This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3459039

Energy-Efficient Ground-Air-Space Vehicular Crowdsensing by Hierarchical Multi-Agent Deep Reinforcement Learning with Diffusion Models

Yinuo Zhao, Chi Harold Liu, *Senior Member, IEEE*, Tianjiao Yi, Guozheng Li, and Dapeng Wu, *Fellow, IEEE*

***Abstract*—The integrated ground-air-space (GAS) communications system can enhance post-disaster rescue and management efforts when traditional networks fail, by navigating unmanned ground vehicles (UGVs) and unmanned arieal vehicles (UAVs) to collaboratively collect sufficient data from point-of-interests (PoIs) in a timely manner. In this paper, we consider the GAS vehicular crowdsensing (VCS) campaign, where UGVs dispatch and callback UAVs periodically across multiple stops in the workzone, to maximize the total collected amount of data, geographic fairness while minimizing the energy consumption simultaneously. Specifically, we propose an energy-efficient, go-directed hierarchical multi-agent deep reinforcement learning (MADRL) method with discrete diffusion models called "gMADRL-VCS", to optimize the high-level goal-conditioned navigation policies of UGVs, and the low-level long-term sensing strategies of UAVs. Extensive experimental results on two real-world datasets in Roma, Italy, and Hong Kong SAR, China show that gMADRL-VCS outperforms baselines in terms of energy efficiency, data collection ratio, energy consumption, and UAV-UGV cooperation factor.**

***Index Terms*—Ground-air-space vehicular crowdsensing, Multi-agent deep reinforcement learning, Diffusion models, Energy-efficiency.**

I. Introduction

The sixth-generation (6G) communications system is expected to provide users with increased capacity, faster data rates, reduced latency, enhanced security, and improved quality of service [1], where the integrated ground-air-space (GAS) wireless networks have the appealing characteristics by using the strong communications and computation capabilities of unmanned aerial vehicles (UAVs), unmanned ground vehicles (UGVs) and low earth orbit (LEO) satellites [2].

The integrated GAS system can facilitate various applications, particularly in disaster response areas where traditional wireless communications systems may fail. As shown in Fig. 1, UGVs can serve as the UAV carrier to dispatch them to collect data from Points of Interest (PoIs, like building damage assessment and trapped individuals' living conditions), which can then transmit to disaster management cloud servers via low earth orbit satellites to aid post-disaster rescue efforts. After a certain period, the UAVs return to their corresponding UGVs to offload the sensory data. Then, each UGV navigates to another stop during which it charges the onboard UAVs to prepare for the next sensing task. In this way, a GAS-enabled vehicular crowdsensing (VCS) campagin is formed.

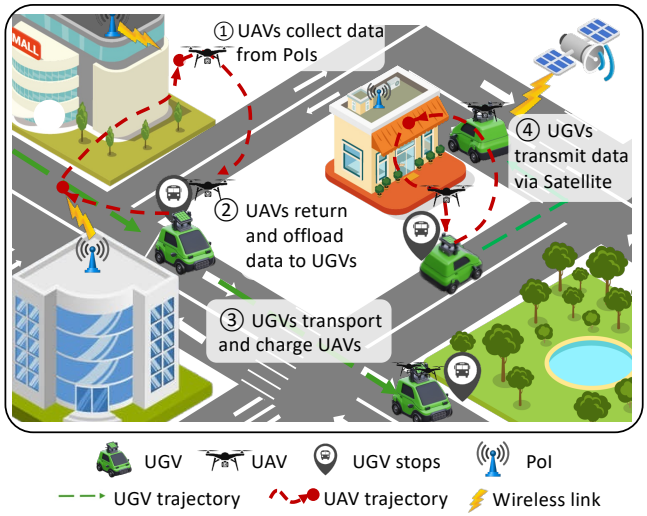


Fig. 1: Considered GAS-VCS compaign in the workzone.

The decision-making process in our considered collaborative GAS-VCS problem is complex with three key challenges identified. First, the uneven data distribution associated with PoIs data and the complicated topology of the underlying road networks make it difficult for UGVs to optimally decide when and where to deploy the carried UAVs for sensing. Second, the cooperation and interaction between UAVs and UGVs needs to be carefully addressed by considering multiple optimization objectives and the exponentially expansion of decision space as the number of agents grows. Finally, the vast workzone can be inefficiently explored with the limited number of UAVs and UGVs pairs, and the collected experience may not bet fully exploited at the beginning of the model training.

Recently, deep reinforcement learning (DRL) achieves great success in playing computer games [3], training autonomous driving policy [4], fine-tuning large language models [5], etc. Some existing studies have focused on solving the above challenges through enhancing generic multi-agent DRL (MARL) models. For example, CQDRL [6] facilitated effective cooperation among agents through a decentralized task assignment algorithm with a learning-based communications channel and a utility mixing network. Guan *et al.* in [7] proposed a MADRL algorithm based on enhanced K-means for UAV trajectory

Y. Zhao, C. H. Liu, T. Yi and G. Li are with School of Computer Science and Technology, Beijing Institute of Technology, China. D. Wu is with City University of Hong Kong, Hong Kong SAR. E-mail: chiliu@bit.edu.cn. Corresponding author: C. H. Liu.

This paper was sponsored by National Key Research Development Program of China (No. 2023YFE0209100) and National Natural Science Foundation of China (No. U21A20519 and U23A20310).

design to support emergency communications in disaster areas. DF-MADDPG [8] jointly optimized the trajectory of UAVs and UGVs with a common objective and reduced the communications overhead with a distributed architecture. GARL [9] considered the topology of road with graph convolutional MADRL. Although previous works tried to obtain efficient routing policy for UAVs and UGVs, none of them can solve all these three challenges together.

In this paper, we explicitly consider a collaborative GAS-VCS system, where UGVs (serving as the UAV carriers) are responsible for transporting UAVs along the roads and charging them, while UAVs are navigated to collect post-disaster data associated with PoIs. Our goal is to improve the overall efficiency by increasing the level of cooperation's between UAVs and UGVs. To this end, we propose a novel goal-directed MADRL framework called "gMADRL-VCS" to learn the route planning policies of UAVs and navigation policies of UGVs simultaneously. Under this framework, a state-conditioned discrete diffusion model is proposed to capture the complex decision process among UGVs considering PoI distributions and road topology. Our contribution is three-fold:

- We propose a hierarchical MADRL framework with diffusion models for GAS-VCS, which includes a goal-directed UGV navigation policy optimization module at the high-level.
- We propose a multi-window goal relabeling module for UGVs to encourage energy-efficient workzone explorations by fully exploiting the collected experience.
- We perform extensive experiments on two real-world datasets in Roma, Italy, and Hong Kong SAR by using real building and road statistics from the OpenStreetMap. We find the most appropriate hyperparameters; ablation study and performance comparisons with four other baselines well justify the effectiveness of gMADRL-VCS.

The remainder of this paper is organized as follows. First, we review the related works in Section II. Then, we present the system model in Section III. Problem definition and formulation are given in Section IV. After, we introduce our proposed method gMADRL-VCS in Section V. Experimental results on two real-world datasets are presented in Section VI. Finally, we conclude the paper in Section VII. Important notations used in this paper are listed in Table. I.

TABLE I: Important notations used in this paper.

Notation	Explanation
t, T	Current timeslot and total timeslots in a task.
τ	UAV one-time sensing period.
$\mathcal{U}, u, U$	UGVs set, index and total number of UGVs.
$\mathcal{V}, v, V$	UAVs set, index and total number of UAVs.
$\mathcal{V}(u)$	A set of UAVs that are carried by a UGV u.
$\mathcal{P}, p, P$	PoIs set, index and total number of PoIs.
$\mathcal{B}, \mathcal{B}'$	UGV stop set, the selected UGV stops.
$\mu_t^{p,u}$	The approximate data transmission rate.
ξ, η, f, ψ	Energy efficiency, data collection ratio, geographic fairness, and UAV-UGV cooperation factor.
$\hat{\boldsymbol{g}}, \boldsymbol{g}_t$	Expected goal, achieved goal.

II. Related Work

A. Ground-Air-Space (GAS) Networks

Equipped with space and aerial platforms deployed at varying altitudes, GAS network features multiple vertical layers and forms an integrated communications 3D structure [10]. It has several advantages such as an integration of frequency bands with the aid of wide-ranging spectrum sharing and providing ubiquitous high-quality connectivity in 6G. However, there is still much work to be done to fully unlock its potentials when considering the dynamic environments, heterogeneous devices and multiple optimization objectives. Cao *et al.* in [11] considered an uplink ground-space communications system, where GAS links were proposed to complement ground-to-satellite links to strengthen the terrestrial communications and save the transmission power. The authors in [12], [13] considered the mm-wave links as the air-to-air and air-to-ground backbones to facilitate a high-capacity yet low-cost GAS architecture. In the integrated GAS system, LEO satellites play a crucial role in transmitting UAV-collected data to the data management servers. In this work, we focus on the efficient data collection in post-disaster scenarios, addressing challenges from ground and air layers. After data collection, various techniques, such as adaptive transmission schemes [14], coding-based multi-path transmission [15], and cooperative HAP and LEO satellite schemes [16] can be utilized to facilitate ground-to-space data transmissions.

Other works focused on resource management issues in GAS networks, such as energy efficiency of UAVs/UGVs [11], [17], spectrum utilization [11], [18] and low latency [19]. In this paper, we aim to optimize the energy efficiency of UAVs while maximizing the data collection ratio and minimizing the energy consumption simultaneously in a limited task duration.

B. Vehicular Crowdsensing (VCS)

VCS uses UAVs and UGVs to offer widespread sensing services for extreme situations like post-disaster rescue and management [20], [21]. Xu *et al.* in [6] proposed a communication-QMIX MADRL solution to solve the task assignment problem in a decentralized manner, to take advantage of redundant computing and network resources of worker devices, thereby reducing the deployment cost. Guan *et al.* in [7] proposed a decentralized K-means enhanced IPPO algorithm to optimize the cooperative trajectory of UAVs in disaster response. However, it did not take the limited energy resources of UAVs into consideration, which is a practical challenge. In [8], they proposed an F-MADDPG based UAV and UGV joint trajectory optimization algorithm to maximize the average spectral efficiency of the emergency network. In order to deploy the UGVs in urban area, Wang *et al.* in [9] proposed a graph convolutional network approach to extract UGV-specific features from stops and facilitate their cooperation to adapt to the changing geometry. All these works cannot be directly leveraged to solve our considered GAS-VCS problem, thus we aim to design a heterogeneous MADRL algorithm to deal with the large action space and can utilize the mobility of UGVs to transport UAVs between different regions.

C. DRL and Generative Models

DRL aims at learning a state-action map function towards maximizing the accumulated reward. It provides an ideal framework for learning long-term routing policies for UAVs and UGVs in GAS-VCS without depending on human supervision. However, the commonly used Gaussian or Categorical kernel for policy distributions and simple multi-layer perception architecture, may not effectively represent complex, goal-conditioned multi-agent policies for UGVs. These standard DRL setups often fail to capture the intricate dynamics and strategic interactions required in multi-agent environments.

Meanwhile, diffusion models [22]–[24] are emerging as a state-of-the-art (SOTA) family of generative models, contributing to their stable training process and strong multi-modal representation ability. They can be utilized to represent complex multi-modal policies and optimized through DRL. For example, Diffusion-QL [25] used a conditional diffusion model to learn the multi-modal policy for robot control with the aid of behavior cloning and Q function estimation. However, it is only compatible with value-based DRL, as the likelihood of diffusion models is intractable. EDP [26] approximated the diffusion policy likelihood from a constructed Gaussian distribution and compatible with various offline DRL methods. However, these works cannot be directly integrated with online DRL methods, since obtaining offline datasets is challenging in our GAS-VCS scenarios. In one concurrent work, Du *et al.* in [27] introduced a diffusion model-based approach called AGOD for generating optimal AI service provider selection decisions. However, it ignored the sparsity nature in large discrete action space. In this paper, we considered to use a more structured categorical corruption process to better represent the navigation policy of UGVs and optimized the decision policies with efficient online MADRL algorithms.

