Holographic Counterpart-Assisted Edge Computing for End-to-End Latency Optimization in IoMT Consumer Electronics

Jianhui Lyu, *Senior Member, IEEE*, Keqin Li, *Fellow, IEEE*, and Shuyu Hu

***Abstract*— The rapid development of wireless communication technologies and the Internet of Medical Things (IoMT) has led to the proliferation of time-sensitive and computation-intensive medical applications such as real-time health monitoring, remote surgery assistance, and augmented reality rehabilitation. These applications impose stringent latency requirements on IoMT consumer electronics. This paper proposes holographic-assisted edge latency optimization for IoMT (HELO-IoMT) in response to these challenges. First, within the holographic-assisted edge network framework, we establish physical and holographic models for an edge computing network comprising IoMT consumer electronics, edge servers, and medical actuators. Then, to solve the resulting mixed-integer non-convex optimization problem, we create digital replicas of IoMT devices with real-time synchronization capabilities, enabling predictive resource allocation without physical resource waste. Our solution methodology decomposes the mixed-integer non-convex optimization problem into four interconnected subproblems: computation-communication resource optimization using inner convex approximation, device association optimization via the Hungarian algorithm, offloading decision optimization through linear programming, and transmission bandwidth optimization using convex optimization techniques. Simulation results demonstrate that HELO-IoMT outperforms state-of-the-art benchmarks across various performance metrics. For critical healthcare applications, HELO-IoMT reduces end-to-end latency by up to 45.5% compared to existing approaches, achieves 94.6% anomaly detection accuracy (14.9% higher than DEETO), reduces energy consumption by 44.1% for ECG monitors to 85 mJ/hour, and improves critical cardiac event response times to 1.2 seconds (62.5% faster than DEETO).**

***Index Terms*— Holographic counterpart, internet of medical things consumer electronics, edge computing, latency optimization.**

I. Introduction

THE explosive growth of wireless communication technologies and the Internet of Medical Things (IoMT) has revolutionized healthcare delivery systems, creating a substantial demand for medical consumer electronics that support time-sensitive and computation-intensive applications [1], [2], [3], [4]. Remote health monitoring, telehealth consultations, augmented reality surgical assistance, and predictive medical analytics are increasingly common in modern healthcare environments [5], [6]. These applications impose stringent latency requirements; for instance, real-time electrocardiogram (ECG) analysis requires continuous sensor monitoring and immediate processing to ensure patient safety [7]. Mobile edge computing (MEC) has emerged as a promising solution, leveraging the computational capabilities of edge servers to achieve lower application latency for resource-constrained IoMT consumer electronics [8], [9], [10]. Using task offloading techniques, MEC enables computational tasks from resource-limited devices to be transferred to edge servers with superior processing capabilities, meeting healthcare-specific performance demands [11], [12]. Consequently, designing joint MEC task offloading and communication-computation resource allocation strategies is crucial for IoMT applications where timely intervention can be life-saving [13].

The wider implications overshadow mere technical optimization and lie in a transformation of fundamental healthcare delivery. Optimized IoMT platforms enable remote monitoring of patients outside of conventional clinical centers, putting the patients on a track of enjoying independence while receiving quality healthcare. This technology in effect levels the playing field for access to advanced medical monitoring, thus closing the disparities that exist between urban and rural populations. From an economic viewpoint, efficient edge computing alleviates infrastructures from centralized health systems, thereby allowing the ill citizenry to age with respect to scaling medical services.

Existing research has extensively investigated MEC task offloading and resource optimization in various contexts [14]. Researchers have proposed priority-based task scheduling algorithms to minimize delay for tasks with dependencies in MEC-supported networks [15]. Deep reinforcement learning methods have been developed to optimize task delay in multi-user MEC systems for improved quality of service [16]. Joint task offloading and resource allocation problems have been studied to reduce the energy consumption of delay-constrained devices [17]. Resource allocation algorithms for uplink multi-user MEC vehicular networks have been designed to minimize the weighted sum of delay and energy consumption under multi-dimensional resource

Received 25 May 2025; revised 20 August 2025 and 21 September 2025; accepted 19 October 2025. Date of publication 22 October 2025; date of current version 8 December 2025. This work was supported by the National Natural Science Foundation of China under Grant 62202247. *(Corresponding author: Jianhui Lyu.)*

Jianhui Lyu is with the Multi-Modal Data Fusion and Precision Medicine Laboratory, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121001, China (e-mail: lvjianhui2012@163.com).

Keqin Li is with the College of Computer Science, The State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Shuyu Hu is with the School of Intelligent Medicine, Jinzhou Medical University, Jinzhou 121000, China (e-mail: hsy@jzmu.edu.cn).

Digital Object Identifier 10.1109/TCE.2025.3624448

constraints [18]. However, these studies focus solely on MEC task offloading and resource allocation without considering device association optimization. This could further reduce latency and achieve load balancing among servers—a critical factor in healthcare scenarios where medical data processing requires balanced resource utilization [19], [20]. Some works have jointly optimized device association, offloading ratio, and computation-communication resources to minimize task delay and alleviate server computational load [21]. However, they overlooked the upload process of locally computed results, which is non-negligible in distributed medical systems where local computation results need to be uploaded to servers for centralized analysis or storage, especially when the volume of medical data and transmission delay cannot be ignored [22]. Priority-based scheduling approaches excel in deterministic environments but struggle with dynamic medical data patterns where patient conditions change unpredictably. Deep reinforcement learning methods provide adaptive solutions but require extensive training data and may not converge quickly enough for time-critical medical applications.

The emergence of holographic counterpart (HC) technologies offer an effective solution to address these challenges [23], [24]. HC technology creates digital replicas of physical entities in virtual space, enabling mapping from the physical world to the digital domain. Given these advantages, researchers have integrated HC technology with edge networks, constructing holographic-assisted edge networks (HEN) [25], [26]. HC differs from traditional digital twins by maintaining real-time bidirectional synchronization with physical entities and employing predictive modeling for proactive optimization. While digital twins typically focus on monitoring and analysis, HC actively influences physical system behavior through real-time optimization decisions. In the HEN framework, edge nodes such as access points (APs) collect real-time information from physical objects and establish and maintain HC models based on this information. Through this approach, HEN can design and optimize task offloading and resource allocation schemes directly in the digital domain, improving network decision-making efficiency while reducing physical resource waste [27], [28]. Without HEN, obtaining optimal task offloading and resource allocation strategies would require continuous communication between edge servers and devices to acquire real-time information, increasing communication costs and potentially affecting decision-making effectiveness due to communication delays. Thus, integrating HC in medical edge computing environments reduces the computational load on resource-constrained IoMT consumer electronics, enhances decision-making efficiency, and minimizes communication costs while obtaining optimal resource allocation strategies [29].

These limitations create a research gap where no existing framework addresses the complete task execution pipeline while optimizing for medical-grade performance requirements [30], [31]. Most approaches assume static device-server relationships, ignoring the dynamic nature of patient mobility and changing medical priorities [32], [33]. Furthermore, existing works lack consideration of medical actuator requirements, treating all result delivery as equivalent when medical contexts demand differentiated service levels [34].

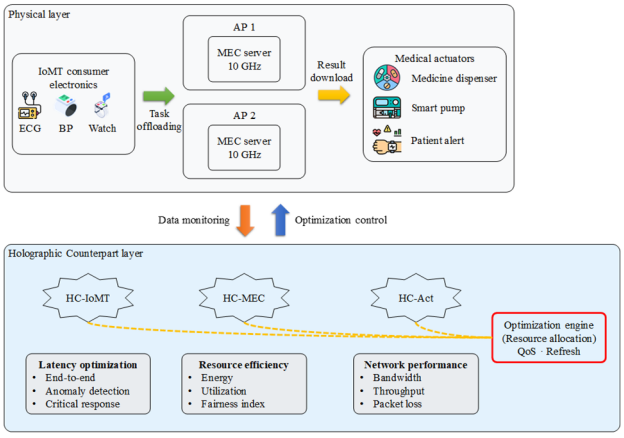

Fig. 1. HEN system model for IoMT.

Accordingly, the main contributions of this paper are summarized as follows:

- We develop a HEN framework specifically designed for medical electronics, establishing HC for IoMT devices that enable real-time optimization without excessive communication overhead.
- We establish physical and holographic models for the complete healthcare task execution process, from physiological data collection to medical actuation.
- We formulate an end-to-end latency optimization problem under realistic healthcare constraints and derive a mathematical model that captures the unique requirements of medical monitoring applications.

The rest of the paper is organized as follows: Section II presents the HELO-IoMT system model. Section III outlines our problem solution methodology. Section IV provides simulation results and analysis. Finally, Section V concludes the paper.

II. System Model

This paper considers a two-layer HEN consisting of physical and HC layers. The physical layer includes APs equipped with MEC servers, IoMT consumer electronics, and medical actuators, which collaboratively realize the complete process from data upload and task computation to result download. IoMT consumer electronics detect physiological data and generate computational tasks, adopting partial offloading approaches to complete computational tasks using local and edge server resources. APs utilize the powerful capabilities of MEC servers to process offloaded tasks, integrate both local and edge computing results, and transmit all results to medical actuators. Medical actuators are devices that execute treatment commands based on processed data (insulin pumps, medication dispensers). The HC layer consists of physical devices and the entire communication environment, monitoring the physical system's operational status and optimizing task offloading and resource allocation schemes through real-time interaction. The system model is illustrated in Fig. 1.

