Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Energy Efficiency Optimization for Dynamic Indoor VLC with RSMA via GenAI-based DRL

Bohao Tang
Chinese Academy of Sciences
Yi Yu
yiyuciomp@163.com
Chinese Academy of Sciences
Wenwen Xie
Jilin University
Weicheng Chu
Chinese Academy of Sciences

Research Article

Keywords: Visible light communication, RSMA, GenAI, DRL, optimization

Posted Date: June 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9832964/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Energy Efficiency Optimization for Dynamic Indoor VLC with RSMA via GenAI-based DRL

Bohao Tang[1,2], Yi Yu[1*], Wenwen Xie[3], Weicheng Chu[1]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, 130033, China.
[2]University of Chinese Academy of Sciences, Beijing, 100190, China.
[3]College of Computer Science and Technology, Jilin University, Changchun, 130012, China.

*Corresponding author(s). E-mail(s): yuyi@ciomp.ac.cn;

Abstract

Visible light communication (VLC) is a promising technology for providing high-speed wireless access in indoor environments due to its dual functionality of illumination and communication. However, conventional multiple access schemes have limited capability in handling interference and improving spectral efficiency in dynamic VLC systems. To address these issues, we consider a dynamic indoor VLC system, where multiple light-emitting diodes (LEDs) deployed on the ceiling provide illumination and serve multiple mobile terrestrial users based on rate-splitting multiple access (RSMA). Considering both communication quality and energy consumption, we formulate an optimization problem that jointly optimizes the precoding matrix and the portion allocation of the common stream to maximize energy efficiency. Due to the coupled optimization variables and dynamic characteristics, we propose an improved deep deterministic policy gradient (IDDPG) algorithm. Specifically, a generative adversarial network (GAN) is introduced to train the critic network, which improves the accuracy of Q-value estimation by learning the distribution characteristics of the state-action value function. Moreover, a prioritized experience replay (PER) mechanism is adopted to improve sample efficiency and accelerate convergence. Simulation results demonstrate that the proposed IDDPG algorithm outperforms two conventional strategies and three deep reinforcement learning (DRL) algorithms in terms of energy efficiency.

Keywords: Visible light communication, RSMA, GenAI, DRL, optimization

1 Introduction

Visible light communication (VLC) is an emerging technology in wireless communication, which provides a promising solution to the scarcity of radio spectrum resources [1]. Compared with radio frequency communication, VLC operates over the visible light band from 380 to 780 nanometers, which offers ultra-high bandwidth potential, does not require spectrum licensing, and is inherently resistant to electromagnetic interference [2]. Moreover, due to the widespread deployment of light-emitting diode (LED) lighting infrastructure, VLC enables the integration of illumination and communication within a unified system architecture. As a result, lighting devices can serve as ubiquitous communication access points, which is why VLC exhibits strong potential in low-latency and high-reliability scenarios, such as indoor positioning [3], smart offices [4], and Industry 4.0 [5]. Consequently, VLC is widely regarded as an important technology for indoor coverage and short-range transmission in future sixth-generation (6G)communication networks [6].

However, due to the rapid growth of Internet-of-Things (IoT) devices, VLC systems face severe challenges in multi-user access under limited resources. Traditional orthogonal multiple access (OMA) partitions resources over time or frequency, which is simple to implement while leading to resource scarcity and limited system throughput in large-scale connection scenarios [7]. To address this limitation, the rate splitting multiple access (RSMA) protocol has been proposed. In this protocol, the transmitter divides each message into common and private parts, and the receivers apply successive interference cancellation for decoding [8, 9]. In this manner, the system enables flexible spatial multiplexing and effective interference management. Therefore, RSMA achieves higher spectral efficiency and stronger robustness than OMA in complex downlink environments of VLC systems.

Although RSMA has been widely investigated in radio frequency communication, it still has considerable challenges in VLC systems, particularly in highly dynamic and practical mobile scenarios. In such scenarios, the time-varying locations of the users cause rapid channel fluctuations and significant system uncertainty. Meanwhile, RSMA introduces coupled resource allocation variables, which include precoding design and common rate allocation, thereby leading to a highly non-convex optimization problem with intricate variable dependencies. Consequently, it becomes challenging to design efficient and robust resource allocation strategies in such dynamic scenarios.

At present, resource allocation for RSMA mainly relies on convex optimization or heuristic search methods [10, 11]. However, these methods require accurate channel information and repeated optimization procedures, which leads to limited adaptability in dynamic environments [12]. In contrast, deep reinforcement learning (DRL) provides a data-driven approach for decision-making in time-varying systems [13]. The agent interacts with the environment and updates the policy according to the received reward, thereby enabling continuous adaptation to channel variations without explicit modeling of system dynamics. However, the performance of actor-critic-based DRL algorithms depends heavily on the accuracy of the critic network in estimating the state-action value function. Inaccurate Q-value estimation introduces bias into the policy update, thereby causing slow convergence, unstable training, and performance oscillation [14].

To address this issue faced by DRL algorithms, generative artificial intelligence (GenAI) provides an effective solution [15]. Specifically, the generative modeling mechanism enables the critic network to learn the distribution characteristics of the state-action-reward relationship. In this manner, the model can better approximate complex nonlinear mappings, thereby reducing function approximation error and improving the accuracy of value estimation. Consequently, the actor network can obtain more reliable gradient information, which leads to improved stability and convergence performance of policy learning.

Therefore, this paper mainly studies the resource allocation in mobile VLC with the RSMA system and proposes a GenAI-based DRL algorithm to solve it. The main contributions of this paper are as follows:

- *Dynamic Indoor VLC System with RSMA*: We consider a dynamic indoor VLC system, where multiple LEDs deployed on the ceiling provide illumination while serving multiple mobile terrestrial users. The mobility of the users leads to time-varying channel conditions, which introduces uncertainty into the system. Moreover, we adopt the RSMA protocol to improve communication efficiency. The considered system is applicable to several scenarios, including indoor smart offices and industrial automation.
- *Optimization Problem Formulation*: Considering both communication quality and energy consumption, we formulate an optimization problem to jointly optimize the precoding matrix and the portion allocation of the common stream for maximizing the energy efficiency. The energy efficiency is defined as the ratio of the achievable communication rate to the total transmit power. The formulated problem involves coupled optimization variables and exhibits dynamic characteristics, which makes it difficult to solve by using conventional optimization methods.
- *GenAI-based DRL Algorithm*: We propose an improved deep deterministic policy gradient (IDDPG) algorithm. Specifically, we introduce a generative adversarial network (GAN) to train the critic network, thereby improving the accuracy of Q-value estimation by capturing the distribution characteristics of the state-action value function. Moreover, we adopt the prioritized experience replay (PER) mechanism to assign higher sampling priority to experiences with larger temporal-difference errors, thereby improving the efficiency of sample usage and accelerating the convergence process.
- *Simulation and Analysis*: We conduct extensive simulations to evaluate the proposed algorithm under different system settings. The results show that the IDDPG algorithm outperforms two conventional strategies and three DRL algorithms in terms of energy efficiency. Furthermore, it achieves faster convergence and more stable training performance among all compared DRL algorithms. Moreover, it maintains superior performance under increasing randomness of user mobility, thereby demonstrating strong adaptability to dynamic VLC environments.

The organization of this paper is as follows: Section 3 presents the system model and formulates the optimization problem. Section 2 show the related works. Section 4 proposes the IDDPG algorithm. Simulation results and analysis are shown in Section 5. Section 6 concludes the paper.

2 Related Work

In this section, we review related work to demonstrate the difference in our work.

Indoor VLC System: Many works have studied the application of OMA in indoor VLC systems due to its ease of implementation. For example, the authors in [16] adopted a dynamic time division multiple access (TDMA) protocol to schedule the communication process in a VLC downlink system. Moreover, in view of the low-pass characteristics of LEDs, the authors in [17] designed an indoor VLC system based on the orthogonal frequency division multiple access (OFDMA) protocol. Although OMA is simple to deploy and avoids inter-user interference, it is only suitable for small-scale scenarios with a small number of users due to its low communication efficiency. In this case, non-orthogonal multiple access (NOMA) has attracted academic attention in VLC applications due to the capability of power-domain multiplexing [18–20], which can improve communication efficiency and communication efficiency in large-scale scenarios. Notably, RSMA combines the features of NOMA and space division multiple access (SDMA) [21], and thus it shows higher flexibility and stronger performance than NOMA. For example, the authors in [21] investigated the application of RSMA in indoor VLC systems, and the simulation results show that it achieves better transmission performance than NOMA and SDMA. Moreover, the authors in [22] examined the feasibility and effectiveness of RSMA in multi-cell VLC systems, which have more complex interference relationships than conventional single-cell systems. However, most existing studies above only consider static or quasi-static environments, and they ignore the impact of user mobility on resource allocation in multiple access schemes.

