Energy-Efficient Beamforming and Adaptive Computational Task Offloading in ISCC Systems

Kai Dong, *Member, IEEE,* Lei Wang, *Graduate Student Member, IEEE,* Sergiy A. Vorobyov, *Fellow, IEEE,* Zhu Han, *Fellow, IEEE,* and Tarik Taleb, *Senior Member, IEEE*

***Abstract*—Integrated sensing, communication, and computation (ISCC) enables next-generation wireless networks to perform environmental perception while processing massive data under stringent quality-of-service (QoS) requirements. Energy consumption is a crucial indicator for the ISCC system design. However, accounting for energy heterogeneity in ISCC system design is an open problem. Specifically, battery-constrained user equipments (UEs) and energy-abundant access points (APs) require fundamentally different energy allocation strategies based on device computational capabilities, battery states, and QoS constraints. In this paper, we introduce a nonconvex energy cost minimization problem by considering a user-specific energy cost ratio coefficient that explicitly balances UE-AP energy consumption according to heterogeneous device energy states. To efficiently address this problem, a double-loop framework combining successive convex approximation and alternating direction method of multipliers is also developed. Numerical results demonstrate that the proposed scheme significantly outperforms the fixed offloading baselines (full offloading, full local and half offloading) in terms of the total energy cost. In particular, the proposed scheme achieves up to $25-47.6\%$ energy cost reduction at moderate latency constraints over fixed offloading baselines, thereby supporting time-sensitive applications. Moreover, this work provides an effective solution for energy-efficient and QoS-aware 6G ISCC systems serving diverse devices with conflicting energy priorities.**

***Index Terms*—Integrated sensing, communication, and computation (ISCC), energy heterogeneity, adaptive task offloading, dual-functional beamforming, multi-access edge computing.**

I. Introduction

INTEGRATED sensing and communication (ISAC) has emerged as one of the candidate revolutionary technologies for next-generation intelligent wireless networks [1]–[3]. It enables the dual functionalities of sensing and wireless communications on a shared hardware platform and spectrum resources. ISAC systems can simultaneously perform environmental perception tasks, such as target detection, localization, and tracking, while maintaining high-quality communication services [4]. Consequently, it has great potential to significantly improve spectral efficiency, reduce deployment costs, and support intelligent and autonomous applications, such as autonomous driving and the Internet of Robotic Things (IoRT) [5]–[7]. Moreover, the growing complexity of sensor integration (e.g., millimeter-wave radar, light detection and ranging (LiDAR), and camera) in these mobile user equipments (UEs) also demands powerful computational capabilities to process the massive amounts of data collected. This necessitates a paradigm shift from ISAC to integrated sensing, communication, and computation (ISCC) to tackle the data-processing tasks [8], [9]. This paradigm shift, however, introduces a critical design challenge: UEs are severely constrained by limited battery capacity and computational resources [10], [11], whereas APs benefit from grid power supply and powerful processors. Consequently, energy-efficient ISCC system design becomes of paramount importance, where the energy consumption must be minimized while satisfying stringent QoS requirements including data processing latency and sensing link quality [9], [12].

ISAC design has attracted increasing attention in recent years, covering dual-functional waveform design [13], [14], performance tradeoff analysis [4], multiple access techniques [2], [15], and artificial intelligence (AI) assisted beamforming optimization across diverse deployment scenarios [16]. Building upon these advances, beamforming optimization for ISAC systems has been extensively studied in unmanned aerial vehicle (UAV)-enabled [17], cell-free massive multiple-input multiple-output (MIMO) [18], distributed [19], and massive MIMO scenarios [20], [21], demonstrating that both communication link quality and sensing signal-to-interference-plus-noise ratio (SINR) must be jointly guaranteed as key performance indicators. In parallel, multi-access edge computing (MEC) has emerged as a promising paradigm to offload intensive computational tasks from resource-constrained UEs to nearby edge servers [22], [23]. While binary offloading schemes [23]–[25] provide simple task allocation decisions, they lack the flexibility to handle diverse energy efficiency and latency requirements under heavy computational burdens. Adjustable offloading schemes [26], [27] offer greater adaptability by continuously tuning the offloading ratio according to real-time resource availability and QoS constraints. However, ISAC-only frameworks leave computational data processing unaddressed, while MEC-only designs ignore the sensing link quality requirements and the dual-functional beamforming constraints that are essential in

This work is partially supported by the Federal Ministry of Research, Technology, and Space (BMFTR), Germany, through the Project 6GEM+ under Grant 16KIS2411; and in part by the European Union's Horizon Europe research and innovation programme under the 6G-Path project (Grant No. 101139172).

Kai Dong, Lei Wang and Sergiy A. Vorobyov are with the Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland. (e-mail: kai.dong@aalto.fi, lei.wang@aalto.fi, sergiy.vorobyov@aalto.fi) (*Corresponding author: Kai Dong*.)

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA. (e-mail: hanzhu22@gmail.com)

Tarik Taleb is with the Faculty of Electrical Engineering and Information Technology, Ruhr University Bochum, 44801 Bochum, Germany. (e-mail: tarik.taleb@rub.de)

ISCC systems. Neither paradigm alone is sufficient to support the emerging requirements of intelligent wireless networks where sensing, communication, and computation are deeply coupled.

By integrating MEC with ISAC, the resulting ISCC paradigm enables distributed UEs to exhibit higher degrees of intelligence and robustness in complex environments [8], [9]. In this context, energy consumption serves as a critical system design indicator, as battery-constrained UEs must sustain prolonged sensing and computation operations while meeting stringent QoS requirements. A distributed edge computing scheme for vehicular ISCC networks was studied in [28] using stochastic geometry with partial task offloading and bandwidth allocation. The energy minimization data processing scheme at the device in [29] jointly determined data offloading ratio, sensing rates and offloading rates, though its single device-server pair limits scalability. An energy-efficient ISCC framework for AI inference on resource-constrained edge devices was investigated in [30], where model partitioning was employed to balance the computational workload across a single device-server pair. Energy minimization in UAV-enabled ISCC systems was investigated in [31] through resource allocation and trajectory design, where the full computational task offloading from the UEs to UAV were considered, and residual computational tasks at the UAV were further offloaded to the AP for execution. The authors in [32] developed an energy-efficient multi-access MEC scheme for a single ISAC device supported by multiple edge servers. Within this framework, sensing data collected at the ISAC device is entirely offloaded to the edge servers, with the objective of maximizing the device's energy efficiency. Given a inference accuracy constraint, the ISCC-based edge AI inference framework including three modes of on-device, on-server, and edge-device cooperation was proposed in [33]. Aiming to minimize the total energy consumption of the edge server and the device for completing the inference task, the best working mode was selected by solving the formulated optimization problem. The authors of [34] investigated a non-orthogonal multiple access-assisted ISCC system where a multi-functional AP simultaneously performs target sensing and provides two-tier task offloading services for multiple edge computing users. Specifically, the AP can further offload part of the received computational workloads to cloudlet servers to balance resource utilization across tiers. The total energy consumption was minimized via joint optimization of transmit beamforming, offloading strategies, and power allocation, while guaranteeing the required sensing performance.

Despite these efforts, a fundamental aspect remains overlooked in practical ISCC deployments, which is the inherent heterogeneity of energy consumption across devices. Specifically, UEs rely on limited batteries to sustain simultaneous sensing, communication, and local computation, while APs benefit from stable grid power supply and abundant computational resources. Moreover, this heterogeneity varies across users, as different UEs operate with distinct battery states, computational capacities, and QoS requirements. This implies that the relative energy cost between UE and AP is inherently asymmetric: when a UE operates with a critically low battery, conserving UE energy becomes paramount. Conversely, when the AP is in heavily loaded conditions, the energy cost at the AP becomes non-negligible. Existing works largely ignore such heterogeneity by minimizing total energy consumption without distinguishing the relative cost of UE versus AP energy consumption across different users, making them unable to adaptively prioritize energy allocation based on per-user device states. This motivates the introduction of a user-specific energy cost ratio coefficient that dynamically captures the individual energy urgency of each UE, enabling the system to adaptively balance local processing and computation offloading according to real-time battery states, computational capacities, and AP load conditions.

The main contributions of this paper are as follows.

- **Energy-heterogeneity-aware problem formulation:** Unlike existing ISCC works that optimize total energy consumption without distinguishing device energy states, a nonconvex total energy cost minimization problem is formulated by introducing a user-specific energy cost ratio coefficient that explicitly captures the relative energy urgency of each UE with respect to the AP. This parameterization enables three practical operation modes: *(i)* battery-critical mode, where UE energy conservation is prioritized through aggressive computation offloading to the AP; *(ii)* AP-constrained mode, where local processing is favored to avoid expensive energy consumption at a heavily loaded or energy-limited AP; and *(iii)* energy-consumption mode, where total energy consumption is minimized without preference between UE and AP energy sources. The problem jointly optimizes adaptive offloading coefficients, dual-functional beamforming vectors, and computational resource allocation, while guaranteeing sensing SINR and latency constraint.
- **Efficient double-loop SCA-ADMM algorithm framework**: To tackle the nonconvexity arising from fractional SINR constraints and latency terms, we develop a double-loop algorithmic framework that integrates successive convex approximation (SCA) with the alternating direction method of multipliers (ADMM). In the outer SCA loop, nonconvex sensing SINR constraints and achievable data rate expressions are linearized via first-order Taylor expansions around feasible points, transforming the problem into a more tractable form at each iteration. The inner ADMM loop then decomposes this reformulated problem into parallel subproblems with efficient updates: *(i)* closed-form updates are derived for the offloading coefficients, while computational resource allocation is obtained by solving biquadratic equations, and *(ii)* the beamforming vectors are obtained via a gradient-descent approach with projection onto the corresponding feasible sets, followed by rank-one recovery.
- **Performance validation**: The simulation results demonstrate that the proposed adaptive offloading scheme significantly outperforms baseline schemes with fixed offloading strategies and an orthogonal frequency-division multiple access (OFDMA)-based baseline in terms of total energy cost and success rate, and exhibits strong

robustness under resource-constrained conditions. Particularly, the proposed scheme achieves up to $25-47.6\%$ energy cost reduction at moderate latency constraints over fixed offloading baselines. The proposed approach provides an effective and practical solution for energy-efficient, QoS-guaranteed 6G ISCC systems serving devices with conflicting energy priorities and diverse computational capabilities.

Paper Organization: We introduce the system model in Section II and the total energy cost minimization formulation in Section III. The problem is then reformulated to a tractable one in Section IV and the SCA-ADMM algorithmic framework is presented in Section V. The performance evaluation, conclusion and discussion are presented in Sections VI and VII, respectively.

Notations: In this paper, scalars, vectors and matrices are represented by normal fonts (e.g., a), bold lowercase letters (e.g., $\mathbf{a}$) and bold uppercase letters (e.g., $\mathbf{A}$), respectively. A complex Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$ is denoted as $\mathcal{CN}(\mathbf{a},\mathbf{B})$. The notations $\mathrm{Tr}\,(\cdot)$ and $\mathrm{rank}\,(\cdot)$ stand for the trace and rank operations, respectively. Moreover, $(\cdot)^H$ denotes the Hermitian transpose operation, while $\Re\{\cdot\}$ takes the real part of a complex number, $\odot$ denotes the Hadamard product, and $\|\mathbf{a}\|_2$ and $\|\mathbf{A}\|_F$ represent the ℓ_2 norm of a vector $\mathbf{a}$ and the Frobenius norm of a matrix $\mathbf{A}$, respectively. Finally, $|\cdot|$ takes the magnitude of a complex number, and $\mathbf{I}$ denotes the identity matrix.

II. System Model

Consider an ISCC framework where an AP equipped with powerful computing resources serves K UEs within its sector coverage with a radius of r_0 in the azimuth plane, as illustrated in Fig. 1. To collaboratively detect a common target located at $[x_0, y_0]$, each UE performs target sensing task via the ISAC system. Meanwhile, a computational task consisting of Q_k bits, $\forall k \in \mathcal{K} = \{1,\cdots,K\}$, collected from multimodal sensing units (e.g., radar, LiDAR, and cameras) need to be processed efficiently within a delay constraint τ_k^{th} to satisfy safety-related QoS requirements. Therefore, an adaptive computational task offloading scheme via the uplink transmission is of paramount importance. Assume that each UE is equipped with two antenna arrays, comprising $M_{\mathrm{u}}^{\mathrm{t}}$ and $M_{\mathrm{u}}^{\mathrm{r}}$ antenna elements for signal transmission and reception, respectively. The AP located at $[x_{\mathrm{a}}, y_{\mathrm{a}}]$ is equipped with M_{a} antenna elements with a deployment height of h_{a}.

A. Communication and Sensing Signal Model

The sensing symbols can be transmitted together with the communication symbols using the same time-frequency resources and waveform at each UE k. The transmitted signal $\mathbf{x}_k \in \mathbb{C}^{M_{\mathrm{u}}^{\mathrm{t}}\times 1}$ from the k-th UE can be represented by

$$\mathbf{x}_k = \mathbf{f}_k s_k^{\mathrm{c}} + \mathbf{u}_k s_k^0, \tag{1}$$

where s_k^{c} is the data symbol for uplink communication with zero mean and unit power, $\mathbf{f}_k \in \mathbb{C}^{M_{\mathrm{u}}^{\mathrm{t}}\times 1}$ denotes corresponding transmit beamforming vector towards the AP, s_k^0 represents the complex Gaussian sensing symbol with zero mean and unit variance, and $\mathbf{u}_k \in \mathbb{C}^{M_{\mathrm{u}}^{\mathrm{t}}\times 1}$ denotes the sensing beamforming vector. Let $\mathbf{F}_k = \mathbf{f}_k\mathbf{f}_k^H$ and $\mathbf{R}_k^0 = \mathbf{u}_k\mathbf{u}_k^H$ represent the communication and sensing beamforming covariance matrices, respectively. Then the following power constraint should hold:

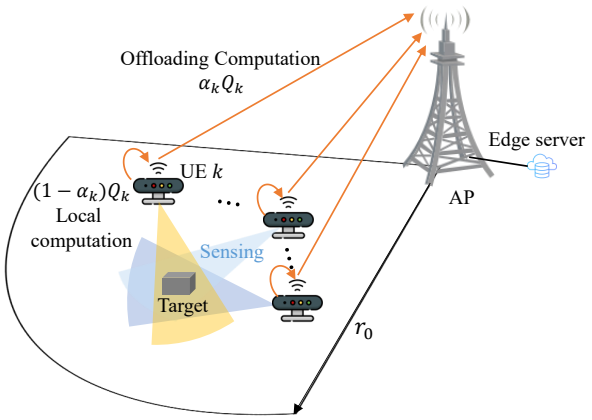


Fig. 1: ISCC system model in an IoT scenario.

$$P_k = \mathrm{Tr}(\mathbf{F}_k + \mathbf{R}_k^0) \le P_{\max}, \tag{2}$$

where $P_{\max}$ is the maximum transmit power budget. Assuming accurate synchronization among the AP and UEs [35], the received signal $\mathbf{y}^{\mathrm{ap}} \in \mathbb{C}^{M_{\mathrm{a}}\times 1}$ at the AP from all the K UEs can be expressed as[1]

$$\mathbf{y}^{\mathrm{ap}} = \sum_{k=1}^{K} \mathbf{H}_k\mathbf{f}_k s_k^{\mathrm{c}} + \sum_{k=1}^{K} \mathbf{H}_k\mathbf{u}_k s_k^0 + \mathbf{n}_1, \tag{3}$$

where $\mathbf{H}_k \in \mathbb{C}^{M_{\mathrm{a}}\times M_{\mathrm{u}}^{\mathrm{t}}}$ denotes the block-fading MIMO channel matrix between the k-th UE and AP [37], $\mathbf{n}_1 \in \mathbb{C}^{M_{\mathrm{a}}\times 1}$ is the Gaussian noise vector, i.e., $\mathbf{n}_1 \sim \mathcal{CN}(\mathbf{0}, \sigma_1^2\mathbf{I}_{M_{\mathrm{a}}})$.