Unlike previous studies mainly focusing either on efficient policies for UAVs or UGVs separately [6], [7] or on complex combined objectives [9], this paper aims to optimize the cooperative long-term goals of UAVs and UGVs for enhanced GAS. Specifically, we introduce a hierarchical MADRL framework that simplifies the joint optimization challenge into a two-level learning process. While previous efforts overlooked road topologies [8], we propose a state-based discrete-diffusion model to better represent the complex routing strategies of UGVs in urban environments. This model utilizes the gradual denoising process and benefits from the inherently flexible conditioning and stable training characteristics of diffusion models. Given the limited amount of training samples at the higher level of our hierarchical framework, on-policy algorithms like TRPO [28] and PPO [29] are not practical due to requirements of large number of samples. Therefore, we propose integrating a multi-window goal relabeling module in our framework to augment the training samples and optimized through off-policy MASAC algorithm to promote energy-efficient exploration in work zones.

III. System Model

We consider an integrated GAS-VCS task where multiple UAVs denoted as $\mathcal{V} \triangleq \{1, \cdots, V\}$ and multiple UGVs denoted as $\mathcal{U} \triangleq \{1, \cdots, U\}$ are deployed in a targeted workzone. UGVs move along the road and UAVs fly in a 2D cartesian coordinate system at a fixed height. Buildings higher than the UAV surveillance altitude are considered as obstacles. Similar to [30], we consider an urban environment where one single omni-directional antenna is mounted on the top of each building. These antennas gather data from sensors hanging in/outside the building and then transmit the data to UAVs. These antennas are referred to as Point-of-Interests (PoIs), denoted as $\mathcal{P} \triangleq \{1, \cdots, P\}$. UGVs are responsible for transporting UAVs between remote regions and charging battery for them. UAVs, equipped with multiple antennas, are responsible for data collections from PoIs and offloading these data to UGVs upon landing. We denote the association between UGVs and UAVs as $\mathcal{V}(u)$, representing the set of UAVs carried by UGV u.

A. GAS Communication Model

We model the ground-to-air data transmission links from PoIs to UAVs as follows. Since UAVs fly through buildings, it is crucial to incorporate both the line-of-sight (LoS) and the non-line-of-sight (NLoS) links into the channel model. Following [31], [32], the path loss for the large-scale fading of the ground-to-air channel from a PoI p to a UAV v at time t can be expressed as:

$$\begin{aligned} l_t[p,v] = 20\log\left(d_t^{p,v}\right) + \left(\eta_{\text{LoS}} - \eta_{\text{NLoS}}\right) p_t^{\text{LoS}}[p,v] \\ + \eta_{\text{NLoS}} + 20\log\left(\frac{4\pi f_c}{c}\right), \end{aligned} \tag{1}$$

where $d_t^{p,v}$ denotes the spatial distance between a PoI p and a UAV v that is calculated by $d_t^{p,v} = \sqrt{\left(x^p - x_t^v\right)^2 + \left(y^p - y_t^v\right)^2 + \left(z^p - z_t^v\right)^2}$, where (x^p, y^p, z^p), (x_t^v, y_t^v, z_t^v) represents the location of the PoI and UAV, respectively. Let η_{LoS} and η_{NLoS} be the additional transmission loss for LoS and NLoS links, respectively. f_c is the carrier frequency. Let $p_t^{\text{LoS}}[p,v]$ be the LoS connectivity probability between a UAV v and a PoI p at timeslot t, which could be expressed by $p_t^{\text{LoS}}[p,v] = \left[1 + a\exp(-b \times (\theta_t^{p,v} - a))\right]^{-1}$, where a and b are environment constants affecting the S-curve parameters that vary according to the building-density of the environment. $\theta_t^{p,v}$ is the elevation angle between a UAV v and a PoI p at timeslot t. Then, we can get the complex channel gain $h_t[p,v]$ as:

$$h_t[p,v] = 10^{-l_t[p,v]/10}\vartheta_t[p,v], \tag{2}$$

where $\vartheta_t[p,v]$ denotes the small-scale channel fading. We apply the F-factor Rician fading with $\mathbb{E}\| \vartheta_t[p,v] \|^2 = 1$ to consider both LoS and NLoS links in data transmission.

Next, we assume that each UAV is equipped with M antennas to receive data from at most M PoIs. To guarantee transmission reliability, a UAV only receives data from PoIs with channel gains higher than a predefined threshold β_{th}. A UAV v receives data from a set of PoIs $\mathcal{P}_t^v$ simultaneously, where:

$$\mathcal{P}_t^v = \{p \in \mathcal{P} | h_t[p,v] \leq h_{\text{th}}\}. \tag{3}$$

To eliminate the interference, we employ linear beamforming [33] from PoIs to a UAV and zero-forcing [34] techniques. Specifically, we consider $\mathbf{W}_t^v = \mathbf{H}_t^v[(\mathbf{H}_t^v)^{\mathrm{H}}\mathbf{H}_t^v]^{-1}$, where $\mathbf{W}_t^v$ denotes the beamforming matrix $[w_{u,\mathcal{P}_t^v(1)_t}, \cdots, w_{u,\mathcal{P}_t^v(M)_t}]$ and $\mathbf{H}_t^v$ represents the corresponding channel matrix between a UAV u and the selected PoIs for sensing; $\mathbf{H}_t^v$ has the same shape $M \times \mathcal{P}_t^v$ with $\mathbf{W}_t^v$. Then, we are able to compute the received signal-to-noise ratio (SNR) from a PoI p to a UAV u as $\frac{\rho_0 \|(w_t^{p,v})^{\mathrm{H}}\mathbf{H}_t^v(p)\|^2}{N_0}$, where ρ_0 denotes the average transmitted power of each PoI and N_0 denotes the power of the white Gaussian noise at each UAV. Following [35], we approximate the data rate $\mu_t^{p,v}$ by applying the Wishart matrix [33] and Jensen's inequality [36] as:

$$\begin{aligned}\mu_t^{p,v} &= W\log\left(1+\frac{\rho_0\|(w_t^{p,v})^{\mathrm{H}}\mathbf{H}_t^v(p)\|^2}{N_0}\right)\\ &\geq W\log\left(1+\frac{\rho_0}{\mathbb{E}\{[((\mathbf{H}_t^v)^{\mathrm{H}}\mathbf{H}_t^v)^{-1}]_{p,p}\}N_0}\right)\\ &= W\log\left(1+\frac{\rho_0 10^{-l_t[p,v]/10}}{\frac{N_0}{M-|\mathcal{P}_t^v|}}\right)\\ &= W\log\left(1+\frac{(M-|\mathcal{P}_t^v|)\rho_0 10^{-l_t[p,v]/10}}{N_0}\right),\end{aligned} \tag{4}$$

where W is the total bandwidth of the channel; Following [35], we use the classical Maximum-Ratio Combining beamforming [37] when only one beam of transmitted data arrives at a UAV, i.e., $|\mathcal{P}_t^v| = 1$. We have:

$$\mu_t^{p,v} \triangleq W\log\left(1+\frac{\Omega_t^v\rho_0 10^{-l_t[p,v]/10}}{N_0}\right), \forall p \in \mathcal{P}_t^v, v \in \mathcal{V}, \tag{5}$$

where $\Omega_t^v = M - |\mathcal{P}_t^v|$ if $|\mathcal{P}_t^v| \geq 2$, else $\Omega_t^v = 1$.

The energy consumption of UAV v at time t, denoted as δe_t^v, includes energy used for movement during data collection and for ascending and descending during the recharge phase. We define this as follows:

B. UAV Energy Consumpition Model

The energy consumption of UAVs results from horizontal movement during data collection and from ascending and descending during the recharge phase. We formulate the energy consumption e_t^{v-} of UAV v at time t as follows:

$$e_t^{v-} = C_1||(x_t, y_t, z_t) - (x_{t-1}, y_{t_1}, z_{t-1})|| + C_2 \tag{6}$$

where C_1, C_2 are constants that depend on the aircraft weight, air density and rotor disc area, as specified in [38]. After recharging, the UAV resumes data collection from its initial location.

IV. Problem Definition and Formulation

A. Problem Definition

In our considered GAS-VCS scenario, UAVs and UGVs work together to collect sensory data in the target work zone and transmit these data back to the LEO satellite to maximize the following performance metrics. First is the **data collection ratio** η, defined as:

$$\eta = 1 - \frac{\sum_p d_T^p}{\sum d_0^p}, \tag{7}$$

where $\sum_p d_T^p$ is the total amount of data after T timeslots, and $\sum_p d_0^p$ denotes the initial data amount of all PoIs.

Second is the **geographic fairness** of collected data, since PoIs may unevenly distributed in the workzone, therefore some far away PoIs may not covered. We use Jain's fairness index [39] to compute it as:

$$f = \frac{\left(\sum_p (d_0^p - d_T^p)/d_0^p\right)^2}{P\sum_p \left((d_0^p - d_T^p)/d_0^p\right)^2}. \tag{8}$$

Next, in order to measure the efficiency of all UAVs to execute the data collection task, we jointly consider data collection ratio and geographic fairness, adding the element of energy consumption, as an integrated performance index, called "**energy efficiency**", as:

$$\xi = \eta \cdot f \cdot \frac{\sum_v (e_0^v + \Delta e_t^v)}{\sum_v (e_0^v + \Delta e_t^v - e_T^v)}, \tag{9}$$

where e_0^v denotes the initial energy reserve of UAV v and e_T^v denotes the remaining energy after T timeslots; Δe_t^v denotes the total amount of energy charged up to timeslot t.

Fourth is the **UAV-UGV cooperation factor** ψ, representing their degree of collaboration as whether or not a UGV is able to transport some UAVs to those areas to receive higher energy efficiency. We first define the total amount of data collected by a UAV v up to timeslot t as:

$$\Delta d_t^v = \sum_{t=1}^{T}\sum_{p\in\mathcal{P}_t^v} \mu_t^{p,v}\delta. \tag{10}$$

Here $\mu_t^{p,v}$ represents the data transmission rate, as in Eqn. (5); δ denotes the length of one timeslot; $\mathcal{P}_t^v$ refers to the set of PoIs with satisfactory quality-of-service (QoS) around v, as in Eqn. (3). We then define the cooperation factor as the bottleneck attained energy efficiency of all UGVs as:

$$\psi = \min_{u\in\mathcal{U}} \frac{\sum_{v\in\mathcal{V}(u)} \Delta d_t^v}{|\mathcal{V}(u)| \cdot \sum_t \kappa_t^u}, \tag{11}$$

where $\mathcal{V}(u)$ refers to the set of UAVs carried by a UGV u, and $|\mathcal{V}(u)|$ specifies the number of UAVs in this set; $\sum_t \kappa_t^u$ denotes the cumulative distance traveled by a UGV u upon the task completion.

B. Problem Formulation

The decision process of UGVs and UAVs is modeled as a decentralized partially observable Markov Decision Process (Dec-POMDP). We formulate it as a two-step framework. At first step, each UGV selects a stop to navigate to. Then, UAVs execute route-planning policies for data collections.