Performance Metrics in Indoor VLC System: Many existing works have focused on optimizing the communication performance of VLC systems. For example, the authors in [23] maximized the signal-to-noise ratio (SNR) of a multi-user indoor VLC system by selecting communication LED nodes from LED array. Moreover, the authors in [24]investigated a NOMA-based indoor VLC system and maximized spectral efficiency and system throughput by dynamically allocating power among users with similar conditions and applying NOMA decoding. The authors in [25] considered a multi-input single-output VLC system and optimized LED beamforming to maximize spectral efficiency under communication quality constraints. In addition, the authors in [26] proposed a UAV-assisted indoor LED deployment scheme to avoid obstacles and optimized the deployment locations of LEDs to maximize the average SNR and reduce the variance of SNR. Similarly, the authors in [27] deployed intelligent reflecting surface (IRS) to mitigate the impact of indoor obstacles and optimized the configuration of IRS and resource allocation to maximize the transmission rate. However, the above studies mainly focus on transmission performance, and they ignore the energy consumption of LEDs, which limits their application in green and energy-efficient scenarios.

Optimization Approach in Indoor VLC System: Many existing works have adopted conventional optimization methods to improve the performance of indoor VLC systems. For example, the authors in [28] proposed a hybrid swarm intelligence algorithm that combines the ocean predator algorithm and the firefly algorithm to optimize the LED deployment, thereby improving both illumination quality and communication

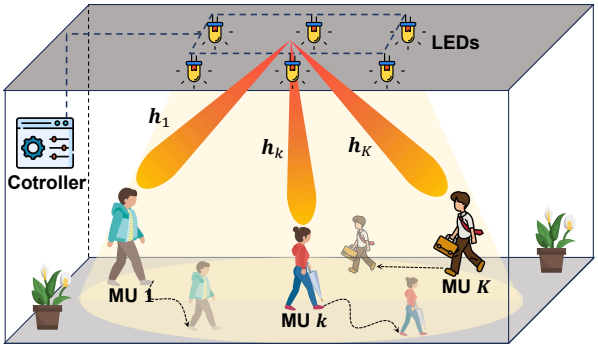

Fig. 1 System model of dynamic indoor VLC system. Specifically, LEDs provide VLC services for multiple mobile users.

performance. Similarly, the authors in [29] optimized beamforming by using a genetic algorithm and a particle swarm optimization algorithm to maximize the achievable data rate. Moreover, the authors in [30] investigated an optical intelligent reflecting surface-assisted VLC system and maximized the sum rate by jointly optimizing active and passive beamforming. To solve this problem, they decomposed the original formulation by using a block coordinate descent method and then applied an iterative algorithm to solve them. However, these methods rely on accurate environmental information and repeated optimization procedures, which leads to performance degradation in dynamic scenarios. Therefore, many recent works have adopted DRL as an alternative solution. For example, the authors in [31] applied the deep deterministic policy gradient (DDPG) algorithm to optimize the performance of a hybrid VLC and millimeter-wave communication system. Moreover, the authors in [32] employed the proximal policy optimization (PPO) algorithm to enhance communication security in indoor VLC systems. Furthermore, the authors in [33] developed an online power allocation strategy based on a multi-agent deep Q-network (DQN) for hybrid radio frequency and VLC networks, thereby maximizing the achievable data rate. Nevertheless, the above DRL algorithms still rely on multilayer perceptron (MLP)-based network structures, which limits their capability to model complex nonlinear relationships in dynamic environments.

3 System Model and Problem Formulation

In this section, we present the details of the system model and formulate an optimization problem that aims to improve the energy efficiency accordingly.

3.1 Overview of System Model

As illustrated in Fig. 1, we consider a dynamic indoor VLC system, where N LEDs denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$ provide VLC services for K MUs denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$. During the VLC process, the intensity modulation and direct detection (IM/DD) technology is adopted for the LEDs to ensure that the emitted

optical intensity is proportional to the amplitude of the transmitted electrical signal. Meanwhile, all MUs are equipped with photodetectors to receive optical signals and recover the transmitted information through electrical conversion and demodulation. We assume that all LEDs are coordinated by a central controller, which has access to complete user information. Moreover, the VLC process in the considered system is managed by the RSMA technology for better communication performance. Specifically, RSMA divides the transmitted information into a common stream and multiple private streams, where the common stream is shared by all users, while each private stream is intended for a specific user. As such, the RSMA technology is capable of improving the spectrum and communication efficiency through the information splitting-based transmission scheme.

To characterize the system dynamics, the entire operation period is divided into T time slots denoted as $\mathcal{T} = \{= 1, 2, \ldots, T\}$ with equal-length t_{d}. Moreover, a three-dimensional (3D) Cartesian coordinate system is adopted to describe the spatial locations of system elements in each time slot, where $\mathbf{q}_n^{\mathrm{L}} = [x_n^{\mathrm{L}}, y_n^{\mathrm{L}}, z_n^{\mathrm{L}}]$ and $\mathbf{q}_k^{\mathrm{M}}[t] = [x_k^{\mathrm{M}}[t], y_k^{\mathrm{M}}[t], 0]$ denote the locations of the n-th LED and k-th MU in time slot t, respectively.

In the following, the models for the channels, transmission process, user mobility, and power consumption related to the system are introduced in detail.

3.2 Channel Model

As studied in [34], the optical wireless channel is typically dominated by line-of-sight (LoS) propagation, since the non-line-of-sight (NLoS) links are negligible compared to the LoS links in the indoor environment. In this case, based on the Lambertian emission model, the channel gain between the n-th LED and the k-th MU in time slot t can be expressed as

$$h_{k,n}[t] = \frac{(m+1)A_{\mathrm{p}}B_{\mathrm{o}}R_{\mathrm{p}}}{2\pi d_{k,n}^2[t]} \cos^m(\phi_{k,n}[t]) \cos(\psi_{k,n}[t]) g(\psi_{k,n}[t]) \tag{1}$$

where m denotes the Lambertian order given by $m = -1/\log_2(\cos(\Phi_{1/2}))$, with $\Phi_{1/2}$ being the semi-angle at half power. Moreover, A_{p} represents the physical area of the photodetector, B_{o} denotes the gain of the optical filter, and R_{p} is the responsivity of the photodetector. In addition, $d_{k,n}[t]$ is the distance between the n-th LED and the k-th MU, while $\phi_{k,n}[t]$ and $\psi_{k,n}[t]$ denote the irradiance angle and incidence angle, respectively. Furthermore, the gain $g(\psi_{k,n}[t])$ of the optical concentrator is defined as

$$g(\psi_{n,k}) = \begin{cases} \frac{I_{\mathrm{r}}^2}{\sin^2(\Psi_{\mathrm{F}})}, & 0 \le \psi_{n,k} \le \Psi_{\mathrm{F}} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where I_{r} is the refractive index and Ψ_{F} denotes the field of view (FOV) of the photodetector.

3.3 Transmission Model

We adopt the RSMA technology to regulate the message transmission in the VLC, which splits each user message into a shared common part and a private part to achieve higher spectrum efficiency and lower inter-user interference. Specifically, the message intended for the k-th MU in time slot t is defined as $G_k[t]$, while the corresponding common part and private part of $G_k[t]$ are represented as $G_k^{\mathrm{c}}[t]$ and $G_k^{\mathrm{p}}[t]$, respectively. Note that all common parts $\{G_k^{\mathrm{c}}[t]\}_{k\in\mathcal{K}}$ are jointly encoded into a common stream denoted by $s_0[t]$, while the private parts $\{G_k^{\mathrm{p}}\}_{k\in\mathcal{K}}$ are independently encoded into K private streams denoted by $\{s_k[t]\}_{k\in\mathcal{K}}$. As such, the transmitted stream in time slot t is represented as $\mathbf{s}[t] = [s_0[t], s_1[t], \ldots, s_K[t]]^T$. Let $\mathbf{W}[t] = [\mathbf{w}_0[t], \mathbf{w}_1[t], \ldots, \mathbf{w}_K[t]]$ denote the precoding matrix in time slot t, where $\mathbf{w}_0[t]$ and $\mathbf{w}_k[t]$ are the precoding vectors for the common stream and k-th private stream with $\mathbf{w}_i[t] = [\mathbf{w}_{1,i}[t], \mathbf{w}_{2,i}[t], \ldots, \mathbf{w}_{N,i}[t]]^T$ and $i \in \{0, 1, \ldots, K\}$. In this case, the transmitted signal of the LEDs in time slot t is given by [35]

$$\mathbf{x}[t] = \underbrace{\mathbf{w}_0 s_0[t]}_{\text{Common}} + \underbrace{\sum\nolimits_{k=1}^{K} \mathbf{w}_k s_k[t]}_{\text{Private}} + \underbrace{\mathbf{d}_{\mathrm{dc}}}_{\text{DC bias}} \tag{3}$$

where $\mathbf{d}_{\mathrm{dc}} = (I_{\mathrm{dc}})_{N\times 1}$ represents the drive current (DC) bias that guarantees the non-negative characteristics of the transmitted signal.

