Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35669

# *A Review of Distributed Large Models in Space-Air-Ground Integrated Networks*

**Haotian Ding**

*School of Telecommunication Engineering, Xidian University, Xi'an, China*
*517210496@qq.com*

***Abstract.*** The space-air-ground integrated network offers a forward-looking direction for alleviating ground computing power bottlenecks. Efficiently deploying large, distributed models, a key technology for 6G ubiquitous intelligence, in highly dynamic, resource-heterogeneous environments holds significant strategic and application value. This paper systematically reviews the current state of research, core challenges, and key technologies for constructing distributed large-scale models in space-air-ground integrated networks. It analyzes critical issues,, including network heterogeneity, on-board resource constraints, and the security-efficiency trade-off, and evaluates the strengths, limitations, and applicability of existing solutions. The study reveals that adaptive model partitioning, lightweight security protocols, and standardized frameworks remain major shortcomings. Future research should focus on lightweight model design and architecture standardization, deep integration of privacy computing with communication, and joint scheduling of communication, sensing, and computing. This paper clarifies current research limitations and future directions, offering references for the large-scale application of space-air-ground integrated computing power networks.

***Keywords:*** Space-Air-Ground Integrated Network, Distributed Large Model, Heterogeneous Collaboration, Federated Learning, On-board Computing Power Optimization

## 1. Introduction

The space-air-ground integrated network, comprising space-based backbone, space-based access, and ground node networks, serves as the core infrastructure for 6G communication-sensing-computing integration and space-ground computing convergence. Extensive research has been conducted on network architecture integration, service adaptation, and scenario applications, with deeply converged space-air-ground integrated low-altitude aerial vehicular networks emerging as a recent focus, covering architecture design, key challenges, and enabling technologies [1, 2]. Despite these advances, existing studies primarily address network-level connectivity and service provisioning, leaving large-scale intelligent computing largely unexplored.

With the advent of the large model era, efficiently deploying distributed large models in space-air-ground integrated networks has become an urgent problem, driven by the need to overcome the computing power bottleneck of ground data centers and achieve ubiquitous computing power sharing. Current research remains fragmented, focusing on isolated aspects such as lightweight

DOI: 10.54254/2755-2721/2026.GL35669

model inference, space-ground federated learning, or on-board resource scheduling, without a systematic synthesis targeting distributed large model construction [3].

To address this gap, this paper systematically review the technical status, core challenges, and key enabling technologies for constructing distributed large models in space-air-ground integrated networks. It analyzes fundamental issues including network heterogeneity and dynamics, strict on-board resource constraints, and the tension between data security and communication efficiency. By comparing the strengths, limitations, and applications of mainstream solutions, this paper identifies key technical approaches, offering theoretical references for future research and engineering implementation.

## 2. Fundamentals of space-air-ground integrated networks and distributed large model architecture

### 2.1. Typical architecture and characteristics of space-air-ground integrated networks

The space-air-ground integrated network consists of three types of heterogeneous nodes: low-orbit satellite constellations, ground base stations, and aerial platforms, achieving seamless global coverage through inter-satellite links and space-ground links [4]. The hyper-converged architecture further connects multi-domain nodes across space, air, ground, and low-altitude vehicles, forming a globally collaborative networking paradigm [2]. Its core characteristics are mainly reflected in three aspects.

Topology dynamics arise from the rapid time-varying network topology caused by the high-speed movement of low-orbit satellites, with significant fluctuations in space-ground link delays and susceptibility to atmospheric interference, rain, and snow. Resource heterogeneity is reflected in the orders-of-magnitude gap between on-board platforms and ground data center nodes in terms of computing power, power consumption, and storage resources [5]. Service complexity requires the network to simultaneously support remote sensing monitoring, emergency communication, global Internet of Things, and other services, imposing strict constraints on network delay, reliability, and transmission bandwidth [6]. The limited hardware of space-based intelligent computing and the complexity of on-orbit operation and maintenance further highlight the inherent challenges of heterogeneous deployment in space-air-ground networks [7].