Using a combining vector $\mathbf{w}_k \in \mathbb{C}^{M_{\mathrm{a}}\times 1}$, the received signal from the k-th UE can be rewritten as

$$y_k^{\mathrm{ap}} = \underbrace{\mathbf{w}_k^H\mathbf{H}_k\mathbf{f}_k s_k^{\mathrm{c}}}_{\text{Communication signal}} + \underbrace{\sum_{j=1,j\neq k}^{K} \mathbf{w}_k^H\mathbf{H}_j\mathbf{f}_j s_j^{\mathrm{c}}}_{\text{Communication interference}} + \underbrace{\sum_{j=1}^{K} \mathbf{w}_k^H\mathbf{H}_j\mathbf{u}_j s_j^0}_{\text{Interference from sensing signals}} + \underbrace{\mathbf{w}_k^H\mathbf{n}_1}_{\text{Noise}}. \tag{4}$$

We assume that the link between the transmit/receive antenna arrays of each UE and the target is under strong line-of-sight propagation conditions [38]. Given a target with an azimuth ψ_k from the local coordinates of the k-th UE, the sensing reflection channel matrix can be expressed as [18]

$$\tilde{\mathbf{H}}_k = \varrho_k\mathbf{a}_r(\psi_k)\mathbf{a}_t^H(\psi_k), \tag{5}$$

where ϱ_k denotes the channel gain involving the radar cross section of the target and two-way path loss by passing the target, $\mathbf{a}_t(\psi_k)$ and $\mathbf{a}_r(\psi_k)$ are the transmit and receive steering

[1]Here we assume that the signal reflections from the target are vanished at the AP because of the long-distance propagation at high radio frequency bands such as millimeter wave (mmWave) or THz frequency bands [36].

vectors, respectively [39]. Note that we focus on a single-target line-of-sight (LoS) sensing model in this work. This assumption is physically justified for the considered cooperative mmWave ISCC system, where severe path loss make the direct LoS the dominant source of target information, allowing heavily attenuated non-LoS (NLoS) components to be treated as background noise. Moreover, this foundational LoS sensing model is also widely adopted in recent ISAC/ISCC literature [18], [31], [34], [38], [40]–[42]. This allows us to tackle the fundamental coupling between beamforming and adaptive computational offloading. Nevertheless, the proposed framework has the potential for generalizations. For instance, in multi-target scenarios (M targets), the sensing constraint can be extended to per-target requirements (i.e., $\gamma^{\mathrm{s}}_{k,m} \geq \gamma^{\mathrm{th}}$), which requires multi-beam designs. Furthermore, in NLoS, the system could leverage resolvable multipath components or cooperative multi-view sensing [43] to extract information from reflected echoes. Exploring these multi-target and NLoS extensions with advanced robust signal processing represents is beyond the scope of this work.

With a beamforming combiner $\tilde{\mathbf{w}}_k \in \mathbb{C}^{M^{\mathrm{r}}_{\mathrm{u}} \times 1}$ at the k-th UE, the received reflected sensing signal can be written as

$$y^{\mathrm{s}}_k = \underbrace{\tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_k \mathbf{u}_k s^0_k}_{\text{Desired sensing signal}} + \underbrace{\sum_{j=1}^{K} \tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_j \mathbf{f}_j s^{\mathrm{c}}_j}_{\text{Interference from other communication signals}} + \underbrace{\sum_{j=1, j\neq k}^{K} \tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_j \mathbf{u}_j s^0_j}_{\text{Interference from other sensing signals}} + \underbrace{\tilde{\mathbf{w}}_k^H \mathbf{n}_2}_{\text{Noise}}, \tag{6}$$

where $\mathbf{n}_2 \in \mathbb{C}^{M^{\mathrm{r}}_{\mathrm{u}} \times 1}$ denotes the Gaussian noise vector, i.e., $\mathbf{n}_2 \sim \mathcal{CN}(\mathbf{0}, \sigma_2^2 \mathbf{I}_{M^{\mathrm{r}}_{\mathrm{u}}})$. It is noted that while all K UEs share a common sensing target, each UE operates independently in a monostatic sensing mode, transmitting its own sensing waveform and receiving the reflected echo via its local receive array. Due to the inherent constraints on UE energy and computational capacity, the target estimation accuracy of a single UE is limited. To address this, each UE independently processes its local sensing data to obtain a local estimate of the target parameters, which is then offloaded to the AP as part of the Q_k-bit computational task. The AP subsequently performs multi-UE estimation fusion (e.g., weighted averaging or over-the-air computation [7]) to synthesize the distributed estimates and improve the overall target estimation accuracy, exploiting the spatial diversity of the K UEs observing the target from diverse angles. Since the sensing waveforms of different UEs are not coherently joint-processed at the local receiver front-ends, the cross-echoes from other UEs act as mutual interference at each UE's local receiver.[2]

Thereby, the SINRs for the communication and sensing links at the k-th UE are given by

$$\gamma^{\mathrm{c}}_k = \frac{|\mathbf{w}_k^H \mathbf{H}_k \mathbf{f}_k|^2}{\sum_{j=1}^{K} |\mathbf{w}_k^H \mathbf{H}_j \mathbf{u}_j|^2 + \sum_{\substack{j=1 \\ j\neq k}}^{K} |\mathbf{w}_k^H \mathbf{H}_j \mathbf{f}_j|^2 + \sigma_1^2 \|\mathbf{w}_k\|_2^2}, \tag{7}$$

$$\gamma^{\mathrm{s}}_k = \frac{|\tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_k \mathbf{u}_k|^2}{\sum_{j=1}^{K} |\tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_j \mathbf{f}_j|^2 + \sum_{\substack{j=1 \\ j\neq k}}^{K} |\tilde{\mathbf{w}}_k^H \tilde{\mathbf{H}}_j \mathbf{u}_j|^2 + \sigma_2^2 \|\tilde{\mathbf{w}}_k\|_2^2}. \tag{8}$$

Note that the sensing SINR γ^s_k in (8) serves as the foundational metric to evaluate the sensing performance, which monotonically dictates higher-level radar metrics such as detection probability and estimation error bounds. More importantly, in the proposed ISCC framework, the sensing and computation phases are intrinsically coupled. The Q_k bits are assumed to be collected from multimodal sensing units (e.g., radar, LiDAR, and cameras) constitute the input to the subsequent computational tasks, reflecting the general nature of the proposed framework in supporting diverse sensing modalities. Among these, the radar sensing link serves as the primary modality for target detection and localization, and the sensing SINR constraint guarantees the quality of the radar-derived data as a critical component of the Q_k-bit computational input, ensuring the semantic effectiveness of the subsequent computational tasks.

Given a pre-allocated bandwidth B to each UE, the achievable data rate for the communication link from the k-th UE to the AP is given by

$$R_k = B \log_2(1 + \gamma^{\mathrm{c}}_k). \tag{9}$$

B. Computational Task Offloading Model

The computational task can be fully executed locally at the UE. Alternatively, it can be offloaded to the edge computing system deployed at the AP via the communication link [26], [44], [45]. However, the computational task needs to be executed within a latency threshold τ^{th}_k to support mobile scenarios. In this paper, we consider a general adaptive offloading framework for processing the computational task of Q_k bits. Specifically, a fraction $\alpha_k \in [0, 1]$ of the Q_k bits from the k-th UE is offloaded to the AP via the communication link while maintaining the target sensing using the ISAC system. Thus, the remaining $(1 - \alpha_k)Q_k$ bits are processed locally at the UE. Note that the computational task is fully executed locally at the k-th UE when $\alpha_k = 0$, and fully offloaded to the AP when $\alpha_k = 1$, which simplifies to the binary offloading model as investigated in [24], [44], [46]. The relevant latencies are then presented.[3]

[2]Extension to cooperative multi-static sensing with coherent signal combining is left for future work.

[3]The computational resource allocation is centrally managed by the AP through signaling exchange. The duration of this signaling exchange is significantly shorter than the overall communication and computation latency, so it is neglected here for brevity. Moreover, the data size of the computation results is generally much smaller than that of the raw offloaded data, and thus, the associated feedback latency to the UEs is ignored for simplicity [47].

1) Local Computing Latency: Assume that ϖ_k denotes the number of CPU cycles required to process 1-bit data at the k-th UE. Let f_k^{u} (in cycles per second) be the local CPU frequency allocated to the computational task. The local computation latency t_k^{local} is then given by [11]

$$t_k^{\mathrm{local}} = \frac{\varpi_k(1-\alpha_k)Q_k}{f_k^{\mathrm{u}}}, \tag{10}$$

where $f_k^{\mathrm{u}}, \forall k \in \mathcal{K}$ is limited by its maximum computational capability C_k^{u}, i.e., $0 \le f_k^{\mathrm{u}} \le C_k^{\mathrm{u}}$.

2) Task Offloading Latency: The task offloading latency comprises the transmission latency t_k^{com} for data offloading and the processing latency t_k^{proc} at the AP, which is given by

$$t_k^{\mathrm{off}} = \underbrace{\frac{\alpha_k Q_k}{R_k}}_{t_k^{\mathrm{com}}} + \underbrace{\frac{\varpi_k \alpha_k Q_k}{f_k^{\mathrm{ap}}}}_{t_k^{\mathrm{proc}}}, \tag{11}$$

where f_k^{ap} denotes the computational resource allocation at the AP for the k-th UE. The total computational resources at the AP are constrained by $\sum_{k=1}^{K} f_k^{\mathrm{ap}} \le C^{\mathrm{ap}}$, where C^{ap} represents its maximum processing capacity.

Therefore, the total latency for the k-th UE under the adaptive task offloading scheme is given by

$$T_k = \max\{t_k^{\mathrm{local}}, t_k^{\mathrm{off}}\}. \tag{12}$$

C. Energy Consumption Model

The energy consumption mainly consists of communication and computation-related items.[4] The sensing-communication related energy consumption at the k-th UE primarily involves the uplink data offloading to the AP and sensing function consumption, which can be expressed as

$$E_k^{\mathrm{isac}} = P_k t_k^{\mathrm{com}} = P_k \frac{\alpha_k Q_k}{R_k}. \tag{13}$$

The energy consumption of the computational task for each UE k includes the components of local processing at the UE and remote execution at the AP.[5] The local processing related energy computation at the k-th UE is given by

$$E_k^{\mathrm{local}} = \xi_k^{\mathrm{u}} \left(f_k^{\mathrm{u}}\right)^3 t_k^{\mathrm{local}} = \xi_k^{\mathrm{u}} \left(f_k^{\mathrm{u}}\right)^2 \varpi_k(1-\alpha_k)Q_k, \tag{14}$$

where ξ_k^{u} denotes the effective capacitance coefficient of the UE's processor that depends on its CPU architecture [12], [44]. Similarly, the corresponding computation energy consumption at the AP for the k-th UE is then given by

$$E_k^{\mathrm{ap}} = \xi^{\mathrm{ap}} \left(f_k^{\mathrm{ap}}\right)^3 t_k^{p} = \xi^{\mathrm{ap}} \left(f_k^{\mathrm{ap}}\right)^2 \varpi_k \alpha_k Q_k, \tag{15}$$

where ξ^{ap} is the effective capacitance coefficient of the AP's processors. Also, the total local energy consumption at the k-th UE is given by

$$E_k^{\mathrm{u}} = E_k^{\mathrm{isac}} + E_k^{\mathrm{local}}. \tag{16}$$

[4]The energy consumption for downlink transmission of the computation results is relatively small compared to those for data offloading and computation, and thus, is ignored here for brevity [46].

[5]The energy consumption for downlink feedback of the computation results from the AP is ignored here, owing to the small data size and short duration.

III. Optimization Problem Formulation

In practice, UEs exhibit different battery levels and computational capabilities, necessitating adaptive energy allocation strategies that can dynamically prioritize either UE battery conservation or AP resource utilization based on real-time system states. To explicitly capture this heterogeneity, we introduce a user-specific energy cost ratio coefficient $\beta_k, \forall k$, that weights AP energy consumption relative to UE energy in the optimization objective. This parameter enables flexible energy allocation according to real-time system conditions: *(i)* when $\beta_k > 1$, the system prioritizes local computation at the UE to avoid expensive AP energy consumption, which is suitable for scenarios with abundant UE battery or costly/congested AP resources; *(ii)* when $\beta_k < 1$, aggressive computation offloading to the AP is favored to conserve critical UE battery, which is suitable for battery-constrained devices or abundant AP energy capacity; *(iii)* when $\beta_k = 1$, the formulation reduces to conventional total energy minimization without cost preference differentiation between UE and AP energy sources. By tuning β_k according to device battery states, computational capacities, and QoS requirements, system operators gain a practical control mechanism to adapt resource allocation policies. Defining the optimization variables set as $\mathcal{A}_k = \{\alpha_k, \mathbf{F}_k, \mathbf{w}_k, \tilde{\mathbf{w}}_k, \mathbf{R}_k^0, f_k^{\mathrm{ap}}, f_k^{\mathrm{u}}\}$, the total energy cost minimization problem can be formulated as

$$
\begin{aligned}
(\mathbf{P1}) \quad \min_{\mathcal{A}_k} \quad & \sum_{k=1}^{K} E_k^{\mathrm{u}} + \beta_k E_k^{\mathrm{ap}} && \text{(17a)}\\
\textit{s.t.} \quad & \gamma_k^{\mathrm{s}} \ge \gamma_{\mathrm{th}}^{\mathrm{s}}, \ \forall k, && \text{(17b)}\\
& T_k \le \tau_k^{\mathrm{th}}, \ \forall k, && \text{(17c)}\\
& 0 \le f_k^{\mathrm{u}} \le C_k^{\mathrm{u}}, \ \forall k, && \text{(17d)}\\
& \sum_{k=1}^{K} f_k^{\mathrm{AP}} \le C^{\mathrm{AP}}, \ f_k^{\mathrm{AP}} \ge 0, \ \forall k, && \text{(17e)}\\
& 0 \le \alpha_k \le 1, \ \forall k, && \text{(17f)}\\
& \mathrm{Tr}(\mathbf{F}_k + \mathbf{R}_k^0) \le P_{\max}, \forall k, && \text{(17g)}\\
& \mathbf{F}_k \succeq \mathbf{0}, \mathbf{R}_k^0 \succeq \mathbf{0} \ \forall k, && \text{(17h)}\\
& \mathrm{rank}\,(\mathbf{F}_k) = 1, \ \mathrm{rank}\left(\mathbf{R}_k^0\right) = 1, \forall k && \text{(17i)}
\end{aligned}
$$

where (17b) represents the sensing SINR constraint.[6]