Then, the MUs utilize the equipped photodiode to convert the optical signal emitted by the LEDs to an electrical signal. Accordingly, the received signal at the k-th user in time slot t is expressed as

$$y_k[t] = \mathbf{h}_k^T[t]\mathbf{x}[t] + \eta_k[t] \tag{4}$$

where $\mathbf{h}_k[t] = [h_{k,1}[t], h_{k,2}[t], \ldots, h_{k,N}[t]]^T$ represents the channel vector from all LEDs to k-th MU in time slot t. Moreover, $\eta_k[t] \sim \mathcal{N}(0, \sigma_k^2[t])$ denotes additive noise.

Subsequently, in the decoding signal process, the MUs can remove the DC part involved in the received signal shown in Eq. (4) by using the alternating current (AC) coupling. After that, they prioritize decoding the common stream s_0 while treating all private streams strictly as additive noise. In this case, the achievable common rate of decoding the common stream for the k-th MU in time slot t is denoted by $R_k^{\mathrm{c}}[t] = \log_2(1 + e\gamma_k^{\mathrm{c}}[t]/2\pi)$, where the signal-to-interference-and-noise ratio (SINR) $\gamma_k^{\mathrm{c}}[t]$ is represented as [36]

$$\gamma_k^{\mathrm{c}}[t] = \frac{|\mathbf{h}_k^T[t]\mathbf{w}_0[t]|^2}{\sum_{j=1}^{K} |\mathbf{h}_k^T[t]\mathbf{w}_j[t]|^2 + \sigma_k^2[t]} \tag{5}$$

Given that s_0 is a shared component among all MUs, it is imperative that every MU be capable of decoding it reliably. To achieve this, the common rate is determined by the minimum achievable common rate among all MUs, and the corresponding

constraints are expressed as [22]

$$R^{\mathrm{c}}[t] = \min_{k \in \mathcal{K}} R_k^{\mathrm{c}}[t] \tag{6a}$$

$$\sum_{k=1}^{K} O_k^{\mathrm{c}}[t] = R^{\mathrm{c}}[t] \tag{6b}$$

where $O_k^{\mathrm{c}}[t]$ is the portion of k-th MU in the common rate $R^{\mathrm{c}}[t]$.

Once the common stream is decoded and removed via successive interference cancellation (SIC), each MU decodes its own private stream with all other private streams treated as noise. In this case, the achievable private rate of k-th MU in time slot t is expressed as $R_k^{\mathrm{p}}[t] = \log_2(1 + e\gamma_k^{\mathrm{p}}[t]/2\pi)$, where SINR $\gamma_k^{\mathrm{p}}[t]$ is given by [22]

$$\gamma_k^{\mathrm{p}}[t] = \frac{|\mathbf{h}_k^{\mathrm{T}}[t]\mathbf{w}_k[t]|^2}{\sum_{j=1, j\neq k}^{K} |\mathbf{h}_k^{\mathrm{T}}[t]\mathbf{w}_j[t]|^2 + \sigma_k^2[t]} \tag{7}$$

Finally, the total achievable rate of the k-th MU in time slot t is obtained by combining the allocated portion of the common rate and the private rate as [37]

$$R_k[t] = O_k^{\mathrm{c}}[t] + R_k^{\mathrm{p}}[t] \tag{8}$$

3.4 Power Consumption

Note that the precoded signal superimposed with a DC bias should remain strictly within the linear dynamic range of the LEDs to avoid nonlinear distortion caused by signal truncation. To achieve this, the precoding matrix needs to satisfy the following constraints, *i.e.*,

$$\|\mathbf{W}(n,:)[t]\|_1 = \sum_{j=0}^{K} |\mathbf{w}_{nj}[t]| \leq \delta \tag{9}$$

where $\mathbf{W}(n,:)[t]$ denotes the n-th row of the precoding matrix $\mathbf{W}$ in time slot t. Moreover, $\delta = \min\{I_{\mathrm{dc}} - I_{\mathrm{L}}, I_{\mathrm{U}} - I_{\mathrm{dc}}\}$ [38], where I_{L} and I_{U} represent the minimum and maximum of the DC, respectively.

The precoding matrix that satisfies the constraint (9) is effective. Then, the power consumption of the LEDs in each time slot depends on the effective precoding matrix, which is denoted as [22]

$$P[t] = P_{\mathrm{d}}[t] + P_{\mathrm{s}} \tag{10}$$

where $P_{\mathrm{d}}[t] = \lambda \operatorname{tr}(\mathbf{W}[t]\mathbf{W}[t]^T)$ is the dynamic transmit power consumption, and λ represents the power amplifier coefficient. Moreover, since $\mathbf{d}_{\mathrm{dc}}$ is assumed to be fixed, $P_{\mathrm{s}} = V_{\mathrm{LED}} 1_N^T \mathbf{d}_{\mathrm{dc}}$ presents the static power consumption, where 1_N is the all-one vector and V_{LED} relates to the forward voltage of the LED.

3.5 Mobility Model

The MUs follow the Gauss-Markov model [39]. In this case, the moving speed $v_k[t]$ and direction $\xi_k[t]$ of the k-th MU in time slot t are expressed as

$$v_k[t] = \zeta_{\mathrm{v}} v_k[t-1] + (1-\zeta_{\mathrm{v}})\mu_{k,\mathrm{v}} + X[t-1]\sqrt{1-\zeta_{\mathrm{v}}^2} \tag{11a}$$

$$\xi_k[t] = \zeta_{\mathrm{d}} \xi_k[t-1] + (1-\zeta_{\mathrm{d}})\mu_{k,\mathrm{d}} + Y[t-1]\sqrt{1-\zeta_{\mathrm{d}}^2} \tag{11b}$$

where $\xi_{\mathrm{v}} \in [0,1]$ and $\xi_{\mathrm{d}} \in [0,1]$ are the parameters that control the randomness of the speed and direction change, respectively. Generally, the higher values of ξ_{v} and ξ_{d} indicate that the volatility of MU mobility is lower. Moreover, $\mu_{k,\mathrm{v}}$ and $\mu_{k,\mathrm{d}}$ represent the mean speed and direction of the k-th MU, respectively. In addition, X and Y denote the random numbers that are sampled from the Gaussian distribution. Negative speeds are clipped to zero, and directions are normalized to $[-\pi, \pi)$.

Based on the Gauss-Markov model above, the location of the k-th MU in the x-axis and y-axis directions in time slot t is given by

$$x_k^{\mathrm{M}}[t] = x_k^{\mathrm{M}}[t-1] + v_k[t]\cos(\xi_k[t])t_{\mathrm{d}} \tag{12a}$$

$$y_k^{\mathrm{M}}[t] = y_k^{\mathrm{M}}[t-1] + v_k[t]\sin(\xi_k[t])t_{\mathrm{d}} \tag{12b}$$

3.6 Problem Formulation and Analysis

Considering that increasing transmission rate helps achieve better communication performance, while reducing transmit power consumption helps save power resources, we take energy efficiency as the optimization objective, which is defined as [22]

$$E[t] = \frac{\sum_{k=1}^{K} R_k[t]}{P_{\mathrm{d}}[t]} \tag{13}$$

where $E[t]$ denotes the energy efficiency in time slot t. Moreover, we select $P_{\mathrm{d}}[t]$ to represent the power consumption of LEDs rather than $P[t]$ since P_{s} is static in our considered system.

Optimizing the energy efficiency depends on two key variables, *i.e.*, the precoding matrix design and portion allocation of the common rate. First, given the mobility of the MUs, the precoding matrix should be adjusted in real time to achieve more substantial rates, while its adjustment also directly affects the transmit power consumption of LEDs according to Eq. (10). Moreover, since the RSMA technology splits the transmitted message into a shared public stream and multiple private streams, the portion of the public rate allocated to each MU directly impacts the actual achievable rate. As such, the optimization variables are defined as: *(i)* $\mathcal{W} = \{\mathbf{W}[t] | t \in \mathcal{T}\}$ represents the precoding matrix set across the operation period. *(ii)* $\mathcal{C} = \{O_k^{\mathrm{c}}[t] | k \in \mathcal{K}, t \in \mathcal{T}\}$ denotes the portion allocation set of the common rate of all MUs across the operation period.

