# Agentic AI for Low-Altitude Semantic Wireless Networks: An Energy Efficient Design

Zhouxiang Zhao, Ran Yi, Yihan Cang, Boyang Jin, Zhaohui Yang, Mingzhe Chen, *Senior Member, IEEE,*
Chongwen Huang, and Zhaoyang Zhang, *Senior Member, IEEE*

***Abstract*—This letter addresses the energy efficiency issue in unmanned aerial vehicle (UAV)-assisted autonomous systems. We propose a framework for an agentic artificial intelligence (AI)-powered low-altitude semantic wireless network, that intelligently orchestrates a sense-communicate-decide-control workflow. A system-wide energy consumption minimization problem is formulated to enhance mission endurance. This problem holistically optimizes key operational variables, including UAV's location, semantic compression ratio, transmit power of the UAV and a mobile base station, and binary decision for AI inference task offloading, under stringent latency and quality-of-service constraints. To tackle the formulated mixed-integer non-convex problem, we develop a low-complexity algorithm which can obtain the globally optimal solution with two-dimensional search. Simulation results validate the effectiveness of our proposed design, demonstrating significant reductions in total energy consumption compared to conventional baseline approaches.**

***Index Terms*—Agentic AI, semantic communications, low-altitude intelligence, energy efficiency.**

## I. Introduction

DRIVEN by the advancements in artificial intelligence (AI), wireless communications, and low-altitude intelligence, autonomous systems featuring unmanned aerial vehicles (UAVs) are becoming integral to a myriad of applications, including intelligent surveillance, disaster response, and logistics [1], [2]. These systems typically operate in a closed loop of sensing, communication, computation, and action. A critical bottleneck in such UAV-assisted networks is the transmission of high-dimensional sensor data, such as real-time video streams, from the UAV to a ground decision-making entity [3]. This process is exceedingly demanding on both bandwidth and UAV's limited energy resources, posing a significant challenge to mission endurance and scalability.

To overcome this limitation, semantic communication has emerged as a transformative paradigm [4]–[6]. Unlike traditional communication systems that focus on bit-level fidelity, semantic communication aims to extract and transmit only the essential, task-relevant meaning embedded within the source data. This approach can drastically reduce the volume of transmit data, thereby enhancing both spectral and energy efficiency. Concurrently, the rise of agentic AI, mostly powered by large AI models, provides the "brain" for these autonomous systems, enabling sophisticated reasoning and decision-making based on the received semantic information [7]. The integration of semantic communication with agentic AI thus paves the way for highly intelligent and efficient low-altitude wireless networks [8].

While the foundational concepts of semantic communication and AI-driven control have been explored, a holistic understanding of the system-level design and resource allocation remains a significant challenge. Prior works have focused on either the semantic codec design [9] or AI task offloading [10], without jointly considering the intricate interplay between the level of semantic compression, physical deployment of the UAV, and strategic offloading of AI inference tasks. For instance, deeper semantic compression reduces transmission energy but increases on-board computation energy. Similarly, offloading the AI inference task from the edge to a powerful cloud server saves local resources but incurs additional communication latency and energy costs. A comprehensive framework that jointly optimizes these highly-coupled variables to enhance the overall system efficiency is therefore urgently needed.

In this letter, we address this gap by designing an energy-efficient agentic AI-assisted semantic wireless network. We consider a collaborative system where a UAV performs semantic sensing, a mobile base station (BS) acts as an edge intelligence hub, and a robot executes control commands. The main contributions of this work are summarized as follows:

- We propose a framework for an agentic AI-assisted low-altitude semantic wireless network that intelligently orchestrates a sense-communicate-decide-act workflow.
- We formulate a system-wide energy consumption minimization problem that jointly optimizes the UAV's three-dimensional (3D) location, the semantic compression ratio, the transmit powers of the UAV and BS, and the binary decision of AI inference task offloading (edge vs. cloud), under latency and quality-of-service (QoS) constraints. To solve the formulated non-convex problem, we develop an efficient algorithm which can find the globally optimal solution.
- Simulation results validate the superiority of the proposed framework in significantly reducing the system's energy consumption compared to baseline approaches.

