Age of Semantics in Cooperative Communications: To Expedite Simulation Towards Real via Offline Reinforcement Learning

Xianfu Chen, *Senior Member, IEEE*, Zhifeng Zhao, *Senior Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, Celimuge Wu, *Senior Member, IEEE*, Honggang Zhang, *Fellow, IEEE*, and Mehdi Bennis, *Fellow, IEEE*

***Abstract*—The age of information metric fails to correctly describe the intrinsic semantics of a status update. In an intelligent reflecting surface-aided cooperative relay communication system, we propose the age of semantics (AoS) for measuring the semantics freshness of status updates. Specifically, we focus on status updates from a source node (SN) to the destination, which is formulated as a Markov decision process. The objective of the SN is to maximize the expected satisfaction of AoS and energy consumption under the maximum transmit power constraint. To seek the optimal control policy, we first derive an online deep actor-critic (DAC) learning scheme under the on-policy temporal difference learning framework. However, implementing the online DAC in practice poses key challenge in infinitely repeated interactions between the SN and the system, which can be dangerous particularly during exploration. We then put forward a novel offline DAC scheme, which estimates the optimal control policy from a previously collected dataset without any further interactions with the system. Numerical experiments verify the theoretical results and show that our offline DAC scheme significantly outperforms the online DAC scheme and the most representative baselines in terms of mean utility, demonstrating strong robustness to dataset quality.**

***Index Terms*—Cooperative communications, information freshness, Markov decision process, offline deep reinforcement learning, semantics.**

Received 29 August 2024; revised 31 March 2025; accepted 1 April 2025. Date of publication 8 April 2025; date of current version 27 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U24A20209, in part by JST ASPIRE under Grant JPMJAP2325, and in part by JSPS Bilateral Joint Research Project under Grant JPJSB120231002. Recommended for acceptance by Dr. Dusit Niyato. *(Corresponding author: Xianfu Chen.)*

Xianfu Chen is with the Shenzhen CyberAray Network Technology Company Ltd., Shenzhen 518100, China (e-mail: xianfu.chen@ieee.org).

Zhifeng Zhao is with the Zhejiang Lab, Zhejiang University, Hangzhou 310027, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: zhaozf@zhejianglab.com).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

Celimuge Wu is with the Meta-Networking Research Center, The University of Electro-Communications, Chofu-shi, Tokyo 182-8585, Japan, and also with the Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu-shi, Tokyo 182-8585, Japan (e-mail: celimuge@uec.ac.jp).

Honggang Zhang is with the Faculty of Data Science, City University of Macau, Taipa 999078, China (e-mail: hgzhang@cityu.edu.mo).

Mehdi Bennis is with the Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland (e-mail: mehdi.bennis@oulu.fi).

Digital Object Identifier 10.1109/TNSE.2025.3558747

I. Introduction

COOPERATIVE relay communications have exhibited high potentials in expanding system coverage and capacity [1]. Recently, hybrid relay systems are emerging to further enhance relaying performance, where intelligent reflecting surfaces (IRSs) are deployed to improve propagation conditions [2]. Specifically, an IRS consists of a large number of passive reflecting elements that adapt the propagation environment by tuning the amplitudes and/or phase-shifts. Without the need of any radio-frequency chains, an IRS is able to achieve cost and energy-efficient communications. In this paper, we study an IRS-aided cooperative relay communication system, where a source node (SN) updates to the destination through sampling the status of an underlying process. Typical scenarios include real-time monitoring in complex smart manufacturing [3] and video analytics in autonomous driving [4], to mention a few, where the fresh information and semantics of the process of interest is crucial from the destination perspective. Let us take vehicle detection and tracking as an illustrative example, where the intelligent camera (IC), i.e., the SN, responds to the remote control unit (RCU), i.e., the destination [5]. Each image captured by the IC can be considered to be composed of the target vehicle part and the background part. To save communication resource, the IC employs a semantic extraction module to compress the image [6], and the compressed data is sent to the RCU. Afterwards, the RCU performs semantic reconstruction to recover the image for inference. With the inference result from the RCU, the IC decides whether or not to, for example, pan, tilt or zoom in/out to capture a new image in a relevant region for high-level applications (e.g., multi-camera traffic surveillance).

A. Related Works and Motivation

Maintaining fresh information of the process of interest at the destination requires the SN to send time-stamped status updates, which motivates the introduction of age of information (AoI) [7], [11]. At the destination, AoI quantifies the time lag since the generation of the most recently received update. In the literature, most efforts have been focused on exploring AoI in single-hop communication systems [7], [8], [9], [10], [11]. It remains a challenge to minimize AoI in cooperative relay communication systems, where path selection and resource constraint noticeably expand the dimensionality. In [12], Talak et al. studied simple

stationary policies to minimize AoI for multi-hop networks under general interference constraints. In [13], Farazi et al. derived lower bounds on peak and average AoI in multi-hop wireless networks with explicit channel contention. In [14], He et al. proposed to minimize the maximum average AoI in a multi-hop Internet-of-Things network, which was formulated as a mixed integer linear programming problem jointly optimizing the beamforming vector and routing. In [15], Lou et al. developed a linearized approximate algorithm and a polynomial time algorithm to optimize AoI in a multi-hop wireless network. In [16], Liu et al. proved that peak/average AoI minimization in multi-path communications was roughly equivalent to minimizing the maximum delay, which was leveraged to design a general approximation solution. However, neglecting the tight coupling between decision-makings and uncertainties leads to suboptimal AoI in dynamic relay communication systems.

In a cooperative relay communication system, uncertainties originate from not only variations during status update transmissions (e.g., temporally changing channel gain and system topology), but also randomness in resource availabilities (e.g., transmit power and computation capability) [17]. Markov decision process (MDP) provides a mathematical framework for controlling actions (i.e., decision-makings) under uncertainties over the time horizon [18]. Using a constrained MDP, Gu et al. investigated the problem of age minimization for a two-hop relay system under a resource budget [19]. In [20], Tripathi et al. converted the AoI optimization into network stability problems, for which the Lyapunov drift was used to find the scheduling and routing policies. The Lyapunov technique does not rely on the MDP statistics but only solves an approximately optimal policy. Reinforcement learning (RL) has been successful in solving an MDP without a priori statistical information [8], [9], [10]. To our best knowledge, there has yet to be a comprehensive attempt to unleash the power of RL for AoI optimization in cooperative relay communication systems.

As we illustrate, this work will go even further and concentrate on the following two aspects.

1) *AoI versus Semantics Freshness:* Previous research has been predominantly directed at optimizing information freshness. There is a huge potential for the SN to boost the resource efficiency of status update transmissions by extracting the intrinsic semantics of the process of interest [5]. In a cooperative relay communication system, the goal is to let the destination promptly grasp the inference (e.g., the inference result from each recovered image for traffic surveillance as in the illustrative example above) from status updates. In line with the definition, even when the destination has a perfect inference of the current process status, AoI continues increasing until a new update is received from the SN. In [21], Sun et al. validated that the expected estimation performance from mean squared error (MSE)-minimum sampling surpasses that of AoI-optimal sampling for a Wiener process with random delay. To address such a shortcoming of AoI, Maatouk et al. introduced the age of incorrect information (AoII), which incorporates the content of status updates into the design of a transmission policy within the MDP framework [22]. However, AoII as in [22], [23] has either of the limitations: a) the threshold-based similarity measure can be insensitive to minor variations in inference results; and b) the assumption that the SN knows both the true and the estimated statuses of the process of interest is infeasible for like image inference in the multi-camera traffic surveillance, where the ground truth is unknown. Despite these efforts, we still lack a metric that reveals the relationship between the semantics of the process of interest at the SN and the timely inference from a status update at the destination.
2) *Online versus Offline RL:* In a simulated system, an RL agent utilizes newly collected interaction experiences to update the control policy parameters. These experiences come from implementing the control policy to be optimized. The repeated alternation between updating the control policy parameters and collecting interaction experiences over the time horizon is considered as online RL training. This falls into the "chicken and egg" paradox, which restricts the application of online RL to a real system. On one hand, the continuous online collection of interaction experiences is extremely challenging. On the other hand, the random actions from the RL agent during trial-and-error exploration are dangerous[1] for operating a system [24], [25]. Accordingly, in a real cooperative relay communication system, the control policy of the SN has to be pre-trained offline. Offline RL training leverages a static dataset, which is composed of a number of interaction experiences (e.g., from historical system operations) [25]. We refer to each interaction experience as a tuple of current system state, action, immediate utility and subsequent system state. Towards this direction, one theme of work was centered on learning the control policy by constraining the feasible action space to the support of the dataset [24], [26], [27], [28]. Another recent theme is aimed at preventing extrapolation errors attributed to out-of-distribution (OOD) actions, which are those that do not appear under the same current system state of an interaction experience in the dataset [29], [30], [31]. However, the state-of-the-art results on offline RL are either not applicable to discrete action settings or are highly sensitive to the dataset quality.

B. Contribution and Structure

In this paper, we address the above challenges by designing an offline RL scheme for the optimization of semantics freshness in an IRS-aided cooperative relay communication system. The SN updates the process of interest to the destination over the infinite time horizon through status sampling. More specifically, the SN compresses status updates using semantic extraction to derive the semantic samples, which are sent to the destination with the help of multiple relay stations (RSs) and an IRS. Accounting for resource constraints, the SN has to learn to choose the sampling

[1] In the example of multi-camera traffic surveillance, the IC might randomly decide not to capture and transmit an image (i.e., a dangerous action), even when the inference result is notably outdated at the RCU. In this case, the presence of a serious traffic accident could be missed detected.

and the RS selection actions with the perception of system uncertainties. We summarize the unique technical contributions from this work as follows.

- We define a novel metric, which is termed as age of semantics (AoS), to connect the semantics of the process of interest at the SN and the timely inference at the destination. In particular, the process status is modelled using a discrete Markov chain, while the inference from reconstructed status update at the destination follows a stochastic process. Different from AoI and AoII, AoS extends information freshness to semantics freshness, which is calculated as the time duration since the perfect inference of the current process status.
- We formulate the problem of semantics freshness optimization as a discrete MDP. The objective is to maximize the expected discounted utility, which weighs AoS and energy consumption at the SN. Without the requirement of system uncertainty statistics, we first propose an online deep actor-critic (DAC) scheme that applies the on-policy temporal difference (TD) method [18]. Through online interactions with the system, the DAC scheme enables the SN to learn to approach the optimal control policy, which maps each system state to a distribution of the sampling and RS selection actions.
- For an IRS-aided cooperative relay cooperative system in practice, continuous online interactions are impossible for the SN, which motivates us to further propose an offline data-driven DAC scheme. Without any interactions with the system, the SN trains the offline DAC scheme using only a previously collected static dataset of interaction experiences. Our proposed offline DAC scheme constructs an augmented loss function that simultaneously lower-bounds the estimated Q-function and penalizes the estimated Q-values of OOD actions. Numerical experiments demonstrate the robustness of our offline DAC scheme to dataset quality by exceeding the online advantage actor critic (A2C) scheme [32] with a small margin, and significantly outperforming the most representative existing offline deep RL scheme, namely, conservative Q-learning (CQL) [30].

The rest of this paper is structured as follows. In the next section, we introduce the system model and elaborate on the AoS metric. In Section III, we formulate the problem of semantics freshness optimization in an IRS-aided cooperative relay system as a discrete MDP and develop an online DAC scheme to approach the optimal control policy. We also analyze the challenges faced by the online DAC scheme. In Section IV, we propose an offline DAC scheme, which enables the SN to learn the control policy from a static dataset of interaction experiences. In Section V, we present the numerical experiments and discuss the evaluation results. Finally, we draw the conclusions in Section VI.