Accordingly, the optimization problem related to the considered system is formulated as follows:

$$\mathcal{P}1: \max_{\mathcal{W},\mathcal{C}} \sum_{t=1}^{T} E[t] \tag{14a}$$

$$\text{s.t.}: \|\mathbf{W}(n,:)[t]\|_1 \leq \delta \quad \forall n \in \mathcal{N},\ t \in \mathcal{T} \tag{14b}$$

$$\sum_{k=1}^{K} O_k^{\mathrm{c}}[t] \leq R_k^{\mathrm{c}}[t] \quad \forall t \in \mathcal{T} \tag{14c}$$

$$R_k[t] \geq R_{\mathrm{th}} \quad \forall k \in \mathcal{K},\ t \in \mathcal{T} \tag{14d}$$

where the constraint (14b) represents that the precoded signal should remain within the linear operation range of the LEDs. Moreover, the constraint (14c) ensures that all mobile users are able to decode the common stream. In addition, the constraint (14d) indicates that the achievable rate of every MU should exceed the threshold R_{th} to keep the quality of the communication process.

As we can see, the optimization problem shows several challenging characteristics. First, given that the precoding matrix and portion allocation of the common rate are coupled, the optimization problem is non-convex. Second, since the locations of MUs vary during the system operation, the optimization variables should be adjusted in each time slot to adapt to the current system states, which means that the optimization problem is dynamic.

4 Proposed Algorithm

In this section, we propose an improved DRL algorithm to solve the optimization problem.

4.1 Motivation of Utilizing DRL Algorithm

The dynamic characteristics of the optimization problem pose critical challenges for the conventional optimization algorithm, while motivating us to select the DRL algorithm to address the optimization problem. Specifically, conventional optimization algorithms typically depend on accurate mathematical models to analyze the relationships between optimization variables and optimization objectives while considering the constraints for finding optimal solutions. This indicates that they need accurate environmental prior knowledge for mathematical modeling, which is challenging in the considered system due to the random mobility of the MUs [40]. Moreover, conventional optimization algorithms lack adaptability, which indicates that they have to utilize mathematical models to recalculate when the environment changes, thereby resulting in higher decision-making latency and computational resource consumption [12].

In contrast, DRL algorithms are inherently designed to handle dynamic and uncertain environments without requiring an explicit mathematical model of the system. First, DRL adopts a trial-and-error learning paradigm for policy learning. Specifically,

the DRL agent interacts with the environment, collects state-action-reward experiences, and gradually learns an optimal policy accordingly, which eliminates the need for accurate prior knowledge of MU mobility patterns. Second, DRL incorporates neural networks as function approximators, which can capture complex nonlinear relationships and generalize across similar states. This enables the agent to adapt to environmental changes in real time without recomputing from scratch. These characteristics above make DRL particularly suitable for solving the dynamic optimization problem in our system [41].

In the following, we present the optimization problem reformulation by using the Markov decision process (MDP) for facilitating DRL application.

4.2 MDP

The MDP provides a mathematical framework for modeling sequential decision making in dynamic environments. It consists of three fundamental elements, *i.e.*, the state space, action space, and reward function. Specifically, the agent observes the current environment state, selects an action according to the current policy, and then receives a reward signal used to evaluate and refine the policy. During the process, the agent aims to learn an optimal policy that maximizes the cumulative reward over time. In the considered system, the controller acts as the agent, and the state space, action space, and reward function are defined as follows.

- **State Space** $\mathcal{S}$: The state captures the essential spatial information of the system elements, including the locations of LEDs and MUs observed by the controller. In this case, the environment state $s[t] \in \mathcal{S}$ in time slot t is represented as

$$s[t] = \left[q_1^{\mathrm{L}}, q_2^{\mathrm{L}} \ldots, q_N^{\mathrm{L}}, q_1^{\mathrm{M}}[t], q_2^{\mathrm{M}}[t] \ldots, q_K^{\mathrm{M}}[t]\right] \tag{15}$$

- **Action Space** $\mathcal{A}$: The action represents the decisions made by the agent, which are equivalent to the optimization variables in the formulated optimization problem. In this case, the action $a[t] \in \mathcal{A}$ in time slot t is given by

$$a[t] = [\mathbf{W}[t], O_1^{\mathrm{c}}[t], O_2^{\mathrm{c}}[t], \ldots, O_K^{\mathrm{c}}[t]] \tag{16}$$

- **Reward Function** $r[t]$: The reward is designed to guide the agent toward maximizing the energy efficiency of the system while satisfying the constraints. In this case, the reward $r[t]$ in time slot t is expressed as

$$r[t] = E[t] + F[t] \tag{17}$$

  where $F[t] = -\rho \sum_{k=1}^{K} \max(0, R_{\mathrm{th}} - R_k[t])$ is the penalty for managing the communication rate constraint with $\rho > 0$.

In the following, we show how to use the DRL algorithm to solve the optimization problem according to this MDP.

4.3 Conventional DDPG

In this part, we present the training process of the conventional DDPG algorithm and clarify its limitations in solving the formulated optimization problem.

4.3.1 Principles of Conventional DDPG

The DDPG is an actor-critic algorithm designed for continuous action spaces and has been widely used for various applications due to its ease of implementation and efficiency [42]. The optimization objective of the DDPG algorithm is to maximize the cumulative reward, which is achieved by updating four neural networks during the training process. Specifically, the actor network $\pi_\theta(s)$ with parameter θ acts as the decision-making engine of the DRL, which maps a state s to a deterministic action a. The critic network $Q_\phi(s,a)$ with parameter ϕ estimates the quality of the actor network, which represents the expected cumulative reward when taking action a in state s. Moreover, the target actor network $\pi_{\theta'}(s)$ and the target critic network $Q_{\phi'}(s,a)$ are copies of the original networks, which provide consistent targets for the temporal-difference (TD) learning. Notably, their parameters θ' and ϕ' are updated slowly to stabilize training.

Different from the online training, DDPG uses an experience buffer $\mathcal{B}$ to store the collected transitions, which preserve the past learning memories for facilitating better training. Specifically, each transition is a tuple $(s[t], a[t], r[t], s'[t])$ collected by the agent during interaction with the environment, where $s'[t]$ represents the next environment state after executing $a[t]$ in the state $s[t]$. At each training step, a mini-batch of transitions is sampled uniformly from $\mathcal{B}$ for network updates. This sampling breaks the temporal correlation between consecutive samples and improves sample efficiency.

Based on the sample transitions from the experience buffer, the critic network is updated by minimizing the TD loss, which is defined as the mean squared error between the predicted Q-value and target Q-value. As such, the loss function of the critic network is given by

$$\mathcal{L}(\phi) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{B}}\left[\left(Q_\phi(s,a) - (r + \gamma Q_{\phi'}(s', \pi_{\theta'}(s')))\right)^2\right] \tag{18}$$

where γ is the discount factor.

Since the actor outputs a deterministic action for each state, its objective is to find a policy that yields the highest action value. In this case, the actor network is updated based on the same mini-batch of experience to maximize the expected value estimated by the critic network. Therefore, the loss function for the actor network is defined as

$$\mathcal{L}(\theta) = \mathbb{E}_{s\sim\mathcal{B}}\left[-Q_\phi(s, \pi_\theta(s))\right] \tag{19}$$

The target networks are updated by using soft updates with a small coefficient $\tau \in (0,1)$ that controls the update rate. The update processes of the target networks

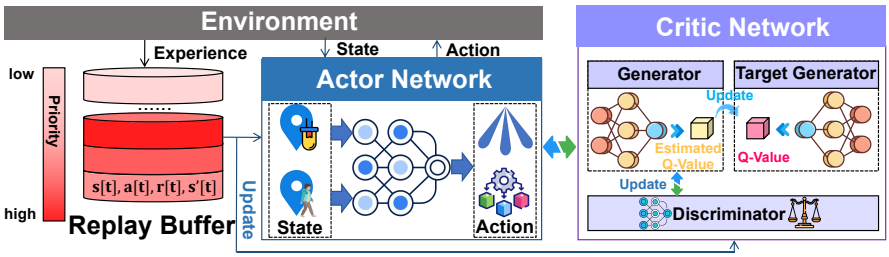

Fig. 2 The framework of the IDDPG algorithm.

are represented as

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \quad \phi' \leftarrow \tau\phi + (1-\tau)\phi' \tag{20}$$

4.3.2 Limitations of DDPG

Despite its simplicity and efficiency, the standard DDPG algorithm suffers from several limitations that may affect performance in complex environments. These issues are discussed as follows.