## II. System Model and Problem Formulation

### A. Network Model

Consider an agentic AI-assisted semantic wireless network comprising a single UAV, a mobile BS, a cloud server, and a robot, as depicted in Fig. 1. The system operates in a low-altitude environment where the UAV is tasked with tracking a moving target. The UAV senses information about the target,

Zhouxiang Zhao, Ran Yi, Boyang Jin, Zhaohui Yang, Chongwen Huang, and Zhaoyang Zhang are with the College of Information Science and Electronic Engineering, Zhejiang University, and also with Zhejiang Provincial Key Laboratory of Info. Proc., Commun. & Netw. (IPCAN), Hangzhou 310027, China (e-mails: {zhouxiangzhao, ranyi, byjin10225, yang_zhaohui, chongwenhuang, ning_ming}@zju.edu.cn).

Yihan Cang is with National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China (e-mail: yhcang@seu.edu.cn).

Mingzhe Chen is with Department of Electrical and Computer Engineering and Institute for Data Science and Computing, University of Miami, Coral Gables, FL 33146, USA (e-mail: mingzhe.chen@miami.edu).

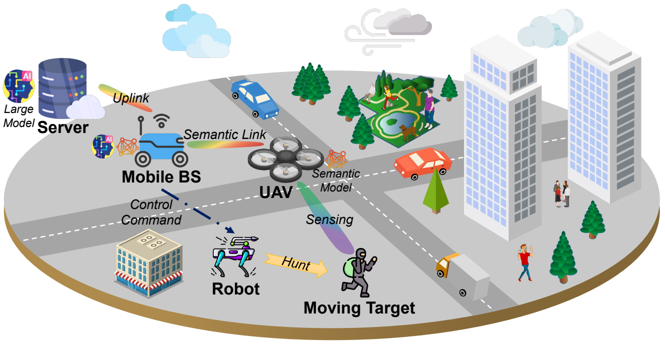

Fig. 1. An illustration of the agentic AI-assisted semantic wireless network.

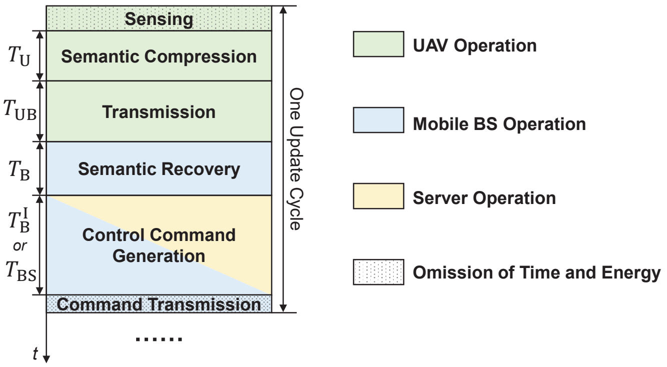

Fig. 2. The operational workflow of the system within a single update cycle.

which is then semantically compressed and transmitted to the mobile BS. We consider a mobile BS to leverage its mobility, thereby maintaining proximity to the UAV and ensuring a robust communication link. Upon receiving the semantic information, the mobile BS first reconstructs the original data using a paired semantic model. Subsequently, it can employ an embodied large AI model to make intelligent decisions and generate control commands based on the reconstructed information. Alternatively, the mobile BS can offload this task by transmitting the reconstructed data to a cloud server, which possesses superior computational resources, for command generation. The final control command is then transmitted to the robot to execute a specific task.

The operational workflow of the system is illustrated in Fig. 2. We define an update cycle as the whole process of handling the data collected in one sensing interval. Each cycle involves several distinct stages: UAV sensing, on-board semantic compression, data transmission to the mobile BS, semantic recovery at the BS, control command generation, and command transmission to the robot. The control command generation can be performed either locally at the mobile BS or remotely at the cloud server. Note that the time and energy for UAV sensing and the final command transmission to the robot are neglected, as they are assumed to be constant or negligible compared to the other components.