For UGVs, their decision process is represented by a tuple $< \mathcal{U}, \mathcal{G}, \mathcal{S}, \mathcal{B}, \mathcal{R} >$, where $\mathcal{G}, \mathcal{S}, \mathcal{B}, \mathcal{R}$ denote the goal, state, action space and the reward function, respectively. At each timeslot t, after all UGVs take actions, they obtain a common

achieved goal $\boldsymbol{g}_t$, defined as the total amount of data from PoIs in the neighborhood of a UGV stop b, as:

$$\boldsymbol{g}_t = \left\{ \sum_{\forall p \in n(b)} d_t^p \Big| b \in \mathcal{B} \right\}, \tag{12}$$

where $n(b)$ denotes a set of PoIs in the neighbourhood around the UGV stop b, defined as:

$$n(b) = \left\{ p \big| \mu_0^{p,b} \geq \mu_{\text{th}}, \forall p \in \mathcal{P} \right\}, \forall b \in \mathcal{B}. \tag{13}$$

Let $\mu_0^{p,b}$ be the data transmission rate from a PoI p to a UAV when it is located at (x^b, y^b, z^v), where (x^b, y^b) is the coordinates of the stop b; and μ_{th} is the predefined threshold that guaranteed the QoS. Similarly, let $\hat{\boldsymbol{g}}$ denote the expected total amount of data around all UGV stops, as $\hat{\boldsymbol{g}} = \boldsymbol{g}_0 \bigodot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a random variable vector sampled from the normal distribution, and $\bigodot$ denotes the element-wise multiplication.

Next, we give the formulation of state $\boldsymbol{s}_t^u$ for each UGV u at timeslot t. It comprises a common and a private observation. The former is the same for every UGV which includes the achieved goal $\boldsymbol{g}_t$, the location of each UGV at timeslot t, and the minimal distance between UGV stops. The latter is a one-hot index vector to uniquely identify a UGV.

The action for each UGV is denoted by b_t^u, which represents the stop it needs to be navigated to. Therefore, the action space is equal to the stop space $\mathcal{B}$.

Finally, the reward function for UGVs is defined as the distance between the expected goal and the achieved goal as:

$$r_t^u = \left(1 - \frac{||\boldsymbol{g}_t - \hat{\boldsymbol{g}}||}{||\boldsymbol{g}_0||}\right) \cdot \frac{\sum_{v \in \mathcal{V}(u)} \left(\Delta d_t^v - \Delta d_{t-\tau}^v\right)}{|\mathcal{V}(u)| \cdot \kappa_t^u}, \forall u \in \mathcal{U}. \tag{14}$$

Here $\Delta d_t^v - \Delta d_{t-\tau}^v$ represents the total amount of data collected by a UAV v during its one-time sensing period τ; κ_t^u denotes the navigation distance for the UGV u at timeslot t. The reward r_t^u encourages UGVs to travel to densely located PoIs that have high amount of remaining data, using the shortest path, aiming to achieve the predefined goal $\hat{\boldsymbol{g}}_t$.

For UAVs, we formulate their decision process as a tuple $< \mathcal{V}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}r >$, where $\mathcal{V}, \mathcal{S}, \mathcal{O}, \mathcal{A}$ are the set of UAVs, global states, local observations, and actions. Following [9], we model the global states as a matrix with four channels. Specifically, the first channel incorporates the location information of obstacles; the second channel includes the remaining data d_t^p for PoIs $p \in \mathcal{P}$ at timeslot t; the third channel is the remaining energy e_t^v for each UAV $v \in \mathcal{V}$; and the last channel shows the location information of UGVs. Since UAVs are not able to access the global states and hence they only make decisions based on their local observations. Specifically, we first model the masked state for a UAV v through masking out PoIs which do not be assigned to a UAV v, in the second channel of global states. A UAV v is assigned to sense a PoI p when its carrier UGV is located at the stop b and $p \in n(b)$, as in Eqn. (13). Then, the local observation $\boldsymbol{o}_t^v$ can be obtained by cropping the masked state matrix around the position of a UAV v within a range, using the UAV's position as the origin.

The action space for UAVs is continuous and consists of the expected locations for sensing. The reward r_t^v for a UAV v is defined as $r_t^v = r_t^{v+} + r_t^{v-}$, where:

$$r_t^{v+} = f(n(b)) \cdot \frac{\Delta d_t^v - \Delta d_{t-1}^v}{e_{t-1}^v - e_t^v}, \quad \forall v \in \mathcal{V}, t, \tag{15}$$

where $f(n(b))$ denotes the geographic fairness over the PoI set $n(b)$; r_t^{v-} is the penalty incurred when UAV v collides with an obstacle.

Our considered GAS-VCS can be modeled as a constrained optimization problem that aims to jointly optimize the policies of UAVs and UGVs, with the goal of maximizing energy efficiency and the cooperation factor. Obviously, it is a NP-hard problem that is challenging to solve considering the vast continuous observation space, the joint behaviour of UAVs and UGVs, and the necessity for long-term sequential decision-making. Therefore, we opt to propose a sub-optimal solution using MADRL methods as follows.

V. Proposed Solution: gMADRL-VCS

We propose an energy-efficient hierarchical MADRL algorithm with diffusion models to jointly optimize the navigation and route planning policies of UGVs and UAVs simultaneously, as shown in Fig. 2.

A. Channel-Aware UGV Stop Selection

Typically, the set of stops $\mathcal{B}$ is too large to solve in a vast workzone. Therefore, we need first to select a subset B' from it. Considering the uneven distribution of PoIs, we partition the entire workzone into regions by solving a channel-aware minimal set coverage problem. Each region consists of one single UGV stop along with multiple PoIs.

Recall that we define a set of PoIs that guarantee the QoS μ_{th} as $n(b)$ around the stop b, in Eqn. (13). Then, the UGV stop selection problem can be converted to the well-known set covering problem (SCP) aiming to find the minimum number of sets (stops) that cover all elements (PoIs):

$$\text{Minimize:} \quad |\mathcal{B}'| \tag{16}$$

$$\text{subject to:} \quad \bigcup_{b \in \mathcal{B}'} n(b) = \mathcal{P}, \quad \forall b \in \mathcal{B}. \tag{17}$$

SCP is a classic NP-hard problem in combinatorial optimization. It can be approximated via linear programming relaxation algorithms or greedy algorithms [40] in polynomial solvable time.

B. Goal-conditioned UGV Policy Optimization by Discrete Diffusion Models

1) Goal-conditioned discrete diffusion model: Inspired by the great success of diffusion models in text-to-image tasks, we propose a goal-conditioned diffusion models as UGV policy generator, and then extend the existing single-agent DRL approach SAC [41] to its multi-age version, called "MASAC". We start from D3PM [42], as a discrete diffusion model designed for image reconstruction. Consider our GAS-VCS scenario where each UGV u selects one stop b_t^u from

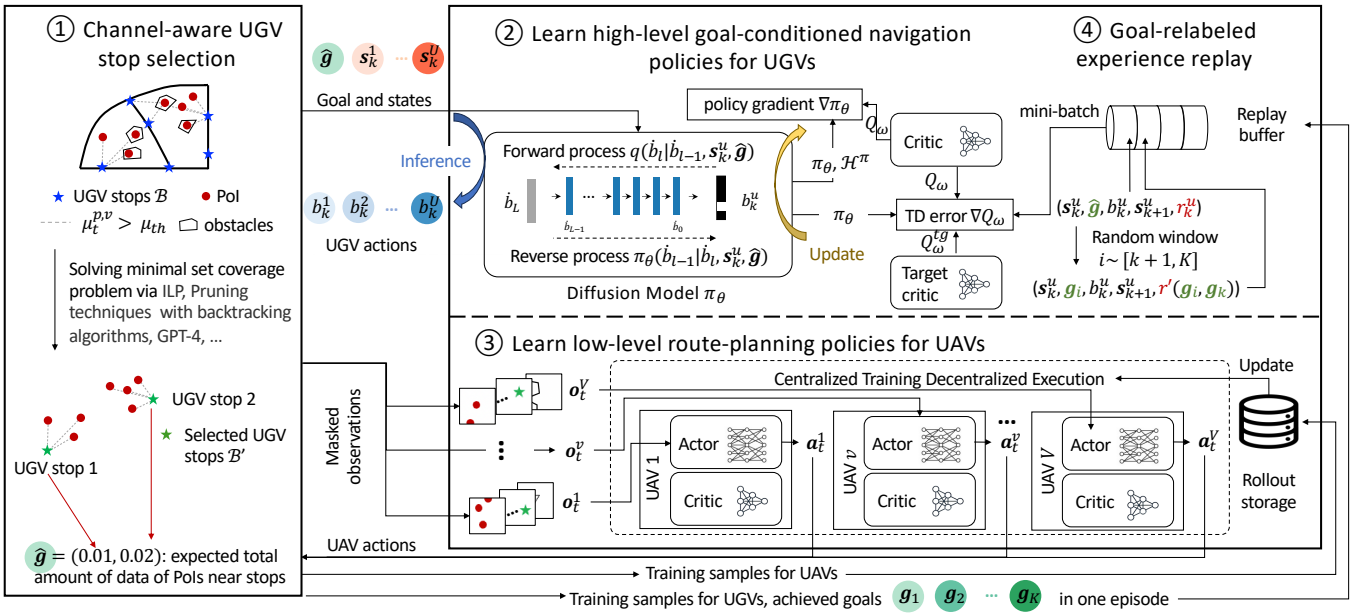


Fig. 2: Proposed solution: gMADRL-VCS.

$\mathcal{B}'$ and navigates to it. The total number of UGV stops considered in our method is represented by $|\mathcal{B}'|$. Rather than using Gaussian kernel, we consider adding uniform noise during the forward process in our goal-conditioned discrete diffusion model. Specifically, in the forward process, we define a discrete transition matrix $[\mathbf{M}_l]_{i,j} = q(b_l = j|b_{l-1} = i)$, representing the transition probability from $b_{l-1} = i$ to $b_l = j$ at denoising step l. The shape of $\mathbf{M}$ is $|\mathcal{B}'| \times |\mathcal{B}'|$. We adopt the uniform distribution as the transition kernel that gradually added to the forward process. Then, the transition matrix can be written as $\mathbf{M}_l = (1-\beta_l)I + \beta_l \mathbf{1}\mathbf{1}^T/|\mathcal{B}'|$, where $\mathbf{1}$ denotes the column vector of all ones and β_l controls the ratio between dumping to a new stop and remaining at the current stop.