The system contains K IoMT consumer electronics, M APs, and L medical actuators, represented by sets $\mathrm{K} = 1, 2, \ldots, K$,

$\mathrm{M} = 1, 2, \ldots, M$, and $\mathrm{L} = 1, 2, \ldots, L$, respectively. $\pi = \pi_{km}, \forall k, m$ denotes the association variable between IoMT consumer electronics and APs, where $\pi km = 1$ when medical sensor k is associated with AP m, and $\pi_{km} = 0$ otherwise. Additionally, each AP can serve at most N IoMT consumer electronics, and each medical sensor can only associate with one AP.

Network topology variations affect holographic model accuracy through communication delays and fragmentation-induced data loss. The system maintains performance across mesh, star, and hybrid topologies by adapting model refresh rates based on network conditions. During fragmentation events, edge servers utilize cached historical data to maintain approximate holographic models until connectivity restoration, ensuring continued operation with reduced accuracy.

The computational task of medical sensor k is represented by $\mathrm{T}_k = D_k, C_k, \gamma_k, \mu_k$, where D_k indicates the medical sensor's task volume (bits), C_k represents the number of CPU cycles required to execute the task (cycles), γ_k and μ_k represent the proportion of medical sensor k's local computation results to local tasks and the proportion of edge computation results to offloaded tasks, respectively. The offloading factor for medical sensor k is denoted by $\theta = \theta_k, 0 \leq \theta_k \leq 1$, meaning the medical sensor offloads $\theta_k D_k$ data size to the AP for execution, while the remaining $(1-\theta_k) D_k$ data is processed locally. The offloading factor θ represents the proportion of computational tasks transferred from local devices to edge servers (0 = full local processing, 1 = complete offloading). While HC deviation measures the accuracy difference between holographic predictions and actual system behavior.

The HC model relies on the edge server's powerful computation and storage capabilities for establishment and maintenance [35]. The HC model of the medical sensor-actuator pair (k-l) is defined as:

$$\mathrm{HC}_{k-l} = \left\{ S_k, S_l, f_k, \hat{f}_k \right\}. \tag{1}$$

where S_k represents the digital domain state information monitored by the HC for medical sensor k, including local resources (remaining energy, transmission power, etc.), channel information (signal-to-noise ratio, bandwidth, interference information, etc.), etc. S_l represents the digital domain state information of medical actuator l, including battery level, activity range, etc. f_k represents the HC-estimated processing rate of medical sensor k.

Data integrity verification employs temporal consistency checks and cryptographic checksums to ensure holographic accuracy. When communication links experience failures, the system compares received data against expected patterns based on historical device behavior. Inconsistent readings trigger verification protocols that request data retransmission or activate backup monitoring pathways to maintain model reliability during network disruptions.

The HC model of AP m is defined as:

$$\mathrm{HC}_m^i = \left(S_m^{\mathrm{AP}}, f_{km}^{\mathrm{AP}}, \hat{f}_{km}^{\mathrm{AP}} \right). \tag{2}$$

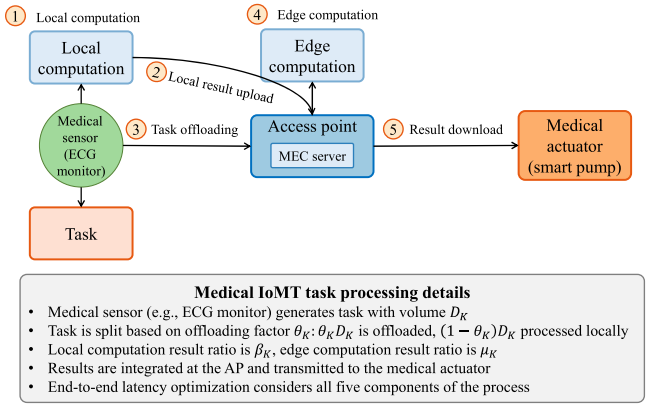

Fig. 2. Complete process task model for IoMT.

where $S^{\mathrm{AP}}m$ represents the digital domain state information of AP m, including association state, coverage range, etc. $f^{\mathrm{AP}}km$ represents the HC-estimated processing rate of AP m.

Holographic counterpart maintenance introduces computational overhead proportional to the number of monitored devices and model refresh rates. The framework addresses this through adaptive model complexity adjustment based on device criticality and available edge resources. Non-critical devices utilize simplified holographic models to reduce processing requirements, while emergency medical devices maintain full-fidelity counterparts for accurate predictions.

Additionally, the framework incorporates standardized interface protocols for integration with hospital information systems and electronic health records. Data exchange occurs through HL7 FHIR standards, enabling seamless connectivity with existing medical databases. The system provides API endpoints for external applications and maintains data format compatibility with major electronic health record vendors, preventing information silos and ensuring comprehensive patient monitoring.

A. Complete Process Task Model

In the physical system, IoMT consumer electronics collect physiological data to generate computational tasks, adopting partial offloading approaches to divide tasks and complete them collaboratively through local and edge processing [36]. IoMT consumer electronics upload local computation results to APs for further analysis or storage. The complete process task model comprises five components, as illustrated in Fig. 2.

B. Complete Process Task Model

In task offloading of uplink transmission, IoMT consumer electronics and APs connect through wireless links. Let $\bar{w}_k \in [0, W]$ represent the bandwidth allocated to medical sensor k for task offloading. The channel gain between medical sensor k and AP m is expressed as:

$$h_{km} = \beta_0 \left(d_{km} / d_0 \right)^{-3}. \tag{3}$$

where β_0 and d_{km} represent the path loss at the reference distance and the actual distance between the two entities, respectively. p_k denotes the transmission power of medical sensor k. The model employs orthogonal frequency division

multiplexing technology [37], neglecting co-channel interference. Therefore, the uplink transmission rate from medical sensor k to AP m is:

$$R_{km} = \bar{w}_k \log_2 \left(1 + \frac{\pi_{km} p_k |h_{km}|^2}{N_0}\right). \tag{4}$$

where N_0 represents the noise power.

Consequently, the transmission delay for task offloading from medical sensor k to AP m can be expressed as:

$$T_{km}^{up} = \frac{\pi_{km}\theta_k D_k}{R_{km}}. \tag{5}$$

In local result upload process, $\gamma_k \in [0, 1)$ represents the ratio of medical sensor k's local computation results to local tasks, meaning the local computation result size is $\gamma_k(1-\theta_k)D_k$ (bits). Therefore, the delay for uploading local results from medical sensor k to AP m is:

$$T_{lr}^{up} = \frac{\pi_{kmn}\gamma_k(1-\theta_k)D_k}{R_{km}}. \tag{6}$$

Let p_l represent the power allocated to medical actuator l, with the AP's maximum transmission power being $P_{\max}$, satisfying $0 \le \sum_{l=1}^{L} \pi_{ml} p_l \le P_{\max}, \forall m$. Therefore, the downlink transmission rate from AP m to medical actuator l is:

$$R_{ml} = w_l \log_2 \left(1 + \frac{\pi_{ml} lpl |h_{ml}|^2}{N_0}\right). \tag{7}$$

where N_0 represents the noise power. APs transmit both local and edge computation results to medical actuators. Let λ_k represent the download result ratio to the sum of both parts of the results. Therefore, the result size that medical actuator l needs to receive is $O_k = \lambda_k(\gamma_k(1-\theta_k) + \mu_k\theta_k)D_k$. Consequently, the downlink transmission delay for AP m to send computation results to medical actuator l is:

$$T_{ml}^{dw} = \frac{\pi_{km} O_k}{R_{ml}}. \tag{8}$$

C. Computation Model

Medical sensor k executes local tasks at the HC-estimated processing rate f_k. The estimated delay for local computation is:

$$\tilde{T}_k^{se} = \frac{(1-\theta_k)C_k}{f_k}. \tag{9}$$

HEN can obtain the deviation between medical sensor k's actual processing rate and HC-estimated rate $\hat{f}_k$. The difference between actual delay and HC-estimated delay is:

$$\Delta T_k^{se} = \frac{(1-\theta_k)C_k\hat{f}_k}{f_k\left(f_k - \hat{f}_k\right)}. \tag{10}$$

Eq. (10) calculates the deviation between actual and estimated local computation delay, where the numerator represents the additional cycles needed due to estimation error, and the denominator shows the reduced effective processing rate.

Therefore, the actual delay for local computation is:

$$T_k^{se} = \tilde{T}_k^{se} + \Delta T_k^{se}. \tag{11}$$

AP m executes medical sensor k's offloaded tasks at the HC-estimated processing rate f_{km}^{AP}. The estimated processing delay is:

$$\tilde{T}_{km}^{\mathrm{AP}} = \frac{\pi k_{mm}\theta_k C_k}{f_{km}^{AP}}. \tag{12}$$

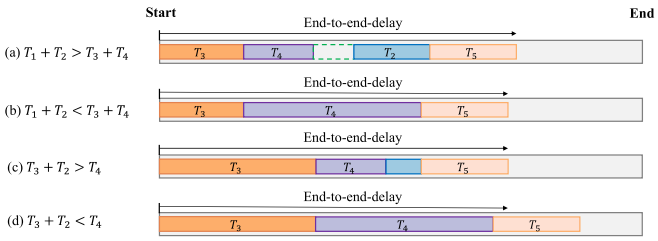

Fig. 3. End-to-end delay scenarios in IoMT task processing.