For analytical tractability, we establish a 3D Cartesian coordinate system where the mobile BS is located at the origin, i.e., $(0,0,0)$. The moving target's location in current time slot is denoted by $(x_\mathrm{T}, y_\mathrm{T}, 0)$. The altitudes of both the mobile BS and the target are considered negligible relative to the UAV's operational altitude. The UAV's location is given by

$$\mathbf{L}_\mathrm{U} = (x_\mathrm{U}, y_\mathrm{U}, H_\mathrm{U}), \tag{1}$$

where $H_\mathrm{U}$ is the UAV's altitude.[1] Consequently, the distance between the UAV and the mobile BS is $d_\mathrm{UB} = \sqrt{x_\mathrm{U}^2 + y_\mathrm{U}^2 + H_\mathrm{U}^2}$, and the distance between the UAV and the moving target is

$$d_\mathrm{UT} = \sqrt{(x_\mathrm{U} - x_\mathrm{T})^2 + (y_\mathrm{U} - y_\mathrm{T})^2 + H_\mathrm{U}^2}. \tag{2}$$

### *B. UAV Sensing, Computation, and Communication Models*

Each update cycle commences with the UAV sensing the target, typically through video capture. The QoS for this sensing task is modeled as an exponentially decaying function of the sensing distance [11]:

$$q = e^{-\xi d_\mathrm{UT}}, \tag{3}$$

where $\xi$ is a parameter characterizing the sensing performance.

Let the raw data captured by the UAV be $\mathcal{D}$, with a total size of $D$ bits. This data undergoes semantic compression. We define the semantic compression ratio as $\rho = C/D$, where $C$ is the size of the compressed data. The computational overhead for compression with respect to $\rho$ is modeled as [12]

$$O(\rho) = -\kappa_1 D \ln(\rho), \tag{4}$$

where $\kappa_1$ is the coefficient representing the computation efficiency of the semantic compression algorithm. The latency required for semantic compression is

$$T_\mathrm{U} = \frac{-\kappa_1 D \ln(\rho)}{f_\mathrm{U}}, \tag{5}$$

where $f_\mathrm{U}$ is the computational capacity (in cycles/s) of the UAV's on-board processor. The corresponding energy consumption for this computation is

$$E_\mathrm{U} = \tau_\mathrm{U} f_\mathrm{U}^3 T_\mathrm{U}, \tag{6}$$

where $\tau_\mathrm{U}$ is the effective switched capacitance coefficient of the UAV's processor.

After compression, the data is transmitted to the mobile BS. We assume the communication channel is dominated by the line-of-sight (LoS) path, which is a reasonable assumption for outdoor, low-altitude scenarios.[2] The channel gain between the UAV and the BS is therefore

$$G(\mathbf{L}_\mathrm{U}) = \frac{G_0}{d_\mathrm{UB}^\alpha}, \tag{7}$$

where $G_0$ is the reference channel gain at a distance of $1$ m, and $\alpha$ is the path-loss exponent. Using the Shannon capacity formula, the transmission latency is

$$T_\mathrm{UB} = \frac{D\rho}{B_\mathrm{U} \log_2\left(1 + \frac{p_\mathrm{U} G(\mathbf{L}_\mathrm{U})}{B_\mathrm{U} N_0}\right)}, \tag{8}$$

[1] The locations of all network components are assumed to be static within a single update cycle due to its short duration.

[2] Specifically, we model a rural environment with a path-loss exponent of $\alpha = 2$ and assume that signal power from antenna sidelobes is negligible.

where $B_{\mathrm{U}}$ is the channel bandwidth, $p_{\mathrm{U}}$ is the UAV's transmit power, and $N_0$ is the noise power spectral density. The energy consumed during this transmission is

$$E_{\mathrm{UB}} = p_{\mathrm{U}} T_{\mathrm{UB}}. \tag{9}$$

Upon reception, the mobile BS performs semantic recovery to reconstruct the original data. The time and energy for this process are modeled symmetrically to the compression phase:

$$T_{\mathrm{B}} = \frac{-\kappa_2 D \ln(\rho)}{f_{\mathrm{B}}}, \tag{10}$$

where $\kappa_2$ and $f_{\mathrm{B}}$ are the semantic recovery efficiency parameter and the computational capacity of the mobile BS, respectively. The energy consumption for semantic recovery is

$$E_{\mathrm{B}} = \tau_{\mathrm{B}} f_{\mathrm{B}}^3 T_{\mathrm{B}}. \tag{11}$$

where $\tau_{\mathrm{B}}$ is the effective switched capacitance coefficient of the mobile BS's processor.

### C. Control Command Generation Model

Based on the recovered information, a control command for the robot is generated. This can occur either at the mobile BS or be offloaded to the cloud server.