### 2.2. Deployment modes of distributed large models in space-air-ground scenarios

Distributed large models are mainly deployed in three modes in space-air-ground integrated scenarios. The centralized deployment places the complete model in a ground data center, with satellites only responsible for data collection and return. This architecture is simple but imposes enormous pressure on space-ground transmission, making it difficult to support tasks with high real-time requirements. The space-ground collaborative distributed deployment adopts model splitting and hybrid parallel strategies to distribute model layers, operators, and parameter fragments across on-board and ground nodes for collaborative training, effectively adapting to scenarios with limited on-board computing power [8]. Federated deployment allows each node to train independently on local data, exchanging only gradients and incremental model parameters. While this protects data privacy, it also introduces additional communication overhead [7]. Therefore, the selection among the three modes must be comprehensively matched based on on-board computing power, link bandwidth, and data security requirements.

DOI: 10.54254/2755-2721/2026.GL35669

## 3. Core challenges in constructing distributed large models in space-air-ground integrated networks

### 3.1. Challenges of network heterogeneity and dynamics

The heterogeneity of nodes and the time-varying topologies in space-air-ground networks pose significant challenges for traditional static distributed training frameworks [5]. Conventional distributed training relies on fixed topologies, stable delays, and clock synchronization. However, the orbital movement of low-orbit satellites leads to link interruptions, bandwidth fluctuations, and delay jitter, which can easily cause unbalanced task scheduling, parameter aggregation timeouts, and unstable training convergence. At the same time, the computing power gap between space and ground nodes is substantial, leading to significant load imbalance and directly restricting the stability and efficiency of cross-domain collaborative training [4]. The hyper-converged space-air-ground network involves more node types and stronger heterogeneity, further amplifying the collaborative difficulties caused by topology dynamics [2].

### 3.2. Challenges of strict on-board resource constraints

On-board platforms face hard constraints on computing power, power consumption, memory, and storage at the quantitative level [6]. The computing power of mainstream on-board processing units is far lower than that of ground servers; typical processors usually do not exceed 1.2 GHz, and FP32 performance is often less than 50 GFLOPS. In addition, power supply is limited by solar illumination cycles, resulting in a significant drop in available computing power in shadow regions. Meanwhile, space radiation and limited heat dissipation conditions make it difficult to support long-term, high-load continuous computing. Directly deploying large billion-parameter models will cause memory overflow, excessive power consumption, and long inference delays. Therefore, lightweight models, fragmented deployment, and computing offloading must be adopted to adapt to on-orbit environments [9]. The scarcity of space-based computing resources and harsh on-orbit operating conditions are the core bottlenecks that restrict the on-board deployment of large distributed models [10].

### 3.3. Challenges in balancing data security and communication efficiency

Data transmitted across space-air-ground scenarios, such as remote sensing data and critical Internet of Things data, are often classified or industry-sensitive, requiring them to be "usable but not visible," with strong privacy protection needs [11]. However, space-ground links have extremely limited bandwidth and round-trip delay jitter of 5–50 ms. Existing federated learning requires uploading model updates in each round, easily leading to training interruptions; differential privacy can protect individual data by adding noise, but significantly amplifies gradient variance, requiring more communication rounds for convergence and further increasing link occupancy; homomorphic encryption imposes extremely high demands on on-board computing power. Therefore, existing solutions face problems such as excessive communication overhead occupying space-ground link bandwidth and encryption/privacy operations consuming substantial on-board computing power, forming an inherent contradiction where improved security reduces efficiency, making it difficult to balance both [3]. The European Union's European Data Space Strategy constructs a cross-domain trusted data sharing framework, but it is mostly oriented toward ground cloud scenarios and does not adapt to the low-power and low-computation constraints on board [12]. Existing lightweight

DOI: 10.54254/2755-2721/2026.GL35669

solutions still struggle to achieve this balance, and designing a secure collaborative protocol adapted to dynamic space-ground links remains an open challenge [3].

## 4. Analysis of key technical systems and typical schemes

### 4.1. Collaborative training technology for heterogeneous space-ground nodes

Space-ground collaborative training lies at the core of current distributed large model construction, mainly based on three technologies: federated learning, model splitting, and hybrid parallelism [8]. For federated learning, the "Star Computing Project" at Guoxing Aerospace has verified a lightweight federated learning architecture in orbit, enabling on-board large-scale model inference and demonstrating the feasibility of federated learning in resource-constrained on-board environments [9]. China Telecom focuses on model splitting and hybrid parallelism, adopting wide-area lossless transmission and collaborative scheduling strategies to significantly improve the efficiency of cross-node joint training by appropriately partitioning model layers between space and ground [8]. Research on on-orbit space AI algorithms is gradually shifting toward federated and multi-agent collaborative architectures, providing new ideas for on-board distributed training [13]. Studies represented by SpaceX Starlink and other research on AI scheduling in space-air-ground networks have produced more theoretical simulation results, but large-scale on-orbit engineering implementation still lags behind [4].