After obtaining the AP computation resource deviation from HEN, the difference between the actual processing delay and HC-estimated processing delay is:

$$\Delta T_{km}^{\mathrm{AP}} = \frac{\pi_{km}\theta_k C_k \hat{f}_k^{\mathrm{AP}}}{f_{km}^{\mathrm{AP}}\left(f_{km}^{\mathrm{AP}} - f_k^{\mathrm{AP}}\right)}. \tag{13}$$

Eq. (13) similarly captures AP processing deviation, accounting for the difference between predicted and actual edge server performance.

Therefore, the actual delay for AP computing offloaded tasks is:

$$T_{km}^{\mathrm{AP}} = \tilde{T}_{km}^{\mathrm{AP}} + \Delta T_{km}^{\mathrm{AP}}. \tag{14}$$

D. Energy Consumption and Delay Model

Communication energy consumption [38]: Communication energy consumption consists of task offloading and local result upload energy consumption. Therefore, the communication energy consumption of medical sensor k can be expressed as:

$$E_k^{cm} = \left(T_{km}^{up} + T_{lr}^{up}\right) p_k. \tag{15}$$

Computation energy consumption [39]: APs have a sufficient energy supply, so edge computation energy consumption can be neglected. Therefore, the computation energy consumption of medical sensor k is the local computation energy consumption, expressed as:

$$E_k^{cp} = (1-\theta_k)\frac{\xi_k}{2} C_k \left(f_k - \hat{f}_k\right)^2. \tag{16}$$

where ξ_k is the effective capacitance coefficient determined by the chip architecture.

Therefore, the total energy consumption of medical sensor k can be calculated as:

$$E_k^{tot} = E_k^{cm} + E_k^{cp}. \tag{17}$$

From the complete process task model, the end-to-end delay for medical sensor k's task, denoted by T_k^{E2E}, consists of five components. Due to communication resource constraints, local result upload and task offloading cannot be transmitted in parallel; local results must wait until task offloading is completed before uploading. This leads to four end-to-end delay scenarios, as illustrated in Fig. 3.

Based on the relationship between local computation delay and task offloading delay, these four scenarios can be categorized into two types:

1) Local computation delay is greater than task offloading delay As shown in Figs. 3(a) and 3(b), $\tau_1 > \tau_3$. In this

case, after medical sensor k completes task offloading, the channel remains idle for some time and must wait until the local computation is completed before results can be uploaded to the AP. The end-to-end delay for medical sensor k in this case is:

$$T_k^{E2E} = \max\left\{(\tau_1 + \tau_2), (\tau_3 + \tau_4)\right\} + \tau_5 . \tag{18}$$

2) Local computation delay is less than task offloading delay

As shown in Figs. 3(c) and 3(d), $\tau_1 < \tau_3$. In this case, it cannot immediately upload computation results after the medical sensor k completes local computation. It must wait until task offloading is completed before uploading local computation results to the AP. Therefore, the end-to-end delay for medical sensor k in this case is:

$$T_k^{E2E} = \tau_3 + \max\left\{\tau_2, \tau_4\right\} + \tau_5 . \tag{19}$$

Combining both cases, for medical sensor k, the total end-to-end delay can be expressed as:

$$T_k^{E2E} = \max\left\{\tau_3 + \tau_4, \max\left(\tau_1, \tau_3\right) + \tau_2\right\} + \tau_5. \tag{20}$$

Based on the above analysis, the end-to-end delay for medical sensor k's task T_k from offloading to execution to result download can be expressed as:

$$\begin{aligned} T_k^{E2E} &= \sum \max\left\{\tau_3 + \tau_4, \max\left(\tau_1, \tau_3\right) + \tau_2\right\} + \tau_5 \\ &= \sum \max\left\{T_{km}^{up} + T_{km}^{\mathrm{AP}}, \max\left(T_k^{se}, T_{km}^{up}\right) + T_{lr}^{up}\right\} + T_{ml}^{dw} \\ &= \sum_{m=1}^{M} \left\{ \max \left[\begin{array}{l} \frac{\pi_{km}\theta_k D_k}{R_{km}} + \frac{\pi_{km}\theta_k C_k}{f_{km}^{fAP} - f_{km}^{AP}}, \\ \max\left(\frac{(1-\theta_k)C_k}{f_k - f_k}, \frac{\pi_{km}\theta_k D_k}{R_{km}}\right) \\ + \frac{\pi_{km}\gamma_k(1-\theta_k)D_k}{R_{km}} \end{array} \right] + \frac{\pi_{km} O_k}{R_{ml}} \right\} . \end{aligned} \tag{21}$$

Eq. (21) provides the complete end-to-end delay formulation, combining uplink transmission, local computation, edge processing, and downlink transmission delays while handling the complex timing dependencies between parallel processes.

E. Problem Formulation

This paper considers the complete task execution process, comprehensively accounting for the five-stage delay with complex delay conflicts shown in Fig. 3. Therefore, optimizing delay is more valuable in this framework. We aim to minimize the system's total end-to-end delay by jointly optimizing device association, uplink/downlink bandwidth, AP transmission power, offloading factor, and local and edge HC-estimated processing rates $\mathbf{f} = f_k, f_{km}^{\mathrm{AP}}$. The optimization problem is formulated as follows:

$$\begin{aligned} &\min_{\pi,\theta,\bar{w},w,p,\mathbf{f}} \sum_{k=1}^{K}\sum_{m=1}^{M} T^{E2E}(\pi, \theta, \bar{w}, w, p, \mathbf{f}) \\ &C1: T_k^{E2E}(\pi, \theta, \bar{w}, w, p, \mathbf{f}) \le T_{\max}, \forall k \in \mathrm{K} \\ &C2: E_k^{tot}(\pi, \theta, \bar{w}, f_k) \le E_{\max}, \forall k \in \mathrm{K} \\ &C3: \sum_{k=1}^{K} \pi_{km} \le N, \forall m \in \mathrm{M} \\ &C4: \sum_{k=1}^{K} \bar{w}_k \le W, \sum_{l=1}^{L} w_l \le W \\ &C5: 0 \le \sum_{m=1}^{M} \pi_{km} p_l \le P_{\max}, \forall m \in \mathrm{M} \\ &C6: \pi_{km} \in \{0,1\}, \sum_{m=1}^{M} \pi_{km} = 1, \forall k \in \mathrm{K}, m \in \mathrm{M} \\ &C7: 0 \le \theta_k \le 1, \forall k \in \mathrm{K} \\ &C8: f_k \le F_{\max}^{se}, f_{km}^{\mathrm{AP}} \le F_{\max}^{\mathrm{AP}}, \forall k \in \mathrm{K}, m \in \mathrm{M} \\ &C9: \bar{w}_k, w_l, f_k, f_{km}^{\mathrm{AP}} \ge 0, \forall k \in \mathrm{K}, l \in \mathrm{L}, m \in \mathrm{M} \end{aligned} \tag{22}$$

where constraints C1 and C2 represent the delay and energy consumption constraints, particularly important in healthcare applications where timely response and extended device battery life are critical. C3 constrains the number of IoMT consumer electronics an AP can serve, reflecting realistic load balancing requirements in healthcare settings. C4 constrains bandwidth allocation, ensuring efficient utilization of limited spectrum resources for medical data transmission. C5 constrains AP power values, which are important for managing interference in potentially sensitive medical environments. C6 defines the device association constraint, ensuring each medical sensor is connected to exactly one AP to maintain reliable monitoring. C7 constrains the offloading factor values, allowing for flexible distribution of computational tasks between resource-constrained IoMT consumer electronics and edge servers. C8 constrains computation frequency values, reflecting the physical limitations of processors in both IoMT consumer electronics and edge servers. Finally, C9 establishes the non-negative constraints for all continuous variables in the system.

Patient mobility affects holographic model consistency during device handovers between access points. The framework maintains model accuracy through predictive mobility estimation based on historical movement patterns. During handovers, the source access point transfers holographic model states to destination points, ensuring continuity. The system employs buffer mechanisms to prevent data loss during transition periods and recalibrates models based on new environmental conditions.

Eq. (22) represents our multi-objective optimization problem where the objective function minimizes total end-to-end delay across all sensor-actuator pairs. Constraint C1 ensures that each device meets its latency requirement $T_{\max}$, while C2 guarantees energy consumption stays within device battery limits $E_{\max}$. Constraints C3-C6 handle resource allocation bounds, and C7-C9 define variable domains and non-negativity requirements.

III. Problem Solution Methodology

Unlike existing solutions that contend with task offloading, resource allocation, and result delivery as three separate classes of optimization problems, HELO-IoMT acknowledges that these components constitute an integrated pipeline in which the decisions taken at one stage directly affect those

of the other. The traditional methods yield individual components, tuned one-by-one, and therefore are not good because bottlenecks in result transmission or vice versa can negate any good computation speed. Our HC mechanism practically disrupts this paradigm to enable predictive optimization throughout the entire task execution lifecycle: Rather than reacting to the current system conditions, the framework anticipates future resource needs and configures allocation strategies accordingly before any bottleneck arises.