II. Models and Assumptions

As illustrated in Fig. 1, we study an IRS-aided cooperative relay communication system, where the SN updates the process of interest to the destination through transmitting the semantic samples over the discrete time slots. The SN cannot reach the destination directly due to the limited coverage and a set $\mathcal{K} = \{1, \ldots, K\}$ of RSs are deployed to extend the communication range of SN.[2] The RSs are half-duplex relays in a decode-forward (DF) mode.[3] All of the SN, the RSs and the destination have a single antenna. In the system, an IRS with I reflecting elements enhances the transmission links from SN to RSs as well as the links from RSs to destination, which can be possibly blocked by the dynamic obstacles. Each IRS element has a smaller size than the wavelength and hence scatters the incoming signal equally in all directions [34].

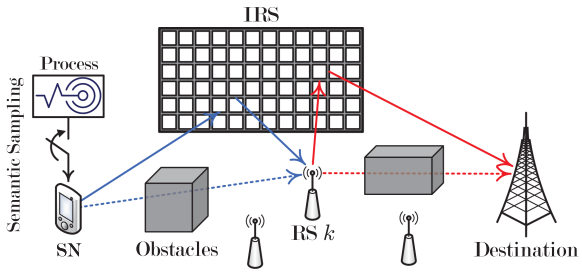

Fig. 1. Illustration of an IRS-aided cooperative relay communication, in which the SN updates the process of interest to the destination through sampling the status over the infinite time horizon.

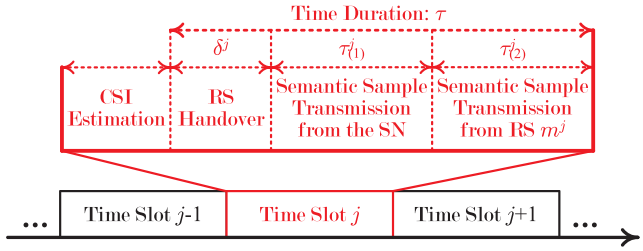

Fig. 2. Time slot structure.

We index each time slot by an integer $j \in \mathbb{N}_+$ and all time slots are assumed to be of equal duration. To facilitate the semantic sample transmissions, a time slot is further divided into four sub-slots as in Fig. 2.

1) *Fixed channel estimation sub-slot* is used to estimate the channel state information (CSI) of all links.
2) *RS handover sub-slot* is triggered for the SN once the selected RS is different from the previously associated one [35].
3) *First flexible transmission sub-slot* is used by the SN to transmit a semantic sample to the selected RS, which is aided by the IRS.
4) *Second dynamic transmission sub-slot* is occupied by the selected RS to decode and forward the received semantic sample to the destination.

[2]In the multi-camera traffic surveillance example, the roadside units in close proximity to the IC work as the RSs to help connect to the RCU.

[3]From the results in [33], DF relaying outperforms amplify-forward relaying in terms of achievable data rate. Consequently, RSs in this work operate in the DF mode.

TABLE I
PRIMARY NOTATIONS USED IN THE PAPER

Notation	Description
$K/\mathcal{K}$	Number/set of RSs
I	Number of reflecting elements of the IRS
n, n^j	Sampling action
m, m^j	RS selection action
y, y^j	SN-RS association state
c, c^j	AoS of the SN
$\mathbf{g}$, $\mathbf{g}^j$	CSI profile for the SN
X^j	Process status at time slot j
$\widehat{X}^j$	Inference at the destination at time slot j
p^j	Transmit power of the SN at time slot j
w	System frequency bandwidth
υ	Data size of a semantic sample
φ	Perfect inference probability
$\mathbf{s}$, $\mathbf{s}^j$	System state
$\mathbf{a}$, $\mathbf{a}^j$	Action of the SN
π	Control policy of the SN
u	Immediate utility of the SN
α	Temperature parameter
γ	Discount factor
V	State-value function of the SN
$\widetilde{V}$	Augmented state-value function of the SN
H	Expected discounted entropy of the control policy
Q	Q-function of the SN
$\widetilde{Q}$	Augmented Q-function of the SN
$\boldsymbol{\theta}$, $\boldsymbol{\theta}^j$	Parameters associated with the deep actor network
$\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^j$, $\boldsymbol{\lambda}^{j,-}$	Parameters associated with the deep critic network
$\mathcal{D}$	Dataset of interaction experiences
$\mathcal{O}^j$	Mini-batch at iteration j

At the end of the time slot, the destination reports the inference result back to the SN when the status update is reconstructed from the received semantic sample. We assume that the inference result is reported using the reverse link via the associated RS, the time consumption of which is negligible because of the small data size [5]. For convenience, we designate τ as the constant sum of the time durations of handover, first flexible transmission and second dynamic transmission sub-slots within a single time slot. Table I lists the primary notations of this paper.

A. IRS-Aided Relaying

Let $n^j \in \{0,1\}$ define the sampling action of the SN at each time slot j, where $n^j = 1$ if the SN decides to sample the process status X^j and otherwise, $n^j = 0$. When $n^j = 1$, a fresh status update is generated and compressed with the semantic extraction, which outputs a semantic sample of size υ. The energy consumed by status sampling and semantic extraction is assumed to be ϱ [37]. Following that, the semantic sample is sent to the destination via at most one selected RS $m^j \in \mathcal{K}$. We denote $m^j \in \widetilde{\mathcal{K}} = \mathcal{K} \cup \{0\}$ as the RS selection action for the SN at a time slot j, where in particular, we let $m^j = 0$ when $n^j = 0$ for notational consistency. Suppose that the status of the process of interest follows a discrete Markov chain, namely, $X^j \in \mathcal{X}$, where $\mathcal{X}$ is a finite state space. For the transmission of a semantic sample, the SN-RS association has to be established. Let $y^j \in \mathcal{K}$ denote the SN-RS association state of the SN at time slot j, we have $y^j = m^j$ if $m^j \in \mathcal{K}$. Otherwise, if $n^j = 0$, the SN-RS association state remains as $y^j = y^{j-1}$.

In this paper, we consider that the RSs transmit with fixed power, while the SN adapts transmit power to the channel conditions [36]. We first concentrate on the second dynamic transmission sub-slot during each time slot j. Let $g^j_{k,(\mathrm{D})} \in \mathbb{C}$ denote the channel from a RS $k \in \mathcal{K}$ to the destination, while the channels between RS k and the IRS as well as between the IRS and the destination are, respectively, denoted by $\mathbf{g}^j_{k,(\mathrm{I})} \in \mathbb{C}^I$ and $\mathbf{g}^j_{(\mathrm{I,D})} \in \mathbb{C}^I$. According to [34], we express the achievable data rate for the IRS-aided uplink from RS k to the destination at slot j as

$$R^j_{k,(\mathrm{D})} = w \cdot \log_2\left(1 + \frac{P_k \cdot \left(\left|g^j_{k,(\mathrm{D})}\right| + \zeta \cdot \sum_{i=1}^{I} \left|\left[\mathbf{g}^j_{k,(\mathrm{I})}\right]_i \cdot \left[\mathbf{g}^j_{(\mathrm{I,D})}\right]_i\right|\right)^2}{w \cdot \sigma^2}\right), \tag{1}$$

where w is the system frequency bandwidth, P_k is the transmit power of RS k, $\zeta \in (0,1]$ is the fixed amplitude reflection coefficient of the IRS, σ^2 is the additive noise power spectral density, and $[\cdot]_i$ denotes the i-th component of a vector. If $n^j = 1$ and $m^j \in \mathcal{K}$, the time consumed by transmitting the semantic sample from the selected RS m^j to the destination can be hence calculated as

$$\tau^j_{(2)} = \frac{\upsilon}{R^j_{m^j,(\mathrm{D})}}. \tag{2}$$

For the first flexible transmission sub-slot during time slot j, we then derive the amount of time used to transmit the semantic sample from the SN to the selected RS m^j as

$$\tau^j_{(1)} = \tau - \tau^j_{(2)} - \delta^j, \tag{3}$$

where $\delta^j = \delta \cdot \mathbb{1}_{\{y^j \neq y^{j-1}\}}$ with δ being the delay during the occurrence of a handover and $\mathbb{1}_{\{\cdot\}}$ denoting an indicator function. Accordingly, the required transmit power by the SN can be deduced as

$$p^j = \frac{w \cdot \sigma^2}{\left(\left|g^j_{m^j}\right| + \zeta \cdot \sum_{i=1}^{I} \left|\left[\mathbf{g}^j_{(\mathrm{I})}\right]_i \cdot \left[\mathbf{g}^j_{(\mathrm{I}),m^j}\right]_i\right|\right)^2} \cdot \left(2^{\frac{\upsilon}{\tau^j_{(1)} \cdot w}} - 1\right), \tag{4}$$

where $\mathbf{g}^j_{(\mathrm{I}),m^j} \in \mathbb{C}^I$ is the channel from the IRS to RS m^j, while $g^j_{m^j} \in \mathbb{C}$ and $\mathbf{g}^j_{(\mathrm{I})} \in \mathbb{C}^I$ are the channels from the SN to RS m^j and the IRS, respectively. We denote by P the maximum transmit power of the SN, which constrains that during each time slot j, $p^j \leq P$.

B. AoS Evolution

After receiving the semantic sample from the SN at the end of a time slot j, the destination reconstructs the status update, which leads to an inference $\widehat{X}^{j+1}$ of the process status X^j. We assume that the destination can only make a perfect inference of X^j, i.e.,

$\widehat{X}^{j+1} = X^j$, with a probability of $\varphi \in [0,1]$ due to resource and knowledge scarcity. That is, there exists a probability of $1-\varphi$ such that $\widehat{X}^{j+1} \in \mathcal{X} \setminus \{X^j\}$. Different from AoI and AoII, which care solely information freshness at the destination, we adopt AoS to quantify the connection between the semantics of the process of interest and the inference on the semantic reconstruction of a status update. More specifically, we define the AoS at the beginning of a time slot j by

$$c^j = \left(j - j^\ell\right) \cdot \mathbb{1}_{\left\{\widehat{X}^j \neq X^j\right\}}, \tag{5}$$

where j^ℓ denotes the last time slot when $\widehat{X}^{j^\ell} = X^{j^\ell}$.

The AoS dynamics of the SN can be analyzed as in the following two different cases.

1) $n^j = 0$: The SN decides not to sample the process status at time slot j. In this case, the destination does not receive any new semantic sample from the SN by the end of time slot j, which indicates that $\widehat{X}^{j+1} = \widehat{X}^j$. At the beginning of the subsequent time slot $j+1$, we arrive at $c^{j+1} = 0$ if the process status switches to $X^{j+1} = \widehat{X}^{j+1}$, and otherwise, $c^{j+1} = \min\{c^j + 1, C\}$. Herein, C reflects the staleness of inference by the destination from the received semantic sample.
2) $n^j = 1$: In this case, the SN samples the process of interest at time slot j. From the received semantic sample at the end of time slot j, the destination makes an inference $\widehat{X}^{j+1} = X^j$ with a probability of φ or $\widehat{X}^{j+1} \in \mathcal{X} \setminus \{X^j\}$ with a probability of $1-\varphi$. At the beginning of the next time slot $j+1$, a) if the process status stays in the same state as at time slot j, namely, $X^{j+1} = X^j$, we let $c^{j+1} = 0$ when $\widehat{X}^{j+1} = X^j$, and $c^{j+1} = \min\{c^j + 1, C\}$ when $\widehat{X}^{j+1} \in \mathcal{X} \setminus \{X^j\}$; and b) if the process status changes to a new state $X^{j+1} \in \mathcal{X} \setminus \{X^j\}$, we set c^{j+1} to either 0 or $\min\{c^j + 1, C\}$, respectively, depending on whether $\widehat{X}^{j+1} = X^{j+1}$ or $\widehat{X}^{j+1} \in \mathcal{X} \setminus \{X^{j+1}\}$.