- *Training Instability*: In the conventional DDPG algorithm, the critic network plays a central role in evaluating the action-value function. However, the critic is prone to overestimation bias. This bias arises because the critic network uses its own estimates to compute the target value during TD learning. As the critic repeatedly updates toward biased targets, the overestimation error accumulates. Consequently, the critic may provide inaccurate Q-value estimates, which mislead the actor network. In this case, the training process becomes unstable, exhibiting large oscillations in the loss function and slow convergence [43].
- *Uniform Random Sampling from Experience Buffer*: The conventional DDPG algorithm samples transitions uniformly from the experience buffer. This simple strategy treats all transitions equally, regardless of their importance. However, not all experiences contribute equally to effective learning. For example, some transitions with high TD error are more informative than others. Uniform sampling may lead to inefficient learning because valuable experiences are reused as frequently as ordinary ones [44].

Accordingly, we propose an IDDPG algorithm with targeted improvements for overcoming the limitations above to improve its performance.

4.4 IDDPG Algorithm

Fig. 2 and Algorithm 1 show the framework and detailed steps of the IDDPG algorithm. Specifically, the IDDPG algorithm utilizes the GAN and PER mechanisms to improve the training stability and learning efficiency, respectively. In the following, the two improvements are detailed.

Algorithm 1 IDDPG

Initialize the actor, generator, and discriminator networks with parameters θ, χ, and ψ
Initialize the target networks as $\theta' \leftarrow \theta$ and $\chi' \leftarrow \chi$
Initialize the replay buffer $\mathcal{B}$ and batch size B
Initialize the number of episodes I and time horizon T
for episode $= 1$ **to** I **do**
    for $t = 1$ **to** T **do**
        Observe the state $s[t]$
        Select the action $a[t]$ according to π_θ
        Receive the reward $r[t]$ and observe the next state $s'[t]$
        Store the transition $\{s[t], a[t], r[t], s'[t]\}$ in the replay buffer $\mathcal{B}$
    end for
    Sample B transitions from $\mathcal{B}$ according to the probability in (27)
    Update θ, χ, and ψ according to (19), (29), and (24) according to the sampled transitions
    Update the target networks according to (20) and (25)
end for
return the actor network

4.4.1 GAN-based Q-value Estimation

GAN is a GenAI model known for generating high-quality and realistic samples through adversarial training [45]. Specifically, it consists of two competing networks. One is a generator network G, which aims to generate realistic synthetic samples, and the other is a discriminator network D, which aims to distinguish between real and fake samples. The optimization objectives of the generator and discriminator are as follows:

$$\max_G \mathbb{E}_{\mathbf{z}\sim P_\mathrm{z}} [\log(1 - D(G(\mathbf{z}))] \tag{21a}$$

$$\min_D \mathbb{E}_{\mathbf{x}\sim P_\mathrm{d}} [\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim P_\mathrm{z}}[\log(1 - D(G(\mathbf{z})))] \tag{21b}$$

where P_z and P_d represent the random noise distribution and real data distribution, respectively. Accordingly, $\mathbf{z}$ and $\mathbf{x}$ are the noise and real data, respectively.

The two networks are trained simultaneously in a minimax game, which is expressed as

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim P_\mathrm{d}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim P_\mathrm{z}} [\log(1 - D(G(\mathbf{z})))] \tag{22}$$

This adversarial mechanism effectively reduces the distribution gap between generated outputs and target data. Inspired by this characteristic, we introduce the GAN framework into the critic network of DDPG to improve the Q-value estimation process.

Specifically, we reformulate the Q-value estimation as an adversarial learning task so that the estimation accuracy can be improved [46]. First, we define a generator network $G_\chi(s,a)$ with parameter χ to produce a Q-value estimate for a given state action pair (s,a). Then, we introduce a target generator network $G_{\chi'}(s,a)$ with parameter χ' as a copy of the generator network so that stable targets can be provided. Moreover, we define a discriminator network $D_\psi(Q)$ with parameter ψ to distinguish the source of the Q-value. The discriminator takes a Q-value as input and outputs the probability that the Q-value is generated by the target generator network rather than the generator network. In the following, we introduce how to update the networks above according to the adversarial learning structure, thereby making the generator network produce Q-value estimates that are close to the output distribution of the target generator network.

First, the generator is updated by minimizing a combined loss, which consists of an adversarial term and a mean-squared error term [47]. The adversarial term encourages the generator to fool the discriminator, and the mean-squared error term ensures the generated Q-value approximates the target Q-value. The loss for the generator is defined as

$$\mathcal{L}(\chi) = \mathbb{E}_{(s,a,r,s')\in\mathcal{B}}\left[-\log(1 - D_\psi(G_\chi(s,a))) + \delta^2\right] \tag{23}$$

where $\delta = \hat{y} - G_\chi(s,a)$ represents the TD error, where $\hat{y} = r + \gamma G_{\chi'}(s', \pi_{\theta'}(s'))$ is the target value computed by using the target generator.

Then, the discriminator is updated to correctly distinguish between the generator output and the target value [47]. Its loss function is defined as

$$\mathcal{L}(\psi) = -\mathbb{E}_{(s,a,r,s')\in\mathcal{B}}\left[\log(D_\psi(\hat{y})) + \log(1 - D_\psi(G_\chi(s,a)))\right] \tag{24}$$

Similar to the conventional DDPG algorithm, the target generator network is updated via soft updates to maintain stability, which is given by

$$\chi' = \tau\chi + (1-\tau)\chi' \tag{25}$$

4.4.2 PER-based Transition Sampling

The core idea of PER is to sample transitions with high priority more frequently so that the agent can learn from informative experiences, thereby improving the learning efficiency [48]. In this mechanism, the priority is measured by the TD error, where a large TD error indicates that the transition is unexpected for the current agent and thus contains more learning information. Therefore, we express the priority of the i-th transition in the experience buffer as

$$p_i = |\delta_i| + \epsilon \tag{26}$$

where δ_i denotes the TD error of the i-th transition. Moreover, ϵ is a small constant, which ensures that transitions with small TD errors can still be sampled.

Then, the sampling probability of the i-th transition is defined based on its priority. Specifically, the probability is set proportional to the priority raised to the power α,

where α controls the strength of prioritization. When $\alpha = 0$, the sampling reduces to the uniform case. Accordingly, the sampling probability $P(i)$ is expressed as

$$P(i) = \frac{p_i^\alpha}{\sum_j p_j^\alpha} \tag{27}$$

However, this prioritization scheme introduces bias in the gradient estimation since the sampling distribution deviates from the uniform distribution. Therefore, the importance sampling weights are introduced to correct this bias so that the update becomes unbiased. Specifically, the weight of the i-th transition is represented as

$$\omega_i = (N_\mathrm{e} \cdot P(i))^{-\beta} \tag{28}$$

where N_e denotes the size of the experience buffer, and β is a hyperparameter that controls the strength of the correction.

To demonstrate the different contributions of the sampled transitions during the learning process, we integrate the importance sampling weights ω into the generator update. Specifically, we weight the TD error of each sampled transition and combine it with the adversarial loss. Accordingly, we express the loss function of the generator network $Q_\chi(s, a)$ as

$$\mathcal{L}(\chi) = \mathbb{E}_{(s,a,r,s')\in\mathcal{B}}\left[-\log(1 - D_\psi(G_\chi(s, a))) + \omega\delta^2\right] \tag{29}$$

4.5 Algorithm Analysis

In this part, we present the design details of the network structures and analyze the computation complexity of the proposed algorithm.

4.5.1 Network Structure

The proposed GAN-DDPG framework consists of three types of neural networks, which include the actor network, generator network, and discriminator network. These networks are modeled as fully connected feedforward neural networks. Specifically, the actor network is a G-layer feedforward neural network, which takes the state as input and outputs the action. Let $D_{\mathrm{i,p}}$ and $D_{\mathrm{o,p}}$ denote the input and output dimensions of the actor network, respectively. Each hidden layer contains F_p neurons, and the ReLU is adopted as the activation function except for the output layer that utilizes the tanh function. Similarly, the generator network is an H-layer feedforward neural network, which outputs the Q-value estimation. Different from the actor network, the input of the generator network is the concatenation of the state, action, and random noise. The input dimension is $D_{\mathrm{i,g}}$, and the output dimension is $D_{\mathrm{o,g}}$. Each hidden layer contains F_g neurons with the ReLU function. Meanwhile, the discriminator network is a K-layer feedforward neural network, which takes the Q-value as input and outputs a probability. Let $D_{\mathrm{i,d}}$ and $D_{\mathrm{o,d}}$ denote the input and output dimensions, respectively. Each hidden layer contains F_d neurons, and the ReLU and sigmoid functions are applied in the hidden and output layers, respectively. In addition, the target networks share the same architecture as the corresponding main networks.