*1) Case 1 (Generate at Mobile BS):* The mobile BS uses its local large AI model for inference. We model the inference latency as being proportional to the original data size:

$$T_{\mathrm{B}}^{\mathrm{I}} = \frac{\kappa_3 D}{f_{\mathrm{B}}}, \tag{12}$$

where $\kappa_3$ is the inference computation efficiency parameter. The energy consumed for task inference by the mobile BS is

$$E_{\mathrm{B}}^{\mathrm{I}} = \tau_{\mathrm{B}} f_{\mathrm{B}}^3 T_{\mathrm{B}}^{\mathrm{I}}. \tag{13}$$

*2) Case 2 (Generate at Cloud Server):* The mobile BS transmits the recovered data to the cloud server. The latency required for this uplink transmission is

$$T_{\mathrm{BS}} = \frac{D}{B_{\mathrm{B}} \log_2\left(1 + \frac{p_{\mathrm{B}} G_{\mathrm{BS}}}{B_{\mathrm{B}} N_0}\right)}, \tag{14}$$

where $B_{\mathrm{B}}$ is the bandwidth for the BS-server link, $p_{\mathrm{B}}$ is the mobile BS's transmit power, and $G_{\mathrm{BS}}$ is the corresponding channel gain. The energy cost for this transmission is

$$E_{\mathrm{BS}} = p_{\mathrm{B}} T_{\mathrm{BS}}. \tag{15}$$

Given the substantial computational power of cloud servers, the inference latency at the server is considered negligible. Similarly, the server's energy consumption is not factored into our model as it is assumed to have a constant power supply.

To distinguish these two cases, we introduce a binary decision variable $a \in \{0, 1\}$. If generation is performed locally at the mobile BS, $a = 1$; otherwise, $a = 0$.

### D. Problem Formulation

The total energy consumption of the system per update cycle is the sum of the energy consumed for UAV computation and transmission, BS computation, and the task-dependent energy for command generation, i.e.,

$$E_{\mathrm{tot}} = E_{\mathrm{U}} + E_{\mathrm{UB}} + E_{\mathrm{B}} + aE_{\mathrm{B}}^{\mathrm{I}} + (1-a)E_{\mathrm{BS}}. \tag{16}$$

Our objective is to minimize the total energy consumption by jointly optimizing the UAV's 3D location, the semantic compression ratio, the transmit powers, and the command generation offloading decision. Mathematically, this optimization problem is formulated as follows:

$$\min_{a, \mathbf{L}_{\mathrm{U}}, \rho, p_{\mathrm{U}}, p_{\mathrm{B}}} \quad E_{\mathrm{tot}} \tag{17}$$

$$\text{s.t.} \quad T_{\mathrm{U}} + T_{\mathrm{UB}} + T_{\mathrm{B}} + aT_{\mathrm{B}}^{\mathrm{I}} + (1-a)T_{\mathrm{BS}} \leq T_{\mathrm{th}}, \tag{17a}$$

$$q \geq q_{\mathrm{th}}, \tag{17b}$$

$$H_{\min} \leq H_{\mathrm{U}} \leq H_{\max}, \tag{17c}$$

$$\rho_{\mathrm{th}} \leq \rho \leq 1, \tag{17d}$$

$$0 < p_{\mathrm{U}} \leq p_{\mathrm{U}}^{\max}, \tag{17e}$$

$$0 < p_{\mathrm{B}} \leq p_{\mathrm{B}}^{\max}, \tag{17f}$$

$$a \in \{0, 1\}. \tag{17g}$$

Constraint (17a) imposes a maximum allowable latency, $T_{\mathrm{th}}$, for the entire update cycle. Constraint (17b) ensures a minimum required sensing QoS, $q_{\mathrm{th}}$. The UAV's altitude is constrained by (17c) to lie within a regulated range $[H_{\min}, H_{\max}]$. The semantic compression ratio is bounded by (17d), where $\rho_{\mathrm{th}}$ is a lower limit to maintain semantic integrity. Constraints (17e) and (17f) define the maximum transmit powers for the UAV and mobile BS, respectively. Finally, (17g) specifies the binary nature of the task offloading variable.

The mixed-integer non-convex problem (17) is NP-hard. In the next section, we exploit the unique structure of problem (17) to find its globally optimal solution via an efficient two-dimensional search.