III. Problem Statement

In this section, we first model the joint process status sampling and RS selection in an IRS-aided cooperative relay communication system as an MDP. The objective of the SN is to maximize the expected discounted utility, which specifies the AoS and energy consumption over the discrete time slots. After that, we discuss the solution under the online deep RL framework and the corresponding challenges.

A. MDP Formulation

Since the system involves multiple RSs, the joint process sampling and RS selection is to determine which RS the SN should use to transmit the semantic sample to the destination during each time slot. Under the MDP, the sampling and RS selection actions are adapted to the system states following a control policy. At the beginning of each time slot j, the system state can be encapsulated as $\mathbf{s}^j = (c^j, \mathbf{g}^j, y^j) \in \mathcal{S}$, where $\mathcal{S}$ represents the finite state space[4] and $\mathbf{g}^j = (\mathbf{g}^j_{(\mathrm{I})}, ((g^j_k, \mathbf{g}^j_{(\mathrm{I}),k}, \mathbf{g}^j_{k,(\mathrm{I})}, g^j_{k,(\mathrm{D})}) : k \in \mathcal{K}), \mathbf{g}^j_{(\mathrm{I},\mathrm{D})})$ represents the CSI profile for the SN. Let π be the stationary control policy of the SN, which maps a system state to a distribution over the actions, namely, $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$, where $\mathcal{A} = \{(0,0)\} \cup (\{1\} \times \mathcal{K})$ denotes the action space. By performing an action $\mathbf{a}^j = (n^j, m^j) \in \mathcal{A}$ selected with the probability of $\pi(\mathbf{s}^j, \mathbf{a}^j)$ at each time slot j, the system state transits from $\mathbf{s}^j$ to $\mathbf{s}^{j+1}$ at the beginning of next time slot $j+1$ with a probability given as

$$\phi\left(\mathbf{s}^{j+1} | \mathbf{s}^j, \mathbf{a}^j\right) = \phi\left(c^{j+1} | c^j, n^j\right) \cdot \phi\left(\mathbf{g}^{j+1}\right) \cdot \phi\left(y^{j+1} | y^j, m^j\right), \tag{6}$$

and the SN realizes an immediate utility

$$u\left(\mathbf{s}^j, \mathbf{a}^j\right) = \exp\left(-\left(\kappa \cdot c^j + \vartheta \cdot \left(\varrho + p^j \cdot \tau^j_{(1)}\right) \cdot \mathbb{1}_{\{n^j = 1\}}\right)\right), \tag{7}$$

where $\sum_{\mathbf{a}^j \in \mathcal{A}} \pi(\mathbf{s}^j, \mathbf{a}^j) = 1$ for each $\mathbf{s}^j \in \mathcal{S}$, ϕ denotes the controlled system state transition probability function, while κ and ϑ are the positive weighting factors. The exponential utility function as in (7) measures the generic satisfaction of the weighted sum of AoS and energy consumption [39].

Executing control policy π across an infinite number of time slots, the state-value function of the SN, which is the expected discounted utility starting from an initial system state $\mathbf{s} = (c, \mathbf{g}, y) \in \mathcal{S}$, can be expressed as

$$V(\mathbf{s}; \pi) = (1-\gamma) \cdot \mathsf{E}_\pi\left[\sum_{t=j}^{\infty} (\gamma)^{t-j} \cdot u\left(\mathbf{s}^t, \mathbf{a}^t\right) | \mathbf{s}^j = \mathbf{s}\right], \tag{8}$$

where $\gamma \in [0,1)$ is the discount factor and the expectation E_π is taken with respect to the probability measure induced by π. As γ approaches 1, the state-value function in (8) also approximates the expected un-discounted utility [40]. This paper chooses to use the discounted criterion due to the favorable mathematical properties [41] and the foresight of system uncertainties [18]. Eventually, the objective of the SN is to find the optimal control policy that maximizes the state-value function.

B. Online DAC Learning and Challenges

This section switches to an augmented state-value function $\widetilde{V}(\mathbf{s}; \pi)$ that combines the state-value function $V(\mathbf{s}; \pi)$ and the expected discounted entropy $H(\pi)$ of the control policy $\pi, \forall \mathbf{s} \in \mathcal{S}$. Namely,

$$\widetilde{V}(\mathbf{s}; \pi) = V(\mathbf{s}; \pi) + \alpha \cdot H(\pi), \tag{9}$$

where the temperature parameter $\alpha \geq 0$ controls the relative strength and

$$H(\pi) = (1-\gamma) \cdot \sum_{t=j}^{\infty} (\gamma)^{t-j} \cdot \sum_{\mathbf{a} \in \mathcal{A}} \left(-\pi(\mathbf{s}^t, \mathbf{a}) \cdot \ln(\pi(\mathbf{s}^t, \mathbf{a}))\right). \tag{10}$$

[4]Although the CSI is generally continuous, we can transform a semi-MDP into a regular discrete MDP with state abstraction [38].

Maximizing the augmented state-value function given by (9) instead of (8) encourages exploring the action space $\mathcal{A}$ adequately and prohibits the early convergence to sub-optimal control policies [42].

By factoring the utility and the entropy at the current time slot in the augmented state-value function (9), we define the augmented Q-function

$$\widetilde{Q}(\mathbf{s},\mathbf{a};\pi) = (1-\gamma)\cdot u(\mathbf{s},\mathbf{a}) + \gamma\cdot\sum_{\mathbf{s}'\in\mathcal{S}}\phi(\mathbf{s}'|\mathbf{s},\mathbf{a})\cdot\widetilde{V}(\mathbf{s}';\pi), \tag{11}$$

where $\mathbf{s}'=(c',\mathbf{g}',y')$ denotes the possible subsequent system state after performing an action $\mathbf{a}\in\mathcal{A}$ under the system state $\mathbf{s}\in\mathcal{S}$. Applying the Bellman equation, we recursively get

$$\widetilde{V}(\mathbf{s};\pi) = \sum_{\mathbf{a}\in\mathcal{A}}\pi(\mathbf{s},\mathbf{a})\cdot\Big(\widetilde{Q}(\mathbf{s},\mathbf{a};\pi)-\alpha\cdot(1-\gamma)\cdot\ln(\pi(\mathbf{s},\mathbf{a}))\Big). \tag{12}$$

Interleaving the policy evaluation and the policy improvement converges to the optimal stationary control policy [43]. However, the extremely large state space $\mathcal{S}$ and the dependence on controlled system state transition probability function ϕ ask for a DAC scheme to learn the optimal control policy. To that end, we approximate the optimal control policy and the optimal augmented Q-function using, respectively, a deep actor network $\pi_{\boldsymbol{\theta}}$ and a deep critic network $\widetilde{Q}_{\boldsymbol{\lambda}}$, where $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are the respective deep neural network parameters.

After performing an action $\mathbf{a}^j\in\mathcal{A}$ under the system state $\mathbf{s}^j\in\mathcal{S}$ following the control policy $\pi_{\boldsymbol{\theta}^j}$ at each time slot j, the SN trains the deep actor network parameters with the purpose of maximizing the augmented state-value function. More specifically, the training of the deep actor network follows (13) shown at the bottom of Page, where $\beta_{\boldsymbol{\theta}}$ is the learning rate, while $\boldsymbol{\theta}^j$ and $\boldsymbol{\lambda}^j$ are the parameters of the deep actor network and the deep critic network at time slot j. For the training of the deep critic network, we follow the standard TD method. In accordance with (11) and (12), the on-policy TD error at each time slot j can be mathematically expressed as (14) shown at the the bottom of this page. for the state transition from $\mathbf{s}^j$ to $\mathbf{s}^{j+1}\in\mathcal{S}$, where action $\mathbf{a}^{j+1}\in\mathcal{A}$ from the control policy $\pi_{\boldsymbol{\theta}^j}$ is performed under $\mathbf{s}^{j+1}$ and $\boldsymbol{\lambda}^{j,-}$ denotes the deep critic network parameters from a previous time slot before slot j. The SN interacts with the system to adapt the deep critic network parameters such that the TD error is as close to 0 as possible. In consequence, the DAC scheme attacks the minimization of $\Gamma_{\boldsymbol{\lambda}^j}(\mathbf{s}^j,\mathbf{a}^j,\mathbf{s}^{j+1},\mathbf{a}^{j+1}) = (\Delta_{\boldsymbol{\lambda}^j}(\mathbf{s}^j,\mathbf{a}^j,\mathbf{s}^{j+1},\mathbf{a}^{j+1}))^2/2$. The rule for updating the deep critic network parameters takes the following form

$$\boldsymbol{\lambda}^{j+1}\leftarrow\boldsymbol{\lambda}^j+\beta_{\boldsymbol{\lambda}}\cdot\nabla_{\boldsymbol{\lambda}^j}\Gamma_{\boldsymbol{\lambda}^j}\left(\mathbf{s}^j,\mathbf{a}^j,\mathbf{s}^{j+1},\mathbf{a}^{j+1}\right), \tag{15}$$

where $\beta_{\boldsymbol{\lambda}}$ is the learning rate. Algorithm 1 briefly summarizes the procedure of the proposed online DAC scheme.

Algorithm 1: Online DAC Scheme for Learning to Optimize Semantics Freshness in IRS-Aided Cooperative Relay Communication Systems.

1: Initialize the deep actor network parameters $\boldsymbol{\theta}^j$ and the deep critic network parameters $\boldsymbol{\lambda}^j$, let $\boldsymbol{\lambda}^{j,-}=\boldsymbol{\lambda}^j$, observe the current system state $\mathbf{s}^j\in\mathcal{S}$, and choose an action $\mathbf{a}^j\in\mathcal{A}$ with probability $\pi_{\boldsymbol{\theta}^j}(\mathbf{s}^j,\mathbf{a}^j)$, for $j=1$.
2: **repeat**
3: The SN performs action $\mathbf{a}^j$ and achieves immediate utility $u(\mathbf{s}^j,\mathbf{a}^j)$.
4: The system transits to the next state $\mathbf{s}^{j+1}\in\mathcal{S}$.
5: With the observation of $\mathbf{s}^{j+1}$, the SN chooses an action $\mathbf{a}^{j+1}\in\mathcal{A}$ with probability $\pi_{\boldsymbol{\theta}^j}(\mathbf{s}^{j+1},\mathbf{a}^{j+1})$.
6: The SN updates the actor network parameters $\boldsymbol{\theta}^{j+1}$ and the critic network parameters $\boldsymbol{\lambda}^{j+1}$ according to (13) and (15), respectively.
7: The SN regularly resets the deep critic network parameters by $\boldsymbol{\lambda}^{j+1,-}=\boldsymbol{\lambda}^{j+1}$, and otherwise $\boldsymbol{\lambda}^{j+1,-}=\boldsymbol{\lambda}^{j,-}$.
8: The system time moves to the next slot $j=j+1$.
9: **until** A predefined stopping condition is satisfied.