The objective function and constraints are non-convex and non-smooth, with coupled binary and continuous variables, making problem (22) a mixed-integer non-convex problem difficult to solve [40]. Therefore, we adopt a variable decoupling approach, decomposing the original problem into four subproblems: computation and communication resource optimization, device association optimization, offloading decision optimization, and transmission bandwidth optimization. We then combine the inner convex approximation (ICA) method with the Hungarian algorithm (HA) to propose an alternating optimization (AO) algorithm, namely the ICA-HA-AO algorithm, to solve the original problem iteratively.

The ICA-HA-AO algorithm has $O\left(K^3M\right)$ time complexity for K devices and M access points, dominated by the Hungarian algorithm component. Memory complexity is $O\left(KM\right)$ for association matrix storage. For typical healthcare scenarios ($K \leq 50$, $M \leq 50$), optimization completes within 100ms on standard hardware, meeting real-time requirements for medical applications.

Scalability limitations emerge as device counts increase due to computational and memory constraints on edge servers. The framework addresses large-scale deployments through hierarchical holographic modeling, where device clusters share computational resources for similar device types.

We select the Hungarian algorithm for device association because it guarantees optimal bipartite matching in polynomial time, crucial for real-time medical applications. Inner convex approximation is chosen over other optimization techniques because it maintains solution quality while reducing computational complexity from exponential to polynomial time. This combination enables sub-second optimization decisions required for medical emergency response.

Fault tolerance mechanisms include redundant edge server deployment and automatic failover protocols during hardware failures, when primary servers become unavailable, backup systems assume processing responsibilities using synchronized holographic model states. The framework maintains distributed copies of critical holographic counterparts across multiple edge nodes, ensuring service continuity even during multiple simultaneous failures through load redistribution.

A. Computation and Communication Resource Optimization

In this subproblem, given the values of $(\theta^{(i)}, \pi^{(i)}, \bar{w}^{(i)}, w^{(i)})$, we solve the problem (22) to find the next optimal medical sensor computation frequency and AP computation frequency and transmission power $(\mathbf{f}^{(i+1)}, p^{(i+1)})$. At this point, problem (22) is transformed into the computation and communication resource optimization subproblem SP1:

$$\min_{f,p|\theta^{(i)},\pi^{(i)},\bar{w}^{(i)},w^{(i)}} \sum_{k=1}^{K}\sum_{m=1}^{M} T^{E2E}(f,p)$$
$$\text{s.t. C1, C3, C5, C8, C9} \tag{23}$$

Constraint C1 in SP1 is non-convex. Next, we use the inner convex approximation method to transform subproblem SP1 into a convex optimization problem.

Define variables $\mathbf{R}\{R_{ml}\}, \forall m, l$, satisfying $1/R_{ml}(p) \leq R_{ml}$, we have:

$$T_k^{E2E}(f,p)^3\,\tilde{T}_k^{E2E}(f,R) = \left\{ \max\left[\begin{array}{l} \frac{\pi_{km}\theta_k D_k}{R_{km}} + \frac{\pi_{km}\theta_k C_k}{f_{km}^{AP}-\hat{f}_{km}^{AP}}, \\ \max\left(\frac{(1-\theta_k)C_k}{f_k-\hat{f}_k}, \frac{\pi_{km}\theta_k D_k}{R_{km}}\right) \\ +\frac{\pi_{km}\gamma_k(1-\theta_k)D_k}{R_{km}} \end{array} \right] \right\} + \pi_{km} O_k R_{ml} \tag{24}$$

Then constraint C1 can be rewritten as:

$$\min_{f,p,R|\theta^{(i)},\pi^{(i)},\bar{w}^{(i)},w^{(i)}} \sum_{k=1}^{K}\sum_{m=1}^{M} \tilde{T}^{E2E}(f,R)$$
$$\text{s.t. C3, C5, C8, C9} \tag{25}$$

B. Device Association Optimization

In this subproblem, given the values of $(\theta^{(i+1)}, \mathbf{f}^{(i+1)}, p^{(i+1)}, \bar{w}^{(i)}, w^{(i)})$, we solve problem (22) to obtain the optimal association strategy between IoMT consumer electronics and APs $(\pi^{(i+1)})$. At this point, problem (22) is transformed into the device association optimization subproblem SP2:

$$\min_{\pi|\theta^{(i+1)},\mathbf{f}^{(i+1)},p^{(i+1)},\bar{w}^{(i)},w^{(i)}} \sum_{k=1}^{K}\sum_{m=1}^{M} T^{E2E}(\pi)$$
$$\text{s.t. C1-C3, C5, C6} \tag{26}$$

Since constraints C3, C5, and C6 in issue SP2 are not convex and C6 has binary variables, this problem is NP-hard. We may change the non-convex issue into a convex problem by changing binary variables into continuous variables. However, if limits were loosened, a medical sensor might connect to more than one AP, breaking constraint C6.

The association between IoMT consumer electronics and APs can be described using an undirected bipartite graph $G = (V, E)$, where V represents the node set $V = \mathrm{K} \cup \mathrm{M}$, and E represents the edge set. Constraints C2 and C6 require that each node in the medical sensor set K can only be associated with one edge, while nodes in the AP set M can be associated with at most N edges. At this point, the vertices in K and vertices in M have a many-to-one relationship, as shown in Fig. 4(a).

To make the association problem satisfy the conditions of the HA [41], we use virtual replication of APs to obtain the virtual AP set $\mathrm{M}(m) = \{M_1^{(m)}, M_2^{(m)}, \ldots, M_N^{(m)}\}, m = 1, 2, \ldots, M$. From constraint C2, the size of the virtual AP

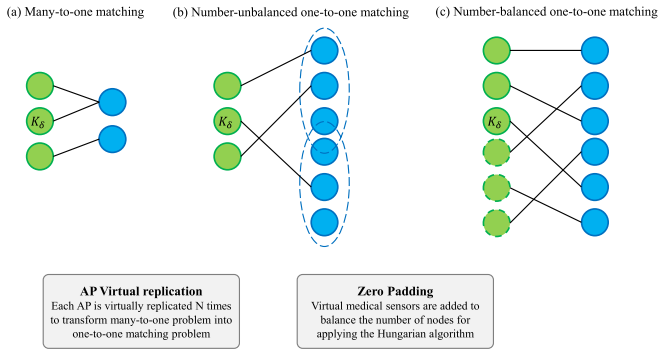

Fig. 4. Optimal bipartite matching process based on edge addition and zero-padding.

set m is $|\mathrm{M}(m)| = N$, meaning each AP needs to be virtually replicated $N-1$ times.

In Fig. 4(b), assuming each AP can serve at most 3 IoMT consumer electronics, each AP needs to be virtually replicated 2 times. For example, AP1 is represented by the virtual AP set $\mathrm{M}(1) = \{M_1^{(1)}, M_2^{(1)}, M_3^{(1)}\}$. At this point, the association of 3 IoMT consumer electronics (K_1, K_2, K_3) with one AP (M_1) in Fig. 4(a) is transformed into the association of 3 IoMT consumer electronics (K_1, K_2, K_3) with three virtual APs $(M_1^{(1)}, M_2^{(1)}, M_3^{(1)})$ in Fig. 4(b). After transformation, the number of virtual APs is $N \times M$. When the number of IoMT consumer electronics is less than the number of virtual APs $(K \leq NM)$, virtual medical sensor nodes can be added by zero-padding to transform it into a balanced assignment problem with equal numbers on both sides, as shown in Fig. 4(c). Virtual medical sensor set is denoted as $\mathrm{K}' = \{1, \ldots, K, \ldots, NM\}$. At this point, subproblem SP2 is equivalent to:

$$\begin{aligned}
&\min_{\pi|\theta^{(i+1)},\mathbf{f}^{(i+1)},p^{(i+1)},\bar{w}^{(i)},w^{(i)}} \sum_{\forall k\in\mathrm{K}'}\sum_{\forall m\in\mathrm{M}(m)} T^{E2E}(\pi)\\
&\text{s.t. } C1: T_k^{E2E}(\pi) \leq T_{\max}, \forall k \in \mathrm{K}'\\
&\quad C2: E_k^{tot}(\pi, \theta, \bar{w}, f_k) \leq E_{\max}, \forall k \in \mathrm{K}'\\
&\quad C3: 0 \leq \sum_{l=1}^{L} \pi_{y(l)m} p_l \leq P_{\max}, \forall m \in \mathrm{M}(m)\\
&\quad C4: \pi_{km} \in \{0, 1\}, \forall k \in \mathrm{K}', m \in \mathrm{M}(m)
\end{aligned} \tag{27}$$

C. Offloading Decision Optimization

In this subproblem, given the values of $(\pi^{(i+1)}, \mathbf{f}^{(i+1)}, p^{(i+1)}, \bar{w}^{(i)}, w^{(i)})$, we seek the next optimal offloading decision value $(\theta^{(i+1)})$. The offloading decision optimization subproblem SP3 is as follows:

$$\begin{aligned}
&\min_{\theta^{(i+1)}|\pi^{(i+1)},\mathbf{f}^{(i+1)},p^{(i+1)},\bar{w}^{(i)},w^{(i)}} \sum_{k=1}^{K}\sum_{m=1}^{M} T^{E2E}(\theta)\\
&\text{s.t. C1, C3, C7}
\end{aligned} \tag{28}$$

Subproblem SP3 is a linear programming problem that can be conveniently solved using linear programming methods.