Remark *1:* The energy cost ratio coefficient β_k introduced in the objective is designed to capture the inherent asymmetry between UE and AP energy consumption. Unlike existing works that treat all devices uniformly in energy minimization, the proposed formulation allows each UE to have its own β_k value reflecting its individual battery state and computational capacity, which is precisely the essence of energy heterogene-

[6]For target tracking or collaborative sensing purpose, the SINR for sensing link is crucial for guaranteeing sensing performance [48].

ity addressed in this work.[7]

The optimization problem $(\mathbf{P1})$ is nonconvex due to the nonconvex expression of SINR constraints in (17b) and the latency constraint in (17c), which makes it intractable to solve $(\mathbf{P1})$ directly using existing algorithms. To address this issue, we first rewrite the latency constraint in (17c) as $t_k^{\text{local}} \le \tau_k^{\text{th}}$ and $t_k^{\text{off}} \le \tau_k^{\text{th}}$. Then we temporarily relax rank-one constraints in (17i) using a semidefinite relaxation (SDR) strategy. Consequently, the original optimization problem $(\mathbf{P1})$ can be reformulated as

$$
\begin{aligned}
(\mathbf{P2}) \quad & \min_{\mathcal{A}_k} \sum_{k=1}^{K} E_k^{\text{u}} + \beta_k E_k^{\text{ap}} && \text{(18a)}\\
& \textit{s.t.} \quad t_k^{\text{local}} \le \tau_k^{\text{th}} && \text{(18b)}\\
& \qquad t_k^{\text{off}} \le \tau_k^{\text{th}} && \text{(18c)}\\
& \qquad \text{(17b)},\ \text{(17d)} - \text{(17h)}. && \text{(18d)}
\end{aligned}
$$

Note that if an optimal solution to $(\mathbf{P2})$ is available and satisfy $\text{rank}(\mathbf{F}_k) = 1, \text{rank}\left(\mathbf{R}_k^0\right) = 1$, we can state that the achieved optimal solution is also optimal for problem $(\mathbf{P1})$. Otherwise, we need to further construct rank-one approximate solution via some existing techniques, such as Gaussian randomization and rank-one approximation [49]–[52], to finally obtain the optimal solution to problem $(\mathbf{P1})$.

IV. Problem Reformulation for $(\mathbf{P2})$ via SCA

After temporarily relaxing the rank-one constraint (17i), problem $(\mathbf{P2})$ remains nonconvex due to: the fractional structure of the sensing SINR constraint in (17b), and the communication latency term t_k^{com} in (18c), which appears in both the objective function (through E_k^{isac}) and the offloading latency constraint, leading to coupling with the communication rate R_k. To tackle this nonconvexity, we employ SCA strategy [53], [54], which iteratively approximates the nonconvex terms via first-order Taylor expansions around feasible points. At each outer iteration $l \ge 1$, the approximation is refined based on the solution from the previous iteration $(l-1)$, ensuring convergence towards a stationary point. This section presents the key linearization steps required for the SCA framework.

A. Approximation of Sensing Constraint in (17b)

We first rewrite the sensing SINR for the k-th UE in fractional form as $\gamma_k^{\text{s}}\left(\mathcal{U}_k^{\text{s}}\right) = \frac{\chi_k^{\text{s}}(\mathcal{U}_k^{\text{s}})}{g_k^{\text{s}}(\mathcal{U}_k^{\text{s}})}$, where $\mathcal{U}_k^{\text{s}} = \left(\mathbf{F}_k, \tilde{\mathbf{w}}_k, \mathbf{R}_k^0\right)$ denotes the feasible expansion point. The sensing SINR constraint in (17b) can then be rewritten as

$$
\mathcal{F}_k(\mathcal{U}_k^{\text{s}}) = \chi_k^{\text{s}}\left(\mathcal{U}_k^{\text{s}}\right) - \gamma_{\text{th}}^{\text{s}}\, g_k^{\text{s}}\left(\mathcal{U}_k^{\text{s}}\right) \ge 0. \tag{19}
$$

[7]In practice, β_k is configured by the network operator at the beginning of each scheduling slot based on real-time device state feedback. To achieve general state-driven adaptivity, a heuristic mapping strategy can be implemented. For instance, β_k can be dynamically scaled to be proportional to the AP's real-time load penalty and positively correlated with the UE's remaining battery level. Therefore, a UE reporting a critically low battery state is inherently assigned a small β_k (e.g., $0 < \beta_k < 1$) to trigger offloading-favorable resource allocation; conversely, when the AP reports high resource utilization, a large β_k (e.g., $\beta_k > 1$) is assigned to penalize excessive AP-side execution and alleviate network congestion. While this heuristic explicitly links β_k to physical system states rather than manual hyperparameter tuning, tracking its continuous dynamic adjustment over long-term temporal horizons is out of the scope of this paper.

Around a feasible point $\mathcal{U}_k^{\text{s}(l)} = \left(\mathbf{F}_k^{(l)}, \tilde{\mathbf{w}}_k^{(l)}, \mathbf{R}_k^{0(l)}\right)$ at the l-th iteration of the SCA loop, $\mathcal{F}(\mathcal{U}_k^{\text{s}})$ in (19) can be linearized as $\bar{\mathcal{F}}_k\left(\mathcal{U}_k^{\text{s}(l)}\right)$, which is given by

$$
\begin{aligned}
\mathcal{F}_k(\mathcal{U}_k^{\text{s}}) \approx\ & \mathcal{F}_k(\mathcal{U}_k^{\text{s}(l)}) + \text{Tr}\left[\left(\nabla_{\mathbf{F}_k}\mathcal{F}_k|_{\mathcal{U}_k^{\text{s}(l)}}\right)^H\left(\mathbf{F}_k - \mathbf{F}_k^{(l)}\right)\right]\\
& + 2\Re\left\{(\nabla_{\tilde{\mathbf{w}}_k}\mathcal{F}_k|_{\mathcal{U}_k^{\text{s}(l)}})^H(\tilde{\mathbf{w}}_k - \tilde{\mathbf{w}}_k^{(l)})\right\}\\
& + \text{Tr}\left[(\nabla_{\mathbf{R}_k^0}\mathcal{F}_k|_{\mathcal{U}_k^{\text{s}(l)}})^H(\mathbf{R}_k^0 - \mathbf{R}_k^{0(l)})\right]\\
=\ & \bar{\mathcal{F}}_k(\mathcal{U}_k^{\text{s}(l)}).
\end{aligned} \tag{20}
$$

The corresponding gradients with respect to the optimization variables are

$$
\nabla_{\mathbf{F}_k}\mathcal{F}_k = -\gamma_{\text{th}}^{\text{s}}\tilde{\mathbf{H}}_k^H\tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^H\tilde{\mathbf{H}}_k, \tag{21}
$$

$$
\nabla_{\mathbf{R}_k^0}\mathcal{F}_k = \tilde{\mathbf{H}}_k^H\tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^H\tilde{\mathbf{H}}_k, \tag{22}
$$

$$
\begin{aligned}
\nabla_{\tilde{\mathbf{w}}_k}\mathcal{F}_k =\ & \tilde{\mathbf{H}}_k\mathbf{R}_k^0\tilde{\mathbf{H}}_k^H\tilde{\mathbf{w}}_k - \gamma_{\text{th}}^{\text{s}}\sum_{j=1}^{K}\tilde{\mathbf{H}}_j\mathbf{F}_j\tilde{\mathbf{H}}_j^H\tilde{\mathbf{w}}_k\\
& - \gamma_{\text{th}}^{\text{s}}\sum_{j=1,j\neq k}^{K}\tilde{\mathbf{H}}_j\mathbf{R}_j^0\tilde{\mathbf{H}}_j^H\tilde{\mathbf{w}}_k - \gamma_{\text{th}}^{\text{s}}\sigma_2^2\tilde{\mathbf{w}}_k.
\end{aligned} \tag{23}
$$

After the linearization of $\mathcal{F}_k(\mathcal{U}_k^{\text{s}})$, the sensing SINR constraint in (17b) becomes tractable.

B. Approximation of Communication Rate R_k in (9)

The communication rate R_k in (9) is neither convex nor concave due to the fractional structure of the communication SINR γ_k^{c} in (7). To obtain a tractable formulation, we employ the SCA technique to linearize R_k via first-order Taylor expansion around a feasible point $\mathcal{U}_k^{\text{c}} = \left(\mathbf{F}_k, \mathbf{w}_k, \mathbf{R}_k^0\right)$ at each SCA iteration l:

$$
\begin{aligned}
R_k(\mathcal{U}_k^{\text{c}}) \approx\ & R_k(\mathcal{U}_k^{\text{c}(l)}) + \text{Tr}\left[(\nabla_{\mathbf{F}_k}R_k|_{\mathcal{U}_k^{\text{c}(l)}})^H(\mathbf{F}_k - \mathbf{F}_k^{(l)})\right]\\
& + 2\Re\left\{(\nabla_{\mathbf{w}_k}R_k|_{\mathcal{U}_k^{\text{c}(l)}})^H(\mathbf{w}_k - \mathbf{w}_k^{(l)})\right\}\\
& + \text{Tr}\left[(\nabla_{\mathbf{R}_k^0}R_k|_{\mathcal{U}_k^{\text{c}(l)}})^H(\mathbf{R}_k^0 - \mathbf{R}_k^{0(l)})\right]\\
=\ & \bar{R}_k(\mathcal{U}_k^{\text{c}(l)}),
\end{aligned} \tag{24}
$$

where $R_k(\mathcal{U}_k^{\text{c}(l)}) = B\log_2(1+\gamma_k^{\text{c}(l)})$ is the rate evaluated at the feasible point. The gradient with respect to each variable can be expressed in the unified form

$$
\nabla_{\mathcal{V}}R_k = \frac{\partial R_k}{\partial\gamma_k^{\text{c}}}\nabla_{\mathcal{V}}\gamma_k^{\text{c}} = \frac{B}{\ln(2)(1+\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}))}\nabla_{\mathcal{V}}\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}), \tag{25}
$$

where $\mathcal{V} \in \{\mathbf{F}_k, \mathbf{w}_k, \mathbf{R}_k^0\}$. To derive these gradients, we first rewrite the communication SINR in (7) in fractional form as $\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}) = \frac{\chi_k^{\text{c}}(\mathcal{U}_k^{\text{c}})}{g_k^{\text{c}}(\mathcal{U}_k^{\text{c}})}$. The gradients of $\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}})$ with respect to each variable are then given by

$$
\nabla_{\mathbf{F}_k}\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}) = \frac{1}{g_k^{\text{c}}(\mathcal{U}_k^{\text{c}})}\mathbf{H}_k^H\mathbf{w}_k\mathbf{w}_k^H\mathbf{H}_k, \tag{26}
$$

$$
\nabla_{\mathbf{w}_k}\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}) = \frac{g_k^{\text{c}}(\mathcal{U}_k^{\text{c}})\nabla_{\mathbf{w}_k}\chi_k^{\text{c}} - \chi_k^{\text{c}}(\mathcal{U}_k^{\text{c}})\nabla_{\mathbf{w}_k}g_k^{\text{c}}(\mathcal{U}_k^{\text{c}})}{(g_k^{\text{c}}(\mathcal{U}_k^{\text{c}}))^2}, \tag{27}
$$

$$
\nabla_{\mathbf{R}_k^0}\gamma_k^{\text{c}}(\mathcal{U}_k^{\text{c}}) = -\frac{\chi_k^{\text{c}}(\mathcal{U}_k^{\text{c}})}{(g_k^{\text{c}}(\mathcal{U}_k^{\text{c}}))^2}\mathbf{H}_k^H\mathbf{w}_k\mathbf{w}_k^H\mathbf{H}_k, \tag{28}
$$

where

$$\nabla_{\mathbf{w}_k}\chi_k^{\mathrm{c}}(\mathcal{U}_k^{\mathrm{c}}) = \mathbf{H}_k\mathbf{F}_k\mathbf{H}_k^H\mathbf{w}_k, \tag{29}$$

$$\nabla_{\mathbf{w}_k}g_k^{\mathrm{c}}(\mathcal{U}_k^{\mathrm{c}}) = \sum_{j\neq k}^{K}\mathbf{H}_j\mathbf{F}_j\mathbf{H}_j^H\mathbf{w}_k + \sum_{j=1}^{K}\mathbf{H}_j\mathbf{R}_j^0\mathbf{H}_j^H\mathbf{w}_k + \sigma_1^2\mathbf{w}_k. \tag{30}$$

The linearized rate $\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)})$ in (24) enables convex reformulation of the constraints and objective terms involving R_k. Therefore, the communication latency can be rewritten as

$$\bar{t}_k^{\mathrm{com}} = \frac{\alpha_k Q_k}{\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)})}, \tag{31}$$

and the energy consumption $\bar{E}_k^{\mathrm{isac}}$ is then given by

$$\bar{E}_k^{\mathrm{isac}} = P_k\bar{t}_k^{\mathrm{com}} = P_k\frac{\alpha_k Q_k}{\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)})}. \tag{32}$$

C. Problem Reformulation with Auxiliary Variables

After linearizing the nonconvex terms via first-order Taylor expansion in the above subsections, the ISAC energy term $\bar{E}_k^{\mathrm{isac}}$ remains nonconvex due to the coupling variables $\mathbf{F}_k$ and $\mathbf{R}_k^0$ existing in both transmit power P_k and communication latency $\bar{t}_k^{\mathrm{com}}$. To eliminate this coupling, we introduce auxiliary variables $\mathbf{V}_{\mathbf{F}_k} = \mathbf{F}_k$ and $\mathbf{V}_{\mathbf{R}_k^0} = \mathbf{R}_k^0$ in the power term, yielding $P_k = \mathrm{Tr}(\mathbf{V}_{\mathbf{F}_k}+\mathbf{V}_{\mathbf{R}_k^0})$. The energy consumption can then be rewritten as $\tilde{E}_k^{\mathrm{isac}} = \mathrm{Tr}(\mathbf{V}_{\mathbf{F}_k}+\mathbf{V}_{\mathbf{R}_k^0})\bar{t}_k^{\mathrm{com}}$, which decouples the variables in the latency term from those in the power term. Consequently, problem $(\mathbf{P2})$ can be reformulated as the following tractable problem

$$\begin{aligned}
(\mathbf{P3}) \quad \min_{\mathcal{A}_k} \quad & \sum_{k=1}^{K}\tilde{E}_k^{\mathrm{u}} + \beta_k E_k^{\mathrm{ap}} && \text{(33a)}\\
s.t. \quad & \bar{\mathcal{F}}_k\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) \geq 0, \ \forall k, && \text{(33b)}\\
& \frac{\alpha_k Q_k}{\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)})} + \frac{\varpi_k\alpha_k Q_k}{f_k^{\mathrm{ap}}} \leq \tau_k^{\mathrm{th}}, \ \forall k && \text{(33c)}\\
& \mathbf{V}_{\mathbf{F}_k} = \mathbf{F}_k, \ \forall k, && \text{(33d)}\\
& \mathbf{V}_{\mathbf{R}_k^0} = \mathbf{R}_k^0 \ \forall k, && \text{(33e)}\\
& (17\mathrm{d}) - (17\mathrm{h}), \text{ and } (18\mathrm{b}), && \text{(33f)}
\end{aligned}$$

where $\tilde{E}_k^{\mathrm{u}} = \tilde{E}_k^{\mathrm{isac}} + E_k^{\mathrm{local}}$.