We denote $\dot{b}$ as the one-hot row vector of b. Then the forward transition probability is given by:

$$q(\dot{b}_l|\dot{b}_{l-1}) = \text{cat}\left(\dot{b}_l; \boldsymbol{p} = \dot{b}_{l-1}\mathbf{M}_l\right), \tag{18}$$

where $\text{cat}\left(\dot{b}; \boldsymbol{p}\right)$ denotes the categorical distribution over the one-hot vector $\dot{b}$ with probabilities given by the row vector $\boldsymbol{p}$, and $\dot{b}_{l-1}\mathbf{M}_l$ can be understood as the j-th row vector taken from $\mathbf{M}_l$ when $b_{l-1} = j$. Starting from the stop $\dot{b}_0$ at denoising step $l = 0$, the l-step marginal distribution in the forward process of diffusion can be computed as:

$$q(\dot{b}_l|\dot{b}_0) = \text{cat}\left(\dot{b}_l; \boldsymbol{p} = \dot{b}_{l-1}\overline{\mathbf{M}}_l\right), \tag{19}$$

with $\overline{\mathbf{M}}_l = \mathbf{M}_l\mathbf{M}_{l-1}\cdots\mathbf{M}_1$. Note that when $L \to \infty$, $q(\dot{b}_L)$ approximates a uniform distribution over the action space $\mathcal{B}'$. Then, the posterior at denoising step $l-1$ with the forward transition probability q is given by:

$$\begin{aligned}\Pr(\dot{b}_{l-1}|\dot{b}_l, \dot{b}_0) &= \frac{q(\dot{b}_l|\dot{b}_{l-1}, \dot{b}_0)q(\dot{b}_{l-1}|\dot{b}_0)}{q(\dot{b}_l|\dot{b}_0)} \\ &= \text{cat}\left(\dot{b}_{l-1}; \boldsymbol{p} = \frac{\dot{b}_l\mathbf{M}_t^T \odot \dot{b}_0\overline{\mathbf{M}}_{l-1}}{\dot{b}_0\overline{\mathbf{M}}_l\dot{b}_l^T}\right).\end{aligned} \tag{20}$$

The reverse diffusion process using the posterior can be expressed as:

$$\begin{aligned}\Pr(\dot{b}_{l-1}|\dot{b}_l) &= \frac{\sum_{\dot{b}_0}\Pr(\dot{b}_{l-1}, \dot{b}_l, \dot{b}_0)}{\Pr(\dot{b}_l)} \\ &= \sum_{\dot{b}_0}\Pr(\dot{b}_{l-1}|\dot{b}_l, \dot{b}_0)\Pr(\dot{b}_0|\dot{b}_l) = \mathbb{E}_{\Pr(\dot{b}_0|\dot{b}_l)}\Pr(\dot{b}_{l-1}|\dot{b}_l, \dot{b}_0).\end{aligned} \tag{21}$$

In practice, we approximate the categorical distribution $\Pr(\dot{b}_0|\dot{b}_l)$ by deep neural networks with parameters θ. After every sensing period at a timeslot, we sample the UGV u's action from the diffusion model π_θ with L denoising steps: $b_t^u \sim \pi_\theta(\dot{b}_0|\dot{b}_L, \hat{\boldsymbol{g}}, \boldsymbol{s}_t^u)$.

2) UGV policy optimization by MASAC in centralized training decentralized execution (CTDE) manner: In the centralized training phase of MASAC, we compute the value of the policy towards the maximum entropy objective. Specifically, the soft Q value network $Q_\omega(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}}, b_t^u)$ predicts the expected discounted returns obtained at timeslot t. It is optimized through minimizing the Bellman equation error as:

$$\mathcal{L}_\omega = \mathbb{E}\left[||r_t^u + \gamma V(\boldsymbol{s}_{t+1}^u, \hat{\boldsymbol{g}}) - Q_\omega(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}}, b_t^u)||\right], \tag{22}$$

where the soft state value function is:

$$V(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}}) = \mathbb{E}_{b_t^u \sim \pi_\theta}\left[Q_\omega(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}}, b_t^u) - \alpha \log \pi_\theta(b_t^u|\boldsymbol{s}_t^u, \hat{\boldsymbol{g}})\right]. \tag{23}$$

The, the policy is updated towards the exponential of the soft Q-function, as:

$$\pi' = \arg\min_\pi D_{KL}\left(\pi(\cdot|\boldsymbol{s}_t^u)\| \frac{\exp(\frac{1}{\alpha}Q^{\pi_{old}}(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}}, \cdot))}{Z^{\pi_{old}}(\boldsymbol{s}_t^u, \hat{\boldsymbol{g}})}\right), \tag{24}$$

where the Kullback-Leibler divergence is chosen to measure the distance between the policy π and the soft Q-function Q. The partition function $Z^{\pi_{old}(\boldsymbol{s}_t^u)}$ normalizes the distribution and can be ignored in the policy updating process. Then,

the parameters in diffusion models can be optimized by minimizing the following objective:

$$\mathcal{L}_\theta = \mathbb{E}_{(\boldsymbol{s}_t,\hat{\boldsymbol{g}})\sim\mathcal{D}_U}\left[\mathbb{E}_{b_t\sim\pi_\theta}[\alpha\log(\pi_\theta(b_t|\boldsymbol{s}_t,\hat{\boldsymbol{g}}))-Q_\omega(\boldsymbol{s}_t,\hat{\boldsymbol{g}},b_t)]\right]. \tag{25}$$

C. Decentralized Sensing Policy Optimization for UAVs

Following the UGV decisions, UAVs are dispatched to different regions to collect data from PoIs within a limited time duration τ. UAV policies are optimized in a CTDE way, as they share the training samples with a common rollout storage. For simplicity, we omit the subscript v and simply denote as $\boldsymbol{o}_t$, $\boldsymbol{a}_t$ and r_t.

Following the backbone of IPPO [43], one policy network π_σ and one value network V_ϕ are used, whose objective function is defined as:

$$\mathcal{L}_\sigma = \mathbb{E}\left[\min(r_t(\sigma)\hat{A}_t, \mathrm{clip}(r_t(\sigma),1-\epsilon,1+\epsilon)\hat{A}_t)\right], \tag{26}$$

where $r_t(\sigma)$ is denoted as the important sampling ratio to measure the action distribution distance between the old and current policy, which is computed as $r_t(\sigma)=\frac{\pi_\sigma(\boldsymbol{a}_t|\boldsymbol{o}_t)}{\pi_{\sigma_{old}}(\boldsymbol{a}_t|\boldsymbol{o}_t)}$. The advantage function $\hat{A}_t$ is calculated via the generalized advantage estimation approach [44]. The clip function is used to limit the range of the policy ratio, preventing excessive updates to the policy. Specifically, if $r_t(\sigma)$ falls outside the range of $[1-\epsilon,1+\epsilon]$, it gets clipped to stay within this range.

The value network V_ϕ is updated by:

$$\begin{aligned}\mathcal{L}_\phi = \mathbb{E}\big[&\max\big((V_\phi(s_t)-\hat{R}_t)^2,\\ &(\mathrm{clip}(V_\phi(\boldsymbol{s}_t),V_{\phi_{old}}(\boldsymbol{s}_t)-\epsilon_1,V_{\phi_{old}}(\boldsymbol{s}_t)+\epsilon_1)-\hat{R}_t)^2\big)\big],\end{aligned} \tag{27}$$

where $\hat{R}_t$ denotes the discounted returns, $V_{\phi_{old}}$ denotes the estimated value of previous models during inference.

D. Algorithm Description

Pseudo-codes of gMADRL-VCS are shown in Algorithm 1, which can be decomposed into three phases: preparation, exploration and exploitation, while the two proceed in alternation. The inputs are total number of UGVs, UAVs and UAV one-time sensing period τ (Line 1). The outputs are UGV policy model π_θ and UAV policy model π_σ (Line 2). First, we specify the 3-D coordinates, PoIs $\mathcal{P}$, UGV stops $\mathcal{B}$, building outlines as obstacles, and the road network topology for UGV navigation (Line 3). During preparation (Line 4-6), we split the workzone into regions and formulate the goal and the action space for UGVs. Specifically, we first obtain the selected UGV stops $\mathcal{B}'$ through solving a channel-aware SCP in Eqn. (16) (Line 4). For every UGV stop $b\in\mathcal{B}'$, it corresponds to a region of several PoIs. Then, we formulate the initial goal g_0 defined in Eqn. (12) and we constraint the action space for UGVs to $\mathcal{B}'$ (Line 5). After, we initialize the parameters of π_θ, Q_ω for UGVs, π_σ, V_ϕ for UAVs (Line 6), and replay buffer $\mathcal{D}_U, \mathcal{D}_V$ as empty (Line 7). During each episode h, we first empty the rollout replay $\mathcal{D}_V$, then set the expected goal value $\hat{\boldsymbol{g}}$ with randomly variable $\boldsymbol{\epsilon}_0$ from Gaussian distribution.

During exploration, for each UGV u, it first obtains the actions by sampling from the discrete diffusion model $\pi_\theta(\dot{b}_0|\dot{b}_L,\hat{\boldsymbol{g}},\boldsymbol{s}_t^u)$ given its state $\boldsymbol{s}_t^u$ and expected goal $\hat{\boldsymbol{g}}$. Then the UGVs navigate to their stops along the shortest path and dispatch their carried UAVs at the stop. After, UAVs perform the sensing tasks within $\min(\tau,T-t)$ timeslots. At each timeslot t, each UAV v independently samples actions $\boldsymbol{a}_t^v\sim\pi_\sigma(\cdot|\boldsymbol{o}_t^v)$. Then, they store their experiences $(\boldsymbol{o}_t^v, \boldsymbol{a}_t^v, r_t^v)$ into a common rollout storage $\mathcal{D}_V$. After a period of τ, UGVs callback UAVs and compute the achieved goal $\boldsymbol{g}_t$ according to Eqn. (12). Then, UGVs store their experiences $(\boldsymbol{s}_k^u, b_k^u, r_k^u, \boldsymbol{s}_{k+1}^u, \hat{\boldsymbol{g}}, \boldsymbol{g}_t)$ into a common replay buffer $\mathcal{D}_U$ (Line 17). Next, UGVs take actions again until task completion.

After exploration, we first perform goal-relabeling in $\mathcal{D}_U$ (Line 19). Specifically, we propose a multi-window goal relabeling method for augmenting the trajectory experience of UGVs, as shown in Algorithm 2. For each UGV u and its corresponding trajectory, we uniformly sample an index i from the next index $k+1$ to the last data sample index K. Then, we replace the expected goal for k-th experience with the achieved

Algorithm 1: gMADRL-VCS

1 **Input**: No. of UGVs, no. of UAVs, UAV one-time sensing period τ;
2 **Outout**: UGV policy π_θ, UAV policy π_σ;
3 **Data**: The location statistics of PoIs $\mathcal{P}$, all UGV stops $\mathcal{B}'$, obstacles, and roads;
/* Preparation */
4 Obtain selected UGV stops $\mathcal{B}'$ by solving SCP in Eqn. (16);
5 Construct the initial goal $\boldsymbol{g}_0$ according to Eqn. (12);
6 Initialize π_θ, Q_ω for UGVs, π_σ, V_ϕ for UAVs;
7 Initialize UGV replay buffer $\mathcal{D}_U=\varnothing$, UAV rollout storage $\mathcal{D}_V=\varnothing$;
8 **for** $h=1,2,...H$ **do**
9 Set $\mathcal{D}_V=\varnothing$, $t=0$, $k=0$;
10 Set goal $\hat{\boldsymbol{g}}=\boldsymbol{g}_0\bigodot\boldsymbol{\epsilon}_0$;
 /* Exploration */
11 **while** $t<T$ **do**
12 UGVs execute actions $b_k^u\sim\pi_\theta(\dot{b}_0|\dot{b}_L,\hat{\boldsymbol{g}},\boldsymbol{s}_k^u)$;
13 **while** $iter<\min(\tau,T-t)$ **do**
14 UAVs execute actions $\boldsymbol{a}_t^v\sim\pi_\sigma(\cdot|\boldsymbol{o}_t^v)$;
15 Store $(\boldsymbol{o}_t^v, \boldsymbol{a}_t^v, r_t^v)$ into $\mathcal{D}_V$ for all $v\in\mathcal{V}$;
16 $t=t+1$;
17 Store $(\boldsymbol{s}_k^u, b_k^u, r_k^u, \boldsymbol{s}_{k+1}^u, \hat{\boldsymbol{g}}, \boldsymbol{g}_k)$ into $\mathcal{D}_U$ for all $u\in\mathcal{U}$;
18 $k=k+1$;
19 Augment $\mathcal{D}_U$ as in Algorithm 2;
 /* Exploitation */
20 $\pi_{\sigma_{old}}=\pi_\sigma$, $V_{\phi_{old}}=V_\phi$;
21 **for** $1,2,..K_v$ **do**
22 Update π_σ by minimizing Eqn. (26);
23 Update V_ϕ by minimizing Eqn. (27);
24 **for** $1,2,..K_u$ **do**
25 Update π_θ by minimizing Eqn. (25);
26 Update Q_ω by minimizing Eqn. (22);

Algorithm 2: Multi-window goal relabeling

Input: Latest experience trajectories for UGVs;
Outout: Augmented experience trajectories;
Denote the length of the trajectory as $K = \lceil T/\tau \rceil$;
for $u = 1, 2, ...U$ **do**
  for $k = 1, 2, ...K-1$ **do**
    Uniformly sample $i \in [k+1, K]$;
    Replace $\hat{\boldsymbol{g}}$ with $\boldsymbol{g}_i$;
    Compute reward r' in Eqn. (14);
    Insert $\left(\boldsymbol{s}_k^u, b_k^u, r', \boldsymbol{s}_{k+1}^u, \boldsymbol{g}_i, \boldsymbol{g}_k\right)$ into $\mathcal{D}_U$.

goal g_i. After, we recompute the reward r' with Eqn. (14) using $\boldsymbol{g}_i$ and $\boldsymbol{g}_k$. Next, we insert the updated experience into the UGV buffer $\mathcal{D}_U$. Then we return to Algorithm 1 to update UGV and UAV policy models, using the experiences in $\mathcal{D}_U$ and $\mathcal{D}_V$, respectively (Line 21-26).