## III. Algorithm Design

### A. Optimal UAV Location

Intuitively, to minimize both energy consumption and latency, the UAV should be positioned as close as possible to the mobile BS, subject to the operational constraints. A shorter distance $d_{\mathrm{UB}}$ enhances the channel gain, which in turn reduces the required transmission power and time. This insight allows us to decouple the UAV placement problem. The optimal UAV location, $\mathbf{L}_{\mathrm{U}}^*$, can be found by solving the following convex optimization problem:

$$\min_{\mathbf{L}_{\mathrm{U}}} \quad d_{\mathrm{UB}} \tag{18}$$

$$\text{s.t.} \quad d_{\mathrm{UT}} \leq \frac{-\ln(q_{\mathrm{th}})}{\xi}, \tag{18a}$$

$$H_{\min} \leq H_{\mathrm{U}} \leq H_{\max}. \tag{18b}$$

Problem (18) seeks to minimize the Euclidean distance to the origin subject to a convex set defined by the intersection of a sphere (representing the QoS constraint) and a slab (representing the altitude constraint). Let $D_{\max} = \frac{-\ln(q_{\mathrm{th}})}{\xi}$

be the maximum permissible distance from the UAV to the moving target. Assuming feasibility, i.e., $H_{\min} \le D_{\max}$, the solution can be derived geometrically.

**Theorem 1** *The optimal UAV altitude is the minimum allowable altitude:*

$$H_{\mathrm{U}}^{*} = H_{\min}. \tag{19}$$

*The optimal horizontal coordinates* $(x_{\mathrm{U}}^{*}, y_{\mathrm{U}}^{*})$ *are determined by one of two cases:*

*1) Case 1 ($x_{\mathrm{T}}^{2} + y_{\mathrm{T}}^{2} \le D_{\max}^{2} - H_{\min}^{2}$): In this case, the target is sufficiently close to the BS. The optimal UAV location is directly above the BS at the minimum altitude:*

$$\mathbf{L}_{\mathrm{U}}^{*} = (0, 0, H_{\min}). \tag{20}$$

*2) Case 2 ($x_{\mathrm{T}}^{2} + y_{\mathrm{T}}^{2} > D_{\max}^{2} - H_{\min}^{2}$): In this case, the UAV must move towards the target to satisfy the QoS constraint. The optimal location lies on the edge of the cylindrical service area in the direction of the target:*

$$x_{\mathrm{U}}^{*} = x_{\mathrm{T}} \left( 1 - \frac{\sqrt{D_{\max}^{2} - H_{\min}^{2}}}{\sqrt{x_{\mathrm{T}}^{2} + y_{\mathrm{T}}^{2}}} \right), \tag{21}$$

$$y_{\mathrm{U}}^{*} = y_{\mathrm{T}} \left( 1 - \frac{\sqrt{D_{\max}^{2} - H_{\min}^{2}}}{\sqrt{x_{\mathrm{T}}^{2} + y_{\mathrm{T}}^{2}}} \right). \tag{22}$$

### *B. Optimal Semantic Compression Ratio and Transmit Power*

*1) Optimal Semantic Compression Ratio with Given Transmit Power:* With the optimal UAV location $\mathbf{L}_{\mathrm{U}}^{*}$, a fixed task offloading indicator $a$ and fixed transmit powers $p_{\mathrm{U}}$ and $p_{\mathrm{B}}$, the subproblem for the semantic compression ratio $\rho$ is

$$\min_{\rho} \quad A\rho - F\ln(\rho) \tag{23}$$

$$\text{s.t.} \quad K_1\rho + K_2\ln(\rho) \le T, \tag{23a}$$

$$\rho_{\mathrm{th}} \le \rho \le 1, \tag{23b}$$

where

$$A \triangleq \frac{p_{\mathrm{U}} D}{B_{\mathrm{U}} \log_2 \left( 1 + \frac{p_{\mathrm{U}} G(\mathbf{L}_{\mathrm{U}}^{*})}{B_{\mathrm{U}} N_0} \right)}, \; F \triangleq D(\tau_{\mathrm{U}} \kappa_1 f_{\mathrm{U}}^{2} + \tau_{\mathrm{B}} \kappa_2 f_{\mathrm{B}}^{2}),$$