The detailed algorithmic framework for solving this problem is presented in the next section.

V. ADMM-based Algorithm Design for $(\mathbf{P3})$

After linearizing the objective and constraints using the SCA technique, the reformulated problem $(\mathbf{P3})$ can be efficiently solved via the ADMM algorithm. This section presents the detailed ADMM framework for problem $(\mathbf{P3})$. The scaled augmented Lagrangian function is defined in (35) shown at the bottom of this page, where $\mathcal{S}_k = \{\varsigma_k^{\mathrm{s}}, \varsigma_k^{\mathrm{off}}, \varsigma_k^{\mathrm{local}}\}$ collects the slack variables for the inequality constraints, $\mathcal{Z}_k = \{\mathbf{Z}_{\mathbf{F}_k}, \mathbf{Z}_{\mathbf{R}_k^0}, z_k^{\mathrm{s}}, z_k^{\mathrm{local}}, z_k^{\mathrm{off}}\}$ collects the corresponding scaled dual variables, and $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_5\}$ denotes the penalty parameters.

A. Update of Variables in $\mathcal{A}_k, \forall k$

The ADMM algorithm solves problem $(\mathbf{P3})$ by iteratively minimizing the augmented Lagrangian function (35) with respect to all variables in $\mathcal{A}_k, \forall k$. At each inner iteration $\iota \leq \hat{L}$, the variables in $\mathcal{A}_k$ are alternatively updated by sequentially solving the following subproblems:

$$\begin{aligned}
\alpha_k^{(\iota+1)} =& \arg\min_{\alpha_k} \tilde{E}_k^{\mathrm{u}(\iota)} + \beta_k E_k^{\mathrm{ap}(\iota)}\\
&+ \frac{\rho_4}{2}(\bar{t}_k^{\mathrm{com}(\iota)} + t_k^{\mathrm{p}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)})^2\\
&+ \frac{\rho_5}{2}\left(t_k^{\mathrm{local}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{local}(\iota)} + z_k^{\mathrm{local}(\iota)}\right)^2, && \text{(36)}\\
f_k^{\mathrm{u}(\iota+1)} =& \arg\min_{f_k^{\mathrm{u}}} \frac{\rho_5}{2}(t_k^{\mathrm{local}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{local}(\iota)} + z_k^{\mathrm{local}(\iota)})^2\\
&+ E_k^{\mathrm{local}(\iota)}, && \text{(37)}\\
f_k^{\mathrm{ap}(\iota+1)} =& \arg\min_{f_k^{\mathrm{u}}} \beta_k E_k^{\mathrm{ap}(\iota)}\\
&+ \frac{\rho_4}{2}(\bar{t}_k^{\mathrm{com}(\iota)} + t_k^{\mathrm{p}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)})^2, && \text{(38)}\\
\mathbf{w}_k^{(\iota+1)} =& \arg\min_{\mathbf{w}_k} \tilde{E}_k^{\mathrm{isac}(\iota)}\\
&+ \frac{\rho_4}{2}(\bar{t}_k^{\mathrm{com}(\iota)} + t_k^{\mathrm{p}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)})^2, && \text{(39)}\\
\tilde{\mathbf{w}}_k^{(\iota+1)} =& \arg\min_{\tilde{\mathbf{w}}_k} \left(-\bar{\mathcal{F}}_k^{(\iota)}\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) + \varsigma_k^{\mathrm{s}(\iota)} + z_k^{\mathrm{s}(\iota)}\right)^2, && \text{(40)}\\
\mathbf{F}_k^{(\iota+1)} =& \arg\min_{\mathbf{F}_k} \tilde{E}_k^{\mathrm{isac}(\iota)} + \frac{\rho_1}{2}\left\|\mathbf{F}_k^{(\iota)} - \mathbf{V}_{\mathbf{F}_k}^{(\iota)} + \mathbf{Z}_{\mathbf{F}_k}^{(\iota)}\right\|_F^2\\
&+ \frac{\rho_4}{2}(\bar{t}_k^{\mathrm{com}(\iota)} + t_k^{\mathrm{p}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)})^2\\
&+ \frac{\rho_3}{2}\left(-\bar{\mathcal{F}}_k^{(\iota)}\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) + \varsigma_k^{\mathrm{s}(\iota)} + z_k^{\mathrm{s}(\iota)}\right)^2, && \text{(41)}\\
\mathbf{R}_k^{0(\iota+1)} =& \arg\min_{\mathbf{R}_k^0} \tilde{E}_k^{\mathrm{isac}(\iota)} + \frac{\rho_2}{2}\left\|\mathbf{R}_k^{0(\iota)} - \mathbf{V}_{\mathbf{R}_k^0}^{(\iota)} + \mathbf{Z}_{\mathbf{R}_k^0}^{(\iota)}\right\|_F^2
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}\left(\mathcal{A}_k, \mathbf{V}_{\mathbf{F}_k}, \mathbf{V}_{\mathbf{R}_k^0}, \mathcal{S}_k, \mathcal{Z}_k, \boldsymbol{\rho}\right) = \sum_{k=1}^{K}\Bigg[& \tilde{E}_k^{\mathrm{u}} + \beta_k E_k^{\mathrm{ap}} + \frac{\rho_1}{2}\left\|\mathbf{F}_k - \mathbf{V}_{\mathbf{F}_k} + \mathbf{Z}_{\mathbf{F}_k}\right\|_F^2 + \frac{\rho_2}{2}\left\|\mathbf{R}_k^0 - \mathbf{V}_{\mathbf{R}_k^0} + \mathbf{Z}_{\mathbf{R}_k^0}\right\|_F^2\\
&+ \frac{\rho_3}{2}\left(-\bar{\mathcal{F}}_k\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) + \varsigma_k^{\mathrm{s}} + z_k^{\mathrm{s}}\right)^2 + \frac{\rho_4}{2}\left(\bar{t}_k^{\mathrm{com}} + t_k^{\mathrm{p}} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}} + z_k^{\mathrm{off}}\right)^2\\
&+ \frac{\rho_5}{2}\left(t_k^{\mathrm{local}} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{local}} + z_k^{\mathrm{local}}\right)^2\Bigg]. \qquad \text{(35)}
\end{aligned}$$

$$\alpha_k^{(\iota+1)} = \frac{\xi_k^{\mathrm{u}} f_k^{\mathrm{u}(\iota)} \varpi_k Q_k - \frac{P_k Q_k}{\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)})} - \beta_k \xi_k^{\mathrm{ap}} f_k^{\mathrm{ap}(\iota)} \varpi_k Q_k - \rho_4 A_k c_{k,1} + \rho_5 D_k^2 + \rho_5 D_k c_{k,2}}{\rho_4 A_k^2 + \rho_5 D_k^2}. \tag{43}$$

$$+ \frac{\rho_4}{2}\left(\bar{t}_k^{\mathrm{com}(\iota)} + t_k^{\mathrm{p}(\iota)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)}\right)^2$$
$$+ \frac{\rho_3}{2}\left(-\bar{\mathcal{F}}_k^{(\iota)}\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) + \varsigma_k^{\mathrm{s}(\iota)} + z_k^{\mathrm{s}(\iota)}\right)^2. \tag{42}$$

Setting the gradient of the objective in (36) with respect to α_k to zero yields a closed-form solution given by (43) shown in the next page, where $A_k = Q_k/\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)}) + \varpi_k Q_k / f_k^{\mathrm{ap}(\iota)}$, $D_k = \varpi_k Q_k / f_k^{\mathrm{u}(\iota)}$, $c_{k,1} = -\tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)} + z_k^{\mathrm{off}(\iota)}$, and $c_{k,2} = -\tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{local}(\iota)} + z_k^{\mathrm{local}(\iota)}$. The resulting $\alpha_k^{(\iota+1)}$ is then projected onto the feasible set

$$\alpha_k^{(\iota+1)} \leftarrow \max\left(0,\ \min(1, \alpha_k^{(\iota+1)})\right),\ \forall k. \tag{44}$$

Similarly, the computational resource variables f_k^{u} and f_k^{ap} can be updated by setting the corresponding derivatives to zero and solving the following quartic equations

$$2\xi_k^{\mathrm{u}} c_{k,3} (f_k^{\mathrm{u}})^4 - \rho_5 c_{k,2} c_{k,3} f_k^{\mathrm{u}} - \rho_5 c_{k,3}^2 = 0, \tag{45}$$
$$2\beta_k \xi_k^{\mathrm{ap}} c_{k,4} (f_k^{\mathrm{ap}})^4 - \rho_4 c_{k,4} c_{k,5} f_k^{\mathrm{ap}} - \rho_4 c_{k,4}^2 = 0, \tag{46}$$

where $c_{k,3} = \varpi_k(1 - \alpha_k^{(\iota+1)}) Q_k$, $c_{k,4} = \varpi_k \alpha_k^{(\iota+1)} Q_k$ and $c_{k,5} = \bar{t}_k^{\mathrm{com}(\iota)} + c_{k,1}$. Then, the updates of $f_k^{\mathrm{u}(\iota+1)}$ and $f_k^{\mathrm{ap}(\iota+1)}$ are obtained by projecting on the constraint (17d) and (17e), respectively.

For each variable $\mathbf{T}_k \in \{\mathbf{w}_k, \tilde{\mathbf{w}}_k, \mathbf{F}_k, \mathbf{R}_k^0\}_k, \forall k$, the update is achieved via gradient descent with the unified form

$$\mathbf{T}_k^{(\iota+1)} = \mathbf{T}_k^{(\iota)} - \eta_{\mathbf{T}_k} \nabla_{\mathbf{T}_k} \bar{\mathcal{L}}\left(\mathbf{T}_k^{(\iota)}\right), \tag{47}$$

where $\nabla_{\mathbf{T}_k} \bar{\mathcal{L}}(\mathbf{T}_k^{(\iota)})$ represents the gradient of the objective function with respect to $\mathbf{T}_k$, as expressed in (39)-(42), and $\eta_{\mathbf{T}_k}$ is the corresponding step size. After each gradient descent step, the updated variable is projected onto the constraint set $\mathcal{C}_{\mathbf{T}_k}$ via $\mathbf{T}_k^{(\iota+1)} \leftarrow \mathcal{P}_{\mathcal{C}_{\mathbf{T}_k}}(\mathbf{T}_k^{(\iota+1)})$ to ensure feasibility. An example for updating $\mathbf{T}_k = \mathbf{F}_k$ is provided in Appendix A.

B. Update of Auxiliary Variables $\mathbf{V}_{\mathbf{F}_k}$ and $\mathbf{V}_{\mathbf{R}_k^0}$

After updating all the primal variables, the updates of the defined auxiliary variables are given by sequentially solving the following optimization problems:

$$\mathbf{V}_{\mathbf{F}_k}^{(\iota+1)} = \arg\min_{\mathbf{V}_{\mathbf{F}_k}} \tilde{E}_k^{\mathrm{isac}(\iota)} + \frac{\rho_1}{2}\left\|\mathbf{F}_k^{(\iota+1)} - \mathbf{V}_{\mathbf{F}_k}^{(\iota)} + \mathbf{Z}_{\mathbf{F}_k}^{(\iota)}\right\|_F^2, \tag{48}$$

$$\mathbf{V}_{\mathbf{R}_k^0}^{(\iota+1)} = \arg\min_{\mathbf{V}_{\mathbf{R}_k^0}} \tilde{E}_k^{\mathrm{isac}(\iota)} + \frac{\rho_2}{2}\left\|\mathbf{R}_k^{0(\iota+1)} - \mathbf{V}_{\mathbf{R}_k^0}^{(\iota)} + \mathbf{Z}_{\mathbf{R}_k^0}^{(\iota)}\right\|_F^2. \tag{49}$$

Similar to the updates of $\mathbf{T}_k, \forall k$, each auxiliary variable is updated via gradient descent followed by projection onto its feasible set. The detailed steps are omitted for brevity.

C. Update of Scaled Dual Variables and Slack Variables

Given the updated primal variables and auxiliary variables, the scaled dual variables in $\mathcal{Z}_k, \forall k$ are updated by

$$\mathbf{Z}_{\mathbf{F}_k}^{(\iota+1)} = \mathbf{Z}_{\mathbf{F}_k}^{(\iota)} + \mathbf{F}_k^{(\iota+1)} - \mathbf{V}_{\mathbf{F}_k}^{(\iota+1)}, \tag{50}$$
$$\mathbf{Z}_{\mathbf{R}_k^0}^{(\iota+1)} = \mathbf{Z}_{\mathbf{R}_k^0}^{(\iota)} + \mathbf{R}_k^{0(\iota+1)} - \mathbf{V}_{\mathbf{R}_k^0}^{(\iota+1)}, \tag{51}$$
$$z_k^{\mathrm{s}(\iota+1)} = z_k^{\mathrm{s}(\iota)} - \bar{\mathcal{F}}_k^{(\iota+1)}\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) + \varsigma_k^{\mathrm{s}(\iota)}, \tag{52}$$
$$z_k^{\mathrm{off}(\iota+1)} = z_k^{\mathrm{off}(\iota)} + \bar{t}_k^{\mathrm{com}(\iota+1)} + t_k^{\mathrm{p}(\iota+1)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{off}(\iota)}, \tag{53}$$
$$z_k^{\mathrm{local}(\iota+1)} = z_k^{\mathrm{local}(\iota)} + t_k^{\mathrm{local}(\iota+1)} - \tau_k^{\mathrm{th}} + \varsigma_k^{\mathrm{local}(\iota+1)}. \tag{54}$$

The slack variables in $\mathcal{S}_k, \forall k$ are updated by

$$\varsigma_k^{\mathrm{s}(\iota+1)} = \max\left(0,\ \bar{\mathcal{F}}_k^{(\iota+1)}\left(\mathcal{U}_k^{\mathrm{s}(l)}\right) - z_k^{\mathrm{s}(\iota+1)}\right), \tag{55}$$
$$\varsigma_k^{\mathrm{off}(\iota+1)} = \max\left(0, \tau_k^{\mathrm{th}} - \bar{t}_k^{\mathrm{com}(\iota+1)} - t_k^{\mathrm{p}(\iota+1)} - z_k^{\mathrm{off}(\iota+1)}\right), \tag{56}$$
$$\varsigma_k^{\mathrm{local}(\iota+1)} = \max\left(0,\ \tau_k^{\mathrm{th}} - t_k^{\mathrm{local}(\iota+1)} - z_k^{\mathrm{local}(\iota+1)}\right). \tag{57}$$

D. Overall Algorithm Summary and Complexity Analysis

1) Convergence Analysis: The proposed double-loop SCA-ADMM algorithm iteratively solves the nonconvex problem $(\mathbf{P1})$ with convergence guarantees. **Outer SCA loop:** At each SCA iteration $l \leq L$, the first-order Taylor expansion around the feasible point transforms the nonconvex sensing SINR constraint into a convex one. Similarly, linearization of the achievable data rate R_k yields a convex upper bound on the communication latency t_k^{com}, and thus, a convex approximation of the original nonconvex objective. The introduction of auxiliary variables $\mathbf{V}_{\mathbf{F}_k}, \mathbf{V}_{\mathbf{R}_k^0}$ further decouples the bilinear energy consumption term, ensuring that the reformulated problem $(\mathbf{P3})$ is tractable at each SCA iteration. By standard SCA convergence theory [53], the convergence of the objective of $(\mathbf{P1})$ is guaranteed. **Inner ADMM loop:** For the l-th SCA iteration, problem $(\mathbf{P3})$ is solved via ADMM. The ADMM convergence [55] guarantees that both primal and dual residuals vanish as the iteration count $\iota \leq \hat{L}$ increases, provided the penalty parameters $\boldsymbol{\rho}$ are properly chosen. Once the ADMM residuals fall below a predefined tolerance threshold, the l-th SCA iteration converges to a stationary point. **Overall procedure:** The optimization variables are set at iteration $(l+1)$ using the updates from iteration l, accelerating convergence of the outer SCA loop. The algorithm terminates when either the relative objective change falls below ϵ_{SCA} or the maximum iteration count L is reached. The complete procedures are presented in **Algorithm 1** (i.e., inner ADMM loop) and **Algorithm 2** (i.e., overall SCA-ADMM framework). **Rank-one recovery:** Although the SDR temporarily removes the rank-one constraint (17i), *Proposition 1* establishes that

Algorithm 1: ADMM algorithm for solving $(\mathbf{P3})$

Input: $\{\mathcal{A}_k^{(0)}, \mathbf{V}_{\mathbf{F}_k}^{(0)}, \mathbf{V}_{\mathbf{R}_k^0}^{(0)}, \mathcal{S}_k^{(0)}, \mathcal{Z}_k^{(0)}\}_k, \forall k, \boldsymbol{\rho}, \hat{L}$.
Output: $\{\mathcal{A}_k^*, \mathbf{V}_{\mathbf{F}_k}^*, \mathbf{V}_{\mathbf{R}_k^0}^*, \mathbf{U}_{\mathbf{R}_k^0}^*, \mathcal{S}_k^*, \mathcal{Z}_k^*\}_k, \forall k$.