E. Complexity Analysis

1) Model Inference: During inference, for UAVs, we adopt convolution neural networks to extract features from the high-dimensional inputs and then feed the features into a MLP network as the policy. Therefore, the computational complexity for UAVs can be represented as:

$$\mathcal{O}\Big(\underbrace{\sum_{i=1}^{H_L} D_{1,i} \cdot D_{2,i} + \sum_{i=1}^{H_C} D_{3,i}^2 \cdot D_{4,i}^2 \cdot D_{5,i} \cdot D_{6,i}}_{o_{uav}}\Big), \quad (28)$$

where H_L is the number of linear layers in the policy network for UAVs and, $D_{1,i}, D_{2,i}$ represent the dimension of input and output features of i-th linear layers. H_C is the number of convolution layers, $D_{3,i}$, $D_{4,i}$, $D_{5,i}$, $D_{6,i}$ represent the size of output feature maps, convolution kernels, input channels, output channels of the i-th convolution layer, respectively.

The computation complexity for UGV policy networks is similar since they both utilize MLPs. However, due to the additional L denoising step in the reverse diffusion process, the computation complexity for UGV network is increased to:

$$\mathcal{O}\Big(\underbrace{(L-1)\sum_{i=1}^{H_U} D_{1,i} \cdot D_{2,i}}_{o_{ugv}}\Big), \quad (29)$$

where H_U is the number of linear layers in the diffusion model for UGVs.

2) Model Training of Algorithm 1: The time complexity is given by $\mathcal{O}(H(To_{uav} + \lceil T/\tau \rceil)o_{ugv})$, where H denotes the training iterations, T denotes episode length. For Algorithm. 2, the time complexity is given by $\mathcal{O}(U(K-1))$, where U denotes the number of UGVs and K denotes the length of the UGV trajectories.

VI. Experimental Results

A. Setup

We use two real-world urban datasets in Roma, Italy, and Hong Kong SAR. The landscape data, including the roads and buildings, were obtained from OpenStreetMap, and then pre-processed including eliminating low-level buildings that do not pose as obstacles for UAVs, distributing UGV stops evenly along roads, and outlining boundaries on the map. The UAVs fly at a fixed height of 10 meters. The height of PoIs matches the height of the buildings on which they are mounted. According to the ITU-R Rec. P.1410-2 model [45], we adopt a Rayleigh distribution for building heights in downtown environment, as utilized in previous research [46], [47].

The UAVs operate at a constant altitude of 10 meters. The height of Points of Interest (PoIs) matches that of the buildings on which they are mounted. According to the ITU-R Rec. P.1410-2 model, we adopt a Rayleigh distribution for building heights, as utilized in previous research.

In Roma, the longitude ranges from 12.4533 to 12.4790 and the latitude ranges from 41.9261 to 41.9415, covering approximately 3.64 square kilometers. We randomly positioned 129 PoIs on top of buildings within this area. In HKSAR, the longitude ranges from 114.1504 to 114.1635 and the latitude ranges from 22.3262 to 22.3351, covering 1.19 square kilometers; and we placed 156 PoIs. Following previous works [48], [49] on the channel model, we choose the simulation parameters as in Table. II.

As shown in Fig. 2, our approach employs three types of neural networks to learn the high-level policy for UGVs. Both the critic network and the target critic network are structured with three linear layers, each containing 128 hidden neurons, and share identical architectures. The actor network within the diffusion model features three linear layers with 128 hidden neurons each. During the denoising process of the diffusion model, the state of each UGV, the time step embeddings, and the previous output from the actor network at the last denoising step are fed into the actor network. For the low-level route-planning policies of UAVs, we employ a network architecture similar to that described in [9]. The discount factors, γ, are set at 0.90 for the Roma dataset and 0.95 for the HKSAR dataset. The learning rates for the actor and critic networks in the UGV models are set at 1e-4 and 1e-3, respectively.

TABLE II: Simulation settings

Notation	Value	Notation	Value	Notation	Value
a, b	9.6,0.16	f_c	3.5GHz	z^v	5m
T	100	W	20MHz	δ	20s
$\eta_{\text{LoS}}, \eta_{\text{NLoS}}$	1dB,20dB	M	8	ρ_0	20dBm

TABLE III: Impact of μ_{th}

Dataset	μ_{th}	ξ	η	f	ψ
Roma	1.5	0.129	0.305	0.329	4.065
	2.0	0.300	0.420	0.436	5.490
	2.5	**0.926**	**0.722**	**0.745**	**7.820**
	3.0	0.777	0.673	0.710	7.014
	3.5	0.645	0.598	0.628	3.691
HKSAR	2.0	0.315	0.358	0.380	5.649
	2.5	0.948	0.718	0.738	6.949
	3.0	**1.342**	**0.830**	**0.866**	**8.689**
	3.5	0.834	0.745	0.776	8.211
	4.0	0.741	0.617	0.654	7.203

TABLE IV: Impact of γ

Dataset	γ	ξ	η	f	ψ
Roma	0.80	0.827	0.680	0.709	5.472
	0.85	0.896	0.706	0.741	**8.995**
	0.90	**0.940**	**0.719**	**0.754**	7.629
	0.95	0.925	0.714	0.752	7.651
	1.00	0.858	0.689	0.714	7.611
HKSAR	0.80	1.286	0.830	0.868	5.790
	0.85	1.297	0.860	0.898	7.257
	0.90	1.363	0.856	0.894	**8.256**
	0.95	**1.376**	**0.858**	**0.897**	7.342
	1.00	1.344	0.843	0.881	5.734

B. Impact of Hyperparameters

We first evaluate the influence of the parameter μ_{th} in gMADRL-VCS. As noted in Eqn. (13) and Eqn. (16), the parameter μ_{th} affects the channel-aware stop selection results and then can influence the overall performance. We set the number of UAVs $V = 4$, the number of UGVs $U = 4$, and the UAV one-time sensing period $\tau = 10$, respectively. As shown in Table III, we identify the optimal values of $\mu_{\text{th}} = 2.5$ and $\mu_{\text{th}} = 3.0$ in Roma and HKSAR, respectively. We observe that all metrics initially increase and then decline as μ_{th} rises. For example, in Roma, the energy efficiency $\xi = 0.926$ when $\mu_{\text{th}} = 2.5$, which is two times higher than that when $\mu_{\text{th}} = 2.0$. This is because a higher μ_{th} indicate that more PoIs around a UGV stop can be collected from, thus in larger regions and fewer selected UGV stops, the diffusion policy model in gMADRL-VCS enables UGVs to remain at one UGV stop consistently. This allows UAVs to thoroughly explore the region, avoiding energy waste from frequent travelling back-and-forth in the workzone. Beyond this threshold, the energy efficiency drops. For example, when μ_{th} increases to 3.5 in Roma, ξ drops 43.4% compared with that when $\mu_{\text{th}} = 2.5$. This is because as excessive number of PoIs are included in $n(b)$, as defined in Eqn. (13), UAVs need to fly across PoI-lacking regions independently. Therefore, within the constraints of limited time slots, the data collection ratio η decreases, and both the energy efficiency and the cooperation factor between UAVs and UGVs experience a decline.

Then, we evaluate the impact of the discount factor γ, which determines how future rewards are taken into account in the learning process. A lower γ biases the agent towards immediate rewards, focusing on short-term gains. As shown in Table. IV, $\gamma = 0.8$ results in the lowest performance across all metrics. While $\gamma = 1$ can also degrade the performance as it complicates the effective computation and optimization of value functions. Therefore, according to the results in Table. IV, we choose $\gamma = 0.90$ in Roma and $\gamma = 0.95$ in HKSAR.

Next, we evaluate the impact of the denoising steps L in the diffusion model. As shown in Eqn. (29), larger denoising steps L in the reverse process of diffusion models increases the computation cost. As shown in Fig. 3, we vary L from 1 to 8 and show its impact on the cooperation factor and model inference time. Notably, we observe the highest UAV-UGV cooperation factor when $L = 5$, with $\xi = 7.820$ and $\xi = 8.689$ in Roma and HKSAR datasets, respectively. While when $L = 5$, it also maintains a reasonable inference time.

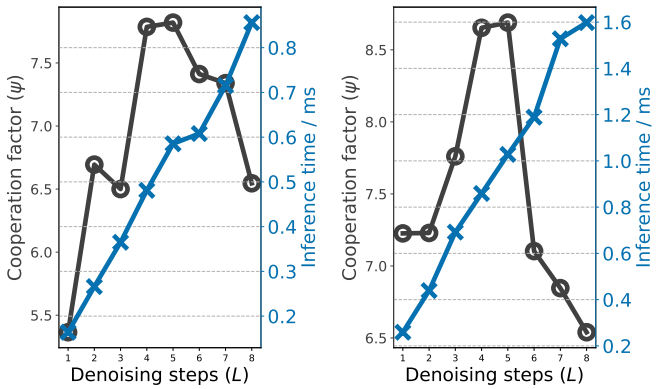


Fig. 3: Impact of denoising steps L in diffusion model.

TABLE V: Ablation study.

	Method	ξ	μ	η	ψ
Roma	**gMADRL-VCS**	**0.926**	**0.722**	**0.745**	**7.820**
	gMADRL-VCS w/o goal	0.734	0.642	0.666	6.247
	gMADRL-VCS w/o diff.	0.722	0.650	0.680	7.015
	gMADRL-VCS w/o goal & diff.	0.660	0.641	0.658	5.598
HKSAR	**gMADRL-VCS**	**1.334**	**0.845**	**0.876**	**8.689**
	gMADRL-VCS w/o goal	1.128	0.780	0.805	8.128
	gMADRL-VCS w/o diff.	1.188	0.803	0.830	7.752
	gMADRL-VCS w/o goal & diff.	1.047	0.771	0.786	7.477

C. Ablation Study

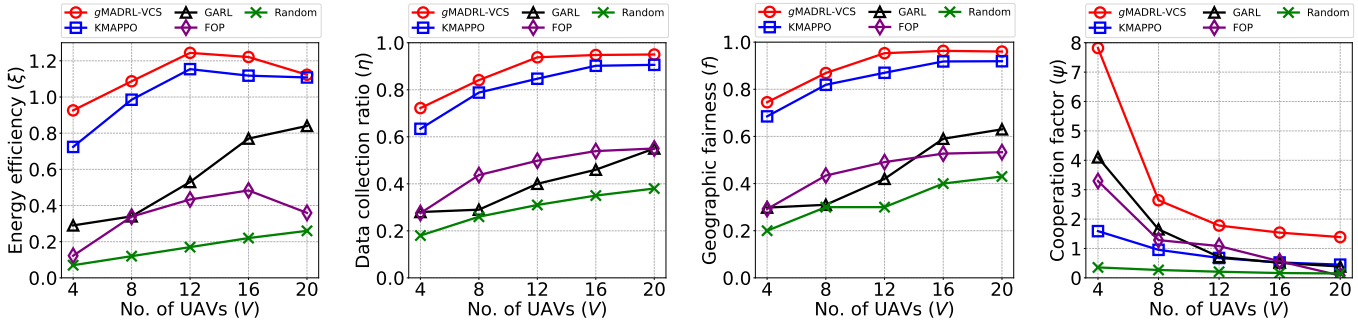
We gradually remove two key modules, the multi-window goal-relabeling and the diffusion model (denoted as diff.). We fix the number of UAVs as 4, the number of UGVs $U = 4$ and the value of UAV one-time sensing period $\tau = 10$. As shown in Table V, both two modules have influence on the energy efficiency ξ and the UAV-UGV cooperation ratio β. For example, in Roma, when removing the multi-window goal-relabeling module, ξ and ψ drop 20.7% and 20.1%, respectively. The multi-window goal-relabeling module modifies the replay buffer $\mathcal{D}$ by replacing the desired goal with the achieved goal from an episode. This process introduces positive training samples with higher reward into the buffer, accelerating the initial learning phase. Thus, this module enables UGVs to learn goal-conditioned policies more effectively, improving the degree of UAV-UGV cooperation and energy efficiency.