$$K_1 \triangleq \frac{D}{B_{\mathrm{U}} \log_2 \left( 1 + \frac{p_{\mathrm{U}} G(\mathbf{L}_{\mathrm{U}}^{*})}{B_{\mathrm{U}} N_0} \right)}, \; K_2 \triangleq -D \left( \frac{\kappa_1}{f_{\mathrm{U}}} + \frac{\kappa_2}{f_{\mathrm{B}}} \right),$$

$$T \triangleq T_{\mathrm{th}} - \left( \frac{a\kappa_3 D}{f_{\mathrm{B}}} + \frac{(1-a)D}{B_{\mathrm{B}} \log_2 \left( 1 + \frac{p_{\mathrm{B}} G_{\mathrm{BS}}}{B_{\mathrm{B}} N_0} \right)} \right).$$

**Theorem 2** *The optimal semantic compression ratio of problem* (23) *is*

$$\rho^{*}(p_{\mathrm{U}}, p_{\mathrm{B}}) = \min\left\{ \max\left\{ \rho_0, \rho_{\mathrm{th}}, \rho_a \right\}, 1, \rho_b \right\}, \tag{24}$$

*where* $\rho_0 = \frac{F}{A}$, $\rho_a < \rho_b$ *are the roots of* $K_1\rho + K_2\ln(\rho) = T$.

**Proof** Problem (23) admits a closed-form solution. The unconstrained minimizer of the objective is $\rho_0$. The latency constraint (23a) defines a feasible interval $[\rho_a, \rho_b]$. These roots can be expressed using the principal ($W_0$) and secondary ($W_{-1}$) branches of the Lambert W function:

$$\rho_a = \frac{K_2}{K_1} W_0 \left( \frac{K_1}{K_2} e^{\frac{T}{K_2}} \right), \rho_b = \frac{K_2}{K_1} W_{-1} \left( \frac{K_1}{K_2} e^{\frac{T}{K_2}} \right). \tag{25}$$

**Algorithm 1** Solving Problem (17) with Optimal Solution

```
Obtain the optimal UAV location $\mathbf{L}_{\mathrm{U}}^{*}$ using Theorem 1.
Obtain the optimal semantic compression ratio $\rho^{*}(p_{\mathrm{U}}, p_{\mathrm{B}})$ using Theorem 2.
for $a \in \{0, 1\}$ do
    Solve problem (26) via two-dimensional search.
    Store the minimum total energy as $V_a$.
end for
Compare $V_0$ and $V_1$ to determine $a^{*}$ and the corresponding optimal variables.
```

The optimal $\rho^{*}$ is then found by projecting the unconstrained minimizer $\rho_0$ onto the final feasible region defined by the intersection of $[\rho_a, \rho_b]$ and $[\rho_{\mathrm{th}}, 1]$, leading to Eq. (24). ■

Crucially, Theorem 2 establishes the optimal semantic compression ratio as an explicit function of the transmit powers.

*2) Two-Dimensional Search on Transmit Powers:* By substituting the analytical solutions for the optimal UAV location from Theorem 1 and the optimal semantic compression ratio from Theorem 2 back into problem (17), the original multi-variable optimization problem is reduced to a two-dimensional problem over the transmit powers for a fixed $a$:

$$\min_{p_{\mathrm{U}}, p_{\mathrm{B}}} \frac{p_{\mathrm{U}} D \rho^{*}(p_{\mathrm{U}}, p_{\mathrm{B}})}{B_{\mathrm{U}} \log \left( 1 + \frac{p_{\mathrm{U}} G(\mathbf{L}_{\mathrm{U}}^{*})}{B_{\mathrm{U}} N_0} \right)} + \frac{p_{\mathrm{B}} D (1-a)}{B_{\mathrm{B}} \log \left( 1 + \frac{p_{\mathrm{B}} G_{\mathrm{BS}}}{B_{\mathrm{B}} N_0} \right)} \tag{26}$$

$$\text{s.t.}\; T_{\mathrm{U}} + T_{\mathrm{UB}} + T_{\mathrm{B}} + aT_{\mathrm{B}}^{\mathrm{I}} + (1-a)T_{\mathrm{BS}} \le T_{\mathrm{th}}, \tag{26a}$$

$$0 < p_{\mathrm{U}} \le p_{\mathrm{U}}^{\max}, \tag{26b}$$

$$0 < p_{\mathrm{B}} \le p_{\mathrm{B}}^{\max}. \tag{26c}$$

Due to the intricate form of $\rho^{*}(p_{\mathrm{U}}, p_{\mathrm{B}})$, which involves the Lambert W function, the objective and constraint functions in problem (26) are highly non-convex, and an analytical solution for the optimal powers is intractable. Therefore, we find the solution by performing a two-dimensional search over the feasible power ranges. This method guarantees global optimality for the reduced problem and remains computationally feasible.