$\iota \leftarrow 0$;
while $\iota < \hat{L}$ **do**
  for $k = 1$ **to** K **do**
    Sequentially update the parameters:
    Update primal variables in $\mathcal{A}_k, \forall k$ by solving (36)-(42), via closed-forms or PGD approach;
    Update auxiliary variables $\mathbf{V}_{\mathbf{F}_k}$ and $\mathbf{V}_{\mathbf{R}_k^0}$ by solving (48)-(49) via PGD;
    Update dual variables according to (50)-(54);
    Update slack variables according to (55)-(57);
  Check convergence criteria:
  Compute objective value, primal and dual residuals.
  if *convergence criteria is met* **then**
    Output the optimal $\{\mathcal{A}_k^{*(l)}, \mathbf{V}_{\mathbf{F}_k}^{*(l)}, \mathbf{V}_{\mathbf{R}_k^0}^{*(l)}, \mathbf{U}_{\mathbf{R}_k^0}^{*(l)}, \mathcal{S}_k^{*(l)}, \mathcal{Z}_k^{*(l)}\}_k$;
    break;
  $\iota \leftarrow \iota + 1$;
l-th SCA iteration is completed.

Algorithm 2: Overall SCA-ADMM algorithm framework for solving $(\mathbf{P1})$

Input: Formulate problem $(\mathbf{P1})$ and define variables $\mathcal{A}_k = \{\alpha_k, \mathbf{f}_k, \mathbf{w}_k, \tilde{\mathbf{w}}_k, \mathbf{u}_k^0, f_k^{\mathrm{ap}}, f_k^{\mathrm{u}}\}, \forall k$.
Output: Optimal $\{\alpha_k^*, \mathbf{f}_k^*, \mathbf{w}_k^*, \tilde{\mathbf{w}}_k^*, \mathbf{u}_k^*, f_k^{\mathrm{ap}*}, f_k^{\mathrm{u}*}\}, \forall k$.
Reformulation of $(\mathbf{P1})$ to $(\mathbf{P2})$ using SDR for (17i);
Implement the SCA approach:
Initialize $\mathcal{A}_k^{(0)}, \forall k$, maximum SCA iterations L.
$l \leftarrow 0$;
while $l < L$ **do**
  Linearize $\mathcal{F}_k(\mathcal{U}_k^{\mathrm{s}})$ in (20) around the feasible point $\mathcal{U}_k^{\mathrm{s}(l)} = \left(\mathbf{F}_k^{(l)}, \tilde{\mathbf{w}}_k^{(l)}, \mathbf{R}_k^{0(l)}\right)$ via Taylor expansion;
  Calculate $\bar{t}_k^{\mathrm{com}(l)}$ in (31) by linearizing R_k in (24);
  Introduce auxiliary variables $\mathbf{V}_{\mathbf{F}_k} = \mathbf{F}_k$ and $\mathbf{V}_{\mathbf{R}_k^0} = \mathbf{R}_k^0$ and rewrite $\bar{E}_k^{\mathrm{com}}$ as $\tilde{E}_k^{\mathrm{isac}}$;
  Problem reformulation from $(\mathbf{P2})$ to $(\mathbf{P3})$ is done;
  Implement the ADMM approach presented in Algorithm 1 for solving $(\mathbf{P3})$;
  Check convergence criteria of the SCA iterations:
  if *convergence criteria of SCA iteration met* **then**
    Output $\{\alpha_k^*, \mathbf{F}_k^*, \mathbf{w}_k^*, \tilde{\mathbf{w}}_k^*, \mathbf{R}_k^{0*}, f_k^{\mathrm{ap}*}, f_k^{\mathrm{u}*}\}, \forall k$;
    break;
  $l \leftarrow l + 1$;
for $k = 1$ **to** K **do**
  Obtain $\mathbf{f}_k^*$ from $\mathbf{F}_k^*, \forall k$; Obtain $\mathbf{u}_k^*$ from $\mathbf{R}_k^{0*}, \forall k$;

the optimal solution $\mathbf{F}_k^*, \forall k$ to the relaxed problem naturally satisfies the rank-one constraint, eliminating the need for costly randomization-based approximations [49]. The beamforming vector is directly recovered via eigenvalue decomposition: $\mathbf{f}_k^* = \sqrt{\lambda_{\max}}\mathbf{v}_{\max}$ [56], where $\lambda_{\max}$ and $\mathbf{v}_{\max}$ are the largest eigenvalue and corresponding eigenvector of $\mathbf{F}_k^*$, respectively. Followed by the similar procedure, we recover the rank-one solution for $\mathbf{u}_k, \forall k$.

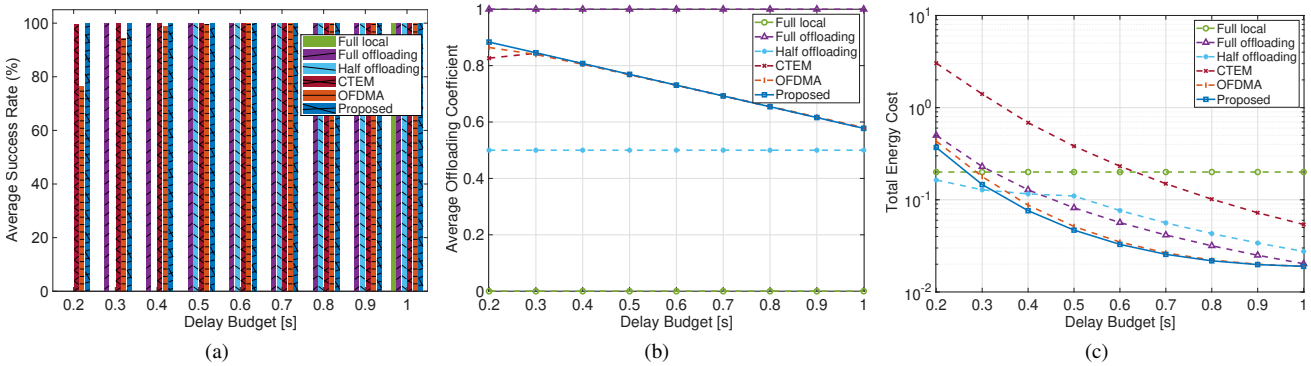
2) Computational Complexity Analysis: The computational complexity per ADMM iteration is dominated by updating the combining vectors $\mathbf{w}_k$ and $\tilde{\mathbf{w}}_k$ across all K UEs. Computing the gradient $\nabla_{\mathbf{w}_k}\gamma_k^{\mathrm{c}}$ requires evaluating interference contributions from all other UEs, necessitating $\mathcal{O}(K)$ matrix-vector products per UE, which yields $\mathcal{O}(K^2(M_{\mathrm{a}}(M_{\mathrm{u}}^{\mathrm{t}})^2 + M_{\mathrm{a}}^2 M_{\mathrm{u}}^{\mathrm{t}}))$ for all K UEs. Similarly, updating $\tilde{\mathbf{w}}_k$ requires $\mathcal{O}(K^2(M_{\mathrm{u}}^{\mathrm{r}}(M_{\mathrm{u}}^{\mathrm{t}})^2 + (M_{\mathrm{u}}^{\mathrm{r}})^2 M_{\mathrm{u}}^{\mathrm{t}}))$. In contrast, the beamforming covariance matrices $\mathbf{F}_k$ and $\mathbf{R}_k^0$ are updated via projected gradient descent with PSD projection but without eigenvalue decomposition during iterations, requiring $\mathcal{O}(K(M_{\mathrm{a}}M_{\mathrm{u}}^{\mathrm{t}} + M_{\mathrm{u}}^{\mathrm{r}}M_{\mathrm{u}}^{\mathrm{t}} + (M_{\mathrm{u}}^{\mathrm{t}})^2))$ for all K UEs. Moreover, each of the auxiliary variables $\mathbf{V}_{\mathbf{F}_k}$ and $\mathbf{V}_{\mathbf{R}_k^0}$ requires $\mathcal{O}(K(M_{\mathrm{u}}^{\mathrm{t}})^2)$. The dual variable updates for matrix variables $\mathbf{Z}_{\mathbf{F}_k}$ and $\mathbf{Z}_{\mathbf{R}_k^0}$ involve matrix additions with complexity $\mathcal{O}(K(M_{\mathrm{u}}^{\mathrm{t}})^2)$, while scalar dual variables, slack variables, and computational resource variables require negligible $\mathcal{O}(K)$ operations. Consolidating these contributions under typical system configurations where $M_{\mathrm{a}} \geq \max\{M_{\mathrm{u}}^{\mathrm{r}}, M_{\mathrm{u}}^{\mathrm{t}}\}$, the per-ADMM-iteration computational complexity simplifies to $\mathcal{O}(K^2 M_{\mathrm{a}}^2 M_{\mathrm{u}}^{\mathrm{t}})$. With a maximum of L outer SCA iterations and $\hat{L}$ inner ADMM iterations per SCA loop, the total iteration complexity is $\mathcal{O}(L\hat{L}K^2 M_{\mathrm{a}}^2 M_{\mathrm{u}}^{\mathrm{t}})$. Once the SCA-ADMM algorithm converged, the rank-one beamforming vectors $\mathbf{f}_k^*$ and $\mathbf{u}_k^*$ are recovered via EVD on $\mathbf{F}_k^*$ and $\mathbf{R}_k^{0*}$, respectively. This requires a complexity of $\mathcal{O}(K(M_{\mathrm{u}}^{\mathrm{t}})^3)$, which is negligible compared to the iteration cost for $L\hat{L} \gg 1$. Therefore, the overall computational complexity of the proposed SCA-ADMM algorithm is $\mathcal{O}(L\hat{L}K^2 M_{\mathrm{a}}^2 M_{\mathrm{u}}^{\mathrm{t}})$.

VI. Numerical Simulations

In this section, we numerically evaluate the performance of the proposed SCA-ADMM algorithm for energy-efficient ISCC systems. The simulation parameters are listed in Table I and serve as default values throughout this section unless otherwise noted. These simulation parameters are selected to represent a practical mmWave-enabled IoT/vehicular ISCC deployment scenario, where battery-constrained devices (e.g., vehicles or robots) can communicate with a nearby MEC-enabled AP. Notably, the uniform setting $\beta_k = \beta, \forall k$ is adopted in the default simulations to isolate the effect of the energy cost ratio on system behavior in a controlled manner, without conflating it with inter-user heterogeneity. Per-user heterogeneous β_k settings are evaluated separately after these comprehensive evaluation. To comprehensively evaluate the proposed adaptive offloading framework, we compare it with five fixed offloading baselines:

(i) **Full local (FL)** ($\alpha_k = 0, \forall k$), i.e., computational tasks are executed locally at the UEs, with no offloading to the AP;

(ii) **Full offloading (FO)** ($\alpha_k = 1, \forall k$), i.e., all computational tasks are offloaded to the AP, with no local processing;



(a) (b) (c)

Fig. 2: Performance evaluation constrained by the latency budget.

(iii) **Half offloading (HO)** ($\alpha_k = 0.5, \forall k$), i.e., exactly half of each computational task is offloaded to the AP, with the remainder processed locally;

(iv) **Conventional total energy minimization (CTEM)** aims to minimize the absolute unweighted sum of energy consumption by permanently setting $\beta_k = 1, \forall k$ (ignores the energy heterogeneity between UEs and the AP), and is solved using the proposed SCA-ADMM framework to serve as an unweighted performance reference;

(v) **OFDMA**-based adaptive Offloading where the total bandwidth B is allocated to K UEs in direct proportion to their pathloss coefficients, i.e., UEs with worse channel conditions receive more bandwidth to compensate for their link quality, and the proposed SCA-ADMM framework is adopted.

The computational complexity of the proposed scheme and all baseline schemes is summarized in Table II. The CTEM, OFDMA, FO, and HO baselines share the same dominant complexity as the proposed scheme, since they all employ the SCA-ADMM framework for joint beamforming and resource allocation optimization. The FL baseline incurs lower complexity by eliminating communication-related optimization variables, but at the cost of significantly degraded performance under heavy computational workloads and constrained resources.

Also, we define a performance evaluation indicator, average success rate (ASR), which is given by

$$\text{ASR} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K}\mathbb{I}_{k,i}}{IK} \times 100\%, \tag{58}$$

where the indicator function $\mathbb{I}_{k,i} = 1$ if the k-th UE in trial i satisfies all constraints, and $\mathbb{I}_{k,i} = 0$ otherwise.