The benefits the goal-conditioned discrete diffusion model introduced are also clear that for example, in HKSAR, the UAV-UGV cooperation factor decreases from 8.689 to 7.752, indicating 10.8% performance drop. This is because our proposal has much better representation ability with the reverse process $q(\dot{b}_{l-1}|\dot{b}_l)$ as defined in Eqn. (21). Thus gMADRL-VCS is able to learn complex multi-modal UGV navigation policies under multi-agent scenarios.

D. Comparing with Four Other Baselines

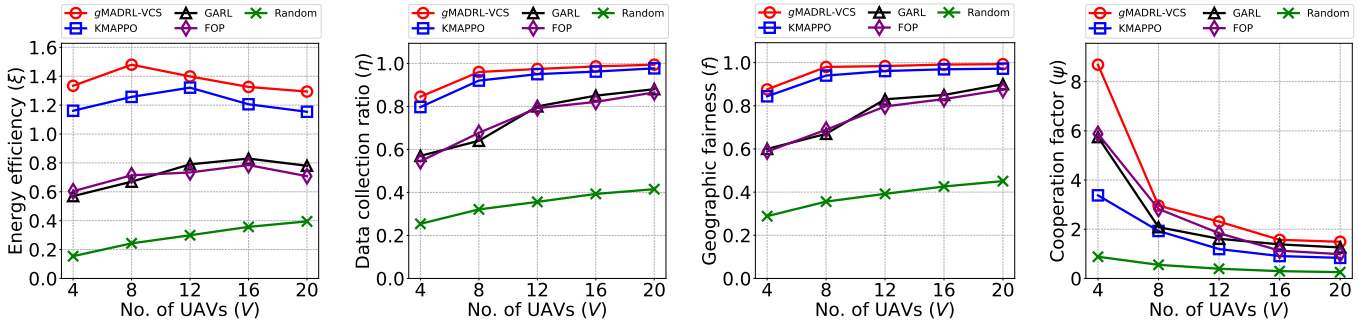
We compare gMADRL-VCS with four baselines, as:

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3459039



(a) ξ (b) η (c) f (d) ψ

Fig. 4: Impact of No. of UAVs V (Roma dataset, $U = 4, \tau = 10$).



(a) ξ (b) η (c) f (d) ψ

Fig. 5: Impact of No. of UAVs V (HKSAR dataset, when $U = 4, \tau = 10$).

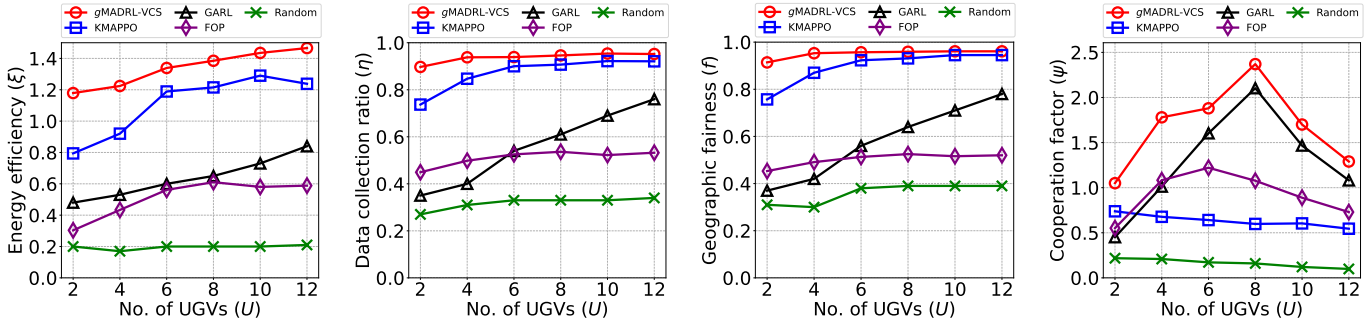
- **KMAPPO** [7]: A cooperative trajectory design method for disaster area emergency communications based on the enhanced K-means and MAPPO algorithm. We consider it as the SOTA approach in VCS by MADRL.
- **GARL** [9]: It is a graph-based MADRL method that jointly optimizes the policy of UAVs and UGVs in air-ground spatial crowdsourcing. We consider it as another SOTA approach in VCS by MADRL.
- **FOP** [50]: It is a SOTA MADRL method that factorizes the optimal joint policy induced by maximum-entropy MARL into individual policies though value decomposition. As it is also unable not handle the large action space in the GAS-VCS, we adopt settings from GARL, allowing FOP to only make decisions for UGVs from nearby stops with action masks.
- **Random**: It controls UAVs and UGVs with actions evenly sampled from their corresponding action space.

1) Impact of No. of UAVs: We set the number of UGVs $U = 4$ and UAV one-time sensing period $\tau = 10$, and investigate the impact of V on four metrics ξ, η, f and ψ in Fig. 4 and Fig. 5 for Roma and HKSAR, respectively. As shown in Fig. 4, we observe that the energy efficiency ξ initially increases and then decreases with more deployed UAVs. This is because more UAVs can collect more data, however too many UAVs may not be necessary and they may waste energy travelling back and forth. Meanwhile, ξ keeps increasing for GARL and Random methods up to $V = 20$ UAVs indicating their inefficiency in handling large workzone, as there are still remaining data uncollected.

Furthermore, gMADRL-VCS achieved the highest energy efficiency and UAV-UGV cooperation factor compared with all other baselines. For example, gMADRL-VCS achieved $\psi = 7.820$ when $V = 4$, nearly 2 times of the second best GARL. This improvement confirms the effectiveness of our proposal by solving the channel-aware UGV stop selection problem in Eqn. (16), where UGVs are able to transport UAVs across distant regions to service those seldom visited PoIs. While in GARL, UGVs can only deploy UAVs at nearby stops. Although KMAPPO also achieved comparable results on data collection ratio and geographic fairness, its UAV-UGV cooperation factor is much lower.

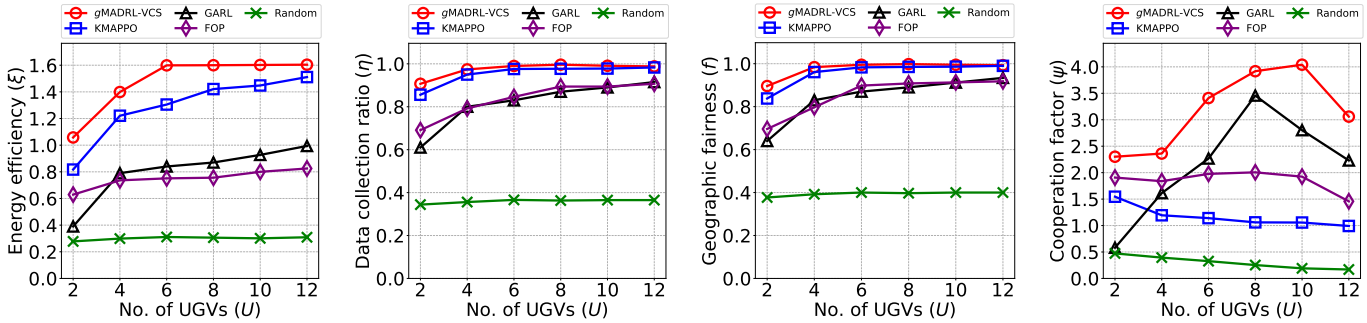
2) Impact of No. of UGVs: We investigate the impact of the number of UGVs by varying $U = 2$ to 12 when fixing the number of UAVs in Fig. 6 and Fig. 7. When $U = 2$, each UGV carries 6 UAVs and when $U = 12$, each carries one. We observe that gMADRL-VCS achieved the best performance; for example, gMADRL-VCS achieved 89.7% data collection ratio with only 2 UGVs in Roma, while GARL only obtains 35.0%, which is 54.7% lower. This confirms our proposed goal-conditioned diffusion models can generate efficient navigation policies for multiple UGVs with the variational reverse decision process, as in Eqn. (21). Also, our proposed multi-window goal-relabeling module enhances

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3459039



(a) ξ (b) η (c) f (d) ψ

Fig. 6: Impact of No. of UGVs U (Roma dataset, when $V = 12, \tau = 10$).



(a) ξ (b) η (c) f (d) ψ

Fig. 7: Impact of No. of UGVs U (HKSAR dataset, when $V = 12, \tau = 10$).

the policy optimization towards achieving goals while reducing navigation distance, as the reward in Eqn. (14).

Furthermore, we observe that the energy efficiency ξ increases with more deployed UGVs for all methods except Random. For example, in Fig. 6a, when $U = 6$, FOP achieves an energy efficiency $\xi = 0.561$ compared to $\xi = 0.303$ when $U = 2$, marking a 85.1% improvement. This improvement is due to FOP's ability to adapt the behavior of UGVs based on the number of UAVs they carry. Consequently, with more UGVs, they can collaboratively deploy UAVs across a larger workzone, enhancing energy efficiency. We also notice that the ξ of Random remains at a low level for all UGVs since no cooperation is enforced between them.

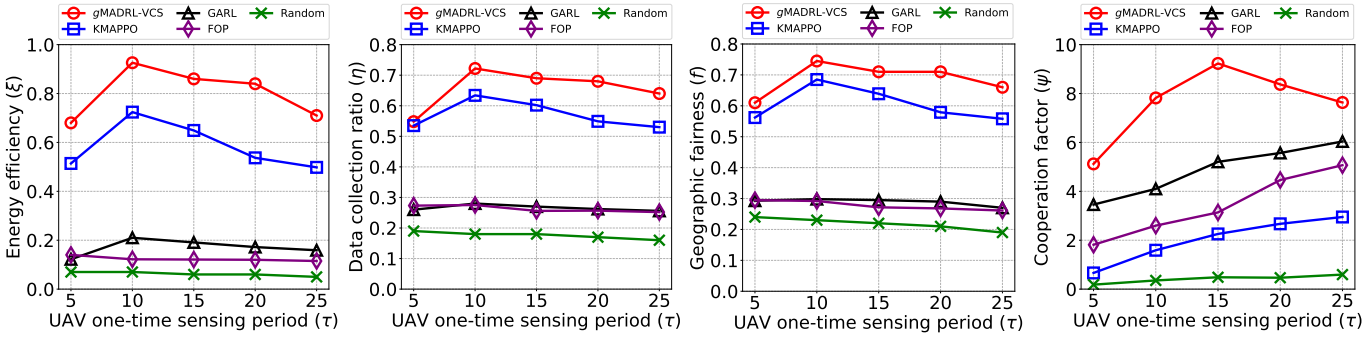
We notice that the UAV-UGV cooperation factor initially goes up but then drops down with more UGVs for both gMADRL-VCS and GARL, as shown in Fig. 6d and Fig. 7d. Both methods are capable of generating cooperative strategies for UGVs. Therefore, with increased number of UGVs, each only needs to travel a short distance to collect data from nearby PoIs. However, with too many UGVs, the amount of data collected by each UGV goes down, leading to a decline in the UAV-UGV cooperation factor ψ.

