Engineering 63 (2026) 278–286

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Wireless Communications—Perspective

Generative and Large AI Models for 6G Wireless Networks: The Optimization Perspective

Yong Zhou [a], Ting Wang [b], Youlong Wu [a], Puyu Cai [c], Fuhui Zhou [d], Yuanming Shi [a,*]

[a] School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
[b] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai 200062, China
[c] Computer Science Department, New York University, New York, NY 10012, USA
[d] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 21 June 2025
Revised 25 February 2026
Accepted 23 March 2026
Available online 10 April 2026

Keywords:
Generative models
Large artificial intelligence models
Sixth-generation wireless networks
Information bottleneck

ABSTRACT

The transition to sixth-generation (6G) wireless networks is expected to introduce increasingly complex network architectures, disruptive wireless technologies, ultra-high network density, and diverse service requirements, necessitating highly efficient algorithm design for large-scale and non-convex network optimization. However, conventional optimization-based algorithms usually require sophisticated mathematical modeling and exhibit high computational complexity, while classic learning-based algorithms often suffer from poor robustness and generalization, as well as a lack of cross-scenario meta-optimization capabilities. In contrast, given their strong reasoning and contextual understanding abilities, generative and large artificial intelligence (AI) models are emerging as promising technologies to overcome these limitations. In this article, we propose the leveraging of generative and large AI models for scalable and generalizable network optimization, with an emphasis on facilitating information compression, beamforming design, and automated optimization for dynamic wireless networks with limited radio resources. We introduce a diffusion-based generation framework to solve multi-objective optimization problems for efficient information compression and transmission. We also present a large AI model-based framework for solving non-convex continuous optimization problems for beamforming design in both cell-free wireless networks and integrated sensing and communication networks. Finally, we propose an innovative large AI model-based framework that can automatically solve mixed-integer nonlinear programming problems for microservice deployment over satellite networks.

© 2026 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sixth-generation (6G) wireless networks are undergoing a transformative evolution, incorporating a broad range of disruptive technologies (e.g., ultra-massive multiple-input multiple-output (mMIMO), terahertz communications, and reconfigurable intelligent surfaces) to deliver unprecedented communication services that will enable 6G to support a wide array of emerging intelligent applications, such as autonomous driving, extended reality, and smart manufacturing [1]. To fulfill this ambitious vision for 6G, it is essential to equip 6G with new capabilities, such as artificial intelligence (AI)-related capabilities, sensing-related capabilities, sustainability, interoperability, coverage, and positioning [2]; it is also crucial to evolve 6G network architectures toward space–air–ground integration with increased flexibility and security. The incorporation of these disruptive technologies, new capabilities, and novel architectures poses significant challenges to network performance optimization, which is essential for meeting the stringent and diverse performance requirements of the emerging applications. However, network performance optimization problems in highly dynamic and increasingly complex 6G networks are typically non-convex, making it difficult to obtain optimal solutions in real time.

Existing solutions to network performance optimization fall into two main categories: optimization-based algorithms [3] and learning-based algorithms [4]. Various optimization-based algorithms—such as approximate message passing, semidefinite programming, and successive convex approximation—have been

* Corresponding author.
E-mail address: shiym@shanghaitech.edu.cn (Y. Shi).

https://doi.org/10.1016/j.eng.2026.03.016
2095-8099/© 2026 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

developed to tackle diverse signal-recovery and wireless resource-allocation problems. Despite achieving good performance and high interpretability through rigorous mathematical modeling and theoretical analysis, optimization-based algorithms have critical limitations in the ever-complex 6G networks. First, an optimization-based algorithm must transform the underlying optimization problem so it has a mathematical structure (e.g., convexity, decomposability, or strong duality) [5], which is problem-specific. In addition, these algorithms rely on an explicitly defined objective function that may not be available in many real-world applications. Second, for most optimization-based algorithms, convergence to a stationary solution typically requires hundreds of iterations, each of which involves dealing with multiple convex sub-problems, leading to high computational complexity [6]. This is a critical issue when the dimension of the optimization variables is high. In addition, because of time-varying channel conditions, optimization algorithms must be repeatedly performed within each coherence period, significantly increasing the computational complexity.

To overcome these challenges, learning-based algorithms that leverage data-driven adaptability are emerging as a powerful alternative. By exploiting the universal approximation property of neural networks, learning-based algorithms can deliver near-optimal solutions, while decreasing the need for expert knowledge for problem transformation and reducing the high computational complexity due to iterative optimizations [4]. Typical examples include algorithm unrolling, graph neural networks (GNNs), deep reinforcement learning (DRL), and federated edge learning [7]. Although learning-based algorithms are promising, their adoption is still hindered by critical technical barriers. First, adherence to resource constraints and quality of service (QoS) requirements is essential in addressing wireless resource-allocation problems. However, existing learning-based algorithms inherently lack explicit mathematical formulations and theoretical guarantees; thus, their data-driven outputs are prone to violating QoS requirements, reducing algorithmic robustness. Second, existing scenario-specific learning-based algorithms often exhibit poor generalization with time-varying network conditions (e.g., channel conditions, user locations, and traffic patterns). When encountering unseen scenarios, their performance is significantly degraded due to their reliance on historical training data, which cannot cover all real-world dynamics. Third, existing learning-based algorithms cannot extract the common principles underlying model construction and optimization strategy design; that is, they lack cross-scenario meta-optimization capabilities. There is a need for robust and generalizable learning-based algorithms to be designed to address these challenges.

Recently, generative and large AI models (LAMs) have demonstrated unique advantages that offer promising new directions for overcoming the limitations of classic learning-based algorithms [8]. In particular, generative AI models can learn patterns from training data and support various content-generation and predictive synthesis applications [9]. For instance, a diffusion model, as a typical generative AI model, offers strong capabilities in modeling complex distributions and generating high-quality solutions. Through a gradual denoising process, a diffusion model enables stable exploration in high-dimensional search spaces, improving the solution quality and robustness for black-box optimization problems. Moreover, LAMs are characterized by their massive scale (e.g., billions to trillions of parameters) and advanced model architectures (e.g., Transformer, diffusion, and Mamba) and are pre-trained on vast datasets using self-supervised learning and extensive computation resources. By learning intricate patterns from vast datasets, LAMs demonstrate emergent cognitive capabilities (e.g., reasoning and contextual understanding) that approach human-level performance; they can also efficiently adapt to specific tasks via few-shot learning. Compared with conventional AI models, LAMs possess high generalizability when considering dynamic wireless networks, exhibiting excellent adaptability to network dynamics without requiring complete retraining and architectural redesign [10]. In addition, by leveraging prompt engineering, mathematical reasoning, and contextual adaptation techniques, LAMs are capable of automatically analyzing and solving various network-resource problems, further reducing dependence on expert knowledge. With these unique advantages, LAMs hold great potential for 6G wireless networks, particularly in enabling intelligent and automated network performance optimization.

