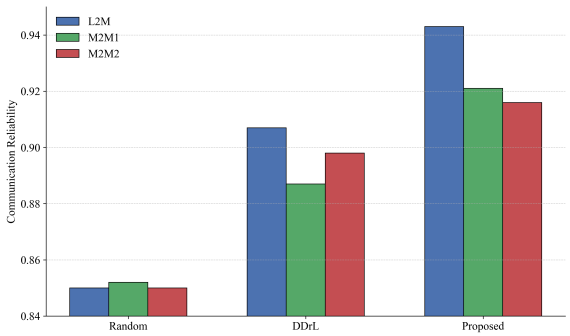

Fig. 2. Communication reliability comparison across methods.

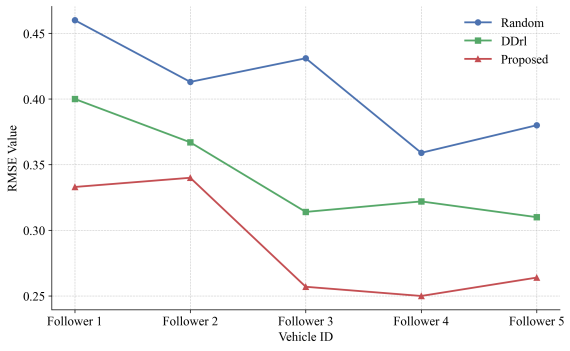

Fig. 3. String stability performance measured by spacing error.

As illustrated in Fig. 2, HDC-PPO achieves an average packet delivery ratio of 0.98, significantly outperforming the

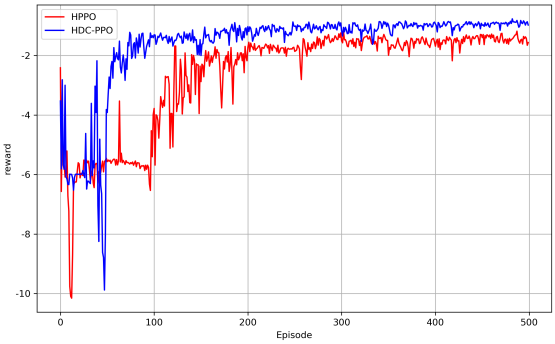

Fig. 4. Hierarchical Reinforcement Learning Framework for Communication Control Joint Optimization.

baseline methods in communication reliability. This improvement can be attributed to the C-VoI metric, which enables the communication layer to prioritize resources based on the control relevance of information, while the hierarchical structure allows coordinated updates across different time scales. For control performance, Fig. 3 shows that HDC-PPO maintains string stability with progressively decreasing spacing errors along the platoon, surpassing both baselines in dynamic stability. These results demonstrate that integrating C-VoI with the hierarchical, dual-layer design effectively enhances both communication and control performance in platooning scenarios.

2) Ablation Experiment: Ablation studies compare HDC-PPO against HPPO without dual-clip. As shown in Fig. 4, HDC-PPO achieves faster convergence and higher rewards, stabilizing after 100 iterations compared to 200+ for alternatives. The dual-clip mechanism ensures stable training where standard methods exhibit oscillations.

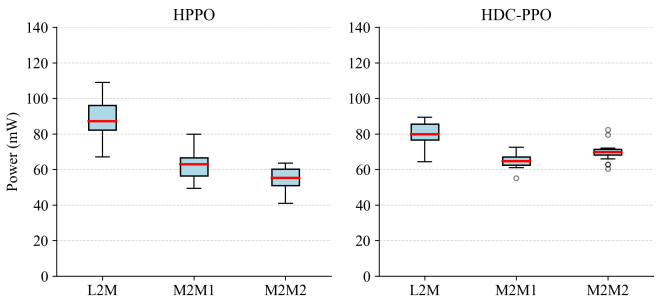

Fig. 5. Reward curves of different agents during training.

As shown in Fig. 5, power allocation analysis reveals HDC-PPO's adaptive capability: high-risk L2M links receive increased power for reliability, while low-risk M2M links operate with reduced power for interference control. This risk-aware allocation outperforms uniform power distribution schemes in both reliability and efficiency.

V. Conclusion

This paper proposes a joint optimization strategy based on HDC-PPO for CVP scenario, achieving cooperative optimization of control and communication through policy updates at different time scales. Specifically, the control module adopts the standard PPO algorithm for policy updates, while the communication module incorporates a Double-Clipped PPO algorithm to suppress unstable updates, thereby enhancing the training stability and robustness of the overall algorithm. To support communication-control joint optimization, the C-VoI metric is introduced, providing a unified metric of control and communication performance. Simulation results demonstrate that, compared with existing joint optimization methods, the proposed approach effectively improves the platoon's string stability and communication efficiency. In addition, ablation studies further validate the effectiveness of introducing the Double-Clipped PPO algorithm.

Future work will extend the framework to more complex multi-lane scenarios and heterogeneous platoons, in order to evaluate its generalization and scalability in complex traffic systems.

References

[1] 3GPP, "Technical Specification Group Radio Access Networks; Study enhancement 3GPP Support for 5G V2X Service," TR 22.886, V15.1.0, Release 15, Mar. 2017.

[2] L. Lei, T. Liu, K. Zheng, and X. Shen, "Multi-Timescale Control and Communications With Deep Reinforcement Learning—Part II: Control-Aware Radio Resource Allocation," *IEEE Internet Things J.*, pp. 1–1, 2024, doi: 10.1109/JIOT.2023.3348594.

[3] T. Liu, L. Lei, K. Zheng, and X. Shen, "Multi-Timescale Control and Communications With Deep Reinforcement Learning—Part I: Communication-Aware Vehicle Control," *IEEE Internet Things J.*, pp. 1–1, 2024, doi: 10.1109/JIOT.2023.3348590.

[4] L. Lei, K. Zheng, J. Mei, and X. Shen, "VoI-Driven Joint Optimization of Control and Communication in Vehicular Digital Twin Network," *IEEE Network*, Dec. 25, 2024.

[5] X. Fu, Q. Yuan, G. Luo, N. Cheng, Y. Li, J. Wang, and J. Liao, "HierNet: A Hierarchical Resource Allocation Method for Vehicle Platooning Networks," *IEEE Internet Things J.*, vol. 11, no. 24, pp. 39579–39592, 2024, doi: 10.1109/JIOT.2024.3444044.

[6] H. Zhang, W. Wang, H. Zhou, Z. Lu, and M. Li, "A Hierarchical DRL Approach for Resource Optimization in Multi-RIS Multi-Operator Networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 4981–4995, Jun. 2025, doi: 10.1109/TWC.2025.3545425.

[7] Y. Xu, K. Zhu, H. Xu, and J. Ji, "Deep Reinforcement Learning for Multi-Objective Resource Allocation in Multi-Platoon Cooperative Vehicular Networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6185–6198, 2023, doi: 10.1109/TWC.2023.3240425.

[8] U.S. Department of Transportation, "NGSIM Next Generation Simulation," 2009.

[9] 3GPP, "Enhanced LTE Support for Aerial Vehicle," document 36.777, Release 15, Dec. 2017.

[10] C. Yu, A. Velu, E. Vinitsky, Y. Wang, and Y. Wu, "The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games," 2021, arXiv:2103.01955. [Online]. Available: https://arxiv.org/abs/2103.01955