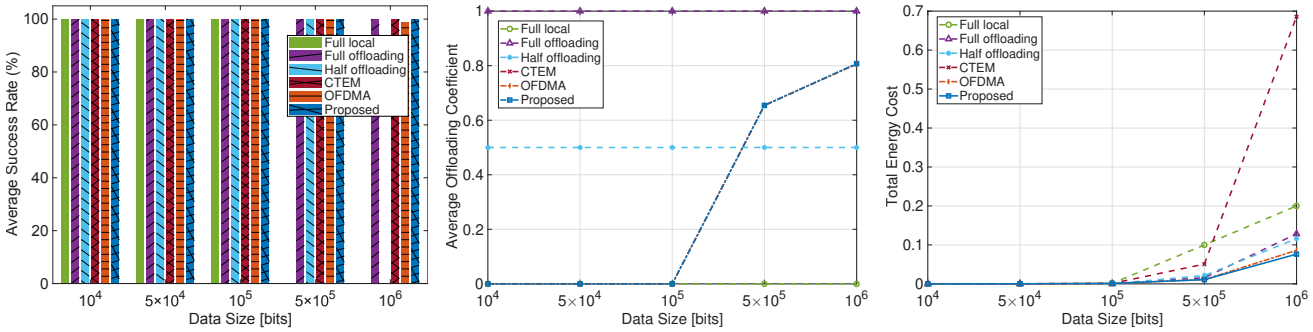
Latency budget is a critical constraint for time-sensitive mobile applications. Fig. 2 illustrates the performance under varying latency budget $\tau_k^{\text{th}} = \tau^{\text{th}}, \forall k$ from 0.2 s to 1.0 s, with computational workload $Q = 10^6, \forall k$ bits and energy cost ratio $\beta = 0.1$. The FL scheme fails to achieve 100% ASR until $\tau^{\text{th}} = 1.0$ s due to limited local computational capacity, as shown in Fig. 2a. Similarly, the FO scheme requires at least $\tau^{\text{th}} = 0.3$ s to reach 100% ASR. The HO baseline achieves 100% ASR at $\tau^{\text{th}} = 0.5$ s. The OFDMA scheme requires $\tau^{\text{th}} \geq 0.6$ s to reach 100% ASR due to the reduced per-UE bandwidth, while the CTEM scheme achieves comparable ASR to the proposed scheme since it uses the same SCA-ADMM framework by setting $\beta_k = 1$. In contrast, the proposed scheme achieves 100% ASR across the entire range $\tau^{\text{th}} \geq 0.2$ s, demonstrating superior robustness under tight latency constraints through adaptive resource allocation and adaptive offloading coefficient adjustment. As the latency budget relaxes from 0.2 s to 1.0 s, the proposed scheme intelligently reduces the average offloading coefficient from 0.88 to 0.58, as illustrated in Fig. 2b. This adaptation reflects the algorithm's ability to shift computation from the AP (expensive due to $\beta < 1$, favoring AP usage) back to local processing as time constraints become less stringent, thereby minimizing total energy cost. The CTEM and OFDMA schemes exhibit similar offloading coefficient trends to the proposed scheme for $\tau^{\text{th}} \geq 0.3$ s, since all three share the same adaptive SCA-ADMM optimization framework. Fig. 2c shows that the proposed scheme consistently achieves the lowest total energy cost among all baselines, reducing from 0.369 to 0.019 as τ^{th} increases from 0.2 s to 1.0 s, demonstrating its adap-

TABLE I: Simulation Parameters

Parameter	Value
Carrier frequency f_{c}	28 GHz
Antenna elements M_{a}, $M_{\text{u}}^{\text{t}}, M_{\text{u}}^{\text{r}}$	64, 64, 64
Coverage radius r_0	100 m
Bandwidth B	20 MHz
Max transmit power $P_{\max}$	10 dBm
Noise power σ_1^2, σ_2^2	-85 dBm
Number of UEs K	2 (default)
Computational workload $Q_k = Q, \forall k$	10^6 bits
UE capacity $C_k^{\text{u}}, \forall k$	1 GHz
AP capacity C^{ap}	10 GHz
Energy coefficient $\xi_k^{\text{u}} = \xi^{\text{ap}}, \forall k$	10^{-28}
CPU cycles per bit $\varpi_k, \forall k$	1,000 cycles/bit
Energy cost ratio $\beta_k = \beta, \forall k$	0.1
Sensing SINR threshold $\gamma_{\text{th}}^{\text{s}}$	0 dB

TABLE II: Computational Complexity Comparison

Scheme	Total Complexity
Proposed	$\mathcal{O}(L\hat{L}K^2M_{\text{a}}^2M_{\text{u}}^{\text{t}})$
CTEM	$\mathcal{O}(L\hat{L}K^2M_{\text{a}}^2M_{\text{u}}^{\text{t}})$
OFDMA	$\mathcal{O}(L\hat{L}K^2M_{\text{a}}^2M_{\text{u}}^{\text{t}})$
FO	$\mathcal{O}(L\hat{L}K^2M_{\text{a}}^2M_{\text{u}}^{\text{t}})$
HO	$\mathcal{O}(L\hat{L}K^2M_{\text{a}}^2M_{\text{u}}^{\text{t}})$
FL	$\mathcal{O}(L\hat{L}K^2M_{\text{r}}^{\text{u}}M_{\text{t}}^{\text{u}}(M_{\text{r}}^{\text{u}}+M_{\text{t}}^{\text{u}}))$



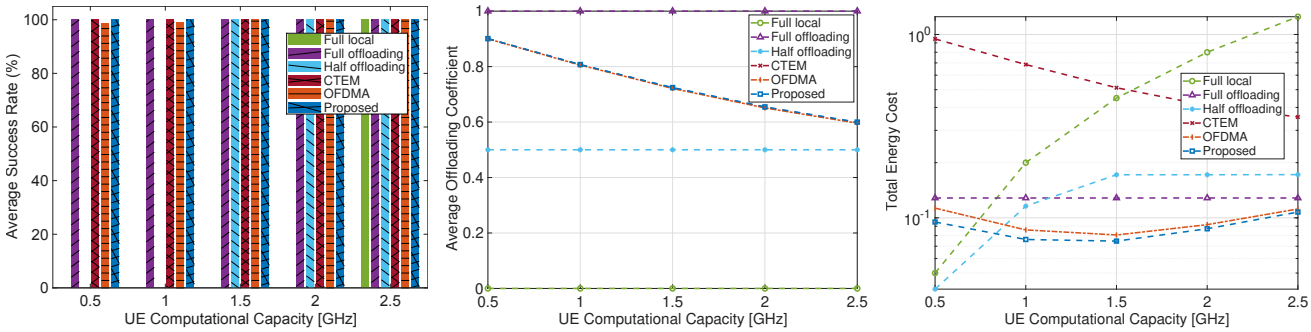
(a) (b) (c)

Fig. 3: Performance evaluation versus the data size Q given $\tau^{\text{th}} = 0.4$ s.

tation to relaxed constraints. Moreover, the proposed scheme achieves approximately 25%-40% energy cost reduction across $\tau^{\text{th}} \in [0.2, 0.9]$ s compared to the FO scheme, and up to 57% savings over HO for $\tau^{\text{th}} = 0.5$ s. Notably, the CTEM scheme incurs significantly higher energy cost (e.g., up to 9.6 times at $\tau^{\text{th}} = 0.3$ s) since it ignores energy heterogeneity and fails to optimally balance UE and AP energy consumption. The OFDMA scheme exhibits higher energy cost than the proposed scheme under tight latency constraints, since the reduced per-UE bandwidth increases transmission latency and forces suboptimal offloading decisions. As the latency budget relaxes, this gap gradually diminishes, confirming that the performance advantage of the proposed full-bandwidth sharing scheme is most pronounced under stringent latency requirements. These results validate the effectiveness of the proposed scheme in jointly optimizing communication, computation, and sensing resources to minimize energy cost while guaranteeing stringent QoS requirements. The ASR curves in Fig. 2a exhibit sharp transitions for the fixed offloading schemes rather than smooth increases, as the algorithm prioritizes energy cost minimization subject to constraint satisfaction, not ASR maximization. When constraints become feasible, multiple UEs may simultaneously transition from infeasible to feasible, causing abrupt ASR jumps. Enhancing the ASR is left for future work. Furthermore, to explicitly validate the practical feasibility of the proposed framework for time-sensitive applications, we evaluated its average execution time for the stringent delay scenario presented in Fig. 2. On a standard simulation platform equipped with an Intel Core Ultra 5 125U processor (1.30 GHz) and 16 GB of RAM running MATLAB R2025b, the average total convergence time of the proposed SCA-ADMM algorithm per scheduling snapshot is approximately 75.07 ms. This average execution time is safely and significantly below the most stringent delay budget ($\tau = 0.2$ s) evaluated in our system. Moreover, since the proposed ADMM framework intrinsically facilitates parallel processing across multiple subproblems, deploying the algorithm on practical APs with dedicated hardware accelerators (e.g., GPUs and FPGAs) would further reduce the actual execution time by orders of magnitude, making it feasible for real-time ISCC deployments.

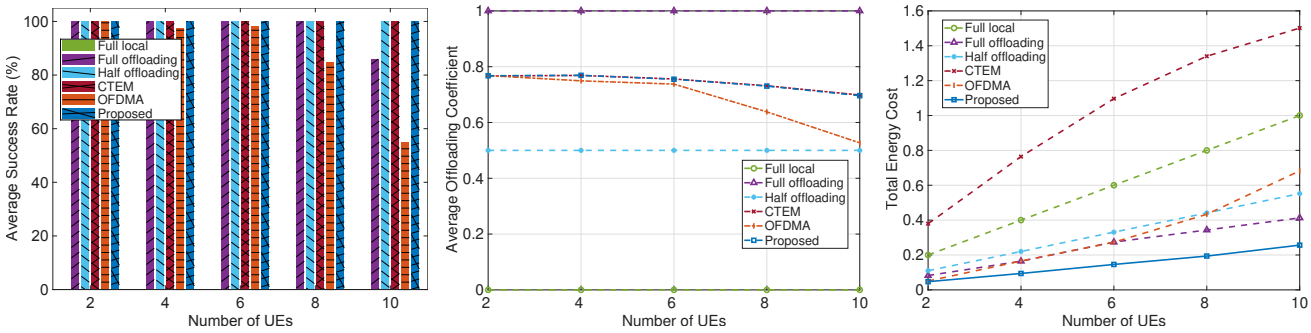
Fig. 3 investigates the system performance versus data size Q with $\tau^{\text{th}} = 0.4$ s. All schemes achieve a 100% success rate for $Q \leq 10^5$ bits, as shown in Fig. 3a. At $Q = 10^6$ bits, the FL scheme fails to satisfy the latency constraint due to limited local computational capacity, while the OFDMA scheme achieves approximately 99% ASR due to its reduced per-UE bandwidth. In contrast, the proposed scheme, and CTEM maintain 100% ASR, with the proposed scheme benefiting from its adaptive offloading strategy. As illustrated in Fig. 3b, the proposed scheme and CTEM exhibit similar offloading behavior, keeping a near-zero offloading coefficient for small data sizes and increasing to approximately 0.81 at $Q = 10^6$ bits to meet the latency constraint. Fig. 3c demonstrates that the proposed scheme consistently achieves the lowest energy cost across all data sizes, while the CTEM scheme incurs dramatically higher energy cost at large data sizes, directly validating the advantage of the energy-heterogeneity-aware formulation over conventional total energy minimization.

When increasing the local computational capacity to 2 GHz, all the schemes achieve a success rate of 100%, as depicted in Fig. 4a. This indicates that fully local processing becomes feasible when sufficient computational resources are available. In contrast, even when subjected to a severely limited computational capacity of 0.5 GHz, our proposed scheme and the CTEM baseline can still maintain a 100% success rate. The OFDMA baseline, however, experiences a slight degradation with a 98.5% success rate under the same condition, as the orthogonal spectrum allocation limits the data transmission rate. This demonstrates that our proposed scheme guarantees ultra-high reliability by intelligently exploiting the powerful computational capacity at the AP via dynamically adjusted offloading coefficients and full-bandwidth sharing, as illustrated in Fig. 4b. The offloading coefficients of the proposed scheme, CTEM, and OFDMA decrease monotonically from approximately 0.9 at 0.5 GHz to 0.6 at 2.5 GHz. This reflects an intelligent shift from AP-centric (i.e., battery-critical mode) to UE-centric (i.e., AP-constrained mode) processing as local computational capacity improves, a dynamic behavior that starkly contrasts with the baseline schemes maintaining fixed offloading strategies (i.e., FO, FL, and HO) regardless of UE capabilities. Most importantly, the total energy cost trends in Fig. 4c explicitly demonstrate multiple advantages of our proposed scheme over conventional baselines. First, while the



(a) (b) (c)

Fig. 4: System performance varying with the maximum UE computational capacity given $\tau^{\mathrm{th}} = 0.4$s.



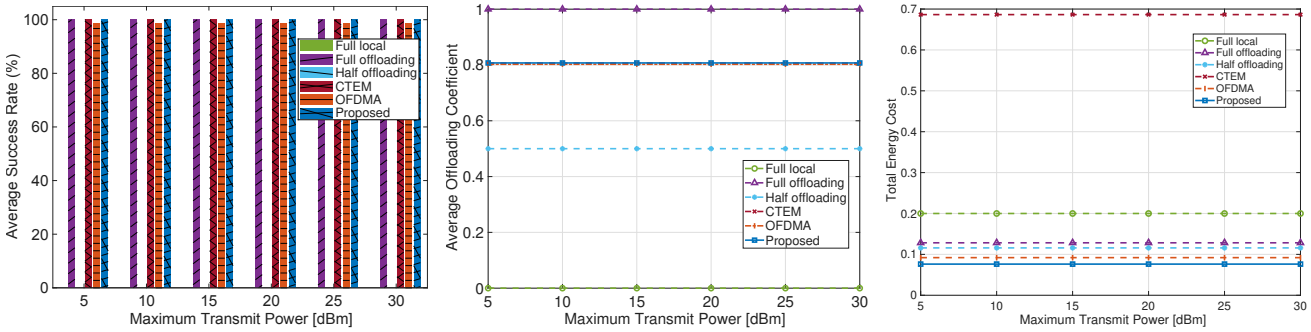
(a) (b) (c)

Fig. 5: System performance affected by the number of UEs given $\tau^{\mathrm{th}} = 0.5$s, $C^{\mathrm{ap}} = 20$ GHz.

advanced CTEM scheme adjusts offloading dynamically, it incurs a substantially higher total energy cost evaluated under the heterogeneous framework. This is because CTEM ($\beta_k = 1$) blindly minimizes the absolute energy consumption without differentiating the energy sources, imposing an excessive energy burden on battery-constrained UEs. Second, our proposed scheme consistently outperforms the OFDMA baseline across all capacity levels. This directly quantifies the significant energy savings achieved by our full-bandwidth sharing and joint beamforming design over orthogonal spectrum allocation. Finally, compared to the fixed FO and HO schemes that lack flexibility, the proposed scheme maintains a remarkably low and stable energy cost profile. These results comprehensively demonstrate that the proposed scheme intelligently adapts to heterogeneous UE computational capabilities and explicitly protects battery-limited devices, achieving superior resource allocation that minimizes the weighted energy cost while guaranteeing QoS requirements.