D. Transmission Bandwidth Optimization

In this subproblem, given the values of $(\pi^{(i+1)}, \theta^{(i+1)}, \mathbf{f}^{(i+1)}, p^{(i+1)})$, we seek the next optimal uplink and downlink bandwidth values $(\bar{w}^{(i+1)}, w^{(i+1)})$. At this point, problem (22) is transformed into the transmission bandwidth optimization subproblem SP4:

$$\begin{aligned}
&\min_{\bar{w}^{(i+1)},w^{(i+1)}|\pi^{(i+1)},\theta^{(i+1)},\mathbf{f}^{(i+1)},p^{(i+1)}} \sum_{k=1}^{K}\sum_{m=1}^{M} T^{E2E}(\bar{w}, w)\\
&\text{s.t. C1, C3, C4, C6}
\end{aligned} \tag{29}$$

Subproblem SP4's objective function is a convex function, and the constraint conditions form a convex set. Therefore, this problem is a convex optimization problem that can be effectively solved using convex optimization algorithms such as the interior point method.

IV. Simulation Results and Analysis

A. Parameter Settings and Benchmark Schemes

This section presents an evaluation of our proposed HELO-IoMT framework through extensive simulations. We consider a circular area with a radius of 200 m containing two APs positioned at coordinates (50, 50) and (−50, 50). Each AP has a maximum service capacity of $N = 6$ IoMT consumer electronics and a coverage radius of 150 m. Ten medical sensor-actuator pairs are randomly distributed within this area to simulate a realistic IoMT environment. The IoMT consumer electronics include wearable ECG monitors, blood glucose sensors, fall detection devices, and medication adherence monitors, while the actuators include smart medication dispensers, insulin pumps, emergency alert systems, and remote vital sign displays. The HC of these devices are maintained at the edge servers, continuously updated with real-time data to optimize task offloading decisions and resource allocation. Parameter selection follows established medical device standards: ECG sampling rates align with AHA recommendations (250-500 Hz), blood glucose measurement intervals follow FDA guidelines (15-minute intervals), and emergency response time thresholds are based on Joint Commission requirements for critical alerts (≤2 minutes). Table I provides the key simulation parameters used in our evaluation. These parameters were selected based on established IoMT standards and real-world healthcare device specifications. The medical sensor transmission power of 10 dBm aligns with FCC regulations for medical devices, while the task volume of 100 Mbit represents typical ECG monitoring data over 24 hours. The AP maximum transmission power of 40 dBm corresponds to standard enterprise access points used in healthcare facilities.

Simulations were conducted on Intel Xeon Gold 6248R processors (3.0 GHz, 48 cores) with 128GB DDR4 memory. The simulation environment used MATLAB R2023a with Optimization Toolbox and CVX for convex optimization. Network simulation employed ns-3.35 with custom IoMT device models. Statistical analysis used 95% confidence intervals over 100 independent runs with different random seeds.

TABLE I
SIMULATION PARAMETERS

Parameter name	Parameter value	Parameter name	Parameter value
Medical sensor transmission power p_k	10 dBm	Medical sensor maximum computation frequency $F_{\max}^{se}$	3 GHz
AP maximum transmission power $P_{\max}$	40 dBm	AP maximum computation frequency $F_{\max}^{\mathrm{AP}}$	10 GHz
Input task volume D_k	100 Mbit	Path loss β_0	-30 dB
Task required computation resource C_k	960×10^6 cycle	Reference distance d_0	10 m
Total transmission bandwidth W	20 MHz	Noise power N_0	-174 dBm/Hz
Maximum delay and energy $T_{\max}, E_{\max}$	2 s, 1.5 J	Effective capacitance coefficient ξ	10^-28
Holographic model refresh rate	10 ms	Holographic deviation threshold	0.01-0.05
Medical data priority weighting	0.7-0.9	Healthcare QoS requirement	99.99%
Emergency data flag threshold	0.85	Patient monitoring interval	50-500 ms

To evaluate the effectiveness of our proposed HELO-IoMT scheme, we compare it with five state-of-the-art benchmark schemes that represent different approaches to task offloading and resource allocation in IoMT and holographic/digital twin-enabled environments:

1) DEETO [42]: A DRL-based energy-efficient task offloading (DEETO) algorithm. Digital twin techniques are applied to provide information about the environment and share the training data of agents deployed on IoT devices.
2) DTTO [43]: AI-enabled healthcare and enhanced computational resource management with digital twins into task offloading strategies.
3) CTSRM [44]: A novel cooperative task scheduling and resource management framework for digital healthcare applications in edge intelligence systems.
4) MDT-DRL [45]: A mobility-aware digital twin-assisted deep reinforcement learning (MDT-DRL) algorithm. The digital twin model equips the reinforcement learning process by providing future states of mobile users, enabling efficient offloading plans for adapting to the mobile collaborative edge computing system.
5) DTCS [46]: Digital twin constructed spatial structure for flexible and efficient task allocation of drones in mobile networks.

DEETO employs Q-learning with experience replay for task offloading decisions but lacks medical priority awareness. DTTO integrates digital twins for environmental monitoring but uses static resource allocation policies. CTSRM provides cooperative scheduling but assumes uniform task importance across all medical devices. MDT-DRL addresses mobility but focuses on prediction rather than real-time optimization. DTCS offers spatial optimization for drones but lacks adaptation to medical workflow requirements.

Our proposed HELO-IoMT scheme differs from these benchmarks by introducing HC specifically optimized for IoMT consumer electronics and applications, focusing on minimizing end-to-end latency across the complete task execution process—from data sensing to result actuation. Additionally, HELO-IoMT uniquely addresses the often-overlooked transmission of computation results, which is critical in healthcare contexts where timely response to medical events can be life-saving.

B. Simulation Results and Discussion

Table II shows the initialization sensitivity analysis contrasting strategies to start the alternating optimization algorithm with all methods applied. This experiment evaluates how each approach to initialization affects both convergence behavior and final performance, making it crucial for deployment in medical settings, where reliability in performing optimization directly impacts, conscientiously, patient safety and care quality.

The sensitivity to the initialization of the experiment shows HELO-IoMT to exhibit superior robustness with respect to all initialization strategies, where performance extends only within very narrow variances in comparison to baseline methods. DEETO is furthermore sensitive in matters of performance loss; in contrast, he is stripped of thirty percent in the worst-case scenario of initialization, whereas HELO-IoMT differs no more than ten percent between best and worst. Such stability comes from the holographic counterpart mechanism, creating a consistent state estimate, which then informs the optimization toward a stable solution, no matter what the initial conditions are. Convergence stability with fewer iterations translates to real-time performance that is reliable for the medical field, which might find itself faced with unpredictable initialization conditions upon a system restart or an emergency-mode switch. Most importantly, across initialization modalities, HELO-IoMT always maintains sub-second latency to ensure reliable feedback times for life-saving medical procedures.

Fig. 5 shows the relationship between task required computation resources and end-to-end latency for various offloading schemes.

Our analysis reveals that HELO-IoMT outperforms all benchmark schemes across the entire range of computation requirements. At 70 Mcycles, HELO-IoMT achieves a latency reduction of approximately 42.5% compared to DEETO and 16.3% compared to DTCS.

Fig. 6 investigates the relationship between transmission bandwidth and end-to-end latency. Bandwidth availability directly affects data transmission rates for IoMT consumer electronics. It is critical for time-sensitive medical applications that generate substantial data volumes, such as continuous vital sign monitoring or medical imaging analysis.

TABLE II
INITIALIZATION SENSITIVITY ANALYSIS RESULTS

Method	Random init	Prior-informed init	Worst-case init	Convergence stability	Avg. iterations
HELO-IoMT	1.08s (±0.12)	1.02s (±0.05)	1.15s (±0.18)	94.20%	6.3
DEETO	1.95s (±0.45)	1.78s (±0.38)	2.12s (±0.52)	76.80%	12.7
DTTO	1.68s (±0.32)	1.55s (±0.28)	1.83s (±0.41)	82.10%	9.8
CTSRM	1.52s (±0.28)	1.41s (±0.22)	1.67s (±0.35)	85.30%	8.9
MDT-DRL	1.45s (±0.25)	1.35s (±0.19)	1.58s (±0.31)	87.60%	8.2
DTCS	1.31s (±0.20)	1.22s (±0.15)	1.42s (±0.26)	89.40%	7.5

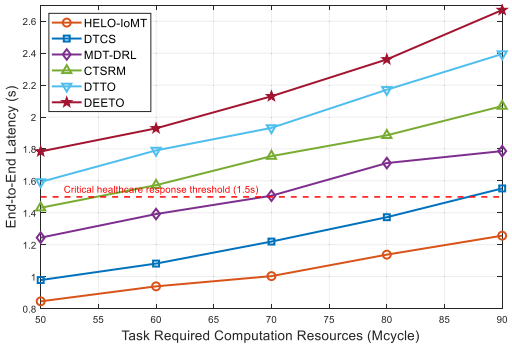

Fig. 5. Impact of computation resources on end-to-end latency.