### *C. Optimal Task Offloading Decision*

The task offloading variable $a$ is binary. Given that our previous steps find the optimal solution for a fixed $a$, we can determine the optimal $a^{*}$ by solving the problem for each case and selecting the one that yields a lower total energy consumption. Let $V_0$ and $V_1$ be the minimum energy values obtained from the two-dimensional search for $a = 0$ and $a = 1$, respectively. The optimal offloading decision is:

$$a^{*} = \begin{cases} 0, & \text{if } V_0 < V_1, \\ 1, & \text{otherwise.} \end{cases} \tag{27}$$

### *D. Overall Algorithm Procedure*

The overall algorithm to solve problem (17) with global optimality is outlined in Algorithm 1.

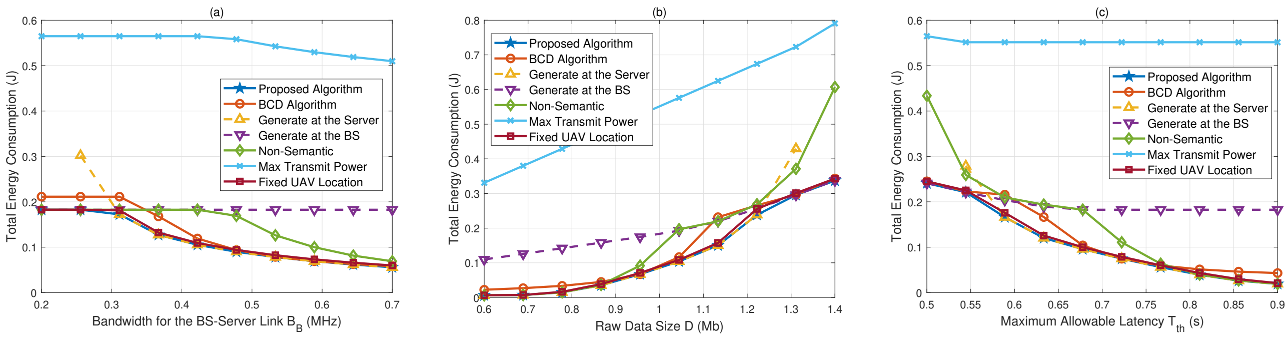

Fig. 3. Total energy consumption versus: (a) Bandwidth for the BS-server link $B_{\mathrm{B}}$, (b) Raw data size $D$, (c) Maximum allowable latency $T_{\mathrm{th}}$.

## IV. Simulation Results

In the simulations, the moving target is located at $(300\,\mathrm{m}$, $100\,\mathrm{m}$, $0\,\mathrm{m})$, and the UAV's operational altitude is constrained to $[40\,\mathrm{m},\ 400\,\mathrm{m}]$. The maximum transmit power for both the UAV and the mobile BS is set to $30\,\mathrm{dBm}$. Unless specified otherwise, the default parameters include a BS-server link bandwidth of $0.5\,\mathrm{MHz}$, a raw data size of $1\,\mathrm{Mb}$, and a maximum allowable latency of $700\,\mathrm{ms}$.

We evaluate the performance of our proposed algorithm against several benchmark schemes: **Generate at the Server/BS**, which consistently offloads the task to the server or processes it at the BS, respectively; **Non-Semantic**, which transmits the raw data without any semantic compression; **Max Transmit Power**, which operates the UAV and BS at their maximum power levels; **Fixed UAV Location**, which places the UAV at $(x_{\mathrm{T}}, y_{\mathrm{T}}, H_{\min})$; and **BCD Algorithm**, which employs block coordinate descent (BCD) to alternately optimize the semantic compression ratio and transmit powers.