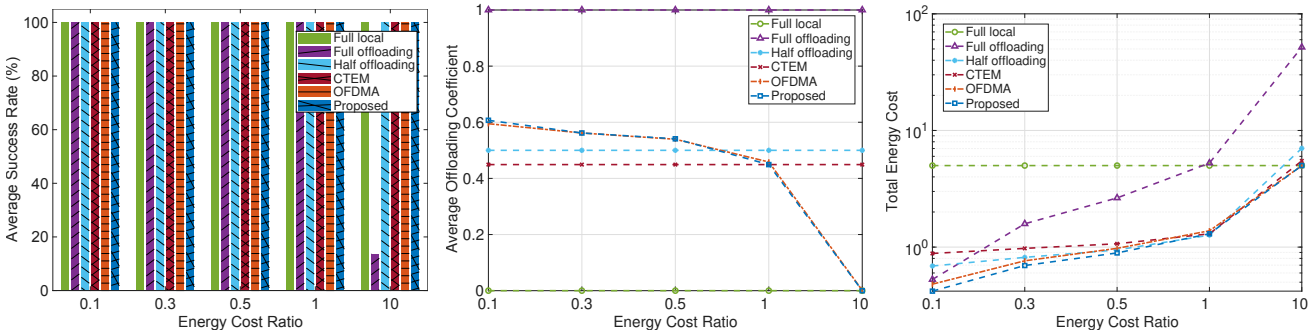
Fig. 5 evaluates the scalability of the proposed scheme for a varying number of collaborative UEs. The FL computation scheme cannot complete the computation task within the given delay budget of 0.5 s because it is limited by local computational capacity, as shown in Fig. 5a. However, the proposed scheme and the CTEM baseline maintain a success rate of 100% across all evaluated user counts, benefiting from the powerful computational capacity at the AP. In contrast, under a congested network condition (e.g., at 10 UEs), the FO scheme begins to experience performance degradation, while the OFDMA baseline performance drops significantly to a success rate of approximately 55%. This highlights the inefficiency of orthogonal spectrum allocation, which severely reduces the per-user bandwidth as the network scales. As illustrated in Fig. 5b, our proposed adaptive scheme maintains a relatively stable offloading coefficient (gradually adjusting from approximately 0.78 to 0.70) to avoid communication congestion, whereas the OFDMA scheme is forced to drastically reduce its offloading ratio. Fig. 5c demonstrates that the total energy cost scales with the number of collaborative UEs. Most importantly, the proposed scheme consistently demonstrates the lowest energy cost across all user counts. For instance, when the number of UEs reaches 10, it achieves a total energy cost of approximately 0.26, significantly outperforming both the basic fixed offloading schemes, CTEM and OFDMA baselines. Specifically, it drastically reduces the cost compared to the OFDMA baseline (costing around 0.68) and the conventional CTEM baseline (costing around 1.50). The dramatic rise in energy cost for CTEM vividly illustrates that ignoring energy heterogeneity (i.e., setting $\beta_k = 1$) forces UEs to aggressively utilize local computation, thereby rapidly consuming their limited battery reserves. The consistent energy advantage and moderate scaling slope validate the proposed scheme's ability to intelligently distribute computational tasks while effectively resolving multi-user resource contention via full-bandwidth sharing and joint beamforming.

Fig. 6 evaluates the system performance as a function of maximum transmit power ranging from 5 dBm to 30



(a) (b) (c)

Fig. 6: Performance evaluation with varying maximum transmit power given $\tau^{\mathrm{th}} = 0.4$ s.



(a) (b) (c)

Fig. 7: Performance evaluation under diverse cost ratio given $\tau^{\mathrm{th}} = 0.2$ s, $C_k^{\mathrm{u}} = 5, \forall k$ GHz and $C^{\mathrm{ap}} = 50$ GHz.

dBm, under a relatively loose latency constraint of 0.4 s. Limited by the local computational capacity, both baseline schemes of FL and HO cannot guarantee the success rate, as illustrated in Fig. 6a. Furthermore, the OFDMA baseline experiences a slight performance degradation at a stringent transmit power limit of 5 dBm due to its restricted per-user bandwidth. In contrast, the proposed scheme, along with the FO and CTEM schemes, successfully maintains a 100% success rate across the power range. As shown in Fig. 6b, the proposed scheme, as well as the advanced CTEM and OFDMA baselines, maintains a stable and high offloading coefficient of approximately 0.81. This adaptive behavior is primarily driven by the latency constraint and the large data size, which naturally favor edge computing. This average offloading coefficient is also entirely consistent with the results observed in previous evaluations (e.g., Fig. 2b) when the delay budget is set to 0.4 s. This indicates that the optimal transmit power lies well below 5 dBm, making the maximum power constraint non-binding. Fig. 6c demonstrates that the total energy cost remains practically constant for all valid schemes as the transmit power limit increases, which is consistent with the non-binding power constraint observation. The proposed scheme consistently achieves the lowest energy cost across the entire power range, outperforming CTEM, OFDMA, and FO (noting that FL and HO fail to provide feasible solutions under this latency constraint). This further confirms the robustness and superior energy cost of the proposed joint offloading and beamforming design.

We also evaluate the system performance by varying the ratio β from 0.1 to 10 in Fig. 7 with the UE and AP computational capacities of 5 GHz and 50 GHz, respectively. As illustrated in Fig. 7a, the proposed scheme, alongside the OFDMA, CTEM, FL, and HO baselines, maintains a 100% success rate across the entire evaluated range of β. However, for the FO scheme, the success rate drops significantly to approximately 13.8% when β reaches 10. This severe degradation occurs because the system becomes over-constrained when attempting to forcefully utilize AP resources despite their heavily penalized energy costs in the objective function. Fig. 7b uncovers the underlying adaptation mechanisms of the optimization frameworks. Notably, the average offloading coefficients of the proposed scheme and the OFDMA baseline decrease progressively from approximately 0.6 to 0 as β increases from 0.1 to 10. This explicitly reflects an intelligent, strategic shift toward local processing (i.e., AP-constrained case) to actively avoid the increasingly expensive AP energy costs. In contrast, the CTEM baseline (which fundamentally lacks energy-heterogeneity awareness by permanently assuming $\beta_k = 1$) exhibits a completely flat offloading curve regardless of the actual β. It inflexibly dictates a fixed offloading proportion around 0.45, failing to adapt to environmental cost variations. The corresponding energy cost is demonstrated in Fig. 7c. Notably, the total energy cost of the CTEM baseline intersects with the proposed scheme exactly at $\beta = 1$,

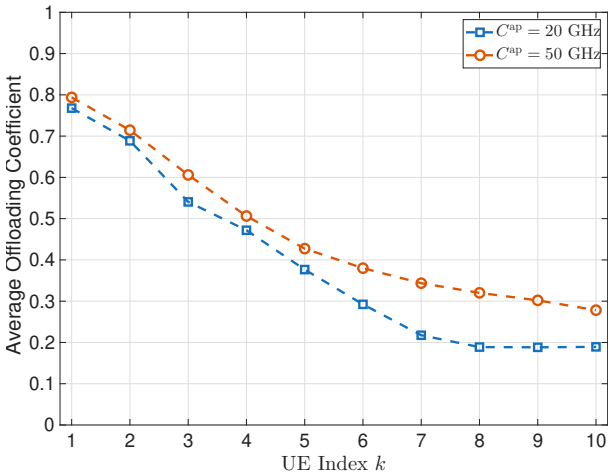


Fig. 8: Average offloading coefficient versus UE index for $K = 10$ UEs with heterogeneous energy cost ratios β_k uniformly spanning $[0.1, 5.0]$ and computational capacities C_k^{u} uniformly spanning $[1.0, 5.0]$ GHz, under two AP computational capacity settings ($C^{\mathrm{ap}} = 20$ GHz and $C^{\mathrm{ap}} = 50$ GHz).

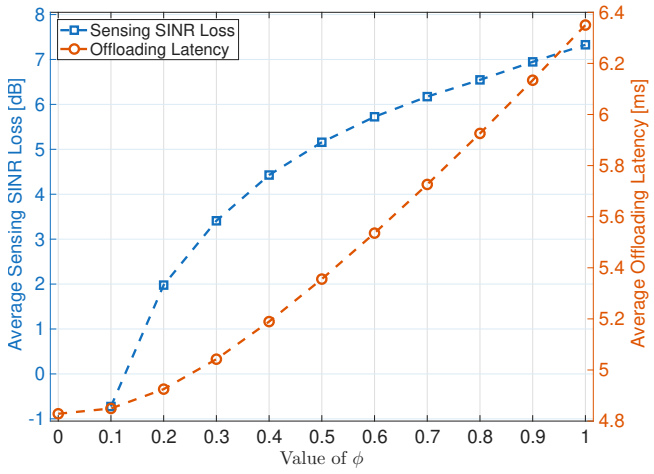


Fig. 9: System performance evaluation under imperfect CSI: Average sensing SINR loss (left axis) and average offloading latency (right axis) versus the channel estimation error bound ϕ.

which rigorously verifies theoretical consistency since CTEM optimizes under the assumption of $\beta_k = 1$. However, as the actual β deviates from 1, CTEM's "blind" fixed strategy incurs distinct performance losses. Specifically, when AP energy cost is cheap ($\beta = 0.1$, i.e., battery-critical mode), CTEM severely under-utilizes the edge server, resulting in a total cost of approximately 0.88 that is more than double that of the proposed scheme with around 0.42 due to excessive local computation. Conversely, when AP energy cost is heavily penalized ($\beta = 10$), the proposed scheme intelligently bounds its cost by shifting entirely to local processing (merging perfectly with the FL). In contrast, the FO scheme fails to avoid the heavily penalized AP, resulting in an exponential cost explosion. Furthermore, the proposed scheme consistently maintains a cost advantage over the OFDMA baseline across all ratios, validating the superior efficiency of full-bandwidth sharing and joint beamforming.

In practice, UEs with critically low battery levels tend to have limited computational resources, while energy-abundant UEs tend to be more capable. To reflect this practical correlation, we assume that $K = 10$ UEs are configured with jointly heterogeneous energy cost ratios and computational capacities. Specifically, Fig. 8 illustrates the average offloading coefficient α_k versus UE index for $K = 10$ UEs under two AP computational capacity settings ($C^{\mathrm{ap}} = 20$ GHz and $C^{\mathrm{ap}} = 50$ GHz), where β_k uniformly spans $[0.1, 5.0]$ and C_k^{u} correspondingly uniformly spans $[1.0, 5.0]$ GHz. For both settings, α_k decreases monotonically as the UE index increases (i.e., as β_k increases and C_k^{u} improves), reflecting the proposed scheme's ability to intelligently shift from AP-centric to UE-centric processing as AP energy becomes more costly and local computational capacity improves. Furthermore, a higher AP computational capacity ($C^{\mathrm{ap}} = 50$ GHz) yields consistently higher offloading coefficients across all UEs, since the increased AP processing capability makes offloading more attractive. These results with $K = 10$ heterogeneous UEs further validate the energy-heterogeneity-aware adaptivity and scalability of the proposed framework to a larger number of users with diverse device states.

Finally, we evaluate the sensitivity of the proposed scheme to practical performance impact under imperfect channel state information (CSI). Specifically we adopt the widely-used bounded channel estimation error model [57], [58], where the actual physical channel $\mathbf{H}_k$ is modeled as $\mathbf{H}_k = \hat{\mathbf{H}}_k + \Delta\mathbf{H}_k, \forall k$, where $\hat{\mathbf{H}}_k$ represents the estimated channel available at the AP, and $\Delta\mathbf{H}_k$ denotes the corresponding estimation error matrix. The channel uncertainty is assumed to be bounded within a continuous spherical region, defined by the uncertainty set of $\Psi_k \triangleq \{\Delta\mathbf{H}_k : \|\Delta\mathbf{H}_k\|_2 \leq \phi\}, \forall k$, where $\phi \geq 0$ specifies the maximum radius of the estimation error bound (i.e., the spectral-norm bound). Based on this worst-case CSI error model, Fig. 9 illustrates the average sensing SINR loss and the average offloading latency versus the error bound ϕ. It can be observed that the sensing SINR loss steadily increases with ϕ due to severe beamforming mismatches, which degrade the main-lobe gain and spatial nulling accuracy. On the other hand, while the offloading latency also increases due to degraded communication rates, the absolute increment is relatively limited (e.g., increasing from approximately 4.83 ms to around 6.35 ms). This demonstrates a certain degree of inherent robustness in the offloading process, primarily benefiting from the large bandwidth available in mmWave systems. It is noted that the analytical frameworks and main simulation results presented in this paper are obtained under the assumptions of perfect CSI and synchronization to establish fundamental performance limits. In practice, as demonstrated in Fig. 9, estimation errors and synchronization offsets can negatively impact both sensing quality and transmission delays. Extending the proposed adaptive offloading scheme to worst-case robust ISCC designs (e.g., incorporating the uncertainty set Ψ_k into the optimization constraints) remains an important and promising direction, which is beyond the main focus of this work.

VII. Conclusion and Discussion

An energy-efficient adaptive computation task offloading scheme for ISCC systems is proposed. The scheme dynamically balances the energy consumption between the UE and AP via an adaptive computational offloading strategy based on the weighted energy cost ratio, which is particularly important in practical scenarios where devices exhibit heterogeneous battery levels, computational capacities, and energy cost priorities. We proposed a SCA-ADMM based algorithmic framework for solving the total energy cost nonconvex minimization problem while guaranteeing the QoS requirements. The closed-form solutions are derived for the optimal offloading coefficient and computational resource allocation at both UE and AP, whereas beamforming vectors are updated via a projected gradient descent approach. Simulation results show that the proposed scheme significantly outperforms baseline schemes with fixed offloading strategies in terms of total energy cost and success rate by dynamically adjusting the offloading coefficient according to the communication and computation resource constraints and energy cost ratio. The proposed approach provides an effective and practical solution for energy-efficient 6G ISCC systems with heterogeneous device capabilities in computation and battery.

Several promising directions remain to extend the proposed ISCC framework. First, developing robust designs against imperfect CSI and synchronization errors is a crucial extension. Second, relaxing the single-target LoS assumption to multi-target scenarios via multi-beam designs is another important direction. This can be achieved by introducing per-target sensing constraint alongside multi-beam transmission designs. Third, extending the system to NLoS or heavily cluttered environments represents a critical challenging direction. When direct LoS paths are severely attenuated or obstructed, future designs will need to advance the physical-layer signal processing to leverage resolvable multipath components. Alternatively, the system could exploit cooperative multi-view sensing among distributed UEs or deployed smart relays to effectively extract target information from reflected echoes. Integrating these advanced sensing paradigms with our adaptive offloading framework will be a key focus of future research.