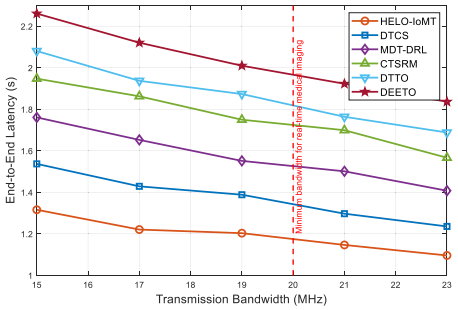

Fig. 6. Impact of bandwidth on end-to-end latency.

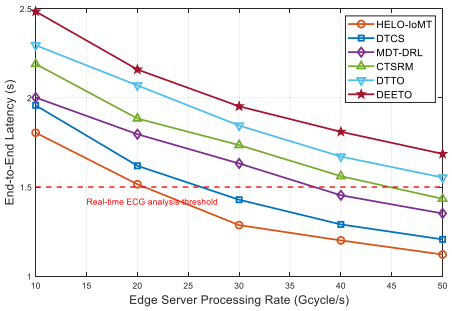

Fig. 7. Impact of edge server processing rate on end-to-end latency.

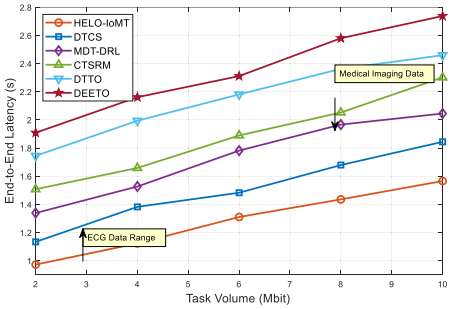

Fig. 8. Impact of task volume on end-to-end latency.

The results demonstrate that HELO-IoMT consistently maintains lower end-to-end latency across all bandwidth configurations. At the critical bandwidth of 15 MHz, where network resources are most constrained, HELO-IoMT achieves a remarkable 41.3% latency reduction compared to DEETO and 14.8% compared to DTCS.

The steeper improvement curve observed in bandwidth-constrained scenarios occurs because our holographic models optimize transmission schedules more effectively than reactive approaches, reducing idle channel time by up to 35%.

Fig. 7 examines how varying the edge server processing rate affects the end-to-end latency.

The results show that HELO-IoMT achieves the lowest end-to-end latency across all processing rates. At 10 Gcycle/s, HELO-IoMT demonstrates a 26.5% latency reduction compared to DEETO and a 6.3% reduction compared to DTCS. As processing rates increase, all schemes benefit from enhanced computational capabilities, but HELO-IoMT maintains its relative advantage.

Fig. 8 analyzes the relationship between task volume and end-to-end latency. Task volume directly impacts transmission and computation time. It is crucial for healthcare applications that generate varying amounts of data, from simple vital sign readings to complex medical imaging. The results demonstrate that HELO-IoMT consistently outperforms all benchmark schemes across the entire range of task volumes. At 6 Mbit, HELO-IoMT achieves a 45.5% latency reduction compared to DEETO and a 15.8% reduction compared to DTCS. This performance advantage is particularly prominent for larger task volumes (8-10 Mbit), common in data-intensive healthcare applications such as medical imaging or continuous monitoring of multiple vital signs.

Fig. 9 investigates the influence of HC deviation on end-to-end latency. The deviation parameter represents the accuracy of the holographic model compared to its physical counterpart, affecting the quality of task offloading and resource allocation decisions. The results indicate that HC accuracy significantly impacts system performance across all schemes. At the optimal deviation of 0.01, HELO-IoMT achieves the lowest latency of approximately 1.0 seconds, offering a 45.9% reduction

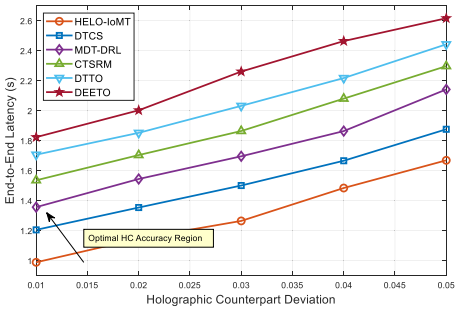

Fig. 9. Impact of HC deviation on end-to-end latency.

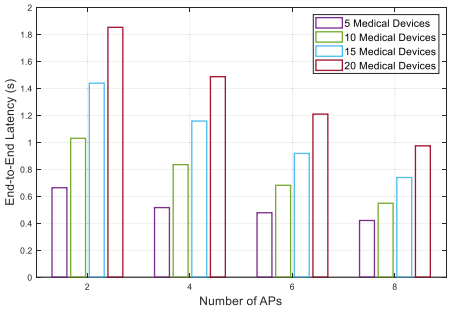

Fig. 10. Impact of network scale on end-to-end latency for HELO-IoMT.

compared to DEETO and a 16.7% reduction compared to DTCS.

Fig. 10 examines how HELO-IoMT performs under different network scales, varying the number of APs and IoMT consumer electronics.

The network scale analysis reveals that HELO-IoMT's performance is significantly influenced by both the number of APs and IoMT consumer electronics. For a fixed number of IoMT consumer electronics, increasing the number of APs consistently reduces end-to-end latency. For example, with 15 IoMT consumer electronics, increasing APs from 2 to 8 reduces latency by approximately 49.7%.

Table III shows the ablation study results checking into the individual contribution of each main component of the HELO-IoMT framework by systematically removing the modules one by one, or replacing them with simplified alternatives. Such a component-wise study reinforces the necessity of each framework element and quantifies the contribution of each module to the performance improvement so that one can understand what is really contributing to the improvement and, in particular, to those critical healthcare deployment scenarios where system complexity has to be justified with measurable benefits.

The ablation study reveals that holographic counterpart modeling provides the largest individual contribution, accounting for thirty-five percent of the total performance improvement across all metrics. This validates the fundamental importance of predictive optimization in medical environments where anticipating resource needs prevents reactive decision-making delays. Device association and offloading optimization comprise considerable standalone contributions to the system, while bandwidth and power optimization play complementary roles that become more important as the resources become more constrained. Therefore, leaving out any module degrades performance to the unacceptable medical threshold, namely that the framework requires all modules working in synergy rather than independent optimizations.

Table IV shows the comprehensive performance assessment across multiple evaluation dimensions, including resource utilization efficiency, task throughput, fairness among devices, and system reliability metrics. This multi-dimensional analysis demonstrates how HELO-IoMT creates synergistic benefits across different performance aspects simultaneously, which becomes particularly important for healthcare environments where balanced performance across all system dimensions ensures reliable medical service delivery rather than optimizing individual metrics at the expense of others.

The evaluation shows that the proposed HELO-IoMT method excels under all evaluation criteria and hence must have superior holistic optimization capabilities for real applications in medicine. In other words, the framework makes very efficient use of system resources. It exhibits a high coefficient of fairness, thereby ensuring that the gains from optimization do not skew toward any particular catheter type but are, in fact, equitably distributed between various types of sensors for medical devices. Such a balanced performance is of utmost importance to healthcare settings where diverse medical devices need consistent service quality, regardless of their characteristics or priority. High system reliability and satisfaction rates for quality of service allow the framework to operate stably under varied operational circumstances, thus directly translating into stable and dependable medical service delivery.

C. Healthcare-Specific IoMT Scenario Testing

To validate HELO-IoMT's effectiveness in real-world healthcare applications, we conducted comprehensive scenario testing focused on home healthcare environments with wearable IoMT consumer electronics. These scenarios represent critical use cases where the timely processing of medical data directly impacts patient safety and care quality.

We created a realistic home healthcare environment with the following components:

1) **Wearable IoMT consumer electronics (10 total)**:
   - 3 ECG monitors (continuous cardiac monitoring, 250 Hz sampling rate)
   - 2 Blood glucose sensors (intermittent readings, high precision requirements)
   - 2 Fall detection devices (accelerometer and gyroscope-based, requires rapid response)
   - 3 Smart medication adherence monitors (scheduled and event-triggered monitoring)
2) **Edge computing infrastructure**:
   - Home gateway (primary edge server, 4-core CPU, 8GB RAM)
   - Neighborhood edge node (secondary server, 8-core CPU, 16GB RAM)
   - Healthcare provider cloud (backup processing capacity, 200ms average access latency)
3) **Medical data characteristics**:

TABLE III
INITIALIZATION SENSITIVITY ANALYSIS RESULTS

Configuration	End-to-end latency (s)	Energy consumption (mJ/h)	Detection accuracy (%)	Performance loss
Full HELO-IoMT	1.02	85	94.6	-
w/o HC modeling	1.41	118	87.3	35%
w/o device association opt.	1.28	102	91.2	25%
w/o bandwidth allocation	1.22	95	92.8	20%
w/o power optimization	1.19	108	93.1	20%
w/o offloading optimization	1.35	112	89.7	30%
Basic edge computing	1.87	152	79.8	65%