Fig. 3 illustrates the total energy consumption as a function of the BS-server link bandwidth, raw data size, and maximum allowable latency. Across all simulated scenarios, the proposed algorithm consistently achieves the lowest energy consumption, which validates its superior performance in joint resource allocation. As depicted in the figures, the 'Max Transmit Power' scheme is uniformly inefficient, underscoring the necessity of adaptive power control. The 'Fixed UAV Location' scheme exhibits a consistent performance gap, which highlights the energy savings gained from optimizing the UAV's deployment. Furthermore, the 'BCD Algorithm' occasionally converges to a local optimum, demonstrating the advantage of our proposed method in attaining a globally optimal solution.

More specifically, Fig. 3(a) shows that as the BS-server bandwidth increases, the energy cost of offloading to the server decreases, making it the preferable strategy. Our algorithm adaptively selects this optimal offloading decision, outperforming the fixed 'Generate at the Server/BS' scheme. As seen in Fig. 3(b) and Fig. 3(c), the 'Non-Semantic' approach performs poorly with large data sizes or stringent latency constraints, due to the excessive energy required for transmitting uncompressed data. This confirms the significant efficiency gains provided by semantic communications.

## V. Conclusion

In this letter, we designed an energy-efficient framework for an agentic AI-assisted low-altitude semantic wireless network. We formulated a comprehensive system-wide energy minimization problem that captures the intricate trade-offs between communication, computation, and sensing performance. By jointly optimizing the UAV's 3D placement, the level of semantic compression, transmit power control, and the AI inference task offloading, our approach aims to enhance the operational endurance of the system. We developed an efficient algorithm to obtain the globally optimal solution for this non-convex optimization problem. Future research could extend this framework to multi-UAV, multi-target scenarios, incorporate dynamic mobility models, and explore reinforcement learning-based methods for real-time resource allocation in more complex and uncertain environments.

## References

[1] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart. 2019.

[2] Z. Zhao, Z. Yang, M. Chen, C. Zhu, W. Xu, Z. Zhang, and K. Huang, "Energy-efficient probabilistic semantic communication over space-air-ground integrated networks," *IEEE Trans. Wireless Commun.*, 2025.

[3] J. Liao, C. Zhan, B. Zeng, and H. Yan, "Energy-efficient optimization for IRS-enabled multiantenna UAV video streaming," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 9522–9535, Mar. 2024.

[4] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 1st Quart. 2023.

[5] X. Luo, R. Gao, H.-H. Chen, S. Chen, Q. Guo, and P. N. Suganthan, "Multimodal and multiuser semantic communications for channel-level information fusion," *IEEE Wireless Commun.*, vol. 31, no. 2, pp. 117–125, Apr. 2024.

[6] Z. Zhao, Z. Yang, Y. Hu, C. Zhu, M. Shikh-Bahaei, W. Xu, Z. Zhang, and K. Huang, "Compression ratio allocation for probabilistic semantic communication with RSMA," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 7304–7318, Sep. 2025.

[7] R. Zhang, S. Tang, Y. Liu, D. Niyato, Z. Xiong, S. Sun, S. Mao, and Z. Han, "Toward agentic AI: generative information retrieval inspired intelligent communications and networking," *arXiv preprint arXiv:2502.16866*, Feb. 2025.

[8] Q. Wei, R. Li, W. Bai, and Z. Han, "Multi-UAV-enabled energy-efficient data delivery for low-altitude economy: Joint coded caching, user grouping, and UAV deployment," *IEEE Internet Things J.*, vol. 12, no. 14, pp. 27 519–27 532, Jul. 2025.

[9] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "SemantiCodec: An ultra low bitrate semantic audio codec for general sound," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 8, pp. 1448–1461, Dec. 2024.

[10] T. X. Tran and D. Pompili, "Joint task offloading and resource allocation for multi-server mobile-edge computing networks," *IEEE Trans. Veh. Tech.*, vol. 68, no. 1, pp. 856–868, Jan. 2019.

[11] S. Zhang, H. Zhang, Z. Han, H. V. Poor, and L. Song, "Age of information in a cellular internet of UAVs: Sensing and communication trade-off design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6578–6592, Oct. 2020.

[12] Y. Yang, J. Zhou, Z. Yang, and M. Shikh-Bahaei, "Fluid antenna-enabled near-field integrated sensing, computing and semantic communication for emerging applications," *IEEE Trans. Cogn. Commun. Netw.*, 2025.