Appendix A

Given the constructed objective function $\bar{\mathcal{L}}\left(\mathbf{F}_k\right)$ expressed in (41), its gradient with respect to $\mathbf{F}_k, \forall k$ is given by

$$\begin{aligned}\nabla_{\mathbf{F}_k}\bar{\mathcal{L}}\left(\mathbf{F}_k\right) =&\beta_k \mathrm{Tr}(\mathbf{V}_{\mathbf{F}_k}+\mathbf{V}_{\mathbf{R}_k^0})\nabla_{\mathbf{F}_k}\bar{t}_k^{\mathrm{com}}\\ &+\rho_1(\mathbf{F}_k-\mathbf{V}_{\mathbf{F}_k}+\mathbf{Z}_{\mathbf{F}_k})\\ &-\rho_3(-\bar{\mathcal{F}}_k\left(\mathcal{U}_k^{\mathrm{s}(l)}\right)+\varsigma_k^{\mathrm{s}}+z_k^{\mathrm{s}})\nabla_{\mathbf{F}_k}\bar{\mathcal{F}}_k(\mathcal{U}_k^{\mathrm{s}(l)})\\ &+\rho_4\left(\bar{t}_k^{\mathrm{com}}+t_k^{\mathrm{p}}-\tau_{\mathrm{th}}+\varsigma_k^{\mathrm{off}}+z_k^{\mathrm{off}}\right)\nabla_{\mathbf{F}_k}\bar{t}_k^{\mathrm{com}},\end{aligned} \tag{59}$$

where $\nabla_{\mathbf{F}_k}\bar{t}_k^{\mathrm{com}}$ and $\nabla_{\mathbf{F}_k}\bar{\mathcal{F}}_k(\mathcal{U}_k^{\mathrm{s}(l)})$ are the corresponding gradients with respect to $\mathbf{F}_k$, which are given by

$$\nabla_{\mathbf{F}_k}\bar{t}_k^{\mathrm{com}}=-\frac{\alpha_k Q_k}{(\bar{R}_k(\mathcal{U}_k^{\mathrm{c}(l)}))^2}\frac{B}{\ln(2)(1+\gamma_k^{\mathrm{c}})}\nabla_{\mathbf{F}_k}\gamma_k^{\mathrm{c}}(\mathcal{U}_k^{\mathrm{c}(l)}), \tag{60}$$

$$\nabla_{\mathbf{F}_k}\bar{\mathcal{F}}_k(\mathcal{U}_k^{\mathrm{s}(l)})=-\gamma_{\mathrm{th}}^{\mathrm{s}}\tilde{\mathbf{H}}_k^H\mathbf{w}_k^{(l)}\mathbf{w}_k^{(l)H}\tilde{\mathbf{H}}_k, \tag{61}$$

respectively. Then $\mathbf{F}_k$ is updated by following the gradient descent approach

$$\mathbf{F}_k^{(\iota+1)}=\mathbf{F}_k^{(\iota)}-\eta_{\mathbf{F}_k}\nabla_{\mathbf{F}_k}\bar{\mathcal{L}}\left(\mathbf{F}_k\right), \tag{62}$$

and then projected onto the set of PSD matrices with trace constraint $\mathbf{F}_k^{(\iota+1)}=\mathcal{P}_{\mathcal{C}_{\mathbf{F}_k}}(\mathbf{F}_k^{(\iota+1)})$ where

$$\mathcal{C}_{\mathbf{F}_k}=\left\{\mathbf{F}_k:\mathbf{F}_k\succeq 0,\ \mathrm{Tr}(\mathbf{R}_k^0+\mathbf{F}_k)\le P_{\max}\right\}. \tag{63}$$

References

[1] S. Lu, F. Liu, Y. Li, K. Zhang, H. Huang, J. Zou, X. Li, Y. Dong, F. Dong, J. Zhu, W. Yuan, Y. Cui, and L. Hanzo, "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19 094–19 120, Jun. 2024.

[2] Y. Liu, T. Huang, F. Liu, D. Ma, W. Huangfu, and Y. C. Eldar, "Next-generation multiple access for integrated sensing and communications," *Proc. IEEE*, vol. 112, no. 9, pp. 1467–1496, Sep. 2024.

[3] X. Zhu, J. Liu, L. Lu, T. Zhang, T. Qiu, C. Wang, and Y. Liu, "Enabling intelligent connectivity: A survey of secure ISAC in 6G networks," *IEEE Commun. Surv. Tutorials*, vol. 27, no. 2, pp. 748–781, Apr. 2024.

[4] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[5] W. Chen, Y. He, G. Yu, J. Wang, and H. Luo, "Sensing framework design and performance optimization with action detection for ISCC," *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 8361–8375, Oct. 2025.

[6] D. Wen, Y. Zhou, X. Li, Y. Shi, K. Huang, and K. B. Letaief, "A survey on integrated sensing, communication, and computation," *IEEE Commun. Surv. Tutorials*, vol. 27, no. 5, pp. 3058–3098, Oct. 2025.

[7] K. Dong, S. A. Vorobyov, H. Yu, and T. Taleb, "Beamforming design for integrated sensing, computation over-the-air, and communication in internet of robotic things," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 32 478–32 489, Oct. 2024.

[8] C. Li, M. Dong, Y. Fu, F. R. Yu, and N. Cheng, "Integrated sensing, communication, and computation for IoV: Challenges and opportunities," *IEEE Commun. Surv. Tutorials*, vol. 28, pp. 1136–1168, Jan. 2026.

[9] J. Zhao, R. Ren, D. Zou, Q. Zhang, and W. Xu, "IoV-oriented integrated sensing, computation, and communication: System design and resource allocation," *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 16 283–16 294, Nov. 2024.

[10] X. Luo, Q. Lin, R. Zhang, H.-H. Chen, X. Wang, and M. Huang, "ISAC–a survey on its layered architecture, technologies, standardizations, prototypes and testbeds," *IEEE Commun. Surv. Tutorials*, vol. 28, pp. 485–526, Jan. 2026.

[11] K. Wang, D. Niyato, W. Chen, and A. Nallanathan, "Task-oriented delay-aware multi-tier computing in cell-free massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2000–2012, Jul. 2023.

[12] C.-H. Hu, Z. Chen, and E. G. Larsson, "Energy-efficient federated edge learning with streaming data: A lyapunov optimization approach," *IEEE Trans. Commun.*, vol. 73, no. 2, pp. 1142–1156, Feb. 2025.

[13] F. Liu, Y.-F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, S. Buzzi, Y. C. Eldar, and S. Jin, "Sensing with communication signals: From information theory to signal processing," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 1–30, Feb. 2026.

[14] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021.

[15] B. Di, H. Zhang, Z. Han, R. Zhang, and L. Song, "Reconfigurable holographic surface: A new paradigm for ultra-massive MIMO," *IEEE Trans. Cognit. Commun. Networking*, vol. 11, no. 6, pp. 3761–3783, Dec. 2025.

[16] M. Vaezi, G. A. A. Baduge, E. Ollila, and S. A. Vorobyov, "A tutorial on AI-empowered integrated sensing and communications," *IEEE Commun. Surv. Tutorials*, vol. 28, pp. 4980–5013, Feb. 2026.

[17] Z. Lyu, G. Zhu, and J. Xu, "Joint maneuver and beamforming design for UAV-enabled integrated sensing and communication," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2424–2440, Apr. 2022.

[18] Z. Behdad, Ö. T. Demir, K. W. Sung, E. Björnson, and C. Cavdar, "Multi-static target detection and power allocation for integrated sensing and communication in cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11 580–11 596, Sep. 2024.

[19] X. Lou, W. Xia, S. Jin, and H. Zhu, "Beamforming optimization in distributed ISAC system with integrated active and passive sensing," *IEEE Trans. Commun.*, vol. 73, no. 3, pp. 1607–1620, Mar. 2025.

[20] B. Liao, H. Q. Ngo, M. Matthaiou, and P. J. Smith, "Power allocation for massive MIMO-ISAC systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 14 232–14 248, Oct. 2024.

[21] N. Xue, X. Mu, Y. Liu, X. Zhang, and Y. Chen, "Hybrid NOMA empowered energy-efficient ISAC," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 3894–3908, May 2025.

[22] Z. Nan, S. Zhou, Y. Jia, and Z. Niu, "Joint task offloading and resource allocation for vehicular edge computing with result feedback delay," *IEEE Trans. Wireless Commun.*, vol. 22, no. 10, pp. 6547–6561, Oct. 2023.

[23] H. Jiang, X. Dai, Z. Xiao, and A. Iyengar, "Joint task offloading and resource allocation for energy-constrained mobile edge computing," *IEEE Trans. Mob. Comput.*, vol. 22, no. 7, pp. 4000–4015, Jul. 2023.

[24] W. Fan, Y. Su, J. Liu, S. Li, W. Huang, F. Wu, and Y. Liu, "Joint task offloading and resource allocation for vehicular edge computing based on V2I and V2V modes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4277–4292, Apr. 2023.

[25] X. Dai, Z. Xiao, H. Jiang, H. Chen, G. Min, S. Dustdar, and J. Cao, "A learning-based approach for vehicle-to-vehicle computation offloading," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7244–7258, Apr. 2023.

[26] G. Interdonato and S. Buzzi, "Joint optimization of uplink power and computational resources in mobile edge computing-enabled cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 72, no. 3, pp. 1804–1820, Mar. 2024.

[27] R. Lin, T. Xie, S. Luo, X. Zhang, Y. Xiao, B. Moran, and M. Zukerman, "Energy-efficient computation offloading in collaborative edge computing," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21 305–21 322, Nov. 2022.

[28] Z. Lin, J. Yang, C. Wu, and P. Chen, "Energy-efficient task offloading for distributed edge computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 14 056–14 061, Sep. 2024.

[29] Z. Zhou, X. Li, G. Zhu, B. Zhou, H. Xing, and K. Huang, "Integrated sensing-communication-computation design for energy efficient data processing," *IEEE Trans. Network Sci. Eng.*, vol. 13, pp. 4172–4186, Sep. 2025.

[30] J. Yao, W. Xu, G. Zhu, K. Huang, and S. Cui, "Energy-efficient edge inference in integrated sensing, communication, and computation networks," *IEEE J. Sel. Areas Commun.*, vol. 43, no. 10, pp. 3580–3595, Oct. 2025.

[31] Y. Zhou, X. Liu, X. Zhai, Q. Zhu, and T. S. Durrani, "UAV-enabled integrated sensing, computing, and communication for internet of things: Joint resource allocation and trajectory design," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12 717–12 727, Apr. 2024.

[32] N. Huang, C. Dou, Y. Wu, L. Qian, and R. Lu, "Energy-efficient integrated sensing and communication: A multi-access edge computing design," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2053–2057, Dec. 2023.

[33] S. Liu, D. Wen, D. Li, Q. Chen, G. Zhu, and Y. Shi, "Energy-efficient optimal mode selection for edge AI inference via integrated sensing-communication-computation," *IEEE Trans. Mob. Comput.*, vol. 23, no. 12, pp. 14 248–14 262, Dec. 2024.

[34] C. Dou, M. Dai, N. Huang, Y. Wu, L. Qian, and T. Q. Quek, "Integrated sensing and two-tier task offloading via non-orthogonal multiple access: An energy-minimization design," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 19 157–19 171, Dec. 2024.

[35] D. Tagliaferri, M. Manzoni, M. Mizmizi, S. Tebaldini, A. V. Monti-Guarnieri, C. M. Prati, and U. Spagnolini, "Cooperative coherent multi-static imaging and phase synchronization in networked sensing," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 10, pp. 2905–2921, Oct. 2024.

[36] X. Li, F. Liu, Z. Zhou, G. Zhu, S. Wang, K. Huang, and Y. Gong, "Integrated sensing, communication, and computation over-the-air: MIMO beamforming design," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5383–5398, Aug. 2023.

[37] Z. He, H. Shen, W. Xu, Y. C. Eldar, and X. You, "MSE-based training and transmission optimization for MIMO ISAC systems," *IEEE Trans. Signal Process.*, vol. 72, pp. 3104–3121, Jun. 2024.

[38] S. Buzzi, C. D'Andrea, and S. Liesegang, "Scalability and implementation aspects of cell-free massive MIMO for ISAC," in *Proc. 19th Int. Symp. Wireless Commun. Syst.*, Rio de Janeiro, Brazil, Jul. 2024, pp. 1–6.

[39] Z. Wu, M. Cui, and L. Dai, "Enabling more users to benefit from near-field communications: From linear to circular array," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 3735–3748, Apr. 2024.

[40] X. Yang, Z. Wei, J. Xu, Y. Fang, H. Wu, and Z. Feng, "Coordinated transmit beamforming for networked ISAC with imperfect CSI and time synchronization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18 019–18 035, Dec. 2024.

[41] G. Cheng, Y. Fang, J. Xu, and D. W. K. Ng, "Optimal coordinated transmit beamforming for networked integrated sensing and communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8200–8214, Aug. 2024.

[42] Z. Ren, L. Qiu, J. Xu, and D. W. K. Ng, "Robust transmit beamforming for secure integrated sensing and communication," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5549–5564, Sep. 2023.

[43] D. T. Bellini, D. Tagliaferri, M. Mizmizi, S. Tebaldini, and U. Spagnolini, "Multi-view integrated imaging and communication," in *Proc. IEEE 5th Int. Symp. Joint Commun. & Sens. (JC&S)*. Oulu, Finland: IEEE, Jan. 2025, pp. 1–6.

[44] T. T. Nguyen, L. B. Le, and Q. Le-Trung, "Computation offloading in MIMO based mobile edge computing systems under perfect and imperfect CSI estimation," *IEEE Trans. Serv. Comput.*, vol. 14, no. 6, pp. 2011–2025, Nov.-Dec. 2021.

[45] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 3, pp. 1657–1681, May 2017.

[46] L. Zhao, E. Zhang, S. Wan, A. Hawbani, A. Y. Al-Dubai, G. Min, and A. Y. Zomaya, "MESON: A mobility-aware dependent task offloading scheme for urban vehicular edge computing," *IEEE Trans. Mob. Comput.*, vol. 23, no. 5, pp. 4259–4272, May 2024.

[47] Z. Yu, Y. Gong, S. Gong, and Y. Guo, "Joint task offloading and resource allocation in UAV-enabled mobile edge computing," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3147–3159, Apr. 2020.

[48] Z. He, W. Xu, H. Shen, D. W. K. Ng, Y. C. Eldar, and X. You, "Full-duplex communication for ISAC: Joint beamforming and power optimization," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, pp. 2920–2936, Sep. 2023.

[49] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[50] K. T. Phan, S. A. Vorobyov, N. D. Sidiropoulos, and C. Tellambura, "Spectrum sharing in wireless networks via QoS-aware secondary multicast beamforming," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2323–2335, Jun. 2009.

[51] A. Khabbazibasmenj, F. Roemer, S. A. Vorobyov, and M. Haardt, "Sum-rate maximization in two-way AF MIMO relaying: Polynomial time solutions to a class of DC programming problems," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5478–5493, Oct. 2012.

[52] Y. Huang, H. Fu, S. A. Vorobyov, and Z.-Q. Luo, "Robust adaptive beamforming via worst-case SINR maximization with nonconvex uncertainty sets," *IEEE Trans. Signal Process.*, vol. 71, pp. 218–232, Jan. 2023.

[53] A. Liu, V. K. Lau, and B. Kananian, "Stochastic successive convex approximation for non-convex constrained stochastic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, Aug. 2019.

[54] X. Mu, Y. Liu, L. Guo, and J. Lin, "Non-orthogonal multiple access for air-to-ground communication," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2934–2949, May 2020.

[55] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[56] Y. Huang and D. P. Palomar, "Rank-constrained separable semidefinite programming with applications to optimal beamforming," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 664–678, Feb. 2010.

[57] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.

[58] Y. Zhang, W. Ni, J. Wang, W. Tang, M. Jia, Y. C. Eldar, and D. Niyato, "Robust transceiver design for covert integrated sensing and communications with imperfect CSI," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 8016–8031, Sep. 2025.