The online implementation of the obtained DAC scheme is displayed in Fig. 3(a), from which we notice that the learning process alternates over discrete time slots between optimizing the control policy and collecting new interaction experiences from the policy. An interaction experience includes the information of current system state, action, immediate utility and subsequent system state. In other words, the control policy improves at each time slot relying on the freshest interaction experience, and simultaneously, each new experience comes from the control policy to be optimized [17], [45]. However, such a "chicken and egg" paradox limits the DAC applicability to the real communication systems [46]. Continuous experience acquisition from online interactions can be expensive, and an inappropriate action can lead to painful consequences, especially during exploration. Though the emerging digital twin technology facilitates virtual simulations to mirror real communication systems, creating a high-fidelity simulator is yet difficult [47].

$$\boldsymbol{\theta}^{j+1}\leftarrow\boldsymbol{\theta}^j+\beta_{\boldsymbol{\theta}}\cdot\nabla_{\boldsymbol{\theta}^j}\left(\left(\widetilde{Q}_{\boldsymbol{\lambda}^j}(\mathbf{s},\mathbf{a})-\alpha\cdot(1-\gamma)\cdot\ln(\pi_{\boldsymbol{\theta}^j}(\mathbf{s},\mathbf{a}))\right)\cdot\ln(\pi_{\boldsymbol{\theta}^j}(\mathbf{s},\mathbf{a}))\right) \tag{13}$$

$$\Delta_{\boldsymbol{\lambda}^j}\left(\mathbf{s}^j,\mathbf{a}^j,\mathbf{s}^{j+1},\mathbf{a}^{j+1}\right) = (1-\gamma)\cdot u\left(\mathbf{s}^j,\mathbf{a}^j\right)+\gamma\cdot\left(\widetilde{Q}_{\boldsymbol{\lambda}^{j,-}}\left(\mathbf{s}^{j+1},\mathbf{a}^{j+1}\right)-\alpha\cdot(1-\gamma)\cdot\ln\left(\pi_{\boldsymbol{\theta}^j}\left(\mathbf{s}^{j+1},\mathbf{a}^{j+1}\right)\right)\right)-\widetilde{Q}_{\boldsymbol{\lambda}^j}\left(\mathbf{s}^j,\mathbf{a}^j\right) \tag{14}$$

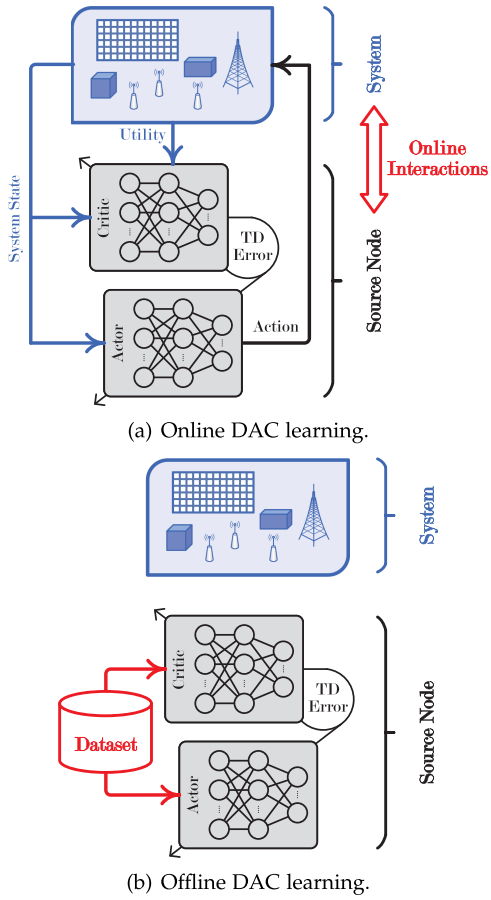

Fig. 3. Implementation comparison between two proposed schemes.

IV. Offline Data-Driven Framework

In this section, we propose an offline DAC scheme that aims to learn the target control policy by leveraging a previously collected static dataset. We designate $\mathcal{D}$ as the dataset consisting of a finite number $|\mathcal{D}|$ of interaction experience tuples, each of which is denoted by $(\mathbf{s}, \mathbf{a}, u(\mathbf{s}, \mathbf{a}), \mathbf{s}') \in \mathcal{D} \subset \mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$. Moreover, we let $\pi_{\mathcal{D}}$ denote the empirical control policy induced from the dataset $\mathcal{D}$. Fig. 3(b) shows the implementation procedure of the offline DAC scheme. In the following discussions, we slightly abuse the notations from previous Section III.

A. Data-Driven Control Policy Learning

The fundamental difference between online DAC learning and offline data-driven control policy learning is the overestimation from extrapolation of OOD actions [25]. From the interactions with the system during online DAC learning, the overestimation can be reduced by exploring actions at each time slot, after which the SN updates the control policy with immediate utility. Learning from a static dataset offline misses the opportunity for the SN to interact with the communication system to collect new experience to correct the control policy. By reformulating the dataset $\mathcal{D}$ as an $|\mathcal{D}|$-time slot trajectory, the online DAC scheme obtained in previous Section III-B can also be implemented in an offline manner to learn the control policy, akin to behaviour cloning [29]. However, the performance from the learned control policy highly relies on dataset quality. To reduce the extrapolation of OOD actions, this section proposes an offline actor-critic scheme that learns the control policy by lower-bounding the Q-function values.

Given the actor network, the training of the critic network is essentially decoupled from the control policy. By restructuring the state-value function as in (8), we define the Q-function of the SN as

$$Q(\mathbf{s}, \mathbf{a}; \pi) = (1-\gamma) \cdot u(\mathbf{s}, \mathbf{a}) + \gamma \cdot \sum_{\mathbf{s}' \in \mathcal{S}} \phi(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \cdot V(\mathbf{s}'; \pi), \tag{16}$$

which describes the expected discounted utility for performing an action $\mathbf{a} \in \mathcal{A}$ under a system state $\mathbf{s} \in \mathcal{S}$ and following the stationary control policy π thereafter. In turn, we get $V(\mathbf{s}; \pi) = \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \cdot Q(\mathbf{s}, \mathbf{a}; \pi)$, with which the Q-function can be reexpressed as the following Bellman equation

$$Q(\mathbf{s}, \mathbf{a}; \pi) = (1-\gamma) \cdot u(\mathbf{s}, \mathbf{a}) + \gamma \cdot \sum_{\mathbf{s}' \in \mathcal{S}} \phi(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \cdot \sum_{\mathbf{a}' \in \mathcal{A}} \pi(\mathbf{s}', \mathbf{a}') \cdot Q(\mathbf{s}', \mathbf{a}'; \pi). \tag{17}$$

Given the dataset $\mathcal{D}$ under the empirical control policy $\pi_{\mathcal{D}}$ in the offline settings, an estimated control policy is evaluated to determine the Q-function values. Since $\mathcal{D}$ may not include all possible interaction experiences in $\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{S}$, the SN trains the critic network by iterating the estimated Q-function in order to minimize the MSE of the Bellman (17), namely, the loss function at each iteration j given by (18), shown at the bottom of the next page, where $\mathcal{O}^j \subset \mathcal{D}$ is a random mini-batch from $\mathcal{D}$ and is of size $|\mathcal{O}^j| = O$, while $\hat{Q}^j$ and $\hat{\pi}^j$ are estimates of the optimal Q-function and the optimal control policy, respectively. With the output of the critic network at iteration j, the estimated control policy $\hat{\pi}$ is improved by training the actor network to maximize

$$f\left(\hat{\pi}; \hat{Q}^j, \mathcal{O}^j\right) = \mathsf{E}_{\{\mathbf{s}:(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}') \in \mathcal{O}^j\}} \left[\sum_{\mathbf{a}' \in \mathcal{A}} \hat{\pi}(\mathbf{s}, \mathbf{a}') \cdot \hat{Q}^j(\mathbf{s}, \mathbf{a}') \right]. \tag{19}$$

It is evident that the above training of critic and actor networks does not circumvent the challenge of action distribution shift [30]. On one hand, as in (18), the Q-function is only trained for a system state and an action that appear in each interaction experience tuple from the dataset $\mathcal{D}$, but the value of the subsequent system state is calculated as the sum of Q-function values weighted by the estimated control policy over all actions. On the other hand, the control policy is trained to maximize the mean state-value in (19), which biases the OOD actions with inaccurately high Q-function values.

To steer the offline control policy learning away from OOD actions, we impose penalties below on the estimated Q-function values during critic network training.

1) The estimated Q-function, with which the estimated control policy improves upon the empirical control policy [48], is minimized together with (18).
2) The Q-function values are regularized to rank the actions appearing in the dataset higher than the OOD actions, which can be interpreted as that the critic network training adjusts the estimated Q-function of an OOD action if the value exceeds that of an action from the dataset by a certain margin [49], [50].

Correspondingly, we recast the loss function given by (18) into an augmented loss function as (20) shown at the bottom of this page, where ω is a control policy that sharpens the empirical control policy $\pi_{\mathcal{D}}$ and will be discussed in details in Section IV-C, while the constant ρ trades off the penalty degree and ν denotes the positive margin. It is worth mentioning that the support of ω caters to $\text{supp}(\omega) \subset \text{supp}(\pi_{\mathcal{D}})$ [51]. For an estimated control policy $\hat{\pi}$, the expected discounted entropy in (10) can be rewritten as the sum of infinite geometric series. That is, we have

$$H(\hat{\pi}) = -\sum_{\mathbf{s}\in\mathcal{S}}\sum_{\mathbf{a}\in\mathcal{A}} \hat{\pi}(\mathbf{s},\mathbf{a}) \cdot \ln(\hat{\pi}(\mathbf{s},\mathbf{a})). \tag{21}$$

With a dataset $\mathcal{D}$ containing a limited number of interaction experience tuples, the empirical control policy $\pi_{\mathcal{D}}$ tends to be deterministic, particularly when the state space is exceptionally large (as observed in numerical experiments). It is obvious that a low-quality dataset hurts actor network training. Therefore, we replace the objective in (19) with (22) shown at the bottom of the page, where the entropy term based on (21) results in a more stochastic control policy.

B. Theoretical Analysis of Bounded Q-Function Estimation

In the previous Section IV-A, the augmented loss function given by (20) includes two penalty terms. The minimization of the first penalty term lower bounds the estimated Q-function [52]. As confirmed by Lemma 1, the second penalty term further improves the lower bound of the estimated Q-function of OOD actions.

Lemma 1: A lower bound of the estimated Q-function of OOD actions can be approximately optimized by the minimization of

$$z(\mathbf{s},\mathbf{a}) = \sum_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \max\Big\{0, \nu + \hat{Q}(\mathbf{s},\mathbf{a}') - \hat{Q}(\mathbf{s},\mathbf{a})\Big\}, \tag{23}$$

for $(\mathbf{s},\mathbf{a})$ appearing in each interaction experience tuple from the static dataset $\mathcal{D}$.