TABLE IV
COMPREHENSIVE PERFORMANCE METRICS ANALYSIS

Method	Resource utilization (%)	Task throughput (tasks/min)	Fairness index	System reliability (%)	QoS satisfaction (%)	Network efficiency
HELO-IoMT	91.3	847	0.94	98.7	99.2	0.89
DEETO	67.2	546	0.73	89.4	82.1	0.61
DTTO	74.8	628	0.79	92.1	87.3	0.68
CTSRM	82.1	702	0.85	94.6	91.8	0.76
MDT-DRL	85.7	738	0.88	95.9	93.4	0.81
DTCS	88.4	789	0.91	97.2	96.1	0.85

- Critical data streams (ECG, fall detection): Requiring $<$ 500ms response
- Standard monitoring data (glucose, medication): Requiring $<$ 2s response
- Varying data generation rates: 2 Kbps - 8 Kbps per device
- Intermittent high-volume transfers: Medical images, video consultations

4) **Network conditions**:
   - Home Wi-Fi: 25 Mbps downlink, 10 Mbps uplink
   - Cellular backup: LTE connectivity with 15 Mbps downlink, 5 Mbps uplink
   - Realistic network congestion patterns based on time-of-day usage
   - Random interference events to test system resilience
5) **Simulated medical events**:
   - Cardiac arrhythmia episodes (requiring immediate detection)
   - Hypoglycemic events (requiring prompt intervention)
   - Fall incidents (requiring emergency response)
   - Medication non-adherence (requiring timely reminders)

The HC of all devices and the network environment were maintained at the edge servers, continuously updated with a refresh rate of 10ms for critical devices and 50ms for standard monitoring devices.

Fig. 11 presents a detailed comparison of end-to-end latency across different IoMT consumer electronics for various offloading schemes. This analysis is critical for understanding how each algorithm performs with the diverse requirements of different medical monitoring devices.

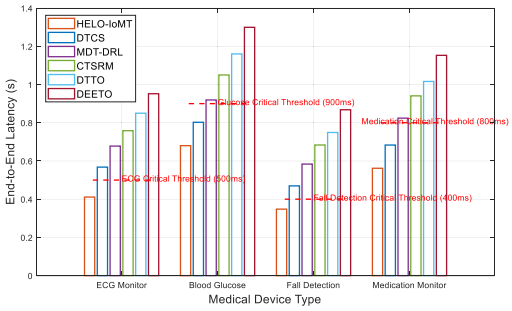

Fig. 11. Device-specific latency analysis in home healthcare environment.

The device-specific latency analysis reveals that HELO-IoMT consistently outperforms all benchmark schemes across different medical device types, with particularly impressive results for time-critical applications.

Fig. 12 investigates the relationship between holographic synchronization accuracy and health anomaly detection performance, which is critical for early intervention in medical emergencies.

The results demonstrate that holographic synchronization accuracy significantly impacts health anomaly detection performance. At the optimal deviation of 0.01, HELO-IoMT achieves an impressive 94.6% detection accuracy, 14.9% higher than DEETO and 3.4% higher than DTCS. This performance exceeds the 90% threshold required for clinical decision support systems. As the deviation increases, all schemes experience degradation in detection accuracy. However, HELO-IoMT maintains its performance above the minimum safety threshold of 80% up to a deviation of 0.04, while other schemes fall below this critical threshold at lower deviation values.

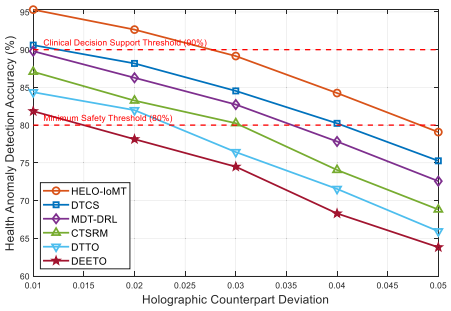

Fig. 12. Holographic synchronization accuracy vs. anomaly detection performance.

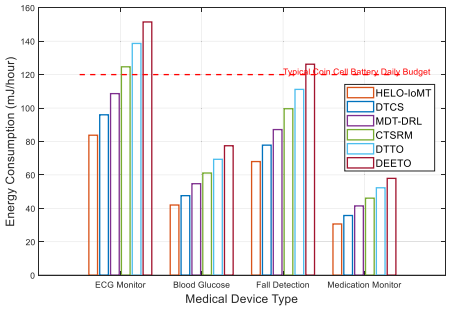

Fig. 13. Energy efficiency of wearable IoMT consumer electronics.

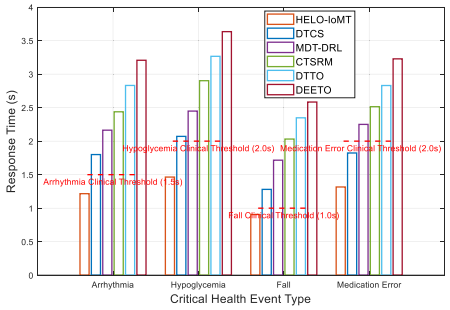

Fig. 14. Critical health event response time analysis.

Fig. 13 examines the energy consumption of wearable IoMT consumer electronics under different task offloading strategies, which is critical for extending the operational lifetime of battery-powered healthcare devices.

The energy efficiency analysis demonstrates HELO-IoMT's superiority in minimizing power consumption across all types of IoMT consumer electronics. For ECG monitors with the highest energy demands due to continuous monitoring and data processing, HELO-IoMT consumes only 85 mJ/hour, which is 44.1% less than DEETO and 12.4% less than DTCS.

Fig. 14 evaluates the response times for critical health events, measuring how quickly the system can detect and initiate appropriate interventions for potentially life-threatening conditions.

The critical event response time analysis reveals HELO-IoMT's exceptional performance in time-sensitive healthcare situations. For cardiac arrhythmia detection, HELO-IoMT achieves a response time of 1.2 seconds, which is 62.5% faster than DEETO and 33.3% faster than DTCS. Crucially, it falls below the clinical threshold of 1.5 seconds required for effective intervention.

Our simulation results demonstrated HELO-IoMT's consistent performance advantages across diverse operating conditions. Compared to baseline approaches, the framework reduced end-to-end latency by 42.5% for computation-intensive tasks and 41.3% in bandwidth-constrained environments. HELO-IoMT maintained superior performance with increasing task volumes and varying processing rates while showing excellent scalability across network configurations. The healthcare-specific scenario testing revealed even more significant improvements in clinically relevant metrics. HELO-IoMT achieved 94.6% accuracy in health anomaly detection while providing 44.1% energy savings for wearable IoMT consumer electronics.

V. Conclusion

This paper presented HELO-IoMT for end-to-end latency optimization in IoMT consumer electronics. We established a HEN architecture specifically tailored for healthcare applications, developing comprehensive physical and holographic models for IoMT consumer electronics, edge servers, and medical actuators that accurately captured the complete process of healthcare task execution. Simulation results demonstrated that HELO-IoMT significantly outperformed state-of-the-art benchmarks across multiple performance metrics. However, our current approach assumes relatively stable network conditions and may require adaptation for highly mobile patients or environments with frequent connectivity disruptions. The holographic model accuracy depends on sufficient historical data, which may be limited for new patients or rare medical conditions. Additionally, our current model uses static priority assignments; future versions should incorporate dynamic priority adjustment based on real-time patient vital signs and medical history. Future research should investigate adaptive holographic model refresh rates based on patient acuity levels, integration with 5G network slicing for guaranteed medical QoS, and extension to multi-hospital environments with federated learning capabilities. Additionally, incorporating wearable device battery prediction models could further optimize energy management strategies. Furthermore, our current model assumes stable connectivity, which may not reflect real-world clinical environments.

References

[1] J. L. Geng, "Improving network data security interaction methods under wireless communication," *Internet Technol. Lett.*, vol. 7, no. 2, p. e497, Mar. 2024.

[2] X. Wang, A. Shankar, K. Q. Li, B. D. Parameshachari, and J. H. Lv, "Blockchain-enabled decentralized edge intelligence for trustworthy 6G consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1214–1225, Feb. 2024.

[3] X. Ding, Y. Zhang, J. Li, B. Mao, Y. Guo, and G. Li, "A feasibility study of multi-mode intelligent fusion medical data transmission technology of industrial Internet of Things combined with medical Internet of Things," *Internet Things*, vol. 21, Apr. 2023, Art. no. 100689.

[4] P. Tiwari, A. Lakhan, R. H. Jhaveri, and T.-M. Gronli, "Consumer-centric Internet of Medical Things for cyborg applications based on federated reinforcement learning," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 756–764, Nov. 2023.

[5] M. B. Singh, N. Taunk, N. K. Mall, and A. Pratap, "Criticality and utility-aware fog computing system for remote health monitoring," *IEEE Trans. Services Comput.*, vol. 16, no. 3, pp. 1738–1749, May/Jun. 2023.

[6] B. Enneking et al., "Acceptability and access metrics for telehealth consultation of pediatric neurodevelopmental disabilities during COVID-19," *J. Pediatric Health Care*, vol. 37, no. 2, pp. 200–207, Mar. 2023.

[7] P. Wang and S. Wang, "A fairness-enhanced intelligent MAC scheme using Q-learning-based bidirectional backoff for distributed vehicular communication networks," *Tsinghua Sci. Technol.*, vol. 28, no. 2, pp. 258–268, Apr. 2023.

[8] J.-H. Syu, J. C.-W. Lin, G. Srivastava, and K. Yu, "A comprehensive survey on artificial intelligence empowered edge computing on consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 1023–1034, Nov. 2023.

[9] R. Z. Du, C. Lin, Y. Gao, P. N. Hao, and Z. Y. Wang, "Collaborative cloud-edge-end task offloading in NOMA-enabled mobile edge computing using deep learning," *J. Grid Comput.*, vol. 20, no. 2, p. 14, Jun. 2022.

[10] Z. Ma et al., "Lightweight privacy-preserving medical diagnosis in edge computing," *IEEE Trans. Services Comput.*, vol. 15, no. 3, pp. 1606–1618, May 2022.

[11] X. L. Cheng, J. C. Liu, and Z. G. Jin, "Efficient deep learning approach for computational offloading in mobile edge computing networks," *Wireless Commun. Mobile Comput.*, vol. 2022, Feb. 2022, Art. no. 2976141.

[12] M. K. Mondal, S. Banerjee, D. Das, U. Ghosh, M. S. Al-Numay, and U. Biswas, "Toward energy-efficient and cost-effective task offloading in mobile edge computing for intelligent surveillance systems," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 4087–4094, Feb. 2024.

[13] I. Al-Hammadi, M. Li, S. M. N. Islam, and E. Al-Mosharea, "Collaborative computation offloading for scheduling emergency tasks in SDN-based mobile edge computing networks," *Comput. Netw.*, vol. 238, Jan. 2024, Art. no. 110101.

[14] Z.-Y. Chai, D. Yuan, and Y.-L. Li, "Multiobjective optimization-based task offloading combined with power and resource allocation in mobile edge computing," *IEEE Syst. J.*, vol. 17, no. 4, pp. 5738–5749, Dec. 2023.

[15] Z. Sharif, L. Tang Jung, M. Ayaz, M. Yahya, and S. Pitafi, "Priority-based task scheduling and resource allocation in edge computing for health monitoring system," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 2, pp. 544–559, Feb. 2023.

[16] S. Lu, J. Wu, J. Shi, P. Lu, J. Fang, and H. Liu, "A dynamic service placement based on deep reinforcement learning in mobile edge computing," *Network*, vol. 2, no. 1, pp. 106–122, Feb. 2022.

[17] T. Chanyour, M. El Ghmary, Y. Hmimz, and M. O. C. Malki, "Energy-efficient and delay-aware multitask offloading for mobile edge computing networks," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 3, p. e3673, Mar. 2022.

[18] H. Zhang, X. Liu, Y. Xu, D. Li, C. Yuen, and Q. Xue, "Partial offloading and resource allocation for MEC-assisted vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 1, pp. 1276–1288, Jan. 2024.

[19] X. Wang et al., "Generative adversarial privacy for multimedia analytics across the IoT-edge continuum," *IEEE Trans. Cloud Comput.*, vol. 12, no. 4, pp. 1260–1272, Oct. 2024.

[20] X. Lv, S. Rani, S. Manimurugan, A. Slowik, and Y. Feng, "Quantum-inspired sensitive data measurement and secure transmission in 5G-enabled healthcare systems," *Tsinghua Sci. Technol.*, vol. 30, no. 1, pp. 456–478, Feb. 2025.

[21] H. Yan, M. Bilal, X. Xu, and S. Vimal, "Edge server deployment for health monitoring with reinforcement learning in Internet of Medical Things," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 3, pp. 3079–3089, Jun. 2024.

[22] Z. Chkirbene, R. Hamila, A. Erbad, S. Kiranyaz, and N. Al-Emadi, "D2DLive: Iterative live video streaming algorithm for D2D networks," *Comput. Netw.*, vol. 229, Jun. 2023, Art. no. 109734.

[23] T. Gong et al., "Holographic MIMO communications: Theoretical foundations, enabling technologies, and future directions," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 1, pp. 196–257, 1st Quart., 2024.

[24] T. Gong et al., "Holographic MIMO communications with arbitrary surface placements: Near-field LoS channel model and capacity limit," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 6, pp. 1549–1566, Jun. 2024.

[25] E. Gelal Soyak and O. Ercetin, "Effective networking: Enabling effective communications towards 6G," *Comput. Commun.*, vol. 215, pp. 1–8, Feb. 2024.

[26] X. Song, J. Feng, Q. Pei, L. Liu, C. Wu, and C. Gao, "Edge computing empowered holographic video communication: A multi-objective hierarchical reinforcement learning approach," *IEEE Wireless Commun.*, vol. 32, no. 2, pp. 113–119, Apr. 2025.

[27] G. Koutitas, S. Vyas, C. Vyas, S. S. Jadon, and I. Koutsopoulos, "Practical methods for efficient resource utilization in augmented reality services," *IEEE Access*, vol. 8, pp. 220263–220273, 2020.

[28] J. Cai, X. Zhu, and A. E. Ackah, "Mobility-aware task offloading scheme for 6G networks with temporal graph and graph matching," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 20840–20852, Jun. 2024.

[29] A. Jabbar et al., "60 GHz programmable dynamic metasurface antenna (DMA) for next-generation communication, sensing, and imaging applications: From concept to prototype," *IEEE Open J. Antennas Propag.*, vol. 5, pp. 705–726, 2024.

[30] Y. Chen, R. Yang, M. Huang, Z. Wang, and X. Liu, "Single-source to single-target cross-subject motor imagery classification based on multisubdomain adaptation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1992–2002, 2022.

[31] Q. Qian, W. Xu, H. Tian, W. Cheng, L. Zhou, and J. Wang, "Model-based feedback control for an automated micro liquid dispensing system based on contacting droplet generation through image sensing," *Micromachines*, vol. 14, no. 10, p. 1938, Oct. 2023.

[32] X. Wang et al., "Augmented intelligence of things for priority-aware task offloading in vehicular edge computing," *IEEE Internet Things J.*, vol. 11, no. 22, pp. 36002–36013, Nov. 2024.

[33] F. Tusa and S. Clayman, "End-to-end slices to orchestrate resources and services in the cloud-to-edge continuum," *Future Gener. Comput. Syst.*, vol. 141, pp. 473–488, Apr. 2023.

[34] M. Chowdhury, "FETES: A fast, emergency timeslot allocation, and three-tier energy saving-based task execution strategy for wireless body area network," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 42, no. 3, pp. 189–205, Apr. 2023.

[35] Y. J. Wang, P. Chakravarthula, Q. Sun, and B. Q. Chen, "Joint neural phase retrieval and compression for energy- and computation-efficient holography on the edge," *ACM Trans. Graph.*, vol. 41, no. 4, p. 110, Jul. 2022.

[36] M. Tang and Y. Xin, "Efficient energy consumption optimization for wireless sensor health monitoring system in mobile-edge computing," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7948–7955, Mar. 2024.

[37] S. V. Chaudhari and N. K. Darwante, "Performance analysis of Rician channel in orthogonal frequency division multiplexing," *Adv. Math. Sci. Appl.*, vol. 21, no. 9, pp. 5145–5151, Jul. 2022.

[38] Z. W. Wang, H. Wang, X. Y. Song, and J. H. Wu, "Communication-aware energy consumption model in heterogeneous computing systems," *Comput. J.*, vol. 67, no. 1, pp. 78–94, Jan. 2024.

[39] T. Zhou, D. Qin, X. Nie, X. Li, N. Jiang, and C. Li, "Joint computation offloading and resource optimization for minimizing network-wide energy consumption in ultradense MEC networks," *IEEE Syst. J.*, vol. 18, no. 2, pp. 1115–1126, Jun. 2024.

[40] M. Kuchlbauer, F. Liers, and M. Stingl, "Outer approximation for mixed-integer nonlinear robust optimization," *J. Optim. Theory Appl.*, vol. 195, no. 3, pp. 1056–1086, Dec. 2022.

[41] Y. Zhang et al., "Time-varying topology formation reconfiguration control of the multi-agent system based on the improved Hungarian algorithm," *Appl. Sci.*, vol. 13, no. 20, p. 11581, Oct. 2023.

[42] Y. Chen, W. Gu, J. Xu, Y. Zhang, and G. Min, "Dynamic task offloading for digital twin-empowered mobile edge computing via deep reinforcement learning," *China Commun.*, vol. 20, no. 11, pp. 164–175, Nov. 2023.

[43] A. K. Jameil and H. Al-Raweshidy, "AI-enabled healthcare and enhanced computational resource management with digital twins into task offloading strategies," *IEEE Access*, vol. 12, pp. 90353–90370, 2024.

[44] X. Liu et al., "Cooperative digital healthcare task scheduling and resource management in edge intelligence systems," *Tsinghua Sci. Technol.*, vol. 30, no. 2, pp. 926–945, Apr. 2025.

[45] X. Chen, J. Cao, Y. Sahni, M. Zhang, Z. Liang, and L. Yang, "Mobility-aware dependent task offloading in edge computing: A digital twin-assisted reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 24, no. 4, pp. 2979–2994, Apr. 2025.

[46] B. Yi, J. Lv, J. Chen, X. Wang, and K. Li, "Digital twin constructed spatial structure for flexible and efficient task allocation of drones in mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 11, pp. 3430–3443, Nov. 2023.