3) Impact of UAV One-time Sensing Period: Next, we investigate the impact of UAV one-time sensing period τ in Roma and HKSAR as shown in Fig. 8 and Fig. 9, respectively. We observe that the energy efficiency ξ first increases and then decreases with longer τ. This is because with longer UAV sensing period, the UAVs have enough time to thoroughly explore the regions to collect data from PoIs. However, too long τ incurs the potential problem of UAV energy waste when flying around and therefore ξ drops.

From Fig. 8b and Fig. 8d, we see that both GARL and FOP reach high UAV-UGV cooperation factors but attain low data collection ratios. This is because both GARL and FOP are limited to move between nearby stops, whereas gMADRL-VCS enables UGVs to travel greater distances to collect more data, which is achieved by our proposed hierarchical framework and a well-designed goal space as outlined in Eqn. (12).

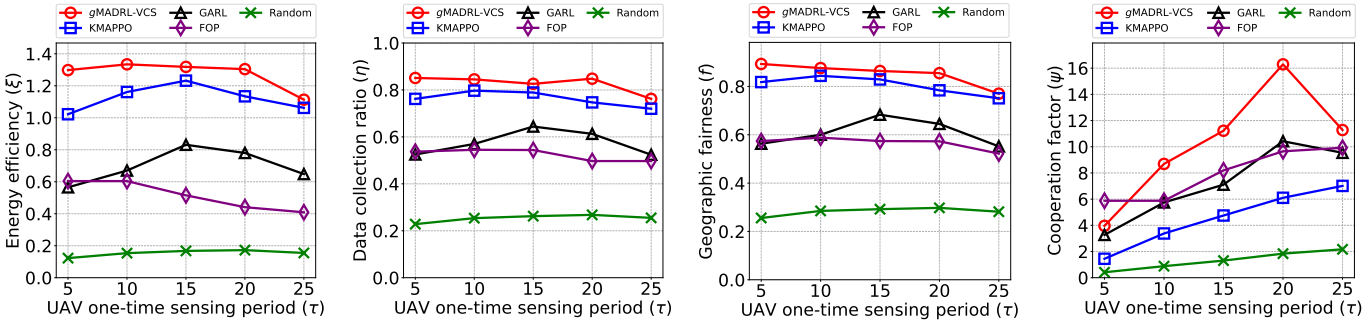
4) Impact of average transmitted power of PoIs: We demonstrate the impact of average transmitted power of PoIs on four metrics in Fig. 10 and Fig. 11 for Roma and HKSAR, respectively. As shown in Eqn. (4), the average transmitted power ρ determines the supported data rate, and then indirectly influences the UAV planning strategies and UGV routing policies. As shown in Fig. 10b, the attained data collection ratio increases with higher ρ, although it slows down as ρ continues to rise, due to the logarithmic influence of the average transmitted power of PoIs on data rate μ. We can observe from Fig. 10c and Fig. 11c that geographic fairness remains consistent regardless of ρ under the Random policy. In contrast, policies learned through other methods can adjust routing based on the average transmitted power of PoIs, leading to increased data fairness as ρ increases.

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3459039



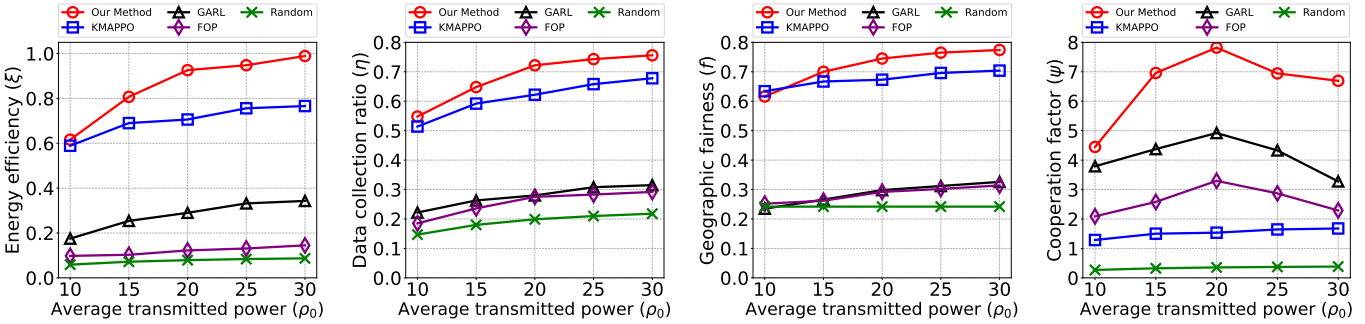
(a) ξ (b) η (c) f (d) ψ

Fig. 8: Impact of length of UAV one-time sensing period τ (Roma dataset, when $U = 4, V = 4$).



(a) ξ (b) η (c) f (d) ψ

Fig. 9: Impact of UAV one-time sensing period τ (HKSAR dataset, when $U = 4, V = 4$).



(a) ξ (b) η (c) f (d) ψ

Fig. 10: Impact of the average transmitted power of PoIs (Roma dataset).

E. UAV-UGV Trajectory Visualization

We visualize the trajectories of UAV deployment, sensing and callback, along with the trajectories of UGVs as the carriers in Fig. 12. The results are obtained by *g*MADRL-VCS in Roma and HKSAR two datasets, with 4 UAVs and UGVs. First, we clearly observe the cooperation between UAVs. For example in HKSAR during $t = [1, 40]$, although three UAVs (represented with red, purple and blue) are deployed from the same stop, they fly towards different directions/regions to collect data. This is due to the adopted CTDE architecture used for MADRL policy optimization, which facilitates multi-agent policy learning through sharing training samples and model parameters. Therefore, *g*MADRL-VCS's policy model enables to efficiently conduct long-term planning for multiple UAVs, aimed at maximizing the designed reward in Eqn. (15). Second, we observe the efficient collaborative behavior among UGVs traversing between stops. For example, in HKSAR dataset when there is little remaining data nearby, after $t = 40$, UGV1 (marked in red) and UGV2 (marked in purple) transported some UAVs to those stops located in the left part of the workzone. Meanwhile, UGV3 (marked in blue) navigates to the right side with abundant PoIs for

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3459039

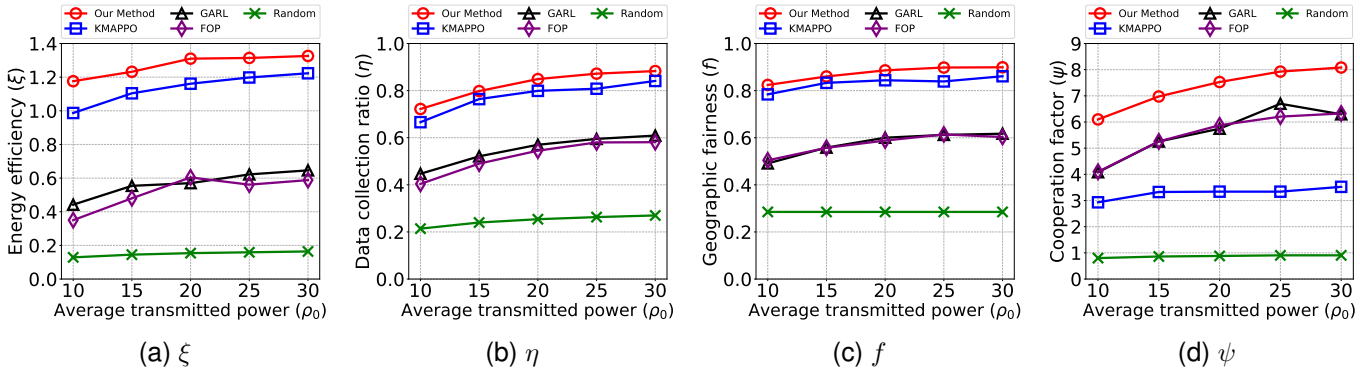


Fig. 11: Impact of the average transmitted power of PoIs (HKSAR dataset).

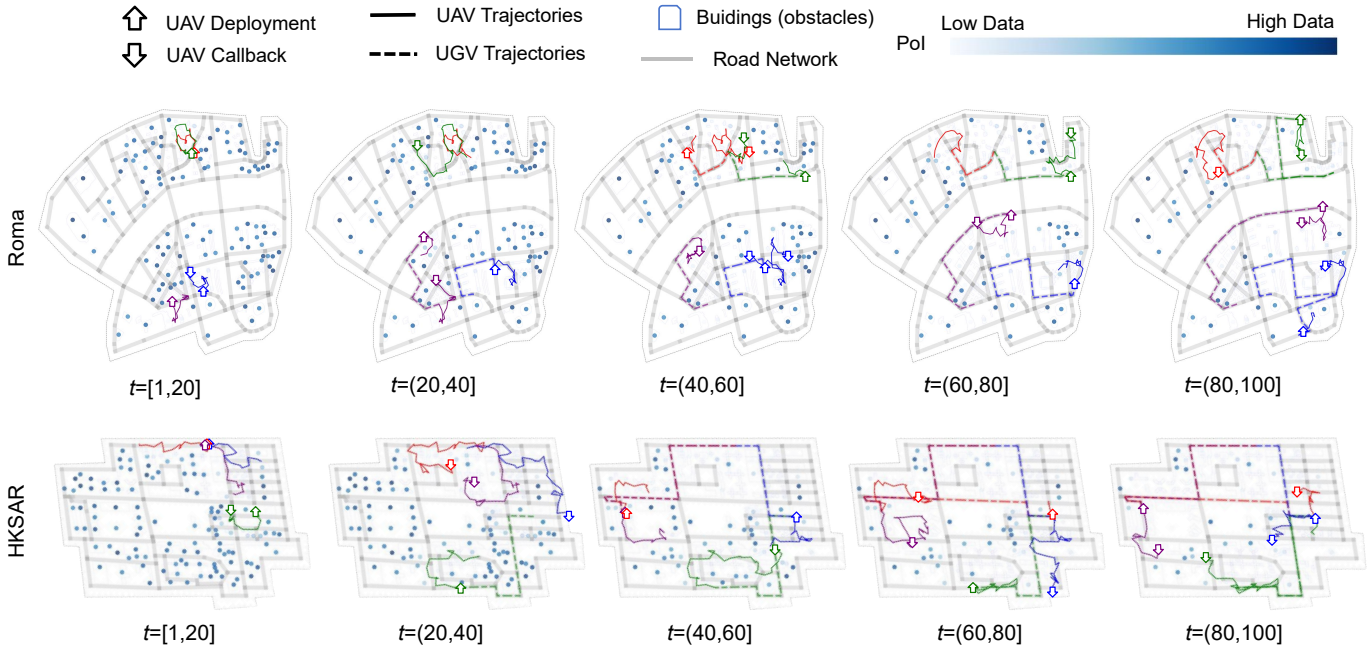


Fig. 12: UAV-UGV trajectory visualization in Roma and HKSAR datasets.

further data collection. Similar findings can be observed in Roma dataset where each UGVs corresponds to a region, dispatching UAVs to collect data from nearby PoIs. The efficiency can be attributed to our proposed highly expressive goal-conditioned discrete diffusion model, generating efficient navigation policies for UGVs. Also, this is further supported with our multi-window goal relabeling method to facilitate the learning process. Finally, we observe a high data collection ratio due to the effective cooperation between UAVs and UGVs. gMADRL-VCS achieved a high data collection ratio in HKSAR (with 89.44%) at timeslot $t = 100$ with only 4 UAVs and UGVs.