Proof: Please refer to Appendix A for the proof details. □

Next, we provide the theoretical support that the minimization of the augmented loss function (20) guarantees the lower-bound of the estimated Q-function. We let $\mathcal{Q}$ be the space of the bounded real-valued functions over $\mathcal{S}\times\mathcal{A}$. For the given target stationary control policy π, we define the Bellman operator $\mathcal{T}^{\pi}$ by

$$\mathcal{T}^{\pi}Q(\mathbf{s},\mathbf{a};\pi) = (1-\gamma)\cdot u(\mathbf{s},\mathbf{a}) + \gamma \cdot \sum_{\mathbf{s}'\in\mathcal{S}} \phi(\mathbf{s}'|\mathbf{s},\mathbf{a}) \cdot \sum_{\mathbf{a}'\in\mathcal{A}} \pi(\mathbf{s}',\mathbf{a}') \cdot Q(\mathbf{s}',\mathbf{a}';\pi), \tag{24}$$

$\forall \mathbf{s}\in\mathcal{S}$ and $\forall \mathbf{a}\in\mathcal{A}$, which is a mapping $\mathcal{T}^{\pi}: \mathcal{Q}\to\mathcal{Q}$. In our offline settings without any interactions with the communication system, (20) uses an estimated Bellman operator $\mathcal{T}^{\hat{\pi}^j}$, where $\hat{\pi}^j$ is an estimate of π based on the dataset $\mathcal{D}$ at each iteration j. Following the martingale concentration inequality [53],

$$\Big|\mathcal{T}^{\pi}Q(\mathbf{s},\mathbf{a};\pi) - \mathcal{T}^{\pi}\hat{Q}(\mathbf{s},\mathbf{a})\Big| \le \frac{\psi}{\sqrt{\max\{1, |\mathcal{D}(\mathbf{s},\mathbf{a})|\}}}, \tag{25}$$

holds with the probability of $1-\epsilon$ for an $\epsilon\in(0,1)$, where ψ is a constant that depends on the statistics of utility realizations in $\mathcal{D}$, while $\mathcal{D}(\mathbf{s},\mathbf{a})\subset\mathcal{D}$ denotes the subset of interaction

$$l\Big(\hat{Q};\hat{\pi}^j,\hat{Q}^j,\mathcal{O}^j\Big) = \frac{1}{2}\cdot \mathsf{E}_{\{(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \left((1-\gamma)\cdot u(\mathbf{s},\mathbf{a}) + \gamma\cdot\sum_{\mathbf{a}'\in\mathcal{A}}\hat{\pi}^j(\mathbf{s}',\mathbf{a}')\cdot\hat{Q}^j(\mathbf{s}',\mathbf{a}')\right)\right)^2\right] \tag{18}$$

$$\begin{aligned} L\Big(\hat{Q};\hat{\pi}^j,\hat{Q}^j,\mathcal{O}^j\Big) &= \frac{1}{2}\cdot \mathsf{E}_{\{(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \left((1-\gamma)\cdot u(\mathbf{s},\mathbf{a}) + \gamma\cdot\sum_{\mathbf{a}'\in\mathcal{A}}\hat{\pi}^j(\mathbf{s}',\mathbf{a}')\cdot\hat{Q}^j(\mathbf{s}',\mathbf{a}')\right)\right)^2\right] \\ &\quad + \rho\cdot \mathsf{E}_{\{(\mathbf{s},\mathbf{a}):(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\sum_{\mathbf{a}'\in\mathcal{A}}\omega(\mathbf{s},\mathbf{a}')\cdot\hat{Q}(\mathbf{s},\mathbf{a}') + \sum_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}}\max\Big\{0,\nu+\hat{Q}(\mathbf{s},\mathbf{a}')-\hat{Q}(\mathbf{s},\mathbf{a})\Big\}\right] \end{aligned} \tag{20}$$

$$F\Big(\hat{\pi};\hat{Q}^j,\mathcal{O}^j\Big) = \mathsf{E}_{\{\mathbf{s}:(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\sum_{\mathbf{a}'\in\mathcal{A}}\hat{\pi}(\mathbf{s},\mathbf{a}')\cdot\Big(\hat{Q}^j(\mathbf{s},\mathbf{a}') - \alpha\cdot\ln(\hat{\pi}(\mathbf{s},\mathbf{a}'))\Big)\right] \tag{22}$$

experience tuples including $(\mathbf{s}, \mathbf{a})$. In the analysis that follows, let $\boldsymbol{\omega}^{\pi} = [\omega_{\mathbf{s}}^{\pi} : \mathbf{s} \in \mathcal{S}]_{|\mathcal{S}|\times 1}$ and $\mathbf{d}^{\pi} = [d_{\mathbf{s}}^{\pi} : \mathbf{s} \in \mathcal{S}]_{|\mathcal{S}|\times 1}$ be two column vectors with $\omega_{\mathbf{s}}^{\pi} = \sum_{\mathbf{a}\in\mathcal{A}}(\pi(\mathbf{s},\mathbf{a}) \cdot (\omega(\mathbf{s},\mathbf{a}) + 1))/\pi_{\mathcal{D}}(\mathbf{s},\mathbf{a})$ and $d_{\mathbf{s}}^{\pi} = \psi \cdot \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{s},\mathbf{a})/\sqrt{\max\{1, |\mathcal{D}(\mathbf{s},\mathbf{a})|\}}$. As one of the major results from this paper, we have Theorem 2 stated as below.

Theorem 2: There exists an $\epsilon \in (0,1)$ such that with the probability $1-\epsilon$, the state-value function under the estimated Q-function $\hat{Q}$ from minimizing the augmented loss function as in (20) satisfies[5]

$$\hat{\mathbf{V}}^{\pi} \leq \mathbf{V}(\pi) - (\mathbf{I} - \gamma \cdot \boldsymbol{\Phi}^{\pi})^{-1} \cdot (\rho \cdot \boldsymbol{\omega}^{\pi} - \mathbf{d}^{\pi}), \tag{26}$$

wherein $\hat{\mathbf{V}}^{\pi} = [\hat{V}^{\pi}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}]_{|\mathcal{S}|\times 1}$ with each $\hat{V}(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{s},\mathbf{a}) \cdot \hat{Q}(\mathbf{s},\mathbf{a})$, $\mathbf{V}(\pi) = [V(\mathbf{s};\pi) : \mathbf{s} \in \mathcal{S}]_{|\mathcal{S}|\times 1}$, $\mathbf{I}$ denotes an $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix, and $\boldsymbol{\Phi}^{\pi}$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with each entry at the position $(e_{\mathbf{s}}, e_{\mathbf{s}'})$ $(1 \leq e_{\mathbf{s}}, e_{\mathbf{s}'} \leq |\mathcal{S}|)$ given by $\boldsymbol{\Phi}^{\pi}_{e_{\mathbf{s}}, e_{\mathbf{s}'}} = \sum_{\mathbf{a}\in\mathcal{A}} \pi(\mathbf{s},\mathbf{a}) \cdot \phi(\mathbf{s}'|\mathbf{s},\mathbf{a})$. If

$$\rho \geq \max_{(\mathbf{s},\mathbf{a})\in\mathcal{S}\times\mathcal{A}} \frac{\psi}{\sqrt{\max\{1, |\mathcal{D}(\mathbf{s},\mathbf{a})|\}}} \cdot \frac{\pi_{\mathcal{D}}(\mathbf{s},\mathbf{a})}{\omega(\mathbf{s},\mathbf{a}) + 1}, \tag{27}$$

then for any system state $\mathbf{s}$ in an interaction experience tuple from the dataset $\mathcal{D}$, $\hat{V}(\mathbf{s}) \leq V(\mathbf{s};\pi)$.

Proof: Please refer to Appendix B for the proof details. □

C. Practical Offline DAC Scheme

Following [29], [55], the optimum ω^* in the augmented loss function (20) can be realized by maximizing

$$b(\omega) = \mathsf{E}_{\{\mathbf{s}:(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{D}\}}\left[\sum_{\mathbf{a}'\in\mathcal{A}} \omega(\mathbf{s},\mathbf{a}') \cdot \hat{Q}(\mathbf{s},\mathbf{a}')\right] - \mathsf{KL}(\omega, \pi_{\mathcal{D}}), \tag{28}$$

where $\mathsf{KL}(\omega, \pi_{\mathcal{D}})$ means the Kullback-Leibler divergence between ω and $\pi_{\mathcal{D}}$. Setting the derivative of $b(\omega)$ with respect to $\hat{Q}(\mathbf{s},\mathbf{a}')$ to 0 yields

$$\omega^*(\mathbf{s},\mathbf{a}') = \pi_{\mathcal{D}}(\mathbf{s},\mathbf{a}') \cdot \exp\left(\hat{Q}(\mathbf{s},\mathbf{a}') - 1\right), \tag{29}$$

for the system state $\mathbf{s}$ in an interaction experience tuple from the dataset $\mathcal{D}$ and each $\mathbf{a}' \in \mathcal{A}$. By substituting (29) back into (28), we thus can rewrite (20) as (30), shown at the bottom of this page, by replacing the first penalty with a softmax value $\ln(\sum_{\mathbf{a}'\in\mathcal{A}} \exp(\hat{Q}(\mathbf{s},\mathbf{a}')))$ [56]. To address the challenge of an extremely large state space, we employ two deep neural networks $\hat{\pi}_{\boldsymbol{\theta}^j}$ and $\hat{Q}_{\boldsymbol{\lambda}^j}$ to model the estimated control policy $\hat{\pi}^j$ and the estimated Q-function $\hat{Q}^j$ at each iteration j, as in the online settings. To be specific, the updating rules for the deep actor network and the deep critic network parameters are given by

$$\boldsymbol{\theta}^{j+1} \leftarrow \boldsymbol{\theta}^j + \beta_{\boldsymbol{\theta}} \cdot \nabla_{\boldsymbol{\theta}^j} F\left(\hat{\pi}_{\boldsymbol{\theta}^j}; \hat{Q}_{\boldsymbol{\lambda}^j}, \mathcal{O}^j\right), \tag{31}$$

and

$$\boldsymbol{\lambda}^{j+1} \leftarrow \boldsymbol{\lambda}^j - \beta_{\boldsymbol{\lambda}} \cdot \nabla_{\boldsymbol{\lambda}^j} L\left(\hat{Q}_{\boldsymbol{\lambda}^j}; \hat{\pi}_{\boldsymbol{\theta}^j}, \hat{Q}_{\boldsymbol{\lambda}^{j,-}}, \mathcal{O}^j\right), \tag{32}$$

respectively, where we choose $\boldsymbol{\lambda}^{j,-}$ to denote the deep critic network parameters from a previous iteration before iteration j and is regularly reset. Algorithm 2 summarizes the implementation procedure of our proposed offline DAC scheme.

Algorithm 2: Offline DAC Scheme for Semantics Freshness Optimization in IRS-Aided Cooperative Relay Communication Systems.

1: initialize the deep actor network parameters $\boldsymbol{\theta}^j$ as well as the deep critic network parameters $\boldsymbol{\lambda}^j$, for $j = 1$.
2: **repeat**
3: Randomly sample a mini-batch $\mathcal{O}^j \subset \mathcal{D}$ of interaction experience tuples.
4: Update the deep actor network parameters $\boldsymbol{\theta}^{j+1}$ according to (31).
5: Update the deep critic network parameters $\boldsymbol{\lambda}^{j+1}$ according to (32).
6: Regularly reset the deep critic network parameters with $\boldsymbol{\lambda}^{j+1,-} = \boldsymbol{\lambda}^{j+1}$, and otherwise $\boldsymbol{\lambda}^{j+1,-} = \boldsymbol{\lambda}^{j,-}$.
7: Set the iteration index $j = j + 1$.
8: **until** A predefined stopping condition is satisfied.

V. Numerical Experiments

In this section, we numerically evaluate the proposed offline DAC scheme by conducting a series of experiments with TensorFlow.