In this article, we advocate the paradigm of leveraging generative AI model- and LAM-enabled network optimization to facilitate information compression, beamforming design, and automated optimization for dynamic wireless networks with limited radio resources. We first elaborate the challenges presented by conventional optimization- and learning-based methods in order to motivate the adoption of generative and large model-enabled methods, and then present the proposed frameworks from an optimization perspective. In Section 2, we introduce a diffusion-based generation framework that exploits the expressive capability of diffusion models to solve multi-objective optimization problems for information compression and transmission. In Section 3, we present an LAM-enabled framework for solving non-convex continuous optimization problems for beamforming design; this is followed by the development of an LAM with low-rank adaptation (LoRA) to cell-free wireless networks and integrated sensing and communication (ISAC) networks. In Section 4, we highlight the unique capability of LAMs to perform automated modeling and optimization; this is followed by the development of an innovative LAM-based framework that applies prompt engineering techniques to solve mixed-integer nonlinear programming (MINLP) problems for microservice deployment in satellite networks. Finally, we conclude this article in Section 5.

2. Generative models for information compression and transmission

This section presents a diffusion-based robust information bottleneck multi-objective Bayesian optimization (DRIB-MOBO) method, which integrates the expressive capacity of diffusion models, the efficiency of black-box optimization in high-dimensional spaces, and the exploratory strength of multi-objective frameworks to increase the efficiency of information compression and transmission.

2.1. Challenges and motivations

In wireless networks with limited bandwidth, information compression is essential for reducing data redundancy and minimizing transmission latency. These are particularly critical in autonomous driving and remote-sensing scenarios, where high-dimensional data must be transmitted and processed in a timely manner. For instance, in satellite systems, directly transmitting raw data from low-Earth orbit (LEO) satellites to ground stations leads to excessive communication overhead and high transmission delay, motivating recent efforts to perform onboard feature compression and compact representation transmission. In particular, the information bottleneck (IB) principle [11] provides an emerging technology for compressing input data while preserving task-relevant information and thus yields a theoretical framework for efficient representation learning. However, IB is essentially an information-compression technique and cannot ensure reliable information transmission over noisy channels. To address this issue, robust IB (RIB) incorporates the transmission rate and

mutual information between transmitted and received features, enabling the joint optimization of accuracy, compression, and robustness in complex environments [12]. Unlike conventional IB, which focuses solely on the trade-off between compression and task relevance, the RIB framework further introduces transmission robustness by optimizing the mutual information between transmitted and received features, thereby enabling reliable and efficient representation learning under noisy and resource-constrained communication environments. This makes RIB a more effective framework for reliable and efficient representation learning in complex communication and perception systems.

RIB can be framed as a multi-objective optimization problem, which is difficult to solve in practical networks due to the inherent complexity of multi-objective optimization and the high dimensionality of real-world data. Conventional multi-objective optimization methods struggle with tackling non-convex and high-dimensional problems that involve implicit objectives and/or complex constraints [13]. Moreover, these methods suffer from high computational cost and slow convergence, and are prone to local optima. Despite offering scalable surrogate modeling to reduce the dependence on the objective function, classic learning-based approaches often fail to effectively address multi-objective conflicts and are highly sensitive to hyperparameter choices in multi-objective optimization. Diffusion models [14] have strong capabilities in modeling complex distributions and generating high-quality solutions, and are well suited for multi-objective optimization with implicit objectives and complex solution spaces [15]. Compared with generative adversarial networks (GANs) and variational autoencoders (VAEs), diffusion models offer more stable training dynamics and controllable sample generation, making them particularly suitable as generative optimizers for black-box and multi-objective optimization tasks where reliable sample generation and stable learning are critical.

2.2. Diffusion-based robust information bottleneck multi-objective Bayesian optimization

To overcome the challenges mentioned above, an RIB problem constitutes a multi-objective black-box optimization problem that can then be dealt with by developing a unified DRIB-MOBO framework. This framework integrates the generative capabilities of a diffusion model with the efficiency of black-box optimization in handling high-dimensional and implicit objective functions. The overall framework is illustrated in Fig. 1, where $P1$ represents a traditional IB problem and $P2$ denotes its robust extension. For clarity, we define the mutual information between the input data and the bottleneck variable as the information complexity, the mutual information between the bottleneck variable and the target features as the utility information, and the mutual information between the bottleneck variable before and after perturbation as the communication rate. The RIB formulation can then be interpreted as a multi-objective optimization problem that simultaneously enhances the utility information and the communication rate under a bounded information-complexity constraint. This joint optimization establishes a practical balance among accuracy, robustness, and bandwidth efficiency, enabling reliable and efficient representation learning in communication-constrained systems. The framework first encodes and maps the input data and target features to reduce modeling complexity. Latin hypercube sampling (LHS) is subsequently adopted for the bottleneck variable initialization, providing space-filling coverage of the high-dimensional search domain and improving the diversity of the initial design points. This initialization provides a well-distributed starting point that approximates a Gaussian distribution, thereby facilitating the subsequent denoising and reconstruction processes. Next, a multi-objective Gaussian process (GP) model is introduced to estimate the mutual information objective, while sample selection is performed under the evaluation of a finite objective function through the acquisition function to improve the optimization efficiency. For the initial bottleneck variable, DRIB-MOBO calculates the multi-objective mutual information for training a multi-objective GP model, which predicts function values for weighting. To improve the sample quality, DRIB-MOBO applies the shifted density estimation method to evaluate candidate fitness and selects top-ranked samples to train the diffusion model. To increase the robustness, DRIB-MOBO adopts adaptive sampling: If the hypervolume growth rate exceeds a threshold, a conditional diffusion model (CDM) generates new samples; otherwise, an alternative method such as a genetic algorithm (GA) is used. This strategy prevents convergence to local optima and improves global exploration.