### 4.2. Joint optimization technology for heterogeneous space-ground resources

Joint optimization of heterogeneous space-ground resources mainly covers three aspects: computing power scheduling, storage allocation, and adaptive power consumption control [5]. Computing power scheduling dynamically splits training tasks by sensing satellite positions, link status, and remaining node computing power in real time, offloading computationally intensive operators to the ground to achieve elastic release of on-board computing resources. Model fragmentation and on-demand parameter loading activate only the necessary parameters during inference or training, effectively reducing peak instantaneous memory occupancy and enabling storage optimization. In terms of power consumption management, training accuracy and iteration frequency are adaptively adjusted based on power supply variations across sunlit and shadow regions, achieving stable control of power output [6]. The space-based distributed operating system developed by Zhejiang Laboratory has realized intelligent scheduling of on-board resources and lightweight adaptation of large models [10]. The resource virtualization, collaborative scheduling, and heterogeneous integration technologies of space-based computing power networks provide systematic support for distributed training in space-air-ground networks [10].

### 4.3. Communication efficiency optimization technology

To address scarce bandwidth and large delay fluctuations in space-ground links, current research mainly adopts three approaches: gradient sparse compression, model quantization, and asynchronous aggregation protocols [3]. Gradient sparse compression uses low-rank decomposition or sparsity pruning to reduce per-round communication volume while keeping gradient components critical for convergence. Model quantization converts 32-bit floating-point parameters to 8-bit integers, greatly cutting transmission and storage overhead with controlled accuracy loss. At the protocol level, an adaptive asynchronous aggregation strategy lets nodes independently upload

DOI: 10.54254/2755-2721/2026.GL35669

updates based on their own computing progress and link status, avoiding link troughs, reducing unnecessary communication, and alleviating space-ground link congestion [7]. The communication-computing joint optimization in 6G integrated communication, sensing, and computing further supports efficient space-ground transmission [14]. However, because experimental settings (model scale, satellite orbit, link model) vary widely, the three optimization approaches lack a standardized evaluation benchmark tailored to space-air-ground integrated scenarios, leaving insufficient data support for engineering selection.

### 4.4. Privacy and security enhancement technology

Lightweight privacy technology adapted to on-board resource constraints is a core research focus, mainly encompassing lightweight differential privacy, simplified homomorphic encryption, and private set intersection, among other directions [12]. The research aims to achieve cross-domain sample alignment and collaborative model updates without leaking raw data, while balancing privacy protection, computational overhead, and communication costs. For example, taking lightweight differential privacy, introducing a sparse noise mechanism that restricts noise addition to key model parameters can further reduce computational overhead. Simplified homomorphic encryption trades off deployability on-board by reducing the multiplication depth of polynomial degrees. The EU data space security specifications provide reference standards for cross-domain data governance in space-air-ground networks, but algorithm tailoring and protocol lightweight adaptation are still required for on-board constrained scenarios [9]. The collaborative, trusted, distributed reasoning technology for space-based intelligent models can effectively address privacy leakage and untrusted reasoning in cross-domain training [14].

## 5. Future directions

### 5.1. Technology development trends

Distributed large models in space-air-ground integrated networks will exhibit three major trends in the future. First, ultra-lightweight models and standardized space-ground collaborative architectures will form dedicated training and inference frameworks. Second, the deep integration of privacy-preserving computing and communication optimization will resolve the inherent tension between security and efficiency. Integrated communication, sensing, and computing will require the unified integration of remote sensing perception, communication transmission, and large model intelligent computing into space-ground resource scheduling systems [3, 4]. Meanwhile, the evolution of hyper-converged space-air-ground networking will drive distributed large models toward multi-domain ubiquitous collaboration [2]. Space-based intelligent computing will advance toward ubiquitous computing power, lightweight models, and collaborative trustworthiness, enabling the large-scale deployment of space-air-ground integrated computing networks [7]. The maturity of 6G integrated communication, sensing, and computing technology will further dissolve the boundaries among communication, sensing, and computing, laying the foundation for the ubiquitous deployment of distributed large models [14].