A. Experimental Configurations and Datasets

We set up an IRS-aided cooperative relay communication system with $K = 5$ RSs. The IRS is deployed to have line-of-sight channels to the SN/RSs/destination. Specifically, we consider a three-dimensional coordinate system, where the locations of the SN, IRS, RSs, and destination are given in Table III. Due to obstacles, there are non-line-of-sight channels between the

[5]For the analysis convenience, the system states, which do not appear in the dateset $\mathcal{D}$, are not excluded from the estimated state-value function. The theoretical analysis still holds by letting the estimated state-values of such states equal to the state-values.

$$L\left(\hat{Q}; \hat{\pi}^j, \mathcal{O}^j\right) = \frac{1}{2} \cdot \mathsf{E}_{\{(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \left((1-\gamma)\cdot u(\mathbf{s},\mathbf{a}) + \gamma \cdot \sum_{\mathbf{a}'\in\mathcal{A}} \hat{\pi}^j(\mathbf{s}',\mathbf{a}') \cdot \hat{Q}^j(\mathbf{s}',\mathbf{a}')\right)\right)^2\right] + \rho \cdot \mathsf{E}_{\{(\mathbf{s},\mathbf{a}):(\mathbf{s},\mathbf{a},u(\mathbf{s},\mathbf{a}),\mathbf{s}')\in\mathcal{O}^j\}}\left[\ln\left(\sum_{\mathbf{a}'\in\mathcal{A}} \exp\left(\hat{Q}(\mathbf{s},\mathbf{a}')\right)\right) + \sum_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \max\left\{0, \nu + \hat{Q}(\mathbf{s},\mathbf{a}') - \hat{Q}(\mathbf{s},\mathbf{a})\right\}\right] \tag{30}$$

TABLE II
PARAMETER VALUES IN EXPERIMENTS

Parameter	Value	Parameter	Value
ζ	1	P	30 dBm
w	10 MHz	κ	$5 \cdot 10^{-2}$
C	30	ϑ	1
υ	$6.2 \cdot 10^6$ bits	σ^2	-174 dBm/Hz
ϱ	0.01 Joule	γ	0.9
τ	0.1 seconds	α	10^{-4}
δ	10^{-3} seconds	ρ	$5 \cdot 10^{-4}$
P_k	30 dBm, $\forall k$	ν	1

TABLE III
LOCATIONS IN THE THREE-DIMENSIONAL COORDINATE SYSTEM

System Node	Coordinate
SN	(0 meters, 0 meters, 0 meters)
IRS	(80 meters, 80 meters, 80 meters)
RS 1	(65 meters, 95 meters, 0 meters)
RS 2	(70 meters, 85 meters, 0 meters)
RS 3	(80 meters, 80 meters, 0 meters)
RS 4	(85 meters, 75 meters, 0 meters)
RS 5	(90 meters, 65 meters, 0 meters)
Destination	(180 meters, 180 meters, 0 meters)

SN and the RSs as well as between the RSs and the destination. Channel gains over the discrete time slots are modeled using the 3GPP Urban Micro scenario [34]. At each time slot, the status of the process of interest at the SN is assumed to be in one of $|\mathcal{X}| = 9$ states, for which we set the probability of remaining in the same state during the next time slot as χ. Then for experimental purpose, the probability of transitioning to another different state is $(1-\chi)/(|\mathcal{X}|-1)$. For both of the online DAC and the offline DAC schemes, the deep actor and the deep critic networks are designed to be with one hidden layer, which contains 64 neurons and uses ReLU as the activation function [57]. As for the output layer, the deep actor network chooses Softmax as the activation function, while the deep critic network selects a linear output layer [58]. Adam is kept as the optimizer throughout all experiments [59]. Other parameter values are listed in Table II.

In addition to the online DAC scheme, we compare the proposed offline DAC scheme with three baselines as well, namely, the A2C scheme [32], the Random scheme and the CQL scheme [30]. Implementing the Random scheme, the SN applies a uniform probability distribution over the sampling and RS selection actions across the time horizon. All datasets in experiments are generated from both the A2C and the Random schemes, and are categorized into the respective Expert Data and Random Data. For each dataset, we collect $|\mathcal{D}| = 2 \cdot 10^5$ interaction experience tuples. The size of each mini-batch is set to be $O = 5 \cdot 10^3$ during the training of our proposed offline DAC scheme.

B. Results and Discussions

1) Convergence Validation: We first evaluate the convergence speed of training the proposed offline DAC scheme using not only the Expert Data but also the Random Data. In the experiment, we assume an IRS with $I = 75$ reflecting elements,

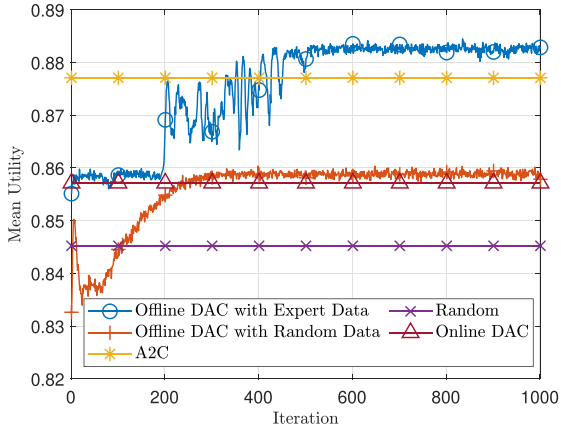

Fig. 4. Illustration of convergence speed of the proposed offline DAC scheme in terms of mean utility: $I = 75$, $\varphi = 0.5$ and $\chi = 0.5$.

while the accurate inference construction probability at the destination and the state-remaining probability for the process at the SN are set to be $\varphi = 0.5$ and $\chi = 0.5$, respectively. We plot the variations in mean utility during the training of our proposed offline DAC scheme in Fig. 4, which shows the mean utility performance from testing the trained A2C, Random and trained online DAC schemes as well. Each point on the curves of the proposed offline DAC scheme corresponds to the mean of 10^5 utility realizations from online testing the deep actor network parameters, which are offline trained at each iteration. The curves clearly tell that the offline training converges within 600 iterations. Besides, the converged offline DAC schemes trained using Expert Data and Random Data outperform the respective A2C and Random schemes. This is attributed to the exploration/exploitation tradeoff during the online A2C learning. It is interesting to see that the proposed offline DAC scheme even trained with Random Data achieves better mean utility performance than the proposed online DAC scheme. Unsurprisingly, the mean utility performance from the online DAC scheme is deteriorated compared to the A2C scheme. The reason is that the online DAC scheme performs on-policy learning (as in (14)) to learn the near-optimal control policy in a more conservative way than the off-policy A2C scheme, which tends to be aggressive and learns directly the optimal control policy [32].

2) Performance Comparison With Baselines: By comparison with the baselines, we then move to demonstrate the performance of our proposed offline DAC scheme in terms of mean AoS, mean energy consumption and mean utility for the SN. The proposed offline DAC and the CQL schemes are trained using both Expert Data and Random Data. We configure a communication system similar as in the previous experiment except that for the process of interest at the SN, the state-remaining probability χ varies between 0.3 and 0.8. The experimental results are exhibited in Figs. 5, 6, and 7, which illustrate, respectively, the mean AoS, the mean energy consumption and the mean utility from all the online and offline schemes.

When trained with Expert Data, it can be observed from the curves in Fig. 7 that the proposed offline DAC scheme achieves the best mean utility performance, while the CQL scheme has

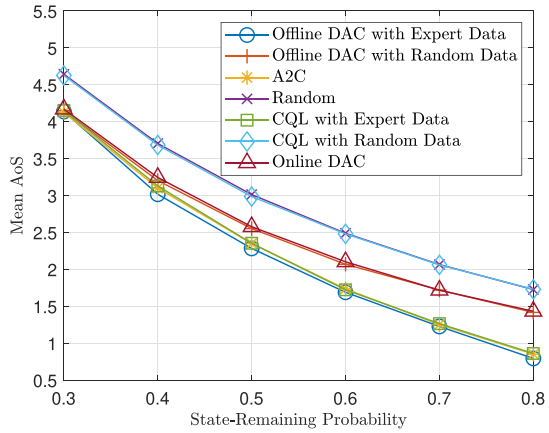

Fig. 5. Mean AoS performance for the SN versus state-remaining probability: $I = 75$ and $\varphi = 0.5$.

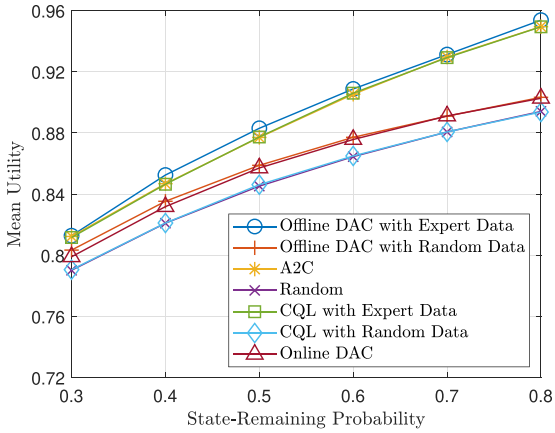

Fig. 7. Mean utility performance for the SN versus state-remaining probability: $I = 75$ and $\varphi = 0.5$.

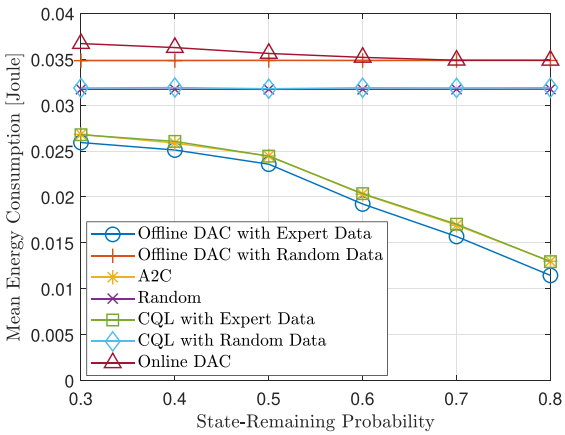

Fig. 6. Mean energy consumption for the SN versus state-remaining probability: $I = 75$ and $\varphi = 0.5$.

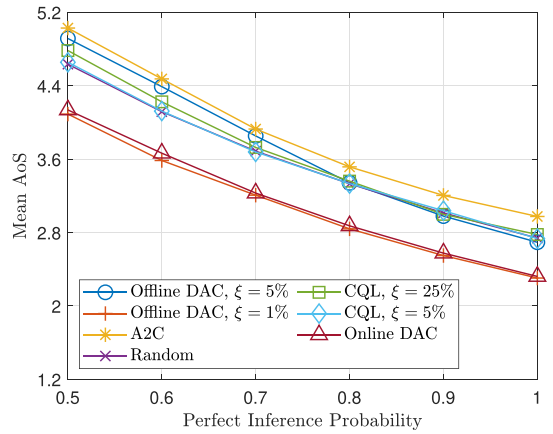

Fig. 8. Mean AoS performance for the SN versus perfect inference probability: $I = 25$ and $\chi = 0.3$.

nearly the same mean utility performance as the A2C scheme. The increase in the state-remaining probability χ increases the chance for the destination to maintain a perfect inference of the process status. As such, the mean AoS performance decreases, and the mean energy consumption of the SN also decreases due to the reduced process sampling frequency, as illustrated in Figs. 5 and 6. The mean energy consumption from the Random scheme remains unchanged, which can be explained by the fully random process status sampling and RS selection. When trained with Random Data, Fig. 7 reveals that our proposed offline DAC scheme slightly outperforms the online DAC scheme in terms of mean utility. Different from the Random scheme, the proposed offline DAC scheme increases the process sampling frequency in order to bring down the mean AoS, but still converges to a random policy, which can be obviously seen from Figs. 5 and 6. That is, the proposed offline DAC scheme fails to dig out the optimal control policy from Random Data. However, the CQL scheme performs worst as the Random scheme. This corroborates that the CQL scheme merely imitates the control policies generating Expert Data and Random Data.

3) Robustness to Dataset Quality: Finally, we carry out an experiment to assess the robustness of our proposed offline DAC scheme to the dataset quality. In this experiment, we blend Expert Data with Random Data and let ξ denote the fraction of Expert Data in a dataset used for training offline schemes. We fix the number of reflecting elements and the state-remaining probability to $I = 25$ and $\chi = 0.3$, respectively. Figs. 8, 9, and 10 depict the mean AoS, the mean energy consumption and the mean utility performance of the SN from all schemes.