During the training phase, the diffusion model learns to predict the noise added at each time step in the forward process. Model parameters are optimized using the mean squared error (MSE) loss to ensure accurate noise estimation. In the reverse diffusion process, a conditional guidance mechanism is introduced to guide sample generation using a weighted gradient, and denoising is gradually performed to generate high-quality samples from noisy data. Samples are denoised step by step to produce high-quality outputs. The weights are computed using an entropy-based method, generating an entropy-weighted gradient that balances and guides multiple optimization objectives effectively. In each round of the generation phase, new solutions satisfying the compressibility constraint are added to the candidate set. The GP model and hypervolume index are then updated to increase accuracy and diversity. Finally, non-dominated sorting is applied to extract the Pareto front, which is decoded to obtain the optimal bottleneck representation, enabling an efficient and controllable solution to the RIB problem.

DRIB-MOBO enhances optimization stability and adaptability, reduces dependence on hyperparameter tuning, and explicitly accommodates multiple competing objectives. By leveraging a generative diffusion-based optimizer, DRIB-MOBO supports robust modeling under implicit constraints, facilitates solution diversity, and ensures convergence in challenging search spaces. Furthermore, this formulation provides a unified and interpretable scheme that offers theoretical insight into the trade-offs between compression, robustness, and predictive performance, while delivering practical scalability and generalization across real-world communication scenarios. By designing the objective functions, DRIB-MOBO integrates a multi-objective Bayesian optimization framework and leverages a diffusion model to generate intermediate representations that approximate the Pareto front; this ensures that the features extracted from the input data retain high task relevance while minimizing redundant information, maintaining transmission accuracy, and achieving robustness and strong generalization under noisy communication conditions.

3. LAMs for beamforming design

This section introduces a unified LAM-enabled optimization framework for beamforming design in multi-antenna wireless networks. By leveraging the reasoning capacity and scalability of LAMs, this framework has the potential to efficiently tackle various high-dimensional and non-convex beamforming design problems. We take two representative network scenarios—cell-free wireless networks and ISAC networks—as examples.

3.1. Challenges and motivations

Air-interface technologies are essential components of wireless communication systems, including beamforming design, channel estimation, power control, and spectrum allocation, where

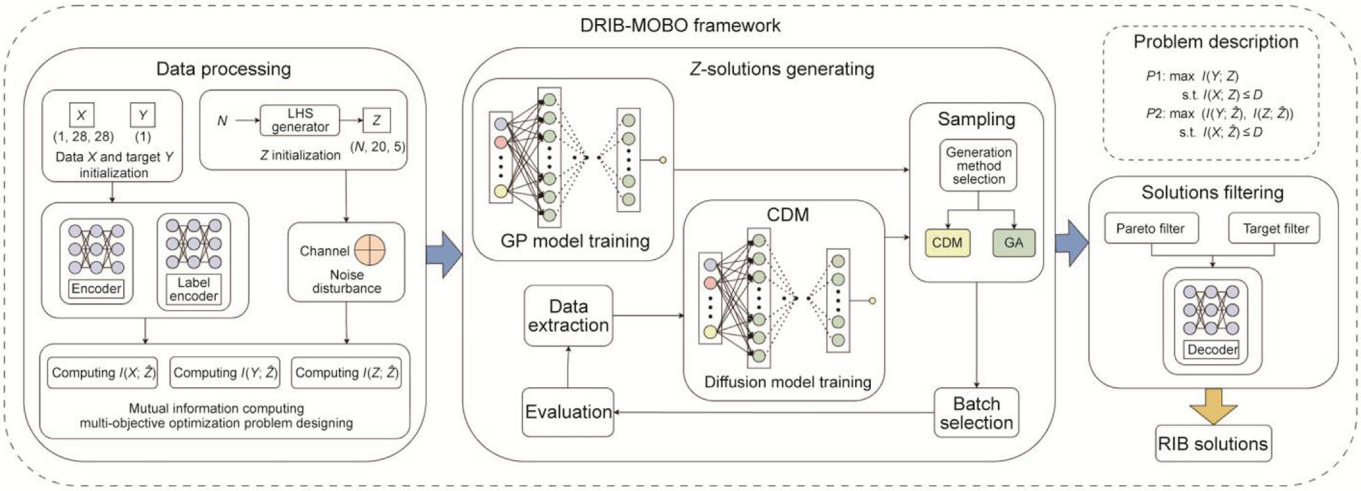

Fig. 1. A diffusion-based generation framework for multi-objective optimization problems. In this framework, X represents the input data and Y denotes the corresponding target feature. N indicates the number of initialization variables generated via Latin hypercube sampling (LHS). The parameters Z and $\hat{Z}$ correspond to the bottleneck variables before and after noise perturbation, respectively. The mutual information measures $I(X;Z)$, $I(Y;Z)$, $I\left(X;\hat{Z}\right)$, $I\left(Y;\hat{Z}\right)$, and $I\left(Z;\hat{Z}\right)$ quantify the deterministic information complexity, normal utility information, information complexity, utility information, and communication rate, respectively. The parameter D specifies a predefined information budget that constrains the maximum allowable information flow from the input to the bottleneck variable. GP: Gaussian process; CDM: conditional diffusion model; GA: genetic algorithm.

beamforming is a key enabling technology to enhance the spectrum and energy efficiency of multi-antenna systems. The design and optimization problems for beamforming design are typically high-dimensional, non-convex, and sensitive to rapidly changing wireless environments, posing significant challenges for conventional optimization-based and learning-based algorithms. On the one hand, while conventional optimization-based algorithms can offer theoretical guarantees under idealized conditions, their scalability and adaptability are often limited in practical large-scale and highly dynamic scenarios. In particular, conventional optimization-based algorithms (e.g., Gram–Schomidt orthogonalization and Riemannian manifolds) for beamforming design tend to exhibit high computational complexity and reduced effectiveness as the number of users and antennas increases or when the channel conditions vary significantly over time. On the other hand, to reduce the overwhelming computational complexity, various learning-based algorithms, such as multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and GNNs, have been developed to learn the direct mappings between instantaneous channel state information (CSI) and beamforming design. However, existing learning-based algorithms for beamforming design are generally tailored to specific optimization problems in specific network scenarios. Small changes in the network density, antenna number, user distribution, and channel characteristics may require model retraining and even redesign, limiting the adaptability of such algorithms in dynamic wireless scenarios. This challenge is further exacerbated by the increasing complexity and heterogeneity of wireless networks. These limitations motivate the need for a unified and adaptable framework that can handle time-varying network parameters.