### 5.2. Future directions

Further research on distributed large models in space-air-ground integrated networks can target four levels: algorithms, systems, evaluation, and standards. Develop adaptive distributed training

DOI: 10.54254/2755-2721/2026.GL35669

frameworks for dynamic topologies. At the system level, address on-board resource bottlenecks and design lightweight, low-power privacy and security protocols. For evaluation, build real testbeds, conduct on-orbit measurements and algorithm iterations, and establish benchmark datasets and performance evaluation frameworks for these scenarios. For standards, promote unified industry specifications for space-ground interfaces, data formats, and collaborative protocols [5, 11]. Building on the 6G integrated computing power network framework, construct a collaborative large-model technology system spanning space and ground, and create standardized, scalable space-ground collaborative training platforms. Furthermore, improve the scheduling mechanism for space-based computing power networks to enhance training collaboration efficiency among heterogeneous nodes [10].

## 6. Conclusion

This paper studies distributed large-model construction in space-air-ground integrated networks. Driven by the national computing power network strategy and 6G convergence requirements, it analyzes ground computing limits and network features, reviews progress and gaps, and identifies three core challenges: heterogeneous collaboration, on-board resource constraints, and the security-efficiency tradeoff. No existing solution is complete. Space-ground collaborative distributed deployment is the most promising, but its model partitioning lacks adaptability to dynamic topologies. Federated deployment protects privacy, but no lightweight scheme reconciles communication overhead with scarce bandwidth. Collaborative training and joint optimization have partial validation, yet communication efficiency and privacy security still need breakthroughs.

This research has certain limitations. It relies on existing literature without simulation verification or on-orbit hardware testing. The algorithms do not fully consider the impact of extreme space environments, including radiation, extreme temperatures, and periodic power supply fluctuations, on training stability. The compatibility of lightweight security protocols with dynamic space-ground links has not been verified through actual measurements, and no unified technical standard has been established. Further research will be carried out in future work.

## References


[1] Gong, L., & Zhuang, Y. (2026). Research on integration and application analysis of space-air-ground integrated networks. *Telecommunications Technology*, 1–3.

[2] Zhou, H., Xue, J., Yuan, Y., et al. (2026). Hyper-converged space-air-ground integrated low-altitude aerial vehicular networks: Architecture, challenges, and technologies. *Fundamental Research*, 1–5.

[3] Li, J., Zhang, Y., & Liu, Q. (2023). Key technical challenges of heterogeneous collaboration in space-air-ground integrated networks. *Journal of Communications, 44*(5), 128.

[4] Zhang, P., Tao, X., & Xu, W. (2023). Dynamic topology control and synchronization protocol for 6G space-air-ground integrated networks. *Journal of Tsinghua University (Natural Science Edition), 63*(8), 1321–1330.

[5] Zhao, J., Liu, M., & Yang, L. (2024). Analysis of the impact of computing power heterogeneity of low-orbit satellite networks on distributed training. *Journal of Astronautics, 45*(1), 110–118.

[6] Sun, Y., Zhao, Y., & Ma, J. (2023). Challenges of AI model deployment under resource constraints of on-board computing platforms. *Acta Aeronautica et Astronautica Sinica, 44*(6), 320-332.

[7] Wu, Z., Fang, H., & Yang, X. (2025). Reflections on the development of space-based intelligent computing. *Space Electronic Technology, 5*(S1), 140–147.

[8] China Telecom. (2024). White paper on wide-area cross-node large model joint training technology [White paper].

[9] European Union Agency for Cybersecurity. (2024). Engineering personal data protection in EU data spaces. *ENISA*.

[10] Zhang, D., Dong, T., Liu, Z., et al. (2025). Review of key technologies for space-based computing power networks. *Space-Air-Ground Integrated Information Network, 6*(4), 9–26.

DOI: 10.54254/2755-2721/2026.GL35669

[11] Guoxing Aerospace. (2026, February 27). Latest progress and roadmap of Star Computing Project. *Guangming Net*. Retrieved March 7, 2026, from https://www.gmw.cn

[12] Zhejiang Laboratory. (2024, November 11). R&D achievements of space-based distributed operating system and Tianhui large model. *Zhejiang Laboratory Official Website*. Retrieved March 7, 2026, from https://www.zhejianglab.com

[13] Duggan, D. (2026). On-orbit space AI: Federated, multi-agent, and collaborative algorithms for satellite constellations. *Algorithms, 19*(7), 318.

[14] Wu, Z., Zhang, H., Ma, X., et al. (2025). Architecture and key technologies of 6G integrated communication, sensing and computing. *Journal of Electronics and Information Technology, 47*(4), 876–887.