It is apparent from Fig. 8 that as the perfect inference probability increases, the mean AoS decreases. The larger the probability of the perfect inference on a reconstructed status update, the destination maintains a fresher inference result of the process status, inspiring the SN to sample the process of interest more frequently. This explains why the mean energy consumptions from the offline DAC scheme trained using datasets with $\xi = 5\%$ Expert Data, the A2C scheme, the CQL scheme and the online DAC scheme increase, as perceived in Fig. 9. When the perfect inference probability is sufficiently large, the SN no longer needs to retain a high sampling frequency for random process status, resulting in the reduction of mean energy consumption. When the fraction of Expert Data is $\xi = 1\%$, it is challenging

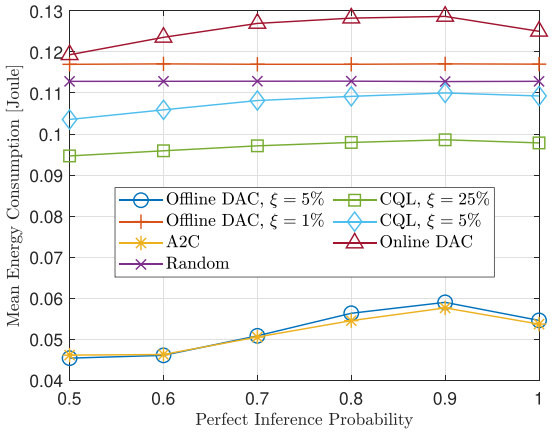

Fig. 9. Mean energy consumption for the SN versus perfect inference probability: $I = 25$ and $\chi = 0.3$.

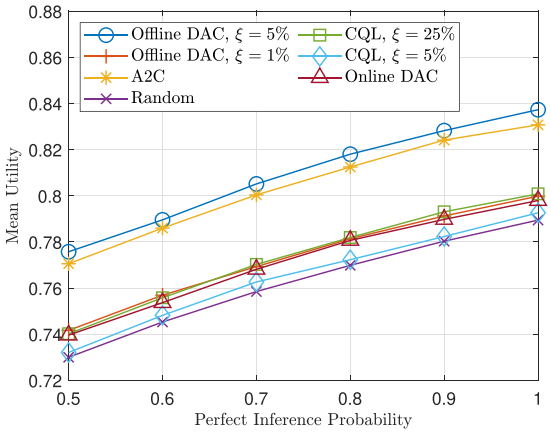

Fig. 10. Mean utility performance for the SN versus perfect inference probability: $I = 25$ and $\chi = 0.3$.

to distinguish the dataset from Random Data. The proposed offline DAC scheme converges to a random RS selection policy, with which the SN consumes the mean energy at a steady level. By comparing Figs. 6 and 9, we also discover that the SN with the Random scheme consumes less energy when the IRS is equipped with a larger number of reflecting elements, justifying the energy efficiency improvements by the IRS. Given the weighting factors, the AoS dominates the utility function value, which conforms the mean utility performance trends in Fig. 10. Last but not least, the mean utility performance from the proposed offline DAC scheme trained using datasets with only $\xi = 5\%$ Expert Data moderately outperforms the A2C scheme and significantly outperforms the CQL baseline trained using datasets with $\xi = 25\%$ Expert Data. From this experiment, the proposed offline DAC scheme exhibits the highly strong robustness to the dataset quality. On the contrary, the CQL scheme is sensitive to the quality of a dataset, which is in line with the findings from [30].

VI. Conclusion and Future Work

In this paper, we propose the notion of AoS to investigate the semantics freshness under the context of an IRS-assisted cooperative relay communication system. Considering the system uncertainties, we formulate the problem of joint process status sampling and RS selection as an MDP, where the objective of the SN is to maximize the expected discounted utility performance over discrete time slots. We first develop an online on-policy DAC scheme to alleviate the dependence on the MDP statistics. To address the "chicken and egg" paradox faced by the online DAC scheme, we then derive an offline DAC scheme. The proposed offline DAC scheme efficiently lower-bounds the estimated Q-function of OOD actions, without any further interactions with the communication system. The accuracy of the proposed studies is theoretically verified. Furthermore, numerical experiments confirm that the proposed offline DAC outperforms state-of-the-art baselines in terms of mean utility and is highly robust to dataset quality.

The theoretical extension of the proposed offline DAC scheme to multi-hop settings is feasible. However, it presents substantial challenges. As discussed in Section I, multi-hop coordination necessitates modeling interactions across sequential RSs. The increase in the number of hops exponentially raises the overheads associated with collecting global system state information and the complexity of selecting a path, thereby complicating control policy learning. In scenarios involving multiple interfering SNs, optimizing semantics freshness can be treated as a multi-agent MDP, where performance depends on the cooperative dynamics among SNs [60]. Building upon the framework proposed in this paper, we suggest that multi-agent offline DAC schemes offer a promising avenue for future research.

Appendix A
Proof of Lemma 1

Using the ranking policy gradient theorem in [50], maximizing the expected discounted utility performance is equivalent to optimizing the ranking control policy, which can be given by

$$\varpi(\mathbf{s},\mathbf{a}) = \prod_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \frac{\exp\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\right)}{1+\exp\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\right)}, \tag{33}$$

for each $(\mathbf{s},\mathbf{a}) \in \mathcal{S}\times\mathcal{A}$. Given the static dataset $\mathcal{D}$ of a finite number of interaction experience tuples, the expected discounted utility maximization is essentially to maximize the log-likelihood of each existing $(\mathbf{s},\mathbf{a})$, following which we have

$$\begin{aligned}
&\max_{\hat{Q}} \ln(\varpi(\mathbf{s},\mathbf{a}))\\
&= \max_{\hat{Q}} \ln\left(\prod_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \frac{\exp\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\right)}{1+\exp\left(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\right)}\right)\\
&= \max_{\hat{Q}} \sum_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \Big(\Big(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\Big) -\\
&\qquad\qquad \ln\Big(1+\exp\Big(\hat{Q}(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a}')\Big)\Big)\Big)\\
&\approx \min_{\hat{Q}} \frac{1}{2}\cdot \sum_{\mathbf{a}'\in\mathcal{A}\setminus\{\mathbf{a}\}} \Big(\ln(4) + \hat{Q}(\mathbf{s},\mathbf{a}') - \hat{Q}(\mathbf{s},\mathbf{a})\Big)
\end{aligned}$$

$$\Leftrightarrow \min_{\hat{Q}} z(\mathbf{s}, \mathbf{a}). \tag{34}$$

From the Q-function definition, we have $|\hat{Q}(\mathbf{s}, \mathbf{a}) - \hat{Q}(\mathbf{s}, \mathbf{a}')| \leq u_{(\max)} \leq 1$, where $u_{(\max)} = \max_{(\mathbf{s},\mathbf{a}) \in \mathcal{S} \times \mathcal{A}} u(\mathbf{s}, \mathbf{a})$. It is easy to find that the ranking policy from (34) is consistent with the policy from minimizing (23). Therefore, we are able to take the minimization of (23) as a surrogate[6] of (34), which completes the proof.

Appendix B
Proof of Theorem 2

By calculating the derivative of the augmented loss function $L(\hat{Q}; \hat{\pi}^j, \mathcal{O}^j)$ given by (20) with respect to $\hat{Q}(\mathbf{s}, \mathbf{a})$ and setting it to 0, we attain

$$\hat{Q}(\mathbf{s}, \mathbf{a}) = \mathcal{T}^{\hat{\pi}^j} \hat{Q}^j(\mathbf{s}, \mathbf{a}) - \rho \cdot \frac{\omega(\mathbf{s}, \mathbf{a}) + 1}{\pi_{\mathcal{D}}(\mathbf{s}, \mathbf{a})}, \tag{35}$$

for any $(\mathbf{s}, \mathbf{a})$ in an interaction experience tuple from $\mathcal{D}$, where the derivation of the second term at the right-hand-side is based on the marginal system state distribution under the empirical control policy $\pi_{\mathcal{D}}$ [54]. It can be noted that the minimization of augmented loss function leads to the upper-bounded estimated Q-function, namely,

$$\hat{Q}(\mathbf{s}, \mathbf{a}) \leq \mathcal{T}^{\hat{\pi}^j} \hat{Q}^j(\mathbf{s}, \mathbf{a}) = \hat{Q}^j(\mathbf{s}, \mathbf{a}). \tag{36}$$

Recall the concentration property as in (25), we hence obtain the following

$$\begin{aligned} \hat{Q}(\mathbf{s}, \mathbf{a}) &= \mathcal{T}^{\hat{\pi}} \hat{Q}(\mathbf{s}, \mathbf{a}) \\ &\leq \mathcal{T}^{\pi} \hat{Q}(\mathbf{s}, \mathbf{a}) - \rho \cdot \frac{\omega(\mathbf{s}, \mathbf{a}) + 1}{\pi_{\mathcal{D}}(\mathbf{s}, \mathbf{a})} \\ &\quad + \frac{\psi}{\sqrt{\max\{1, |\mathcal{D}(\mathbf{s}, \mathbf{a})|\}}}, \end{aligned} \tag{37}$$

with the probability $1 - \epsilon$, which indicates that the estimated state-value $\hat{V}(\mathbf{s})$ from the estimated Q-function $\hat{Q}(\mathbf{s}, \mathbf{a})$ and the control policy π fulfills

$$(\mathbf{I} - \gamma \cdot \mathbf{\Phi}^{\pi}) \cdot \hat{\mathbf{V}}^{\pi} \leq (1 - \gamma) \cdot \mathbf{u} - \rho \cdot \boldsymbol{\omega}^{\pi} + \mathbf{d}^{\pi}, \tag{38}$$

where $\mathbf{u} = [u_{\mathbf{s}}^{\pi} : \mathbf{s} \in \mathcal{S}]_{|\mathcal{S}| \times 1}$ with each $u_{\mathbf{s}}^{\pi} = \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{s}, \mathbf{a}) \cdot u(\mathbf{s}, \mathbf{a})$. For $\gamma < 1$, the spectral radius of $\gamma \cdot \mathbf{\Phi}^{\pi}$ is smaller than 1, hence the inverse of $(\mathbf{I} - \gamma \cdot \mathbf{\Phi}^{\pi})$ exists. By multiplying $(\mathbf{I} - \gamma \cdot \mathbf{\Phi}^{\pi})^{-1}$ with both sides of (38), we acquire (26) by taking into account that $\mathbf{V}(\pi) = (1 - \gamma) \cdot (\mathbf{I} - \gamma \cdot \mathbf{\Phi}^{\pi})^{-1} \cdot \mathbf{u}$. Consequently, the penalty constant ρ chosen according to (27) prevents the extrapolation error of the estimated Q-function. This completes the proof of Theorem 2.

References

[1] R. Wang and V. K. N. Lau, "Delay-aware two-hop cooperative relay communications via approximate MDP and stochastic learning," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7645–7670, Nov. 2013.

[2] Z. Chen, M.-M. Zhao, K. Xu, Y. Cai, and M.-J. Zhao, "Intelligent reflecting surface assisted full-duplex relay systems: Deployment design and beamforming optimization," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4493–4508, Jul. 2024.

[3] M. Noor-A-Rahim et al., "Towards industry 5.0: Intelligent reflecting surface in smart manufacturing," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 72–78, Oct. 2022.

[4] J. Lin, P. Yang, N. Zhang, F. Lyu, X. Chen, and L. Yu, "Low-latency edge video analytics for on-road perception of autonomous ground vehicles," *IEEE Trans. Ind. Inform.*, vol. 19, no. 2, pp. 1512–1523, Feb. 2023.

[5] K. Du et al., "Server-driven video streaming for deep learning inference," in *Proc. Annu. Conf. ACM Special Int. Group Data Commun. Appl., Technol., Archit., Protoc. Comput. Commun.*, Aug. 2020, pp. 557–570.