With billions of parameters and extensive pretraining, LAMs have demonstrated strong capabilities in abstraction, reasoning, and pattern understanding across a broad class of downstream tasks. Their ability to generalize beyond specific training distributions and adapt to new objectives with minimal task-specific design has enabled progress in scientific computing, robotics, and decision-making. These developments have motivated efforts to apply a unified modeling paradigm to wireless networks, which often involve diverse optimization problems and dynamic network settings. For instance, the use of LAMs for channel prediction in Refs. [16,17] demonstrates their potential to handle high-dimensional wireless data and generalize to varying scenarios. Compared with optimization-based algorithms that rely on explicit mathematical modeling and incur high computational complexity, LAMs offer a data-driven modeling paradigm that enables low-complexity wireless air-interface designs by leveraging pretrained semantic representations and attention-based reasoning. In contrast to conventional learning-based algorithms, which are typically tailored to specific network configurations, LAMs exhibit great adaptability to network reconfigurations, without the requirement of model retraining and architectural redesign. However, research on LAMs for air-interface technologies is still at an early stage and is encountering critical challenges when dealing with more complex beamforming design tasks. In addition, existing studies on LAMs for air-interface design often retain simple MLP-based output heads [18], leading to limited architectural flexibility and poor generalization.

3.2. Beamforming design for cell-free wireless networks and ISAC networks

In this subsection, we present a unified beamforming framework that integrates an LAM with modular scenario-specific components, as shown in Fig. 2. Unlike existing learning-based methods, which often require the design of dedicated models for each specific task and network configuration, our approach adopts a shared LAM backbone as a general reasoning core. Task adaptability is achieved by plugging in lightweight and modular encoders and output heads tailored to each scenario. The framework comprises three components: scenario-specific feature encoders, a shared LAM backbone, and scenario-specific output heads. This allows the same backbone to be applied to different network tasks without retraining or architectural redesign [19]. In the following, we present the framework's applications in two representative beamforming design scenarios: cell-free wireless networks and ISAC networks.

3.2.1. Cell-free wireless networks

Cell-free wireless networks are pivotal for enabling seamless connectivity and achieving high-spectrum efficiency through

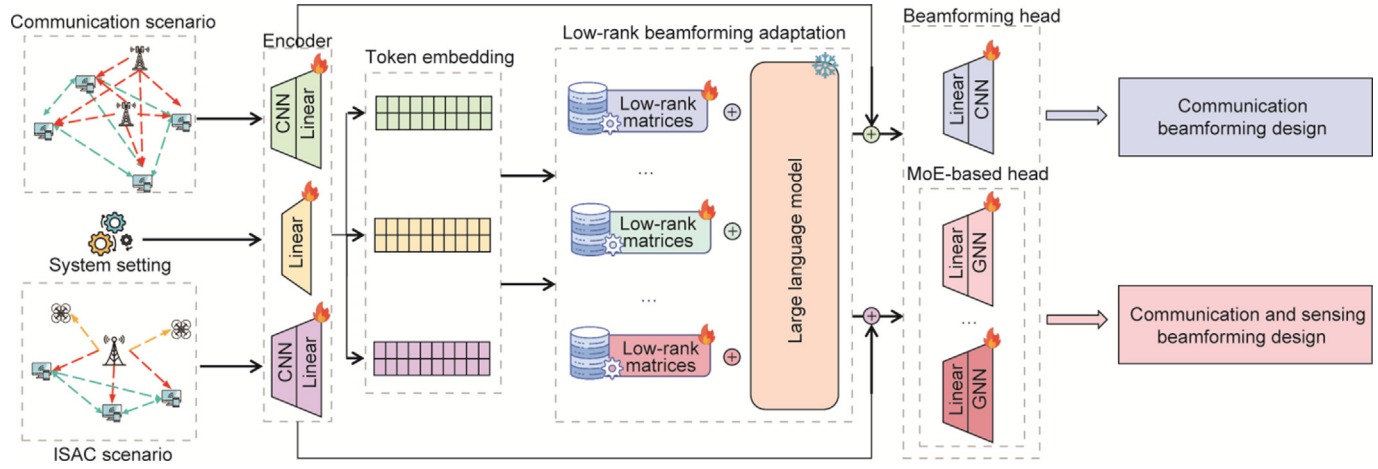

Fig. 2. An LAM-enabled beamforming design for cell-free wireless networks and ISAC networks. MoE: mixture-of-experts.

user-centric cooperation among distributed access points (APs). In order for a typical cell-free network that consists of multiple APs to be able to collaboratively serve geographically dispersed mobile users, it is essential to maximize the achievable sum communication rate for all users, while accounting for various practical constraints (e.g., the limited radio resources and maximum transmit power of each AP), time-varying channel characteristics (e.g., Rayleigh, Rician, and Nakagami fading), and dynamic network topologies. To facilitate beamforming design under such a cell-free network, we present an LAM-based framework that consists of three key components: an encoder, a low-rank beamforming adaptation module, and a CNN-based beamforming head, as shown in the upper part of Fig. 2. The encoder takes the global instantaneous CSI as input, structured as a user-by-antenna matrix, and extracts spatial features using convolutional layers. These features are then projected into the LAM-compatible embedding space through a linear transformation. In addition to the CSI, the channel index is encoded via a linear layer and incorporated into the model input, enabling the LAM to distinguish between different channel conditions during inference. To ensure compatibility with respect to the time-varying user and AP counts, we define a loosely bounded maximum input size and apply zero-padding to maintain consistent input formatting without retraining. In addition, the LAM backbone is integrated with LoRA modules, allowing efficient fine-tuning through a small number of trainable parameters while preserving the generalization capacity of the pretrained model [20]. The contextual embeddings produced by the LAM are concatenated with the original CSI features extracted by the encoder to support beamforming decisions. To generate beamforming outputs, we employ a CNN-based head that combines the contextual embeddings from the LAM with spatial CSI features along the channel dimension, enabling the model to jointly leverage global contextual reasoning and local spatial representations. The CNN head provides a structurally compatible output, maintaining consistent output dimensions and supporting flexible adaptation to different user and AP configurations without architectural modification. The framework can be trained in an unsupervised manner to maximize the network performance (e.g., the sum communication rate), enabling scalable and deployment-adaptive beamforming design for cell-free wireless networks.