VII. Discussion and Conclusion

In this paper, we consider a novel Ground-Air-Space Vehicular Crowdsensing (GAS-VCS) campaign where UGVs dispatch a group of UAVs to collect sensory data from PoIs and call them back to transport to another region until the task completion. We proposed gMADRL-VCS, a hierarchical MADRL method with diffusion models, to maximize the overall energy efficiency. Specifically, at the high-level, a goal-conditioned discrete diffusion model is proposed to generate representative navigation policies for UGVs, and then the policy model is optimized through MASAC algorithm. A multi-window goal-relabeling mechanism is then proposed to improve the training performance. At the low-level, we optimize the route-planning policy for UAVs based on IPPO in a CTDE manner. We conducted extensive experiments on two real-world datasets in Roma, Italy, and Hong Kong SAR, China. We found the most appropriate hyperparmaters and show the benefits of our proposal by performing ablation study and comparing with four other baselines.

In our future work, we aim to enhance this GAS-VCS system by integrating additional data types, such as language commands, to align with human instructions. Additionally, we will implement our method on actual UGV-UAV platforms and deploy the system for environmental monitoring and disaster response to gather essential data.

REFERENCES

[1] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6g wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 957–975, 2020.

[2] H. Chen, M. Xiao, and Z. Pang, "Satellite-based computing networks with federated learning," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 78–84, 2022.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in *IEEE ITSC'19*, 2019, pp. 2765–2771.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," vol. 35, 2022, pp. 27 730–27 744.

[6] C. Xu and W. Song, "Decentralized task assignment for mobile crowdsensing with multi-agent deep reinforcement learning," *IEEE Internet of Things Journal*, 2023.

[7] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun, and H. Gao, "Cooperative uav trajectory design for disaster area emergency communications: A multi-agent ppo method," *IEEE Internet of Things Journal*, vol. 11, pp. 8848–8859, 2024.

[8] S. Wu, W. Xu, F. Wang, G. Li, and M. Pan, "Distributed federated deep reinforcement learning based trajectory optimization for air-ground cooperative emergency networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 9107–9112, 2022.

[9] Y. Wang, J. Wu, X. Hua, C. H. Liu, G. Li, J. Zhao, Y. Yuan, and G. Wang, "Air-ground spatial crowdsourcing with uav carriers by geometric graph convolutional multi-agent deep reinforcement learning," in *IEEE ICDE'23*, 2023, pp. 1790–1802.

[10] J. Cui, S. X. Ng, D. Liu, J. Zhang, A. Nallanathan, and L. Hanzo, "Multiobjective optimization for integrated ground-air-space networks: Current research and future challenges," *IEEE Vehicular Technology Magazine*, vol. 16, no. 3, pp. 88–98, 2021.

[11] X. Cao, B. Yang, C. Yuen, and Z. Han, "Hap-reserved communications in space-air-ground integrated networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 8286–8291, 2021.

[12] X. Huang, J. A. Zhang, R. P. Liu, Y. J. Guo, and L. Hanzo, "Airplane-aided integrated networking for 6g wireless: Will it work?" *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 84–91, 2019.

[13] M. Hu, T. Zhang, S. Wang, G. Li, Y. Chen, Q. Li, and G. Chen, "Integrated robotics networks with co-optimization of drone placement and air-ground communications," in *IEEE VTC2023-Fall*, 2023, pp. 1–5.

[14] Y. Ma, T. Lv, T. Li, G. Pan, Y. Chen, and M.-S. Alouini, "Effect of strong time-varying transmission distance on leo satellite-terrestrial deliveries," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9781–9793, 2022.

[15] M. Ouyang, R. Zhang, B. Wang, J. Liu, T. Huang, L. Liu, J. Tong, N. Xin, and F. R. Yu, "Network coding-based multi-path transmission for leo satellite networks with domain cluster," *IEEE Internet of Things Journal*, 2024.

[16] Z. Jia, M. Sheng, J. Li, and Z. Han, "Toward data collection and transmission in 6g space–air–ground integrated networks: Cooperative hap and leo satellite schemes," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10 516–10 528, 2021.

[17] Y. Gong, H. Yao, D. Wu, W. Yuan, T. Dong, and F. R. Yu, "Computation offloading for rechargeable users in space-air-ground networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 3, pp. 3805–3818, 2022.

[18] R. Han, H. Li, E. J. Knoblock, M. R. Gasper, and R. D. Apaza, "Joint velocity and spectrum optimization in urban air transportation system via multi-agent deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 9770–9782, 2023.

[19] A. Asheralieva, D. Niyato, and X. Wei, "Ultra-reliable low-latency slicing in space-air-ground multi-access edge computing networks for next-generation internet of things and mobile applications," *IEEE Internet of Things Journal*, vol. 11, no. 3, pp. 3956–3978, 2024.

[20] C. H. Liu, Z. Dai, H. Yang, and J. Tang, "Multi-task-oriented vehicular crowdsensing: A deep learning approach," in *IEEE INFOCOM'20*, 2020, pp. 1123–1132.

[21] C. H. Liu, Z. Chen, and Y. Zhan, "Energy-efficient distributed mobile crowd sensing: A deep learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1262–1276, 2019.

[22] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR'21*, 2021.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," vol. 33, 2020, pp. 6840–6851.

[24] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *ICLR'21*, 2021.

[25] Z. Wang, J. J. Hunt, and M. Zhou, "Diffusion policies as an expressive policy class for offline reinforcement learning," in *ICLR'21*, 2023.

[26] B. Kang, X. Ma, C. Du, T. Pang, and S. Yan, "Efficient diffusion policies for offline reinforcement learning," *ICLR'24*, vol. 36, 2024.

[27] H. Du, Z. Li, D. Niyato, J. Kang, Z. Xiong, H. Huang, and S. Mao, "Diffusion-based reinforcement learning for edge-enabled ai-generated content services," *IEEE Transactions on Mobile Computing*, pp. 1–16, 2024.

[28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[30] M. Pinem, P. M. Sihombing, M. Zulfin, S. P. Panjaitan, H. H. Rangkuti, and M. A. Siregar, "Implementation of outdoor to indoor path loss model at 1.8 ghz and 2.1 ghz with a transmitter placed on top of the building," in *2022 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, 2022, pp. 111–116.

[31] Y. Wang, Z. Gao, J. Zhang, X. Cao, D. Zheng, Y. Gao, D. W. K. Ng, and M. Di Renzo, "Trajectory design for uav-based internet of things data collection: A deep reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3899–3912, 2021.

[32] X. Qin, Z. Song, Y. Hao, and X. Sun, "Joint resource allocation and trajectory optimization for multi-uav-assisted multi-access mobile edge computing," *IEEE Wireless Communications Letters*, vol. 10, no. 7, pp. 1400–1404, 2021.

[33] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser mimo systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, 2013.

[34] H. Huh, A. M. Tulino, and G. Caire, "Network mimo with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2911–2934, 2011.

[35] Y. Ye, H. Wang, C. H. Liu, Z. Dai, G. Li, G. Wang, and J. Tang, "Qoi-aware mobile crowdsensing for metaverse by multi-agent deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 3, pp. 783–798, 2024.

[36] K. Zhang, F. He, Z. Zhang, X. Lin, and M. Li, "Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 121, p. 102861, 2020.

[37] T. K. Lo, "Maximum ratio transmission," in *IEEE INFOCOM'99*, vol. 2, 1999, pp. 1310–1314.

[38] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE transactions on wireless communications*, vol. 18, no. 4, pp. 2329–2345, 2019.

[39] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

[40] F. J. Vasko, Y. Lu, and K. Zyma, "What is the best greedy-like heuristic for the weighted set covering problem?" *Operations Research Letters*, vol. 44, no. 3, pp. 366–369, 2016.

[41] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *ICML'18*, 2018, pp. 1861–1870.

[42] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," vol. 34, 2021, pp. 17 981–17 993.

[43] C. S. de Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. Torr, M. Sun, and S. Whiteson, "Is independent learning all you need in the starcraft multi-agent challenge?" *arXiv preprint arXiv:2011.09533*, 2020.

[44] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[45] P. Series, "Propagation data and prediction methods for the design of terrestrial troadband millimetric radio access systems," *Int. Telecommun. Union, Rep. ITU-R Rec*, pp. 1410–2, 2003.
[46] Z. Cui, K. Guan, C. Briso-Rodríguez, B. Ai, and Z. Zhong, "Frequency-dependent line-of-sight probability modeling in built-up environments," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 699–709, 2019.
[47] L. Zhou, Z. Yang, G. Zhao, S. Zhou, and C.-X. Wang, "Propagation characteristics of air-to-air channels in urban environments," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–6.
[48] S. Mokhtari, N. Nouri, J. Abouei, A. Avokh, and K. N. Plataniotis, "Relaying data with joint optimization of energy and delay in cluster-based uav-assisted vanets," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 24 541–24 559, 2022.
[49] A. Saif, K. Dimyati, K. A. Noordin, N. S. M. Shah, S. Alsamhi, Q. Abdullah, and N. Farah, "Distributed clustering for user devices under uav coverage area during disaster recovery," in *2021 IEEE International Conference in Power Engineering Application (ICPEA)*, 2021, pp. 143–148.
[50] T. Zhang, Y. Li, C. Wang, G. Xie, and Z. Lu, "Fop: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning," in *ICML*, 2021, pp. 12 491–12 500.

Yinuo Zhao receives a BEng degree in Software Engineering from the Beijing Institute of Technology, China, in 2019. She is currently working towards the Ph.D. degree under the supervision of Prof. Chi Harold Liu at the School of Computer Science and Technology at the Beijing Institute of Technology, China. She is now working on the problems of mobile crowdsensing and deep reinforcement learning.

Chi Harold Liu (SM'15) receives a Ph.D. degree in Electronic Engineering from Imperial College, UK in 2010, and a B.Eng. degree in Electronic and Information Engineering from Tsinghua University, China in 2006. He is currently a Full Professor and Vice Dean at the School of Computer Science and Technology, Beijing Institute of Technology, China. He has worked for IBM Research - China and Deutsche Telekom Laboratories. His current research interests include mobile crowdsensing by deep learning. He received the IBM First Plateau Invention Achievement Award in 2012, ACM SigKDD'21 Best Paper Runner-up Award, and ACM MobiCom'21 Best Community Paper Runner-up Award. He serves as the Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING. He is a senior member of IEEE, and a Fellow of IET and British Computer Society.

Tianjiao Yi is currently an undergraduate student under the supervision of Prof. Chi Harold Liu at the School of Computer Science and Technology at the Beijing Institute of Technology, China. He is now working on the problems of mobile crowdsensing and deep reinforcement learning.

Guozheng Li received his Ph.D. degree in Computer Science from the School of EECS, Peking University, in 2021. He is an assistant professor at the School of Computer Science and Technology, Beijing Institute of Technology, China. His primary research interests include information visualization, especially hierarchical data visualization, and visualization authoring.

Dapeng Wu (F'13) received a Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA in 2003. He is currently Yeung Kin Man Chair Professor of Network Science, and Chair Professor of Data Engineering at the Department of Computer Science, City University of Hong Kong. Previously, he was on the faculty of University of Florida, Gainesville, FL, USA and was the director of NSF Center for Big Learning, USA. He has served as Editor in Chief of IEEE Transactions on Network Science and Engineering, Editor-at-Large for IEEE Open Journal of the Communications Society. He was elected as a Distinguished Lecturer by IEEE Vehicular Technology Society in 2016, and an IEEE Fellow.