Table 1 System Parameters

Notion	Parameter	Value
m	Lambertian order	1
Ψ_{F}	FoV semi-angle	60°
A_{p}	Photodetector area	$10^{-4}\mathrm{m}^2$
B_{o}	Optical filter gain	1
R_{p}	Photodetector responsivity	0.4A/W
$\Phi_{1/2}$	Semi-angle at half power	60°
I_{r}	Refractive index	1.5
λ	Power amplifier coefficient	5
I_{dc}	DC bias	0.5A
I_{L}	Minimum DC	0.1A
I_{U}	Maximum DC	1A
R_{th}	Rate threshold	0.4bps/Hz

4.5.2 Computation Complexity

The computation complexity of the proposed IDDPG algorithm consists of the following two parts:

- *Complexity of Network Update*: Assume that the training process is carried out over I episodes with T time steps per episode. According to the structure of feedforward neural networks, the overall computational complexity of the GAN-DDPG algorithm can be expressed as $\mathcal{O}(IT(D_{\mathrm{i,p}}F_{\mathrm{p}} + (G-2)F_{\mathrm{p}}^2 + F_{\mathrm{p}}D_{\mathrm{o,p}} + D_{\mathrm{i,g}}F_{\mathrm{g}} + (H-2)F_{\mathrm{g}}^2 + F_{\mathrm{g}}D_{\mathrm{o,g}} + D_{\mathrm{i,d}}F_{\mathrm{d}} + (K-2)F_{\mathrm{d}}^2 + F_{\mathrm{d}}D_{\mathrm{o,d}}))$
- *Complexity of Transition Sampling*: The PER mechanism introduces additional computational overhead. Specifically, the prioritized sampling and priority updating operations require $\mathcal{O}(\log N_{\mathrm{e}})$ complexity for each transition. Therefore, the overall complexity introduced by PER can be expressed as $\mathcal{O}\left(IT\log N_{\mathrm{e}}\right)$.

Consequently, the total computational complexity of the proposed GAN-DDPG algorithm with PER is given by $\mathcal{O}(IT(D_{\mathrm{i,p}}F_{\mathrm{p}} + (G-2)F_{\mathrm{p}}^2 + F_{\mathrm{p}}D_{\mathrm{o,p}} + D_{\mathrm{i,g}}F_{\mathrm{g}} + (H-2)F_{\mathrm{g}}^2 + F_{\mathrm{g}}D_{\mathrm{o,g}} + D_{\mathrm{i,d}}F_{\mathrm{d}} + (K-2)F_{\mathrm{d}}^2 + F_{\mathrm{d}}D_{\mathrm{o,d}}) + IT\log N_{\mathrm{e}})$.

5 Simulation

In this section, we conduct simulations to validate the performance of the IDDPG algorithm to improve energy efficiency in the dynamic VLC system.

5.1 Simulation Setup

In the simulation environment, the size of the room is $5 \times 5 \times 3\mathrm{m}^3$, where there are four LEDs and three mobile users. All LEDs are installed on the ceiling of the room and their locations are $[1.5, 1.5, 3.0]$, $[3.5, 1.5, 3.0]$, $[1.5, 3.5, 3.0]$, and $[3.5, 3.5, 3.0]$, respectively. Moreover, the initial locations of the three mobile terrestrial users are

$[1.0, 1.0, 0]$, $[4.5, 4.5, 0]$, and $[3.0, 1.0, 0]$, respectively. Other parameters are shown in Table 1.

For the network configurations of the IDDPG algorithm, we set the actor network and generator network to have two hidden layers with 256 neurons in each layer. Moreover, we set the discriminator network to have one hidden layer with 128 neurons. This difference in the network structure is due to the distinct functional roles of these networks. Specifically, the actor network and generator network aim to learn complex mappings from states to actions and from noise to synthetic samples, respectively, which require sufficient representation capability. Therefore, deeper and wider network structures are adopted for these two networks. However, the discriminator network aims to distinguish real samples from generated samples, which is essentially a binary classification task. Due to this simpler objective, a relatively shallow network structure is sufficient to achieve reliable discrimination performance.

5.2 Comparison Algorithms

We compare the proposed IDDPG algorithm with two conventional strategies.

- *Uniform Distribution (UD)*: The transmit power of each stream for one LED is equal, while the portion of each user in the common rate is also equal in this strategy.
- *Random Distribution (RD)*: The precoding matrix and portion allocation of the common stream are set randomly in this strategy.

Moreover, we consider several DRL algorithms as comparison methods, which include the DDPG, twin delayed deep deterministic policy gradient (TD3) [49], and PPO [50].

- *DDPG*: The original DDPG algorithm is adopted to validate the effectiveness of the proposed improvements.
- *TD3*: This algorithm is the improved version of the DDPG algorithm. Specifically, it improves the performance of DDPG by introducing two critic networks to reduce the overestimation bias of the Q-value. Moreover, it updates the actor network in a delayed manner and adds noise to the target policy, thereby enhancing training stability.
- *PPO*: This algorithm is a policy gradient method based on stochastic policies. It constrains the policy update by a clipped surrogate objective, thereby avoiding large policy deviations and improving training stability.

5.3 Optimization Results

In this part, we first present the optimization performance of different algorithms and then demonstrate the convergence performance of the proposed algorithm.

5.3.1 Algorithm Performance Evaluation

Fig. 3 shows the energy efficiency value obtained by different algorithms. First, we observe that the UD and RD strategies exhibit poor performance due to the lack of

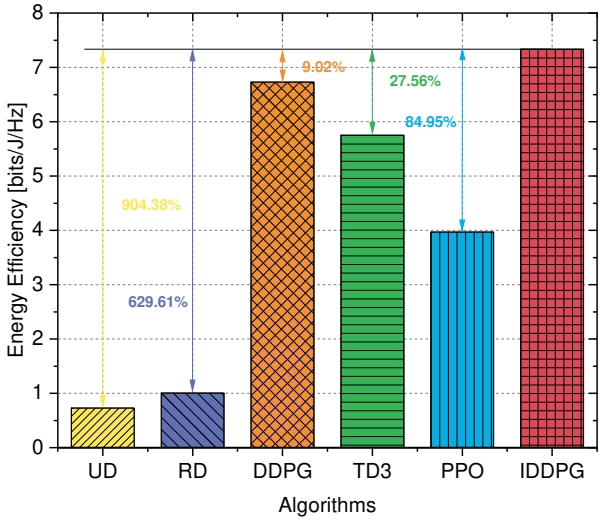

Fig. 3 Energy efficiency of different algorithms.

awareness and adaptability to dynamic environments. In particular, the UD strategy adopts a fixed precoding matrix and a static portion allocation scheme for the common stream, and thus it cannot respond to severe variations in optical channel states caused by user mobility. Consequently, its performance is even inferior to that of the RD strategy, which introduces random perturbations. This result indicates conventional strategies are insufficient for dynamic VLC systems. In contrast, the DRL algorithms achieve significantly better performance due to their strong capability in adaptive learning. Among these algorithms, the performance of PPO is unsatisfactory. This is because PPO relies on an online policy update mechanism, which has low sample efficiency and requires more interactions to approach a high-quality policy in rapidly varying environments. It is worth noting that the IDDPG algorithm achieves the best energy efficiency. Simulation data shows that IDDPG significantly improves energy efficiency by 904.38%, 629.61%, 84.95%, 27.56%, and 9.02% compared to UD, RD, PPO, TD3, and the baseline DDPG algorithm, respectively. This result strongly demonstrates that IDDPG has better robustness in handling resource scheduling problems in complex dynamic VLC environments and can more accurately match real-time changing communication needs.

5.3.2 Convergence Performance

In this part, we begin by analyzing the parameter settings on the algorithm performance, and then compare with other DRL algorithms.

1) Convergence Performance under Different Parameters: First, Fig. 4 shows the reward curves of the IDDPG algorithm under different learning rates (LRs). We observe that when LR$= 3\times10^{-4}$, the IDDPG algorithm achieves a rapid increase in the early training stage and maintains a high and stable average reward in the later stage. In contrast, a smaller learning rate leads to slow parameter updates, which makes the

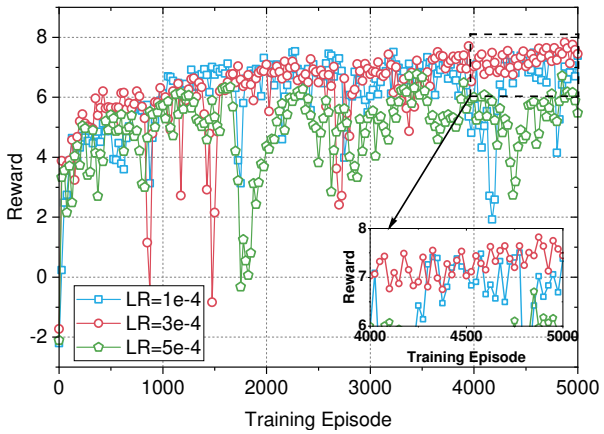

Fig. 4 Reward convergence of IDDPG algorithm under different learning rates.

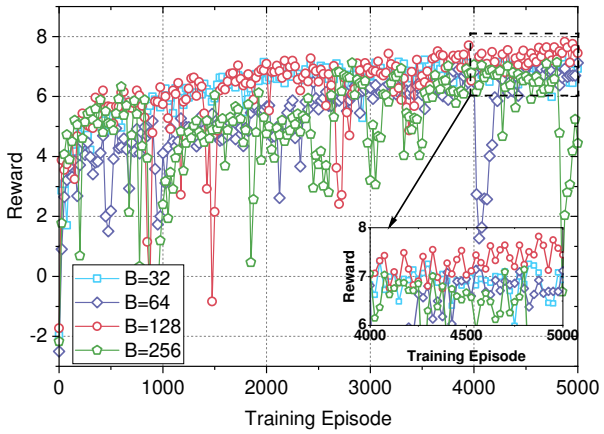

Fig. 5 Reward convergence of IDDPG algorithm under different mini-batch sizes.

DRL agent prone to local optima. On the other hand, a larger learning rate accelerates the initial convergence but introduces significant oscillations in performance. Therefore, we set the learning rate of the IDDPG algorithm to 3×10^{-4}.

Second, Fig. 5 shows the impact of the mini-batch size B on the training performance. When B is small, such as $B = 32$ or $B = 64$, the gradient estimation suffers from high variance due to limited sample diversity in each update, thereby causing frequent fluctuations in the reward curve and preventing convergence to a high value. However, when B increases to 256, the performance degrades and the training becomes less stable. This phenomenon occurs because an excessively large batch size reduces the stochasticity of the gradient update, which weakens the exploration capability of the agent in continuous action spaces. Consequently, it becomes difficult to identify the optimal solutions in dynamic environments. Therefore, we set the mini-batch size of the IDDPG algorithm to 128.

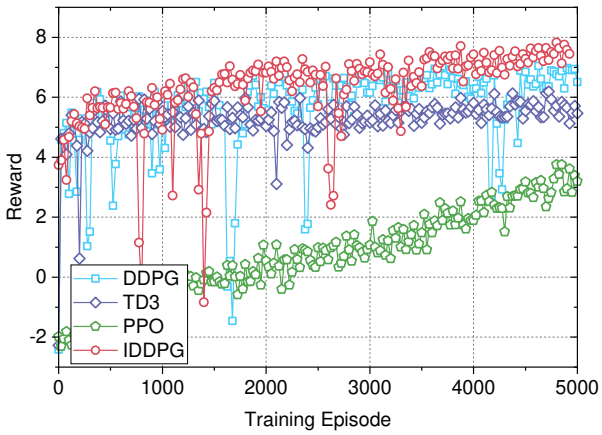

Fig. 6 Reward convergence of different algorithms.

2) Convergence Performance of Different Algorithms: Fig. 6 shows the reward curves of different algorithms during the training process. We observe that the proposed IDDPG algorithm outperforms DDPG, TD3, and PPO in terms of both convergence speed and the final steady-state reward value. This performance gain arises from two key improvements to the DDPG algorithm. First, we introduce a GAN-based critic network to improve the accuracy of value estimation. The generative adversarial mechanism enables the network to capture the distribution characteristics of the Q-value in complex dynamic VLC environments, thereby reducing the estimation bias in conventional critic networks and providing more reliable guidance for the policy update of the actor network. Second, we adopt the PER mechanism to replace uniform random sampling. This mechanism assigns higher sampling priority to experiences with larger temporal-difference errors, thereby improving the efficiency of sample usage and accelerating convergence. Consequently, the IDDPG algorithm can achieve faster and more stable convergence in dynamic scenarios, and thus it can effectively identify a high-quality resource allocation scheme.

5.3.3 Impact of User Mobility

Fig. 7 shows the impact of the parameters $\zeta_{\mathrm{v}}/\zeta_{\mathrm{d}}$ on the energy efficiency of different algorithms under the Gauss–Markov model. In this model, $\zeta_{\mathrm{v}}/\zeta_{\mathrm{d}}$ determines the dependence of the motion state of the user on previous states, where a larger value corresponds to lower randomness and stronger predictability of the trajectory. We observe that when $\zeta_{\mathrm{v}}/\zeta_{\mathrm{d}}$ increases from 0.94 to 0.96, the energy efficiency of the DRL algorithms shows an increasing trend. This phenomenon occurs because reduced randomness strengthens the temporal correlation of VLC channel states, thereby enabling the agent to learn the evolution pattern of the environment more accurately and to generate more effective decisions.

Notably, the proposed IDDPG algorithm consistently outperforms all comparison algorithms under different levels of randomness. This advantage arises because the GAN-based critic network introduces a generative adversarial learning mechanism to

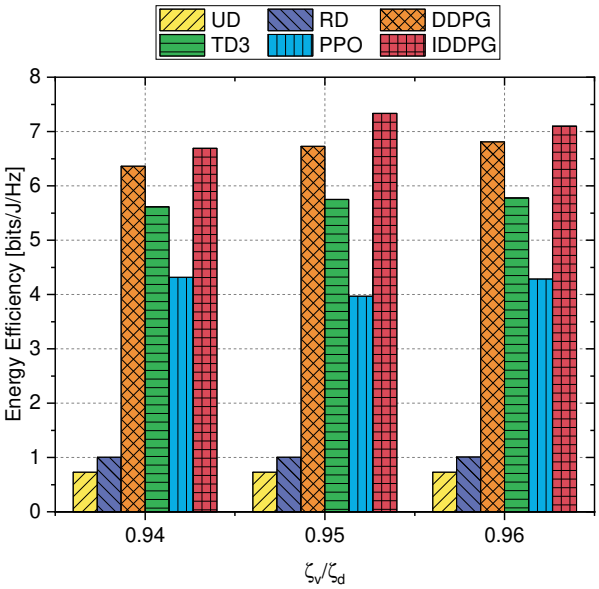

Fig. 7 Energy efficiency obtained by different algorithms under different randomness levels of MU mobility.

capture richer statistical characteristics of the Q-value, thereby reducing the estimation bias in conventional critic networks and providing more accurate value estimation than the critic networks of DDPG and TD3. In contrast, conventional UD and RD strategies consistently exhibit low energy efficiency and are insensitive to parameter changes due to a lack of awareness of movement characteristics. Simulation results strongly demonstrate that IDDPG possesses strong environmental adaptability when handling dynamic VLC scenarios with varying degrees of randomness.

6 Conclusion

This paper has investigated a dynamic indoor VLC system with RSMA, where multiple LEDs provide illumination and communication services for mobile users under time-varying channel conditions. An energy efficiency maximization problem has been formulated by jointly optimizing the precoding design and common stream splitting. To solve the optimization problem, an IDDPG algorithm has been proposed, where the GAN model and PER mechanism have been integrated to improve the stability and effectiveness of the decision-making process in complex VLC scenarios. Simulation results have shown that the proposed method achieves higher energy efficiency compared to conventional strategies and existing DRL algorithms. In particular, the IDDPG algorithm has maintained reliable performance under increasing user mobility, which indicates its adaptability to dynamic indoor environments.

References

[1] Imran, A. *et al.* Machine learning and deep learning in VLC systems: A comprehensive survey. *IEEE Open J. Commun. Soc.* **6**, 7955–8006 (2025).

[2] Palitharathna, K. W. S., Skouroumounis, C. & Krikidis, I. Liquid lens-based imaging receiver for MIMO VLC systems. *IEEE Trans. Commun.* **73**, 11663–11678 (2025).

[3] Luo, J., Fan, L. & Li, H. Indoor positioning systems based on visible light communication: State of the art. *IEEE Commun. Surv. Tutorials* **19**, 2871–2893 (2017).

[4] Xu, B., Hussain, B., Wang, Y., Cheng, H. C. & Yue, C. P. Smart home control system using VLC and bluetooth enabled AC light bulb for 3D indoor localization with centimeter-level precision. *Sensors* **22** (2022).

[5] Harada, R. & Popoola, W. O. Experimental demonstration of integrated VLC-RF system for secure industrial IoT applications. *IEEE Internet Things J.* **13**, 7217–7227 (2026).

[6] Nguyen, H., Imran, A. & Jang, Y. M. Survey of next-generation optical wireless communication technologies for 6G and beyond 6G. *ICT Express* **11**, 576–589 (2025).

[7] Abuajwa, O., Muhammad, S., Ambak, Z., Mitani, S. & Tiang, J. Energy efficiency techniques in non-orthogonal multiple access (NOMA) in 5G network: An overview and outlook. *Wirel. Pers. Commun.* **140**, 971–1013 (2025).

[8] Mao, Y. *et al.* Rate-splitting multiple access: Fundamentals, survey, and future research trends. *IEEE Commun. Surv. Tutorials* **24**, 2073–2126 (2022).

[9] Yang, Z., Chen, M., Zhang, Z. & Huang, C. Energy efficient semantic communication over wireless networks with rate splitting. *IEEE J. Sel. Areas Commun.* **41**, 1484–1495 (2023).

[10] Asif, M. *et al.* Robust beamforming optimization for STAR-RIS empowered multi-user RSMA under hardware imperfections and channel uncertainty. *IEEE Trans. Commun.* **74**, 7667–7680 (2026).

[11] Katwe, M. V., Deshpande, R., Singh, K., Ku, M. & Clerckx, B. Rsma-enabled aerial RIS-aided MU-MIMO system for improved spectral-efficient URLLC. *IEEE Trans. Veh. Technol.* **74**, 3110–3127 (2025).

[12] Zhang, C. *et al.* Multi-objective aerial collaborative secure communication optimization via generative diffusion model-enabled deep reinforcement learning. *IEEE Trans. Mob. Comput.* **24**, 3041–3058 (2025).

[13] Wang, X. *et al.* Deep reinforcement learning: A survey. *IEEE Trans. Neural Networks Learn. Syst.* **35**, 5064–5078 (2024).

[14] Zhang, Y. *et al.* A perspective of Q-value estimation on offline-to-online reinforcement learning. *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI* 16908–16916 (2024).

[15] Sun, G. *et al.* Generative AI for deep reinforcement learning: Framework, analysis, and use cases. *IEEE Wirel. Commun.* **32**, 186–195 (2025).

[16] Abdelhady, A. M., Amin, O., Chaaban, A., Shihada, B. & Alouini, M. Downlink resource allocation for dynamic TDMA-based VLC systems. *IEEE Trans. Wirel. Commun.* **18**, 108–120 (2019).

[17] Dong, Y. *et al.* Resource allocation in multi-user MIMO-OFDMA VLC systems over low-pass LED channels. *IEEE Photonics Technol. Lett.* **37**, 633–636 (2025).

[18] Liu, Z., Yang, F., Sun, S., Song, J. & Han, Z. NOMA-based multi-cell VLC systems with optical intelligent reflecting surface: Resource allocation for energy efficiency maximization. *IEEE J. Sel. Areas Commun.* **43**, 1691–1705 (2025).

[19] Sun, Y. *et al.* Enhanced coati optimization algorithm and its application in power allocation for NOMA-VLC systems. *Opt. Commun.* **579**, 131577 (2025).

[20] Sharma, A., Singh, K., Bhatia, V., Li, C. & Ding, Z. Ostar-RIS-assisted secure and high-coverage NOMA transmission for multi-user VLC systems. *IEEE Trans. Veh. Technol.* **75**, 4651–4664 (2026).

[21] Naser, S. *et al.* Rate-splitting multiple access: Unifying NOMA and SDMA in miso vlc channels. *IEEE Open J. Veh. Technol.* **1**, 393–413 (2020).

[22] Xing, F., He, S., Leung, V. C. M. & Yin, H. Energy efficiency optimization for rate-splitting multiple access-based indoor visible light communication networks. *IEEE J. Sel. Areas Commun.* **40**, 1706–1720 (2022).

[23] Yilmaz, A. F. & Duman, Ç. Optimizing SNR for indoor visible light communication by using TLBO algorithm. *Ad Hoc Networks* **158**, 103480 (2024).

[24] Feng, R. *et al.* A hybrid framework based on cluster reforming of indoor massive MIMO VLC networks. *IEEE Trans. Green Commun. Netw.* **10**, 2503–2515 (2026).

[25] Liu, Q. *et al.* Spectral efficiency optimization for optical IRS assisted multi-user MISO beamforming-based indoor visible light communication systems. *Opt. Commun.* **591**, 132005 (2025).

[26] Singh, A., Salameh, H. B., Ayyash, M. & Elgala, H. Optimal UAV deployment and led power allocation strategies for VLC-enabled indoor communication networks.

Phys. Commun. **73**, 102884 (2025).

[27] Liu, H. *et al.* Downlink resource allocation for IRS-aided indoor VLC considering link blockage. *Comput. Netw.* **278**, 112092 (2026).

[28] Benayad, A. *et al.* LEDs placement in indoor VLC system based on an improved marine predator algorithm. *J. Opt.* 1–34 (2025).

[29] Abdeljabar, S., Eltokhey, M. W. & Alouini, M.-S. Sum rate and fairness optimization in RIS-assisted VLC systems. *IEEE Open J. Commun. Soc.* **5**, 2555–2566 (2024).

[30] Liu, Z., Yang, F., Song, J. & Han, Z. NOMA-based MISO visible light communication systems with optical intelligent reflecting surface: Joint active and passive beamforming design. *IEEE Internet Things J.* **11**, 18753–18767 (2024).

[31] Saifaldeen, D. A., Al-Baseer, A. M., Ciftler, B. S., Abdallah, M. M. & Qaraqe, K. A. DRL-based IRS-assisted secure hybrid visible light and mmwave communications. *IEEE Open J. Commun. Soc.* **5**, 3007–3020 (2024).

[32] Guo, Y. *et al.* Secrecy energy efficiency maximization in IRS-assisted VLC MISO networks with rsma: A DS-PPO approach. *IEEE Trans. Wirel. Commun.* **24**, 6475–6489 (2025).

[33] Ciftler, B. S., Abdallah, M., Alwarafy, A. & Hamdi, M. DQN-based multi-user power allocation for hybrid RF/VLC networks. *IEEE International Conference on Communications (ICC)* 1–6 (2021).

[34] Ma, S. *et al.* Optimal power allocation for mobile users in non-orthogonal multiple access visible light communication networks. *IEEE Trans. Commun.* **67**, 2233–2244 (2019).

[35] Hu, J. *et al.* Resource efficient and robust RSMA for visible light communications. *IEEE Photonics J.* **16**, 1–12 (2024).

[36] Girdher, A., Gupta, N. & Bansal, A. Fairness-aware energy harvesting in RSMA-aided MU-MISO VLC network. *IEEE Commun. Lett.* **28**, 1062–1066 (2024).

[37] Naser, S. A., Sofotasios, P. C., Muhaidat, S. & Al-Qutayri, M. Rate-splitting multiple access for indoor visible light communication networks. *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)* 1–7 (2021).

[38] Qiu, Z., Mao, Y., Ma, S. & Clerckx, B. Robust max–min fair beamforming design for rate splitting multiple access-aided visible light communications. *IEEE Internet Things J.* **12**, 10043–10057 (2025).

[39] He, R., Ai, B., Stüber, G. L. & Zhong, Z. Non-stationary mobile-to-mobile channel modeling using the Gauss-Markov mobility model. *9th International Conference*

on Wireless Communications and Signal Processing (WCSP) 1–6 (2017).

[40] Guo, S. & Zhao, X. Multi-agent deep reinforcement learning based transmission latency minimization for delay-sensitive cognitive satellite-UAV networks. *IEEE Trans. Commun.* **71**, 131–144 (2023).

[41] Zhao, W. *et al.* A survey on DRL-based UAV communications and networking: DRL fundamentals, applications and implementations. *IEEE Commun. Surv. Tutorials* **28**, 3911–3941 (2026).

[42] Lillicrap, T. P. *et al.* Continuous control with deep reinforcement learning. *arXiv* (2015).

[43] Zhang, Z., Chen, J., Chen, Z. & Li, W. Asynchronous episodic deep deterministic policy gradient: Toward continuous control in computationally complex environments. *IEEE Trans. Cybern.* **51**, 604–613 (2021).

[44] Liu, X. *et al.* Value distribution DDPG with dual-prioritized experience replay for coordinated control of coal-fired power generation systems. *IEEE Trans. Ind. Inform.* **20**, 8181–8194 (2024).

[45] Goodfellow, I. J. *et al.* Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).

[46] Hua, Y., Li, R., Zhao, Z., Chen, X. & Zhang, H. GAN-powered deep distributional reinforcement learning for resource management in network slicing. *IEEE J. Sel. Areas Commun.* **38**, 334–349 (2020).

[47] Zhang, J. *et al.* Generative-adversarial-network-enhanced DRL for ISAC with double active RISs. *IEEE Internet Things J.* **12**, 13487–13499 (2025).

[48] Hou, Y., Liu, L., Wei, Q., Xu, X. & Chen, C. A novel DDPG method with prioritized experience replay. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 316–321 (2017).

[49] Fujimoto, S., Hoof, H. & Meger, D. Addressing function approximation error in actor-critic methods. *International conference on machine learning (ICML)* 1587–1596 (2018).

[50] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv* **abs/1707.06347** (2017).