3.2.2. ISAC networks

As a typical use case of 6G, ISAC unifies the target sensing and user communication capabilities via shared spectrum and hardware resources to increase the network efficiency and promote new applications. Unlike communication-only scenarios, sensing-centric ISAC focuses on generating a beamforming matrix to improve the sensing performance (e.g., to closely match a predefined radar beampattern), while ensuring communication quality (e.g., satisfying the per-user signal-to-interference-plus-noise ratio (SINR) threshold) and diverse practical constraints (e.g., rank-one constraints on individual user beamforming matrices, positive semi-definiteness of the joint transmit covariance matrix). To enable beamforming design under such an ISAC scenario, we advocate an LAM-based framework consisting of an encoder, a low-rank beamforming adaptation module, and a GNN-based beamforming head, as shown in the lower part of Fig. 2. By processing global instantaneous CSI in the user-by-antenna format, the encoder leverages a CNN to distill spatial features. A linear projection maps these distilled features onto the task-specific embedding space of the LAM. In addition to the CSI, system-level information such as user SINR requirements and radar angular parameters can be projected via dedicated linear layers and concatenated with processed CSI features at the input stage. To ensure compatibility across deployments with varying user counts, we prescribe a flexibly constrained maximum input dimension and employ zero-padding to sustain uniform input structuring without retuning. The LAM incorporates LoRA modules to facilitate parameter-efficient refinement with dedicated task-specific weights while retaining the pretrained model's generalization capability. The resulting contextual embeddings capture both the spatial characteristics of the CSI and the sensing–communication constraints, and are combined with the original CSI features for beamforming design. To construct beamforming matrices, we employ a GNN-based head that integrates the LAM-generated contextual embeddings and CSI features through channel-dimensional concatenation. A heterogeneous graph is constructed, where each user is represented as a node containing local CSI-derived features and all radar targets are collectively modeled as a single node encoding the angular configuration. Edges are formed between user nodes to capture mutual interference and between the radar node and user nodes to reflect sensing–communication coupling. To improve adaptability under diverse SINR constraints, a mixture-of-experts (MoE) structure is incorporated into the GNN-based output head [21]. Each expert corresponds to a distinct GNN output head, and a gating mechanism selects the top-k expert branches based on the input SINR requirements, allowing the model to activate different output pathways tailored to varying communication scenarios. The framework is first trained in a supervised manner by minimizing the MSE between the predicted and ground-truth beamforming matrices for both communication users and radar targets. As this stage does not strictly enforce practical constraints, the

second-stage fine-tuning is performed by incorporating penalty terms to improve constraint satisfaction, where each expert is further optimized under corresponding conditions. The MoE-based GNN head enables adaptive beamforming design under varying SINR constraints, where each expert is pretrained under a representative SINR regime and the gating mechanism selects or combines experts according to the current SINR requirement, thereby balancing the sensing accuracy and communication reliability across different operation points.

This unified LAM-based framework enables generalization across different beamforming tasks through a shared backbone and modular scenario-specific designs. By integrating CNN and GNN output heads for the cell-free and ISAC networks, respectively, the framework supports efficient adaptation without retraining and maintains strong performance under varying system conditions. In particular, the CNN and GNN modules act as scenario-specific output heads in the unified LAM framework, where the CNN efficiently captures local spatial correlations in cell-free beamforming [22], while the GNN models the relational dependencies between users and radar nodes in ISAC networks [23].

4. LAMs for automated optimization

In this section, we propose an effective LAM-enabled framework to automatically address resource optimization problems in complex wireless networks. By underscoring the limitations of traditional manual modeling techniques and conventional learning-based algorithms, we highlight the unique capability of LAMs to perform automated modeling and optimization. This paradigm shift enables a more generalizable and intelligent solution framework tailored for 6G wireless networks.

4.1. Challenges and motivations

Wireless resource optimization is a pivotal enabler for improving spectrum and energy efficiency while ensuring QoS in highly dynamic and increasingly complex wireless communication networks, such as integrated space–air–ground networks [24] and massive multiple-input multiple-output (MIMO) networks. However, attaining optimal resource allocation in such prominent networks presents significant challenges, as it typically requires solving non-deterministic polynomial-time hard (NP-hard) MINLP problems [25]. On the one hand, model-based methods, such as branch-and-bound and convex relaxation, require strong expertise and substantial labor costs. First, constructing complex mathematical formulations tailored to specific scenarios—such as building an accurate model for a massive MIMO network to capture the relationship among the high-dimensional antenna array responses, time-varying CSI, and achievable performance—demands deep expertise in both wireless communications and optimization theory, resulting in substantial labor costs and prolonged development cycles. Second, the implementation of model-based optimization algorithms often involves cumbersome design efforts. For example, the effectiveness of branch-and-bound methods heavily depends on customized branching and pruning strategies tailored to the problem structure. More critically, when the network architectures change, the original system models (e.g., path loss formulas and interference constraints) become invalid. Consequently, a complete redevelopment of the modeling and solving process is required. This strong coupling between optimization strategies and specific system characteristics severely limits the scalability and cross-scenario transferability of model-based methods. On the other hand, learning-based alternatives including reinforcement learning and supervised learning have the potential to alleviate the aforementioned limitations, but they also introduce new bottlenecks [26]. More specifically, reinforcement learning requires the scene-specific design of Markov decision processes (MDPs) and training strategies. In addition, policy exploration in reinforcement learning often suffers from slow convergence and high training costs, leading to an exponential increase in engineering deployment expenses. While supervised learning can establish end-to-end mappings based on historical data, the performance heavily relies on the quality and completeness of labeled data. In emerging application scenarios (e.g., integrated space–air–ground networks) where historical data are scarce, the generalization ability of supervised models is significantly degraded [27]. Moreover, existing learning-based methods are unable to abstract the common principles underlying model construction and optimization strategy design as human experts can do; that is, they lack cross-scenario meta-optimization capabilities [28]. These fundamental challenges have fueled the need for generalized and automated optimization methods.

Recent studies have shown that LAMs are not only capable of interpreting the semantic descriptions of complex optimization problems but can also autonomously derive mathematical models and generate feasible solutions. This capability for automated modeling effectively reduces dependence on expert knowledge, enabling optimization systems to adapt to emerging network architectures and evolving service demands. To address the data-dependency issues inherent in conventional learning-based methods, LAMs leverage their pretrained domain knowledge and reasoning capabilities to facilitate rapid transfer learning even with minimal sample data. Furthermore, the natural-language-based interaction mechanism significantly increases the interpretability of the optimization process, allowing human experts to remain involved in the decision-making loop. These characteristics position LAMs as a transformative tool for dealing with the multidimensional, large-scale optimization challenges prevalent in wireless networks. With its powerful pretrained knowledge, the LAM holds promise to serve as the foundation model through which to achieve a so-called “one model for all tasks” with even better performance and stronger generalization.

Although LAMs offer a new paradigm for automated modeling, their application in wireless resource optimization still presents several challenges. First, LAMs are prone to hallucinations in mathematical modeling, potentially generating incorrect constraints. Second, the lack of domain-specific guidance often results in the generation of infeasible strategies from an engineering perspective. Existing studies have attempted to address these issues through fine-tuning, which involves supervised training with large quantities of paired samples comprising problem descriptions, mathematical models, and optimization strategies. However, in multi-objective and multi-constraint wireless network optimization scenarios, the scarcity of high-quality labeled data severely limits the generalization capability of such fine-tuned models, leading to a failure to overcome the fundamental shortcomings faced by traditional learning-based methods. Therefore, there is an urgent need for a more generalizable, interpretable, and automated optimization framework that can effectively guide LAMs to generate high-quality solutions under diverse wireless communication scenarios, without requiring extensive labeled data and manual intervention.

4.2. Automated optimization for microservice deployment

In this subsection, we present an LAM-based optimization framework that applies prompt engineering techniques to open-source LAMs for automated resource allocation. The general solving pipeline of our framework follows a structured sequence: Starting from a natural-language problem description, the LAM first extracts key domain semantics and generates mathematical

optimization formulations tailored to the problem scenario; it then automatically produces solver-ready code, which is subsequently executed by industry-standard solvers. To increase robustness and modeling accuracy, especially in scenarios lacking large-scale labeled data, our framework can incorporate domain knowledge—such as canonical mathematical modeling patterns and common network-optimization paradigms—through structured prompt design and constraint templates, effectively guiding the model to produce valid formulations. This mitigates common hallucinations in mathematical reasoning and improves the consistency between generated constraints and actual system behaviors. In addition, logical verification is embedded to check and refine the output before solver execution. Finally, based on the solver feedback, the LAM iteratively refines the formulation or parameters to improve solution quality. By incorporating domain knowledge injection and logical verification mechanisms, our framework systematically regulates the LAM's problem-solving process, enabling zero-shot optimization modeling without a reliance on extensive labeled datasets. Through replacing conventional manual modeling approaches, this framework opens a new technological pathway toward enabling autonomous resource optimization.

As illustrated in Fig. 3, the LAM-based framework introduces an LAM-driven solving pipeline tailored to the unique challenges of wireless resource optimization. Wireless networks such as satellite constellations [29–31] exhibit highly dynamic and resource-constrained environments. System model descriptions in this context often include limited link capacities, diverse propagation delays, and service demands represented as dependency-aware microservice graphs. To bridge the gap between such domain-specific semantics and mathematical optimization, our framework initiates the process by extracting structured representations from raw communication system inputs. These include key parameters, such as bandwidth budgets, satellite connectivity graphs, link delay matrices, and workload graphs, which are then transformed into optimization-ready mathematical formulations using a retrieval-augmented generation (RAG) mechanism. We design a structured knowledge base containing standard optimization templates for wireless resource optimization and use a dense retriever to query this knowledge base for relevant models. These models are then used in conjunction with task-specific inputs to generate optimized mathematical formulations through the generative LAM encoder–decoder. More specifically, a dense retriever first queries a structured knowledge base consisting of canonical modeling templates for communication-constrained optimization problems. Retrieved candidates—such as mathematical representations of link capacity constraints, microservice placement formulations, and delay minimization objectives—are then fused with task-specific inputs using a generative LAM encoder–decoder, enabling accurate translation of system semantics into solver-compatible formulations. This ensures that the optimization logic adheres to core communication-centric objectives, such as minimizing end-to-end latency and balancing network loads. Building on this semantic-to-mathematical abstraction, the reasoning and generalization capabilities of LAMs are leveraged to automatically generate solver-ready code compatible with industry-standard engines (e.g., CPLEX). The generated code encapsulates decision variables (e.g., service-to-node mappings), objective functions (e.g., latency-aware deployment cost), and constraints (e.g., bandwidth limits and node capacities) in a programmatic format. Our framework can address a variety of constraints commonly found in wireless resource-optimization problems, such as bandwidth limitations, computational capacity restrictions, latency requirements, and resource capacity limits. These constraints are transformed into mixed-integer combinatorial optimization problems, which are then automatically modeled and encoded by our framework. The resulting models are processed using industry-standard solvers, which compute optimal solutions through our framework's automated optimization pipeline. Beyond static formulation, the framework supports a feedback-driven refinement loop in which solver outputs are continuously monitored for infeasibility or suboptimality and the model is adaptively revised to improve performance. If an infeasible solution is identified, the system adjusts the optimization model through a feedback mechanism and regenerates the solution to correct the violations. This iterative adaptation is particularly valuable in communication systems where environmental dynamics can rapidly degrade the performance of static models. Through this end-to-end,

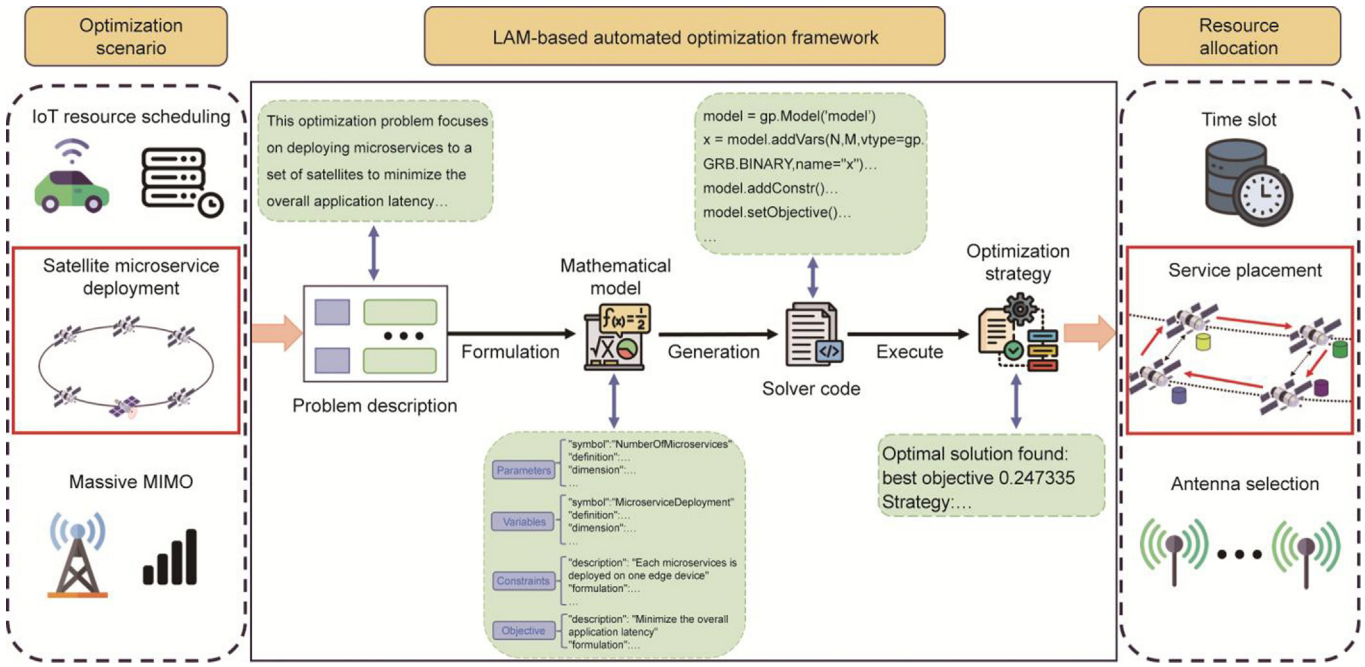

Fig. 3. An LAM-based automated optimization framework. IoT: Internet of Things.

LAM-enabled pipeline, the framework automates the complex modeling and optimization process, significantly reducing reliance on human experts. It offers a scalable, intelligent, and automated solution paradigm for managing network resources in large-scale wireless infrastructures, enabling real-time, adaptive resource allocation across communication systems.

To evaluate the performance of our LAM-based optimization framework, we focus on a microservice deployment problem over a LEO satellite network. The objective is to minimize the total end-to-end latency of all applications while satisfying resource and deployment constraints. Each application consists of microservices structured as a directed acyclic graph, where each microservice must be placed on one of its candidate satellites. The total latency includes the local processing delay, transmission delay between dependent microservices, and propagation delay across inter-satellite links. The deployment must ensure that each satellite's computing capacities and communication workloads are not exceeded. This problem is challenging due to dynamic link characteristics, limited onboard resources, and strict latency requirements. We evaluate the LAM-based automated optimization framework in a Walker-Delta LEO satellite constellation, which consists of six orbital planes with 8, 10, and 12 satellites per plane. To account for real-world situations, small variations are added to each satellite, reflecting the dynamic and uncertain nature of satellite networks' positioning and communications. We also assume the satellites have diverse computational resources.

To validate the effectiveness of the framework, three baseline algorithms are considered: ① a random algorithm that randomly assigns microservices to satellites under resource constraints, ② a heuristic algorithm that selects satellites with the most residual capacity, and ③ an optimal algorithm that utilizes an exhaustive search with branch-and-bound pruning to find the optimal solution. In contrast to these manually designed approaches, the LAM-based automated optimization framework powered by GPT-4 automatically interprets natural-language descriptions of deployment tasks and generates optimized placement decisions without human-crafted heuristics or explicit search strategies. We evaluate the system-level performance in terms of the total application latency across different constellation sizes. The results are summarized in Fig. 4. As can be observed, the LAM-based method consistently outperforms both the random and the heuristic algorithms. This case study demonstrates that LAMs can serve as general-purpose optimizers across both classical and emerging network scenarios, highlighting their potential as foundation models for automated and scalable network resource optimization. Besides microservice deployment, our framework is applicable to various wireless network resource-optimization problems that can be modeled as mixed-integer combinatorial optimization problems, such as spectrum allocation and bandwidth optimization.

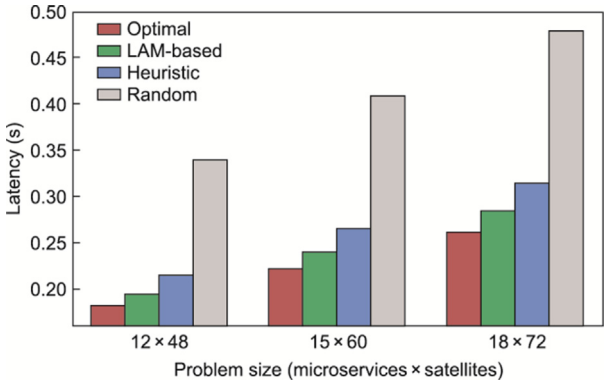

Fig. 4. Total latency of random, heuristic, LAM-based, and optimal algorithms.

5. Conclusions

In this article, we investigated the role of generative AI models and LAMs in revolutionizing network performance optimization in 6G networks, with specific emphasis on three essential examples: information compression, beamforming design, and network resource allocation. To enable effective and robust information compression, we proposed a diffusion-based generation framework, harnessing the expressive power of diffusion models to facilitate the learning of rich intermediate representations that make it possible to tackle various black-box and multi-objective optimization problems. In addition, we presented a unified and generalizable LAM-based beamforming design framework featuring a shared backbone and modular scenario-specific design, which can address non-convex continuous optimization problems. Furthermore, we developed an LAM-enabled framework to automate the modeling and solving of MINLP problems for network resource allocation without requiring extensive labeled data and manual intervention. We hope this article will serve as a useful guideline for leveraging generative AI models and LAMs to shape the future of 6G networks.

CRediT authorship contribution statement

Yong Zhou: Writing – original draft, Methodology, Conceptualization. **Ting Wang:** Writing – review & editing, Data curation, Conceptualization. **Youlong Wu:** Writing – review & editing, Methodology, Investigation. **Puyu Cai:** Writing – review & editing, Validation, Formal analysis. **Fuhui Zhou:** Writing – review & editing, Project administration, Conceptualization. **Yuanming Shi:** Writing – review & editing, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Yong Zhou acknowledges the funding support in part from the Natural Science Foundation of Shanghai (23ZR1442800) and the Science and Technology Commission Foundation of Shanghai (25DP1501900 and 25DP1502000). Ting Wang acknowledges the funding support in part from the Shanghai Special Fund for Promoting High-Quality Industrial Development (2025030702) and the Key Program of National Natural Science Foundation of China (62432007). Yuanming Shi acknowledges the funding support in part from the National Natural Science Foundation of China (62522117), the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (BK 2024CSJZN00303), and the Science and Technology Commission Foundation of Shanghai (25DP1500100).

References

[1] Letaief KB, Shi Y, Lu J, Lu J. Edge artificial intelligence for 6G: vision, enabling technologies, and applications. IEEE J Sel Areas Commun 2022;40(1):5–36.
[2] M.2160: Framework and overall objectives of the future development of IMT for 2030 and beyond. Geneva: Radio Communication Division of the International Telecommunication Union; 2023.
[3] Luo ZQ, Yu W. An introduction to convex optimization for communications and signal processing. IEEE J Sel Areas Commun 2006;24(8):1426–38.
[4] Shi Y, Lian L, Shi Y, Wang Z, Zhou Y, Fu L, et al. Machine learning for large-scale optimization in 6G wireless networks. IEEE Commun Surv Tutor 2023;25(4):2088–132.

[5] Liu YF, Chang TH, Hong M, Wu Z, So AMC, Jorswieck EA, et al. A survey of recent advances in optimization methods for wireless communications. IEEE J Sel Areas Commun 2024;42(11):2992.
[6] Zhou Y, Zou Y, Wu Y, Shi Y, Zhang J. Machine learning for low latency communications. London: Elsevier; 2024.
[7] Tao M, Zhou Y, Shi Y, Lu J, Cui S, Lu J, et al. Federated edge learning for 6G: foundations, methodologies, and applications. Proc IEEE 2025;113(9):1075–113.
[8] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021. arXiv:2108.07258.
[9] Liu Y, Du H, Niyato D, Kang J, Xiong Z, Kim DI, et al. Deep generative model and its applications in efficient wireless network management: a tutorial and case study. IEEE Wirel Commun 2024;31(4):199–207.
[10] Wang Z, Shi Y, Zhou Y, Zhu J, Letaief KB. Edge large AI models: revolutionizing 6G networks. IEEE Commun Mag 2025;63(10):36–42.
[11] Tishby N, Pereira FC, Bialek W. The information bottleneck method. 2000. arXiv:physics/0004057.
[12] Xie S, Ma S, Ding M, Shi Y, Tang M, Wu Y. Robust information bottleneck for task-oriented communication with digital modulation. IEEE J Sel Areas Commun 2023;41(8):2577–91.
[13] Miettinen K. Nonlinear multiobjective optimization. Boston: Springer; 1999.
[14] Ho J, Jain AN, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems; 2020 Dec 6–12; online. Red Hook: Curran Associates Inc.; 2020.
[15] Li B, Di Z, Lu Y, Qian H, Wang F, Yang P, et al. Expensive multi-objective Bayesian optimization based on diffusion models In: Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence; 2025 Feb 25–Mar 4; Philadelphia, PA, USA. Association for the Advancement of Artificial Intelligence; 2024. p.27063–71.
[16] Liu B, Liu X, Gao S, Cheng X, Yang L. LLM4CP: adapting large language models for channel prediction. J Commun Inf Netw 2024;9(2):113–25.
[17] Fan S, Liu Z, Gu X, Li H. Csi-LLM: a novel downlink channel prediction method aligned with LLM pre-training. In: Proceedings of IEEE Wireless Communications and Networking Conference; 2025 Mar 24–27; Milan, Italy; New York City: IEEE; 2025.
[18] Zheng T, Dai L. Large language model enabled multi-task physical layer network. IEEE Trans Commun 2026;74(1):307–21.
[19] Wu D, Wang X, Qiao Y, Wang Z, Jiang J, Cui S, et al. NetLLM: adapting large language models for networking. In: Proceedings of ACM SIGCOMM 2024 Conference; 2024 Aug 4–8; Sydney, NSW, Australia. New York City: Association for Computing Machinery; 2024. p. 661–78.
[20] Wang Z, Zhou Y, Shi Y, Letaief KB. Federated fine-tuning for pre-trained foundation models over wireless networks. IEEE Trans Wirel Commun 2025;24(4):3450–64.
[21] Zhao Y, Zhou Y, Wang Z, Shi Y, Cheng N, Zhou H. Learning to beamform for integrated sensing and communication: a graph neural network with implicit projection approach. IEEE Trans Wirel Commun 2025;24(7):5931–45.
[22] Chen G, Wang Z, Jia Y, Huang Y, Yang L. An efficient architecture search for scalable beamforming design in cell-free systems. IEEE Trans Veh Technol 2024;73(7):10241–53.
[23] Lian L, Bai C, Xu Y, Dong H, Cheng R, Zhang S. Learning to beamform for cooperative localization and communication:a link heterogeneous GNN-based approach. IEEE Trans Wirel Commun 2025;25:6177–90.
[24] Abdelsadek MY, Chaudhry AU, Darwish T, Erdogan E, Karabulut-Kurt G, Madoery PG, et al. Future space networks: toward the next giant leap for humankind. IEEE Trans Commun 2023;71(2):949–1007.
[25] Shi Y, Zhou Y, Wen D, Wu Y, Jiang C, Letaief KB. Task-oriented communications for 6G: vision, principles, and technologies. IEEE Wirel Commun 2023;30(3):78–85.
[26] Goldfeld Z, Polyanskiy Y. The information bottleneck problem and its applications in machine learning. IEEE J Sel Areas Inf Theory 2020;1(1):19–38.
[27] Yang P, Wen D, Zeng Q, Zhou Y, Wang T, Cai H, et al. Over-the-air computation empowered vertically split inference. IEEE Trans Wirel Commun 2024;23(12):19634–48.
[28] Wang Z, Zhao Y, Zhou Y, Shi Y, Jiang C, Letaief KB. Over-the-air computation for 6G: foundations, technologies, and applications. IEEE Internet Things J 2024;11(14):24634–58.
[29] Shi Y, Zeng L, Zhu J, Zhou Y, Jiang C, Letaief KB. Satellite federated edge learning: architecture design and convergence analysis. IEEE Trans Wirel Commun 2024;23(10):15212–29.
[30] Zhu J, Shi Y, Zhou Y, Jiang C, Kuang L. Hierarchical learning and computing over space–ground integrated networks. IEEE Trans Mobile Comput 2025;24(10):10423–40.
[31] Yang P, Wang T, Cai H, Shi Y, Jiang C, Kuang L. Brain-inspired decentralized satellite learning in space computing power networks. IEEE Trans Mobile Comput 2025;24(12):12935–49.