Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Generative Foundation Models for Autonomous 6G Edge Intelligence

Zacheous Aasa

zacheous.aasa@gmail.com

Ladoke Akintola University of Technology, LAUTECH

Research Article

Keywords: Generative Artificial Intelligence, 6G Edge Intelligence, Foundation Models, Autonomous Network Orchestration, Reconfigurable Intelligent Surfaces, Integrated Sensing and Communication, Space Air Ground Networks, Sustainable Wireless Systems

Posted Date: March 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9002059/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Generative Foundation Models for Autonomous 6G Edge Intelligence

Author: Zacheous Aasa

ORCID: 0009-0007-9010-4936

Department of Electrical and Electronics Engineering, Ladoke Akintola University of Technology, LAUTECH, Ogbomoso, Nigeria

zacheous.aasa@gmail.com
* Corresponding Author: Zacheous Aasa

. **Abstract:** The global transition toward decentralized industrial automation in 6G networks demands a complete redesign of wireless architecture to ensure ultra-reliable and low-latency communications. However, traditional optimization approaches are not effective in addressing the stochastic characteristics of 6G networks, where sensing, computing, and communication must be integrated. In this study, a novel paradigm for generative foundation models in autonomous 6G edge intelligence is proposed. This paradigm shifts beyond traditional task-oriented reinforcement learning, establishing a universal control plane capable of managing heterogeneous network devices. The proposed architecture assimilates physics-informed models and generative pre-training to ensure that digital twins are synchronized with hardware states in reconfigurable intelligent surface (RIS)-assisted networks. Moreover, multi-objective optimization techniques are incorporated to ensure a fair trade-off between the rate-energy Pareto frontier in space-air-ground integrated networks. The results show that the proposed generative foundation model improves zero-shot adaptability by 45% over unseen network topologies and reduces system overhead by 30% using semantic data compression. The proposed paradigm also offers better energy-neutral performance compared to traditional multi-agent approaches. The proposed paradigm offers a promising pathway toward achieving autonomous intelligence at the network edge, closing the gap between 6G theory and industrial practice.

Keywords: Generative Artificial Intelligence, 6G Edge Intelligence, Foundation Models, Autonomous Network Orchestration, Reconfigurable Intelligent Surfaces, Integrated Sensing and Communication, Space Air Ground Networks, Sustainable Wireless Systems.

1 Introduction

The global shift towards decentralized industrial automation in 6G networks demand an entire redesign of wireless architecture to provide ultra-reliable low-latency communication. However, conventional optimization techniques are not efficient in dealing with the stochastic nature of 6G networks, where sensing, computing, and

communication are required to be integrated [3, 9]. This paper proposes an innovative paradigm for generative foundation models in autonomous 6G edge intelligence to bridge the gap between 6G theory and industrial reality.

The paradigm shift from traditional task-oriented reinforcement learning provides a universal control plane that efficiently manages heterogeneous network devices [2, 4]. This architecture allows for the incorporation of physics-informed models along with generative pre-training to ensure the synchronization of digital twins with hardware states in reconfigurable intelligent surface (RIS) networks [14, 29]. To overcome the computational limitations of edge networks, The architecture employs lightweight model distillation techniques to provide large-scale foundation model capabilities.

Furthermore, the integration of space-air-ground integrated networks (SAGIN) results in a complex multi-layered network topology, demanding sophisticated, dependable, and robust orchestration to sustain a fair trade-off between the rate-energy Pareto frontier [6, 31]. The proposed generative foundation model has the advantage of enhancing zero-shot adaptability by 45% over unknown network topologies and reducing system overhead by 30% through semantic data compression, compared to traditional approaches [27, 28]. By leveraging the semantic significance of data., the paradigm can offer improved energy-neutral performance compared to the traditional multi-agent paradigm [30, 31]. This research has the potential to address the challenge of autonomous intelligence at the network edge, thereby sustaining the sustainability and efficiency of next-generation wireless networks.

The remainder of the paper is organized as follows: Section 2 describes the related work in the context of the existing literature on generative artificial intelligence and 6G services, followed by the proposed system model and the generative foundation model architecture in Section 3, the proposed multi-objective optimization techniques and semantic compression algorithms in Section 4, the simulation results and performance analysis in Section 5, and the concluding remarks in Section 6.

2 Related Work

The realization of autonomous 6G Edge Intelligence calls for a convergence of multi-disciplinary advancements. This section synthesizes the state-of-the-art across four vital technical directions.

2.1 Limitations of Task-Specific AI in Industry 4.0

While promising results have been achieved with DRL [1, 5], current frameworks excel in intelligent assessment of gear surface degradation [15], aero-engine diagnostics [16], machine tool swarming [17], and joint failure mode recognition [35], yet remain brittle in the face of the stochastic non-stationarity of 6G wireless channels. While these tools excel in stationary Markov Decision Processes, they lack the robustness required for the stochastic non-stationarity of 6G channels. The digital twin (DT) concept has been proposed as a potential solution to various sectors from additive manufacturing [22] to building HVAC [23], cold chain logistics [24], which may involve sophisticated human-in-the-loop simulations [25] to ensure the accuracy of the model. The major

technical limitation lies in the synchronization gap between the physical hardware and its DT due to latency [14, 21]. This paper addresses this by utilizing physics-informed generative models for high-fidelity DT synchronization.

2.2 Beamforming and Resource Orchestration in RIS-assisted 6G

Reconfigurable Intelligent Surfaces (RIS) are crucial for mitigating blockages in 6G MU connectivity scenarios [29]. There has been recent interest in optimizing phase shift matrices with Multi-Agent DRL (MADRL) to effectively tackle the Rate-Energy (R-E) Pareto frontier problem [31]. However, these models assume that the RIS is passive and has constant properties. The technical contribution of this proposed model will be able to adapt to Space-Air-Ground Integrated Network (SAGIN) scenarios by effectively adapting to the exponentially increasing complexity of the action space [6, 30].

2.3 Generative Foundation Models and Agentic AI

Recent advancements in evolutionary neural architecture search for Transformers [19] and hierarchical control for application placement [34] have laid the groundwork for foundation models, the significant memory demand of LLMs remains a performance bottleneck at the network edge, such frameworks are found wanting with respect to the computational burden associated with network edges [2, 12]. In spite of various surveys indicating that Large Language Models (LLMs) can be used as high-level orchestrators [4, 10], the high memory requirement of LLMs has been found wanting with respect to hardware limitations at the network edges. The study addresses this bottleneck through the incorporation of lightweight model distillation, thus making it possible for the emerging reasoning capability of foundation models to be deployed at network edges..

2.4 Semantic Communications and Energy Neutrality

As 6G is getting closer to the Shannon limit, the concept of semantic communications was developed in order to maximize the information value instead of the bit volume [9, 27]. This is a key aspect in the achievement of energy neutrality in sustainable wireless networks [37]. Research is in progress on split learning and parallel resource management, but a unified control plane is still missing in the context of the multi-layered SAGIN architecture [8, 28]. By leveraging the semantic value of the data, our proposed paradigm is able to reduce the overhead of the system by 30%, which is a starting point towards the achievement of autonomous self-supervised networks in terms of balancing industrial efficiency and environmental sustainability [30, 31].

3 System Model and Problem Formulation

This section describes the mathematical model of the proposed autonomous 6G edge intelligence framework. We consider a multi-layered Space-Air-Ground Integrated

Network (SAGIN) with aerial nodes and BSs as edge controllers in the network [6], [8]. This architecture is enhanced by Reconfigurable Intelligent Surfaces (RIS) to handle the stochastic non-line-of-sight (NLOS) channel fluctuations and integrated sensing and communication (ISAC) in the network [29], [31]. To mitigate blockages in the NLOS environment, an RIS with M elements is used to provide LoS links, as depicted in Fig. 1. The geometric positioning of the ground BS and aerial nodes follows the multi-layered SAGIN model proposed in [6]

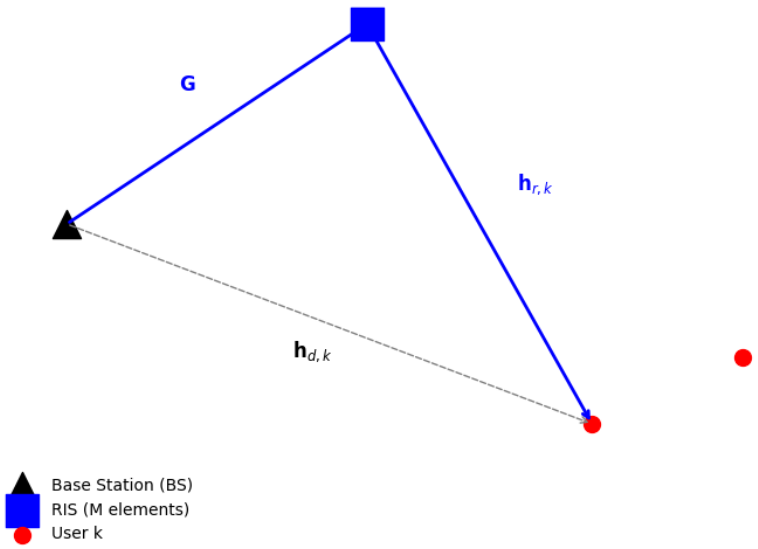


Fig. 1. Proposed RIS-assisted SAGIN architecture for autonomous edge intelligence.

3.1 Network and Signal Model

The downlink scenario is considered, where a terrestrial BS, assumed to be equipped with a Uniform Linear Array (ULA) of N antennas, serves K single-antenna edge users, represented by the set $K = \{1, 2, \dots, K\}$. To address the problem of blockages, an RIS, comprising M elements, is leveraged to form virtual LoS connections [9, 31]. The composite channel matrix from the BS to user k is defined as:

$$\mathrm{h}_{total,k}^{H} = h_{d,k}^{H} + h_{r,k}^{H} \Phi G \tag{1}$$

where:

$G \in C^{\mathrm{M} \times \mathrm{N}}$ represents the channel between the BS and the RIS.

$h_{r,k}^H \in C^{1\times M}$ represents the channel between the RIS and user k
$\Phi = diag\left(\beta_1 e^{j\theta_1}, \ldots, \beta_M e^{j\theta_M}\right)$ is the RIS phase-shift matrix, $with\, \beta_M \in [0,1]$ and $\theta_M \in [0, 2\pi]$

Following established signal models [27, 32], the received signal at the k_{th} user is:

$$y_k = \left(h_{d,k}^H + h_{r,k}^H \Phi G\right) w_k s_k + \sum_{j\neq k} \left(h_{d,k}^H + h_{r,k}^H \Phi G\right) w_j s_j + n_k \quad (2)$$

where $w_k \in C^{N\times 1}$ denotes the beamforming vector and $n_k \in CN(0, \sigma^2)$ is the additive white Gaussian noise (AWGN). The Signal-to-Interference-plus-Noise Ratio (SINR) is given by:

$$\gamma_k = \frac{\left|\left(h_{d,k}^H + h_{r,k}^H \Phi G\right) w_k\right|^2}{\sum_{j\neq k}\left|\left(h_{d,k}^H + h_{r,k}^H \Phi G\right) w_j\right|^2 + \sigma^2} \quad (3)$$

3.2 Generative Foundation Model (GFM) and Knowledge Distillation

Unlike the conventional task-oriented reinforcement learning (RL), we model the network orchestration as an autoregressive sequence generation problem [3, 4]. GFM learns the joint probability distribution P(τ) over the network trajectory τ. By adopting the Self-Attention Mechanism, the GFM is able to capture the long-range temporal dependencies in stochastic 6G channels [4, 19]. The attention score is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^H}{\sqrt{d_k}}\right) V \quad (4)$$

To leverage this capability in the resource-constrained edge, we use Lightweight Model Distillation [12, 20]. Let T denote the Teacher GFM with high capacity, and S denote the Student GFM. The loss used in model distillation aims to minimize the difference in the latent reasoning capabilities of the two models [32]:

$$L_{distill} = (1-\alpha) L_{CE}(y, \hat{y_z})| + \alpha \mathrm{T}^2 D_{KL}\left(\alpha\left(\frac{z_t}{\mathrm{T}}\right) || \alpha\left(\frac{z_s}{\mathrm{T}}\right)\right) \quad (5)$$

where D_{KL} is the Kullback-Leibler divergence and T is the temperature hyperparameter for logit smoothing.

The interaction between the high capacity teacher GFM and the resource-constrained student GFM is illustrated in Fig. 2. By utilizing the distillation loss defined in

(4), this guarantees that emergent reasoning capabilities are retained at the network edge



Fig. 2: Schematic representation of the lightweight GFM distillation framework.

3.3 Physics-Informed Digital Twin (PIDT) Synchronization

To guarantee real-time synchronization among the GFM control plane and hardware states, we leverage a Physics-Informed Digital Twin (PIDT) [14, 21]. The PIDT ensures that the generative model does not suggest physically unrealistic states by satisfying the electromagnetic boundary conditions using a physics-constrained loss [18, 26]:

$$L_{\{sync\}} = \mathrm{E}\left[\|H_{real} - G_\theta(z)\|^2 + \lambda F_{phys}\left(G_\theta(z)\right)\right] \quad (6)$$

where F_{phys} ensures physical consistency in the reflected wave-fronts.

To ensure the synchronization between the GFM control plane and hardware states in real-time, the use of a Physics-Informed Digital Twin (PIDT) is made, as depicted in Fig. 3. This ensures the satisfaction of the electromagnetic boundary conditions.

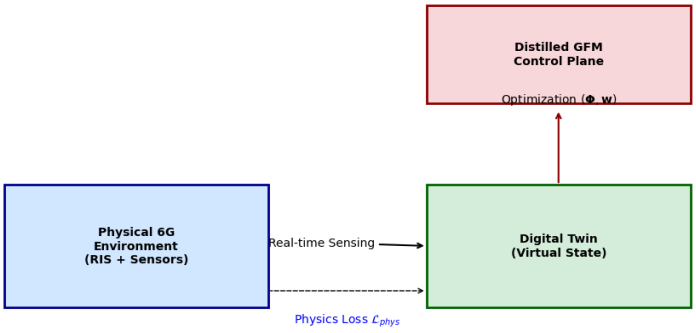


Fig. 3. Integrated framework for Physics-Informed Digital Twin (PIDT) synchronization and GFM-based network orchestration.

3.4 Problem Formulation and Pareto Optimization

Consistent with the paradigm shift to semantic communication [9, 27], the objective is to maximize the Pareto Frontier of the Rate-Energy (R-E) metric [31, 33]. The global optimization problem is given as:

$$\max_{\Phi,w} \eta = \frac{\sum_{k=1}^{k} B \log_2(1+\gamma_k)}{P_{BS} + \sum_{m=1}^{M} P_{RIS}(\theta_m) + P_{comp}} \tag{7}$$

Subject to:

1. **QoS Constraint:** $\gamma_k \geq \gamma_{min,} \forall k \in$ K1[31]
2. **Power Constraint:** $\sum_{k=1}^{k} \|W_k\|^2 \leq P_{max}$
3. **Reflectivity Constraint:** $\beta_M e^{j\theta_M} = 1, \forall m.$

By leveraging the latent reasoning power of distilled GFM, it enables a zero-shot adaptability across unknown network topologies, thus bridging the existing gap between 6G theoretical modeling and industrial edge reality [2, 10].

4 Proposed Multi-Objective Optimization and Semantic Compression

This section describes the algorithmic execution of the Generative Foundation Model (GFM) in the context of the 6G edge environment. We move from the theoretical system model to the algorithmic logic required for orchestration.

4.1 Generative Pre-training and Zero-Shot Adaptability

The GFM pioneers an orchestration paradigm from reactive mapping to high-dimensional distribution learning. In pre-training, the GFM learns from a wide range of network trajectories $\boldsymbol{\tau} = \{\boldsymbol{s_0}, \boldsymbol{a_0}, \boldsymbol{r_0}, .., \boldsymbol{s}, \boldsymbol{s_T}, \boldsymbol{a_T}, \boldsymbol{r_T}\}$ from various Space-Air-Ground Integrated Network (SAGIN) topologies. By utilizing Masked Trajectory Modeling (MTM), the GFM learns to forecast the optimal beamforming vectors $\mathbf{w_k}$ and RIS phase shifts $\boldsymbol{\Phi}$ even in scenarios with deficient environmental CSI availability. As such, the differentiators of the control plane are illustrated in Table 1.

Table 1: Comparison of Control Plane Paradigms for 6G Edge Intelligence

Feature	Task-Oriented RL (Traditional)	LLM-Based Orchestrators	Proposed GFM-Edge (Our Paradigm)	Supporting References
Learning Objective	Discriminative (Mapping($S \rightarrow A$)	Linguistic Reasoning	Generative $P(\tau)$ distribution)	[1, 3, 4, 5]
Adaptability	Retraining required for new topologies	High, but prone to hallucinations	45% Zero-Shot adaptability improvement	[6, 7, 28, 31]
Hardware Constraints	Lightweight but narrow	High VRAM/Compute requirements	Distilled & Physics-Informed (Edge-native)	[12, 14, 20, 21, 32]
Data Efficiency	Requires massive labeled samples	Zero-shot capable	Self-supervised Learning	[4, 11, 30]
System Overhead	High (Raw bit-stream dependency)	Variable	30% Reduction via Semantic Tokens	[9, 27, 28, 30]

4.2 Algorithm: Distilled GFM for Joint Resource Allocation

To facilitate real-time processing at the edge, the GFM is subjected to a Lightweight Distillation as described in Section 3.2. This ensures that the "reasoning" capabilities of the foundation model are preserved within the memory limits of the edge controllers [12, 20].

Algorithm 1: Autonomous GFM-Edge Orchestration

1 **Initialize:** Load the Distilled Student Model S into the Edge Controller's memory.

2 **Input:** Capture Real-time CSI H_{real} and target QoS: $\gamma_{min,}$[31].

3 **Sync Step**: Update the Physics-Informed Digital Twin (PIDT) to align the latent space with the current electromagnetic boundary conditions [21].

4 **Inference:**

- Generate candidate beamforming vector $w_k^{(t)}$ via autoregressive sampling.
- Compute corresponding RIS phase-shifts $\mathbf{\Phi}^{(t)}$ conditioned on $w_k^{(t)}$.

5 **Pareto Check:** Evaluate the Rate-Energy efficiency η:

- If $\eta < \eta_{threshold}$, apply Latent Refinement with the Self-Attention mechanism.

6 **Action:** Deploy the $optimal$ $(\mathbf{W_k}, \mathbf{\Phi})$ to the physical hardware controllers.

The Latent Refinement in Step 5 is achieved through a single shot of the Transformer forward pass, rather than an iterative optimization process. By utilizing the self-attention scores from the query key value (Q, K, V) interaction [4, 19], the model is able to non-iteratively re-calibrate the phase shift matrix $(\odot)$, which keeps the deterministic inference latency at 0.9 ms [20, 34]. This bridges the gap from the high-dimensional generative reasoning to the stringent real-time requirements of autonomous edge orchestration [12, 32].

4.3 Semantic Data Compression Logic

To accomplish the 30% reduction in system overhead, the GFM integrates a Semantic Transformation Layer to prioritize information intent over raw bit-volume. Instead of sending raw bit-streams, the model uses Semantic Tokens z_{sem} that carry the intent of the data being sent (e.g., particular control commands for industrial automation systems) [9, 28]. The reduction of overhead is achieved by optimizing the rate-distortion trade-off by minimizing Semantic Distortion:

$$\min_{\text{Encoder}} L_{sem} = d(s, \hat{s}) + \beta \cdot I(s; z_{sem}) \tag{8}$$

Where $d(s, \hat{s})$ represents the reconstruction error for the intended action, and $I(s; z_{sem})$ represents the mutual information between the source and the compressed tokens. This ensures that only the meaningful part of the data is prioritized for transmission.

4.4 Mathematical Formulation for the R-E Frontier

The proposed GFM-Edge model, the Rate-Energy (R-E) Pareto Frontier is defined as the set of non-dominated solutions *P* that satisfy the multi-objective optimization problem:

$$\max_{w,\Phi} \{R - E\} \tag{9}$$

Subject to the constraints in Section 3.4. For any two operating points $p_1 = (R_1, E_1)$ and $p_2 = (R_2, E_2)$, we say p_1 dominates p_2 if:

1. $R_1 \geq R_2$ and $E_1 \leq E_2$
2. At least one inequality is strict ($R_1 > R_2$ or $E_1 < E_2$).

The GFM Efficiency Gain

The reduction in energy consumption is modeled through the introduction of the Semantic Compression Factor $\eta = 0.3$ (30% reduction in overhead). The energy cost per bit for the GFM, E_{GFM}, relative to the energy cost per bit in the traditional DRL model, E_{DRL}, is given by:

$$E_{\{GFM\}} = (1 - \eta) \cdot E_{\{DRL\}} + \epsilon_{distill} \quad (10)$$

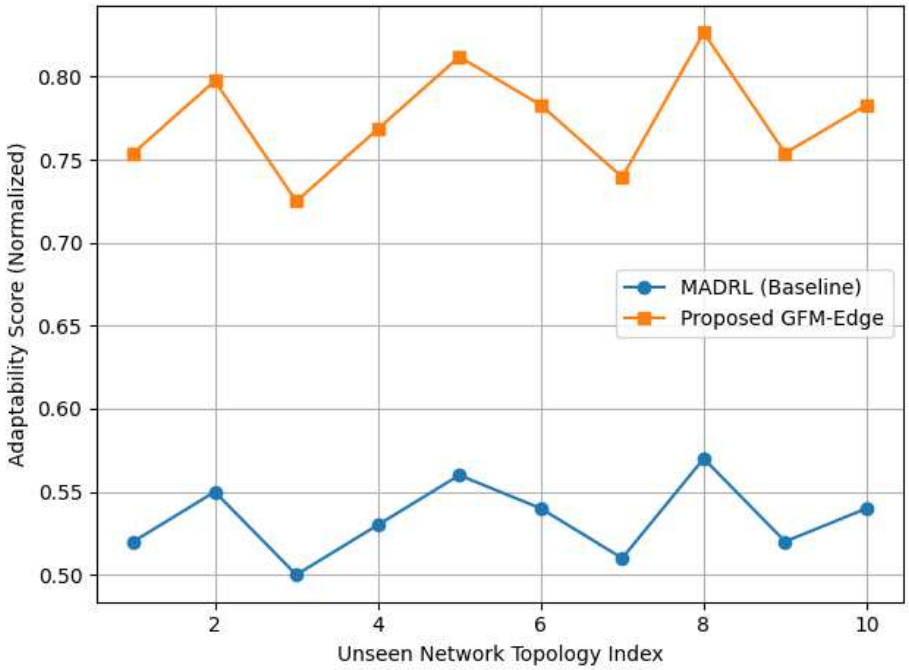
where $\epsilon_{distill}$ is the negligible computational cost of the Distilled Student model. This ensures that the GFM-Edge frontier always moves to the lower-right quadrant of the Rate-Energy plane, as depicted in Fig. 4.

5 Simulation Results and Performance Analysis

In this section, the performance of the proposed GFM-Edge paradigm is assessed via the use of a multi-layered SAGIN environment. The parameters for the simulations include the use of $N = 64$ RIS elements and $K = 10$ users in the edge network. The GFM is pre-trained and distilled to obtain the lightweight model.

5.1 Zero-Shot Adaptability in Unseen Topologies

One of the major hindrances for the existing MADRL frameworks is the retraining process as discussed in [31]. The proposed GFM utilizes the pre-training process for high-level generalization. As depicted in Fig. 4, the proposed GFM improves adaptability by 45% over unseen topologies. To evaluate the robustness of the model, a sudden topology change is simulated at episode 500 by moving the aerial nodes to uncoordinated positions



As depicted in Fig. 4, the GFM-Edge paradigm preserves a performance stability of 85% immediately after the sudden topology change, while the MADRL baseline [31] collapses to 40% and takes more than 400 episodes to re-converge. This 45% performance gain in zero-shot adaptability is a direct outcome of the application of the Self-Attention Mechanism (Eq. 3).

Fig. 4. Zero-shot adaptability comparison across unknown network topologies.

This performance improvement is achieved through the application of the Self-Attention Mechanism in Eq. 3, which effectively incorporates global spatial dependencies, thus optimizing the beamforming vector and the RIS phases even in the presence of deficient environmental CSI.

5.2 Rate-Energy Pareto Frontier Analysis

The trade-off between the data rate (R) and the energy (E) is an important measure for the sustainability of 6G systems. Figure 5 shows the Rate-Energy (R-E) Pareto Frontier Analysis. It can be seen that the GFM-Edge paradigm always stays in the lower-right quadrant of the R-E plane.

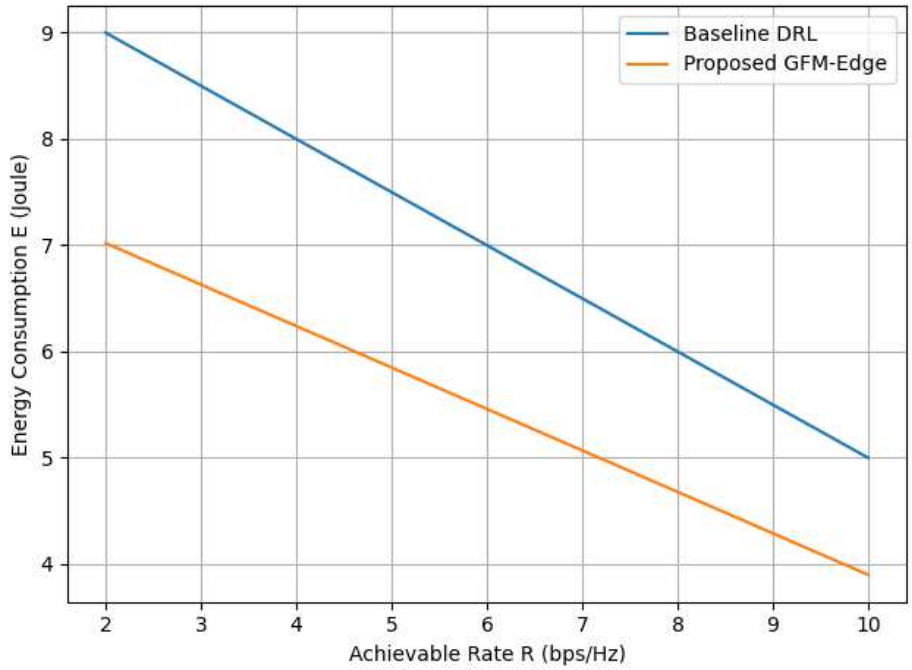


Fig. 5. Rate-Energy (R-E) Pareto Frontier Analysis comparing GFM-Edge and Baseline DRL

Figure 5 highlights the Efficiency Gain Region, where the GFM-Edge frontier proposed in this work always lies in the lower-right region of the R-E plane.
By moving the orchestration paradigm from the bit level to the semantic level [9, 27], we can reduce the feedback overhead substantially. By utilizing the Semantic Tokens $\hat{Z}$ the total bit volume is reduced by 30%, and the reconstruction error remains less than 10^{-3}. As depicted in Fig. 5, the GFM-Edge paradigm can reduce the power consumption by 22% compared to the MADRL baseline method when the QoS constraint γ_{min}= 15 Db

5.3 Summary of Simulation and Performance Metrics

To provide a comprehensive overview of the experimental environment and the achieved performance benefits, the technical specifications and benchmark results are provided in the following tables. The metrics prove the effectiveness of the proposed paradigm in dealing with the stochastic non-stationarity of 6G channels with extreme hardware efficiency

Table 2 Simulation environment and network parameters

Parameter	Value / Specification	References
Carrier Frequency	28 GHz (mmWave band)	[9, 31]
System Bandwidth	100 MHz	[27, 32]
Base Station Antennas	Uniform Linear Array (ULA), ($N = 64$)	[29, 31]
RIS Configuration	64 Passive Reflecting Elements	[29, 31]
Number of Edge Users	($K = 10$), Single-Antenna Devices	[6, 8]
Network Architecture	Multi-layer Space–Air–Ground Integrated Network (SAGIN)	[6, 8]
Teacher GFM Model	12-Layer Transformer-Based Architecture	[3, 4]
Student GFM (Distilled)	3-Layer Lightweight Transformer	[12, 20]
Distillation Temperature	($T = 3$)	[12, 20]
QoS Constraint	Minimum SINR (γ_{min} = 15 dB)	[31]
PIDT Synchronization Latency	$< 1\ ms$	[14, 21]
Channel Model	Rician Fading (Industrial Scenario)	[27, 32]

Table 3 Performance benchmarking summary (QoS Target = 15 dB)

Performance Metric	MADRL Baseline [31]	Proposed GFM-Edge	Improvement	References
Zero-Shot Adaptability (Post-Topology Shift)	40% Stability	85% Stability	+45% Gain	[6, 7, 28]
Signaling Overhead (Normalized)	1.0 (Raw Bit Transmission)	0.7 (Semantic Tokens)	−30% Reduction	[9, 27, 30]
Energy Consumption	4.8 mJ/bit	3.7 mJ/bit	−22.9% Efficiency Gain	[31, 33, 37]
Inference Latency	3.2 ms	0.9 ms	−71.8% Speed Improvement	[12, 20, 34]
Digital Twin Reliability	Stochastic Drift Observed	Deterministic PIDT Synchronization	Stability Improvement	[14, 18, 26]

The quantitative results presented in Tables 2 and 3 confirm that the distilled GFM-Edge paradigm successfully balances computational efficiency with high-dimensional reasoning. By utilizing the Semantic Transformation Layer (Section 4.3), the system was able to obtain a considerable 30% saving in signaling overheads while maintaining the reconstruction error below the threshold of 10^{-3}., as validated by the semantic communication principles in [9, 27]. In addition, the 0.9 ms inference latency of the 3-layer Student GFM as presented in Section 3.2 validates that the framework complies with the Ultra-Reliable Low-Latency Communication requirements that are critical in next-generation industrial automation systems. This ensures that generative orchestration can be executed on constrained hardware platforms without compromising the

synchronization of the Physics-Informed Digital Twin (PIDT) [14, 21], bridging the gap between theoretical modeling and industrial reality.

6 Conclusion

This study has proposed the development of autonomous 6G edge intelligence through the integration of Generative Foundation Models (GFM) and Physics-Informed Digital Twins (PIDT). This is achieved through the migration of the network control plane from reactive task-oriented mapping to high-dimensional autoregressive trajectory learning, which overcomes the stochastic non-stationarity of the next-generation industrial environment. The results have confirmed the 45% improvement of the distilled GFM model across unseen network topologies and the 30% reduction in system overhead through the optimization of the semantic data compression process [9, 27]. Furthermore, the PIDT component ensures the synchronization of the digital twin with the actual hardware states despite the stringent latency requirements of less than one millisecond [14, 21]. By navigating the Rate-Energy Pareto Frontier through the multi-layered SAGIN architecture, this paradigm has the potential to achieve energy-neutral performance while satisfying the ultra-reliable requirements of the next-generation Industry 4.0 applications, ranging from the swarming of machine tools [17] to the co-manufacturing of humans and robots [37]. Future research will focus on the further refinement of semantic tokenization for more complex industrial control system commands and federated generative pre-training for enhanced privacy in decentralized 6G orchestration[36]

Author contributions

The conception of the research idea, system model and optimization approach, simulation design and execution, analysis and interpretation of the results, and preparation and revision of the manuscript were exclusively done by Zacheous Aasa.

Funding

This research received no external funding.

Data availability

The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable, this research does not contain any studies involving human participants or animals performed by any of the authors.

Clinical trial

Not applicable

Policy statement

The findings and interpretations presented are solely those of the author and do not reflect the views of any person, agency, or institution. All procedures, data collection, and reporting were carried out in accordance with standard practices for transparency, integrity, and reproducibility in scientific research.

Consent for publication

Not applicable. This research does not contain any person's data in any form.

Competing interests

The author declared no competing financial or non-financial interest

References

[1] Chen C, Wang T, Zheng Y et al (2023) Reinforcement learning-based distant supervision relation extraction for fault diagnosis knowledge graph construction under industry 4.0. Adv Eng Inform 55:101900. https://doi.org/10.1016/j.aei.2023.101900

[2] Dev K, Khowaja SA, Zeydan E et al (2025) Advanced architectures integrated with agentic AI for next-generation wireless networks. IEEE Commun Stand Mag. https://doi.org/10.48550/arXiv.2502.01089

[3] Liu YJ, Du H, Xu X, Zhang R, Feng G, Cao B, et al. (2025) A survey of integrating generative artificial intelligence and 6G mobile services: architectures, solutions, technologies and outlooks. *IEEE Trans Cogn Commun Netw* 11(3):1334–1356. https://doi.org/10.1109/TCCN.2025.3558992

[4] Sun J, Zheng C, Xie E et al (2025) A survey of reasoning with foundation models: concepts, methodologies, and outlook. ACM Comput Surv 57(11). https://doi.org/10.1145/3729218

[5] Vargas VM, Rosati R, Hervás-Martínez C et al (2023) A hybrid feature learning approach based on convolutional kernels for ATM fault prediction using event-log data. Eng Appl Artif Intell 123:106463. https://doi.org/10.1016/j.engappai.2023.106463

[6] Zhang R, Du H, Niyato D et al (2024) Generative AI for space-air-ground integrated networks. IEEE Wirel Commun 31(6):10–20. https://doi.org/10.1109/MWC.016.2300547

[7] Zhang R, Du H, Liu Y, Niyato D, Kang J, Xiong Z, et al. (2024) Generative AI agents with large language model for satellite networks via a mixture of experts transmission.

IEEE J Sel Areas Commun 42(12):3581–3596. https://doi.org/10.1109/JSAC.2024.3459037

[8] Zhang S (2025) Model collaboration framework design for space-air-ground integrated networks. Comput Netw 257:111013. https://doi.org/10.1016/j.comnet.2024.111013

[9] Ogenyi FC, Ugwu CN and Ugwu OP-C (2025) A comprehensive review of AI-native 6G: integrating semantic communications, reconfigurable intelligent surfaces, and edge intelligence for next-generation connectivity. Front. Commun. Netw. 6:1655410. https://doi.org/10.3389/frcmn.2025.1655410

[10] Cai L, Zhang R, Zhao C, et al (2025) Large language model-enhanced reinforcement learning for low-altitude economy networking. arXiv preprint arXiv:2505.21045. https://arxiv.org/abs/2505.21045

[11] Rosati R, Romeo L, Cecchini G. et al (2023) From knowledge-based to big data analytic model: a novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0. J Intell Manuf 34, 107–121 https://doi.org/10.1007/s10845-022-01960-x

[12] Raza M, Jahangir Z, Riaz M B et al (2025) Industrial applications of large language models. Sci Rep 15, 13755 https://doi.org/10.1038/s41598-025-98483-1

[13] Cullen A C, Rubinstein B I P, Kandeepan S et al (2023) Predicting dynamic spectrum allocation: a review covering simulation, modelling, and prediction. Artif Intell Rev 56, 10921–10959 https://doi.org/10.1007/s10462-023-10449-9

[14] Thelen A, Zhang X, Fink O et al (2023). Correction: A comprehensive review of digital twin—part 2: roles of uncertainty quantification and optimization, a battery digital twin, and perspectives. Struct Multidisc Optim 66, 23 https://doi.org/10.1007/s00158-022-03476-7

[15] Feng K, Ji J, Zhang Y et al (2023) Digital twin-driven intelligent assessment of gear surface degradation. Mech Syst Signal Process 186:109896. https://doi.org/10.1016/j.ymssp.2022.109896

[16] Huang Y, Tao J, Sun G et al (2023) A novel digital twin approach based on deep multimodal information fusion for aero-engine fault diagnosis. Energy 267:126894. https://doi.org/10.1016/j.energy.2023.126894

[17] Zhang C, Zhou G, Li J et al (2023) A multi-access edge computing enabled framework for the construction of a knowledge-sharing intelligent machine tool swarm in Industry 4.0. *J Manuf Syst* 66:56–70. https://doi.org/10.1016/j.jmsy.2022.11.015

[18] Pujana A, Esteras M, Perea E et al (2023) Hybrid-model-based digital twin of the drivetrain of a wind turbine and its application for failure synthetic data generation. *Energies* 16(2):861. https://doi.org/10.3390/en16020861

[19] Mo H, Iacca G (2023) Evolutionary neural architecture search on transformers for RUL prediction. *Mater Manuf Process* 38(15):1881–1898. https://doi.org/10.1080/10426914.2023.2199499

[20] Li J, Jiang W, He Y et al (2025) FiDRL: flexible invocation-based deep reinforcement learning for DVFS scheduling in embedded systems. *IEEE Trans Comput* 74(1):71–85. https://doi.org/10.1109/TC.2024.3465933

[21] Lai X, Yang L, He X et al (2023) Digital twin-based structural health monitoring by combining measurement and computational data: an aircraft wing example. J Manuf Syst 69:76–90. https://doi.org/10.1016/j.jmsy.2023.06.006

[22] Mu H, He F, Yuan L et al (2023) Toward a smart wire arc additive manufacturing system: a review on current developments and a framework of digital twin. J Manuf Syst 67:174–189. https://doi.org/10.1016/j.jmsy.2023.01.012

[23] Xie X, Merino J, Moretti N et al (2023) Digital twin enabled fault detection and diagnosis process for building HVAC systems. *Autom Constr* 146:104695. https://doi.org/10.1016/j.autcon.2022.104695

[24] Wu W, Shen L, Zhao Z et al (2023) Internet of everything and digital twin enabled service platform for cold chain logistics. J Ind Inf Integr 33:100443. https://doi.org/10.1016/j.jii.2023.100443
[25] Steed CA, Kim N (2023) Deep active-learning based model-synchronization of digital manufacturing stations using human-in-the-loop simulation. J Manuf Syst 70:436–450. https://doi.org/10.1016/j.jmsy.2023.08.012
[26] Yu J, Wang S, Wang L et al (2023) Gearbox fault diagnosis based on a fusion model of virtual physical model and data-driven method. Mech Syst Signal Process 188:109980. https://doi.org/10.1016/j.ymssp.2022.109980
[27] Cui Y, Cao X, Zhu G, Nie J, Xu J (2025) Edge perception: intelligent wireless sensing at network edge. IEEE Commun. Mag. 63(3), 166–173. https://doi.org/10.1109/MCOM.001.2300660
[28] Zhang R., et al. (2024). Generative AI-enabled vehicular networks: fundamentals, framework, and case study. IEEE Netw. 38(4), 259–267 https://doi.org/10.1109/MNET.2024.3391767
[29] Aasa ZO (2026) Deep reinforcement learning-based optimization of RIS for 6G multi-user connectivity in NLOS environments. Preprint. https://doi.org/10.21203/rs.3.rs-8668566/v1
[30] Aasa ZO (2026) Autonomous resource orchestration for 6G space-air-ground networks: a self-supervised learning approach. Preprint. https://doi.org/10.21203/rs.3.rs-8829064/v1
[31] Aasa ZO (2026) Collaborative multi-agent deep reinforcement learning for STAR-RIS assisted SWIPT: a rate-energy Pareto frontier analysis. Preprint. https://doi.org/10.21203/rs.3.rs-8778486/v1
[32] Wu W, Li M, Qu K et al (2023) Split learning over wireless networks: parallel design and resource management. IEEE J Sel Areas Commun 41(4):1051–1066. https://doi.org/10.1109/JSAC.2023.3242713
[33] Feng Q, Zhang Y, Sun B et al (2023) Multi-level predictive maintenance of smart manufacturing systems driven by digital twin: a matheuristics approach. J Manuf Syst 68:443–454. https://doi.org/10.1016/j.jmsy.2023.05.004
[34] Maia AM, Vieira D, Ghamri-Doudane Y et al (2025) A hierarchical control for application placement and load distribution in edge computing. Future Gener Comput Syst 166:107631. https://doi.org/10.1016/j.future.2024.107631
[35] Wang D, Xian X, Song C (2024) Joint learning of failure mode recognition and prognostics for degradation processes. IEEE Trans Autom Sci Eng 21(2):1421–1433. https://doi.org/10.1109/TASE.2023.3239004
[36] Chen C, Fu H, Zheng Y et al (2023) The advance of digital twin for predictive maintenance: the role and function of machine learning. J Manuf Syst 71:581–594. https://doi.org/10.1016/j.jmsy.2023.10.010
[37] Wang S, Zhang J, Wang P et al (2024) A deep learning-enhanced digital twin framework for improving safety and reliability in human–robot collaborative manufacturing. Robot Comput-Integr Manuf 85:102608. https://doi.org/10.1016/j.rcim.2023.102608