[6] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Aug. 2022.

[7] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2731–2735.

[8] X. Chen et al., "Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2268–2281, Apr. 2020.

[9] X. Chen et al., "Information freshness-aware task offloading in air-ground integrated edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 243–258, Jan. 2022.

[10] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in RF-powered communication systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4747–4760, Aug. 2020.

[11] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[12] R. Talak, S. Karaman, and E. Modiano, "Minimizing age-of-information in multi-hop wireless networks," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 2017, pp. 486–493.

[13] S. Farazi, A. G. Klein, J. A. McNeill, and D. Richard Brown, "On the age of information in multi-source multi-hop wireless status update networks," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun.*, Kalamata, Greece, Jun. 2018, pp. 1–5.

[14] T. He, K.-W. Chin, Z. Zhang, T. Liu, and J. Wen, "Optimizing information freshness in RF-powered multi-hop wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7135–7147, Sep. 2022.

[15] J. Lou, X. Yuan, P. Sigdel, X. Qin, S. Kompella, and N.-F. Tzeng, "Age of information optimization in multi-channel based multi-hop wireless networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 5719–5732, Oct. 2023.

[16] Q. Liu, H. Zeng, and M. Chen, "Minimizing AoI with throughput requirements in multi-path network communication," *IEEE/ACM Trans. Netw.*, vol. 30, no. 3, pp. 1203–1216, Jun. 2022.

[17] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, "Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4005–4018, Jun. 2019.

[18] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[19] Y. Gu, Q. Wang, H. Chen, Y. Li, and B. Vucetic, "Optimizing information freshness in two-hop status update systems under a resource constraint," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1380–1392, May 2021.

[20] V. Tripathi, R. Talak, and E. Modiano, "Information freshness in multihop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 31, no. 2, pp. 784–799, Apr. 2023.

[21] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, Feb. 2020.

[22] A. Maatouk, S. Kriouile, M. Assaad, and A. Ephremides, "The age of incorrect information: A new performance metric for status updates," *IEEE/ACM Trans. Netw.*, vol. 28, no. 5, pp. 2215–2228, Oct. 2020.

[23] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2621–2635, Apr. 2023.

[24] Z. Xu, K. Wu, W. Zhang, J. Tang, Y. Wang, and G. Xue, "PnP-DRL: A plug-and-play deep reinforcement learning approach for experience-driven networking," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2476–2486, Aug. 2021.

[6] The reason of using (23) rather than (34) in the augmented loss function is to stabilize the training process. In the numerical experiments, the estimated Q-function parameters are randomly initialized.

[25] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," Accessed on: Apr. 15, 2024. [Online]. Available: https://arxiv.org/pdf/2005.01643.pdf
[26] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, Jun. 2019, pp. 2052–2062.
[27] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient online reinforcement learning with offline data," in *Proc. Int. Conf. Mach. Learn.*, Honolulu, Hawaii, USA, Jul. 2023, pp. 1577–1594.
[28] N. Siegel et al., "Keep doing what worked: Behavior modelling priors for offline reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, Apr. 2020.
[29] Z. Wang et al., "Critic regularized regression," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 7768–7778.
[30] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 1179–1191.
[31] J. Liu et al., "Beyond OOD state actions: Supported cross-domain offline reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, Vancouver, Canada, Feb. 2024, pp. 13945–13953.
[32] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, Jul. 2016, pp. 1928–1937.
[33] G. Farhadi and N. C. Beaulieu, "On the ergodic capacity of multi-hop wireless relaying systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2286–2291, May 2009.
[34] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.
[35] S. Cho, E. W. Jang, and J. M. Cioffi, "Handover in multihop cellular networks," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 64–73, Jul. 2009.
[36] N. Zlatanov and R. Schober, "Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2816–2840, May 2013.
[37] S. Wang et al., "Distributed reinforcement learning for age of information minimization in real-time IoT systems," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 501–515, Apr. 2022.
[38] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artif. Intell.*, vol. 112, no. 1–2, pp. 181–211, Aug. 1999.
[39] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Mar./Apr. 2010.
[40] S. Mahadevan, "Sensitive discount optimality: Unifying discounted and average reward reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Bari, Italy, Jul. 1996, pp. 328–336.
[41] J. N. Tsitsiklis, "NP-hardness of checking the unichain condition in average cost MDPs," *Oper. Res. Lett.*, vol. 35, no. 3, pp. 319–323, May 2007.
[42] B. Eysenbach and S. Levine, "Maximum entropy RL (provably) solves some robust RL problems," in *Proc. Int. Conf. Learn. Representations*, Apr. 2022.
[43] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 1861–1870.
[44] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.
[45] X. Chen et al., "Multi-tenant cross-slice resource orchestration: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2377–2392, Oct. 2019.
[46] G. Dulac-Arnold et al., "Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis," *Mach. Learn.*, vol. 110, pp. 2419–2468, Apr. 2021.
[47] Y. Wu, K. Zhang, and Y. Zhang, "Digital twin networks: A survey," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 13789–13804, Sep. 2021.
[48] A. Abdolmaleki, J. T. Springenberg, N. Heess, Y. Tassa, and R. Munos, "Maximum a posteriori policy optimisation," in *Proc. Int. Conf. Learn. Representations*, Vancouver, BC, Canada, Apr./May 2018.
[49] D. Su, J. Ooi, T. Lu, D. Schuurmans, and C. Boutilier, "ConQUR: Mitigating delusional bias in deep Q-learning," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 9187–9195.
[50] K. Lin and J. Zhou, "Ranking policy gradient," in *Proc. Int. Conf. Learn. Representations*, Apr. 2020.
[51] S. Rezaeifar et al., "Offline reinforcement learning as anti-exploration," in *Proc. AAAI Conf. Artif. Intell.*, Feb./Mar. 2022, pp. 8106–8114.
[52] B. O'Donoghue, "Variational Bayesian reinforcement learning with regret bounds," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 28208–28221.
[53] Y. Min, J. He, T. Wang, and Q. Gu, "Learning stochastic shortest path with linear function approximation," in *Proc. Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 2022, pp. 15584–15629.
[54] A. Sharma, R. Ahmad, and C. Finn, "A state-distribution matching approach to non-episodic reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Baltimore, MD, USA, Jul. 2022, pp. 19645–19657.
[55] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 2772–2782.
[56] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. Int. Conf. Mach. Learn.*, Sydney, Australia, Aug. 2017, pp. 1352–1361.
[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 807–814.
[58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, May 2015.
[60] X. Chen, C. Wu, Z. Liu, N. Zhang, and Y. Ji, "Computation offloading in beyond 5G networks: A distributed learning framework and applications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 56–62, Apr. 2021.

Xianfu Chen (Senior Member, IEEE) received the Ph.D. (with Hons.) degree from Zhejiang University, Hangzhou, China, in 2012. In 2012, he joined the VTT Technical Research Centre of Finland, Oulu, Finland, as a Research Scientist and also as a Senior Scientist from 2013 to 2023. He is currently the Chief Research Engineer with the Shenzhen CyberAray Network Technology Company Ltd., Shenzhen, China. His research interests include various aspects of wireless communications and networking, with emphasis on human-level and artificial intelligence for resource awareness in next-generation communication networks. He was the recipient of the 2021 IEEE Communications Society Outstanding Paper Award, and 2021 IEEE Internet of Things Journal Best Paper Award. He is an Editor ofIEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, an Academic Editor of *Wireless Communications and Mobile Computing*, and an Associate Editor of *China Communications.*

Zhifeng Zhao (Senior Member, IEEE) received the Ph.D. degree in communication and information system from the PLA University of Science and Technology, Nanjing, China, in 2002. From 2002 to 2004, he was a Postdoctoral Researcher with Zhejiang University, China, where his works were focused on multimedia next generation networks and soft-switch technology for energy efficiency. From 2005 to 2006, he was a Senior Researcher with the PLA University of Science and Technology, where he performed research and development on advanced energy-efficient wireless router, ad hoc network simulator, and cognitive mesh networking test bed. He is currently the Director of the Research Development Department with Zhejiang Lab, and he is also an Associate Professor with the College of Information Science and Electronic Engineering with Zhejiang University, Hangzhou, China. His research interests include cognitive radio, wireless multihop networks (ad hoc, mesh and wireless sensor networks), wireless multimedia networks, and green communications. Dr. Zhao was the Symposium Co-Chair of the ChinaCom 2009 and 2010, and the TPC Co-Chair of the IEEE ISCIT 2010.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University in 2004. He is currently a Professor and Earle C. Williams Eminent Scholar, and Director of the Wireless Engineering Research and Education Center with Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grid. He is the Editor-in-Chief of IEEE Transactions on Cognitive Communications and Networking, a member-at-large on the Board of Governors of IEEE Communications Society, and Vice President of Technical Activities of IEEE Council on Radio Frequency Identification. He was the recipient of several journal paper awards and service awards from the IEEE, and he is a co-recipient of the Best Paper/Demo Awards of several conferences.

Celimuge Wu (Senior Member, IEEE) received the Ph.D. degree from The University of Electro-Communications, Japan, Chofu, Japan. He is currently a Professor and the Director of Meta-Networking Research Center, The University of Electro-Communications. His research interests include Vehicular Networks, Semantic Communications, Edge Computing, IoT, and AI for Wireless Networking and Computing. He was an Associate Editor of IEEE Transactions on Cognitive Communications and Networking, IEEE Transactions on Network Science and Engineering, and IEEE Transactions on Green Communications and Networking. He is the Vice Chair (Asia Pacific) of IEEE Technical Committee on Big Data. He is the recipient of the 2021 IEEE Communications Society Outstanding Paper Award, 2021 IEEE Internet of Things Journal Best Paper Award, IEEE Computer Society 2020 Best Paper Award and IEEE Computer Society 2019 Best Paper Award Runner-Up. He is an IEEE Vehicular Technology Society Distinguished Lecturer.

Honggang Zhang (Fellow, IEEE) is currently a Professor with the Faculty of Data Science, City University of Macau, Macau, China. He was the founding Chief Managing Editor of Intelligent Computing, a Science Partner Journal, and also a Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He was an Honorary Visiting Professor with the University of York, York, U.K., and an International Chair Professor of Excellence with the Université Européenne de Bretagne and Supélec, France. He has coauthored and edited two books: *Cognitive Communications: Distributed Artificial Intelligence (DAI), Regulatory Policy & Economics, Implementation* (John Wiley & Sons) and *Green Communications: Theoretical Fundamentals, Algorithms and Applications* (CRC Press), respectively. His research interests include cognitive radio networks, semantic communications, green communications, machine learning, artificial intelligence, intelligent computing, and Internet of Intelligence. He is the corecipient of the 2021 IEEE Communications Society Outstanding Paper Award and the 2021 IEEE Internet of Things Journal Best Paper Award. He was the leading Guest Editor for the Special Issues on Green Communications of the *IEEE Communications Magazine*. He was the Series Editor for the *IEEE Communications Magazine* (Green Communications and Computing Networks Series) from 2015 to 2018 and the Chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is the Associate Editor-in-Chief of *China Communications*.

Mehdi Bennis (Fellow, IEEE) is currently a Professor with the Centre for Wireless Communications, University of Oulu, Finland. He has coauthored one book and published more than 200 research papers in international conferences, journals and book chapters. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning in 5G networks and beyond. He was the recipient of several awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, 2016 Best Tutorial Prize from the IEEE Communications Society, 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks, all-University of Oulu award for research, 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, 2021 IEEE Communications Society Outstanding Paper Award, and 2021 IEEE Internet of Things Journal Best Paper Award. He is an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS.